亚马逊云科技

Amazon EMR 版本指南

Amazon EMR

Amazon EMR: Amazon EMR 版本指南

Amazon Web Services 文档中描述的 Amazon Web Services 服务或功能可能因区域而异。要查看适用于中国区域的差异，请参阅 中国的 Amazon Web Services 服务入门 (PDF)。

Table of Contents

关于 Amazon EMR 发行版

Amazon EMR 发行版是一组来自大数据生态系统的开源应用程序。每个发行版由您在创建集群时选择让 Amazon EMR 安装和配置的各个大数据应用程序、组件和功能组成。应用程序是使用基于 Apache BigTop（与 Hadoop 生态系统关联的开源项目）的系统打包的。本指南提供 Amazon EMR 发行版中所含应用程序的信息。

有关 Amazon EMR 入门和使用的更多信息，请参阅 Amazon EMR 管理指南。

启动集群时，有多个 Amazon EMR 发行版可供选择。这允许您测试和使用满足您解决方案兼容性需求的应用程序版本。可使用发行版标注指定版本。发行版标注的格式是 `emr-x.x.x`。For example, `emr-6.14.0`.

您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发行版 5.18.0 开始）附带的准确版本的库和依赖项的任务代码。有关更多信息，请参阅使用 Amazon EMR 项目存储库检查依赖项。

订阅 RSS 源，通过 https://docs.amazonaws.cn/emr/latest/ReleaseGuide/amazon-emr-release-notes.rss 获取 Amazon EMR 发布说明，以便在新的 Amazon EMR 发行版可用时接收更新。

最新发行版详细信息，包括 Amazon EMR 6.x 系列和 5.x 系列的应用程序版本、发布说明、组件和配置分类：

- Amazon EMR 发行版 6.14.0
- Amazon EMR 版本 5.36.1

> **Note**
>
> 从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

最新 Amazon EMR 发行版的发布说明和所有发行版的历史记录：

- 新增功能
- 发布说明的 Amazon EMR 存档

每个 Amazon EMR 发行版中的应用程序版本的全面历史记录：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

每个 Amazon EMR 发行版的详细信息 和发行版系列之间的差异（如果适用）：

- Amazon EMR 6.x 发行版
- Amazon EMR 5.x 发行版
- Amazon EMR 4.x 发行版
- Amazon EMR 2.x 和 3.x AMI 版本

Amazon EMR 6.x 发行版

本部分内容涵盖每个 Amazon EMR 6.x 发行版中可用的应用程序版本、发布说明、组件版本和配置分类。

启动集群时，有多个 Amazon EMR 发行版可供选择。这允许您测试和使用满足您解决方案兼容性需求的应用程序版本。可使用发行版标注指定版本。发行版标注的格式是 `emr-x.x.x`。 For example, `emr-6.14.0`.

从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

有关每个 Amazon EMR 6.x 发行版中的应用程序版本的综合表格，请参阅Amazon EMR 6.x 发行版中的应用程序版本。

主题

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 发行版 6.14.0
- Amazon EMR 版本 6.13.0
- Amazon EMR 版本 6.12.0
- Amazon EMR 版本 6.11.1
- Amazon EMR 版本 6.11.0
- Amazon EMR 版本 6.10.1
- Amazon EMR 版本 6.10.0

- Amazon EMR 版本 6.9.1
- Amazon EMR 发行版 6.9.0
- Amazon EMR 版本 6.8.1
- Amazon EMR 发行版 6.8.0
- Amazon EMR 发行版 6.7.0
- Amazon EMR 发行版 6.6.0
- Amazon EMR 发行版 6.5.0
- Amazon EMR 发行版 6.4.0
- Amazon EMR 发行版 6.3.1
- Amazon EMR 发行版 6.3.0
- Amazon EMR 发行版 6.2.1
- Amazon EMR 发行版 6.2.0
- Amazon EMR 发行版 6.1.1
- Amazon EMR 发行版 6.1.0
- Amazon EMR 发行版 6.0.1
- Amazon EMR 发行版 6.0.0

Amazon EMR 6.x 发行版中的应用程序版本

有关列出每个 Amazon EMR 6.x 发行版中可用的应用程序版本的综合表格，请在浏览器打开 Amazon EMR 6.x 发行版中的应用程序版本。

Amazon EMR 发行版 6.14.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.14.0	emr-6.13.0	emr-6.12.0	emr-6.11.1
Amazon SDK for Java	1.12.543	1.12.513	1.12.490	1.12.446
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.4.0	2.4.0	2.4.0	2.2.0
Flink	1.17.1-amzn-0	1.17.0	1.17.0	1.16.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.17	2.4.17	2.4.17	2.4.15
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.13.1-amzn-2	0.13.1-amzn-1	0.13.1-amzn-0	0.13.0-amzn-0

	emr-6.14.0	emr-6.13.0	emr-6.12.0	emr-6.11.1
Hue	4.11.0	4.11.0	4.11.0	4.11.0
Iceberg	1.3.1-amzn-0	1.3.0-amzn-1	1.3.0-amzn-0	1.2.0-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.5.0	1.5.0	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.3	5.1.3	5.1.3	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.281	0.281	0.281	0.279
Spark	3.4.1	3.4.1	3.4.0	3.3.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.11.0	2.11.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	422	414	414	410
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.14.0 的信息。更改与 6.13.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.14.0 supports Apache Spark 3.4.1, Apache Spark RAPIDS 23.06.0-amzn-2, Flink 1.17.1, Iceberg 1.3.1, and Trino 422.
- Amazon EMR 托管式自动扩缩功能 现已在 `ap-southeast-3` 亚太地区（雅加达）区域开放，可用于您使用 Amazon EMR 6.14.0 及更高版本创建的集群。

更改、增强功能和解决的问题

- 6.14.0 发行版通过在 Amazon EC2 上运行的 Amazon EMR 来优化日志管理。因此，您可能会看到集群日志的存储成本略有降低。
- 6.14.0 发行版改进了扩展工作流，以满足 Amazon EBS 卷大小差异很大的不同核心实例需求。此改进仅适用于核心节点；任务节点的缩减操作不受影响。
- 6.14.0 发行版改进了 Amazon EMR 与 Apache Hadoop YARN ResourceManager and HDFS NameNode 等开源应用程序交互的方式。此改进降低了集群扩展导致操作延迟的风险，并减少了由于开源应用程序连接问题导致的启动故障。
- 6.14.0 发行版优化了集群启动时的应用程序安装。此改进缩短了某些 Amazon EMR 应用程序组合的集群启动时间。
- 6.14.0 发行版修复了在具有自定义域的 VPC 上运行的集群遇到核心节点或任务节点重启时，集群的缩减操作可能会停滞的问题。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 906.0	4.14.322	2023 年 9 月 11 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.4.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
delta-standalone-connectors	0.6.0	Delta Connectors 提供不同的运行时，将 Delta Lake 与 Flink、Hive 和 Presto 等引擎集成。
emr-ddb	5.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.7.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.11.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env

组件	版本	描述
emr-s3-dist-cp	2.28.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.7.0	EMR S3 Select 连接器
emr-wal-cli	1.1.0	用于 emrwal 列表/删除的 cli。
emrfs	2.59.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.17.1-amzn-0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.17.1-amzn-0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-hdfs-journalnode	3.3.3-amzn-6	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.17-amzn-2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.17-amzn-2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.17-amzn-2	HBase 命令行客户端。
hbase-rest-server	2.4.17-amzn-2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.17-amzn-2	用于向 HBase 提供 Thrift 终端节点的服务。

组件	版本	描述
hbase-operator-tools	2.4.17-amzn-2	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-7	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-7	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-7	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-7	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-7	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-7	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-7	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.13.1-amzn-2	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.13.1-amzn-2	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.13.1-amzn-2	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.13.1-amzn-2	用于运行 Spark 以及 Hudi 的捆绑库。

组件	版本	描述
hue-server	4.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	1.3.1-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.5.0	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.7.0	开源计算机视觉库。
phoenix-library	5.1.3	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.3	Apache Phoenix-Connectors for Spark-3

组件	版本	描述
phoenix-query-server	5.1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.281-amzn-2	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.281-amzn-2	用于执行查询的各个部分的服务。
presto-client	0.281-amzn-2	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	422-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	422-amzn-0	用于执行查询的各个部分的服务。
trino-client	422-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.4.1-amzn-1	Spark 命令行客户端。
spark-history-server	3.4.1-amzn-1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	3.4.1-amzn-1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.4.1-amzn-1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	23.06.0-amzn-2	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-5	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-5	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.14.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-java-home	更改 Hadoop 的 KMS java 主页	Not available.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.

分类	描述	重新配置操作
delta-defaults	更改 Delta 的 delta-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-lakeformat ion	更改 Presto 的 lakeforma tion.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.

分类	描述	重新配置操作
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.

分类	描述	重新配置操作
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

6.14.0 发行版的更改日志和发布说明

日期	事件	描述
2023-11-02	部署完成	Amazon EMR 6.14.0 已全面部署到所有支持的区域
2023-10-10	文档发布	首次发布 Amazon EMR 6.14.0 发布说明
2023-10-04	首次发布	Amazon EMR 6.14.0 首次部署到部分商业区域

Amazon EMR 版本 6.13.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.13.0	emr-6.12.0	emr-6.11.1	emr-6.11.0
Amazon SDK for Java	1.12.513	1.12.490	1.12.446	1.12.446
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.4.0	2.4.0	2.2.0	2.2.0
Flink	1.17.0	1.17.0	1.16.0	1.16.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.17	2.4.17	2.4.15	2.4.15
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.13.1-amzn-1	0.13.1-amzn-0	0.13.0-amzn-0	0.13.0-amzn-0
Hue	4.11.0	4.11.0	4.11.0	4.11.0
Iceberg	1.3.0-amzn-1	1.3.0-amzn-0	1.2.0-amzn-0	1.2.0-amzn-0

	emr-6.13.0	emr-6.12.0	emr-6.11.1	emr-6.11.0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.5.0	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.3	5.1.3	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.281	0.281	0.279	0.279
Spark	3.4.1	3.4.0	3.3.2	3.3.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.11.0	2.11.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	414	414	410	410
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.13.0 的信息。更改与 6.12.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.13.0 supports Apache Spark 3.4.1, Apache Spark RAPIDS 23.06.0-amzn-1, CUDA Toolkit 11.8.0, and JupyterHub 1.5.0.

更改、增强功能和解决的问题

- 6.13.0 版本改进了 Amazon EMR 日志管理进程守护程序，以确保在发出集群终止命令时，所有日志都定期上传到 Amazon S3。这有助于更快地终止集群。
- 6.13.0 版本增强了 Amazon EMR 日志管理功能，确保所有日志文件一致而及时地上传到 Amazon S3。这尤其有利于长期运行的 EMR 集群。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 808.0	4.14.320	2023 年 8 月 24 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.4.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。

组件	版本	描述
delta-standalone-connectors	0.6.0	Delta Connectors 提供不同的运行时，将 Delta Lake 与 Flink、Hive 和 Presto 等引擎集成。
emr-ddb	5.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.6.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.10.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.27.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.6.0	EMR S3 Select 连接器
emr-wal-cli	1.1.0	用于 emrwal 列表/删除的 cli。
emrfs	2.58.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.17.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.17.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-5	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	3.3.3-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.17-amzn-1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.17-amzn-1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.17-amzn-1	HBase 命令行客户端。
hbase-rest-server	2.4.17-amzn-1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.17-amzn-1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.17-amzn-1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-6	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-6	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-6	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-6	Hive 命令行客户端。

组件	版本	描述
hive-hbase	3.1.3-amzn-6	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-6	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-6	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.13.1-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.13.1-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.13.1-amzn-1	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.13.1-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	1.3.0-amzn-1	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.5.0	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.7.0	开源计算机视觉库。
phoenix-library	5.1.3	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.3	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.281-amzn-1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.281-amzn-1	用于执行查询的各个部分的服务。
presto-client	0.281-amzn-1	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。

组件	版本	描述
trino-coordinator	414-amzn-1	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	414-amzn-1	用于执行查询的各个部分的服务。
trino-client	414-amzn-1	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.4.1-amzn-0	Spark 命令行客户端。
spark-history-server	3.4.1-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.4.1-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.4.1-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	23.06.0-amzn-1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-4	tez YARN 应用程序和库。

组件	版本	描述
tez-on-worker	0.10.2-amzn-4	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.13.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.

分类	描述	重新配置操作
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.

分类	描述	重新配置操作
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.

分类	描述	重新配置操作
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-java-home	更改 Hadoop 的 KMS java 主页	Not available.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-def aults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhu b_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.js on 文件中的值。	Not available.

分类	描述	重新配置操作
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.

分类	描述	重新配置操作
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-lakeformation	更改 Presto 的 lakeformation.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.

分类	描述	重新配置操作
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.13.0 的更改日志和发布说明

日期	事件	描述
2023-09-23	部署完成	Amazon EMR 6.13.0 已全面部署到所有支持的区域

日期	事件	描述
2023-09-12	文档发布	Amazon EMR 6.13.0 发布说明首次发布
2023-09-01	首次发布	Amazon EMR 6.13.0 首次面向部分商业区域部署

Amazon EMR 版本 6.12.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.12.0	emr-6.11.1	emr-6.11.0	emr-6.10.1
Amazon SDK for Java	1.12.490	1.12.446	1.12.446	1.12.397

	emr-6.12.0	emr-6.11.1	emr-6.11.0	emr-6.10.1
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.4.0	2.2.0	2.2.0	2.2.0
Flink	1.17.0	1.16.0	1.16.0	1.16.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.17	2.4.15	2.4.15	2.4.15
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.13.1-amzn-0	0.13.0-amzn-0	0.13.0-amzn-0	0.12.2-amzn-0
Hue	4.11.0	4.11.0	4.11.0	4.10.0
Iceberg	1.3.0-amzn-0	1.2.0-amzn-0	1.2.0-amzn-0	1.1.0-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.5.0
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.3	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0

	emr-6.12.0	emr-6.11.1	emr-6.11.0	emr-6.10.1
Presto	0.281	0.279	0.279	0.278
Spark	3.4.0	3.3.2	3.3.2	3.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.11.0	2.11.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	414	410	410	403
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.12.0 的信息。更改与 6.11.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.12.0 supports Apache Spark 3.4.0, Apache Spark RAPIDS 23.06.0-amzn-0, CUDA 11.8.0, Apache Hudi 0.13.1-amzn-0, Apache Iceberg 1.3.0-amzn-0, Trino 414, and PrestoDB 0.281.
- Amazon EMR 发布 6.12.0 及更高版本支持 LDAP 通过 HiveServer2（HS2）、Trino、Presto 和 Hue 与 Apache Livy、Apache Live 集成。您还可以在使用 6.12.0 或更高版本的 EMR 集群上安装 Apache Spark 和 Apache Hadoop，并将它们配置为使用 LDAP。有关更多信息，请参阅使用 Active Directory 或 LDAP 服务器通过 Amazon EMR 进行身份验证。

更改、增强功能和解决的问题

- Amazon EMR 6.12.0 及更高版本为 Flink 提供 Java 11 运行时系统支持。有关更多信息，请参阅将 Flink 配置为使用 Java 11 运行。

- Amazon EMR 6.12.0 默认支持所有搭载 Amazon Corretto 8 的应用程序，但 Trino 除外。对于 Trino，Amazon EMR 从Amazon EMR 版本 6.9.0 开始默认支持 Amazon Corretto 17。Amazon EMR 还支持某些搭载 Amazon Corretto 11 和 17 的应用程序。下表列出了这些应用程序。如果要更改集群上的默认 JVM，请按照在集群上运行的每个应用程序的 配置应用程序来使用特定 Java 虚拟机 中的说明进行操作。一个集群只能使用一个 Java 运行时系统版本。Amazon EMR 不支持在同一集群的不同运行时系统版本上运行不同的节点或应用程序。

  虽然 Amazon EMR 在 Apache Spark、Apache Hadoop 和 Apache Hive 上同时支持 Amazon Corretto 11 和 17，但当您使用这些版本的 Corretto 时，某些工作负载的性能可能会下降。我们建议您在更改默认值之前先测试工作负载。

  Amazon EMR 6.12 中应用程序的默认 Java 版本

应用程序	Java/Amazon Corretto 版本（默认为粗体）
Delta	17、11、8
Flink	11、8
Ganglia	8
HBase	11、8
HCatalog	17、11、8
Hadoop	17、11、8
Hive	17、11、8
Hudi	17、11、8
Iceberg	17、11、8
Livy	17、11、8
Oozie	17、11、8
Phoenix	8
PrestoDB	8

应用程序	Java/Amazon Corretto 版本（默认为粗体）
Spark	17、11、8
Spark RAPIDS	17、11、8
Sqoop	8
Tez	17、11、8
Trino	17
Zeppelin	8
Pig	8
Zookeeper	8

- 6.12.0 版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 6.12.0 版本修复了一个问题，即当处于正常停用状态的核心节点在完全停用之前出于任何原因变得运行不正常时，集群的缩减操作可能会停滞不前。
- 6.12.0 版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 6.12.0 版本通过提高记录实例状态变化的速度，提高了 Amazon EMR 运行状况监控服务的性能和效率。这一改进降低了运行多个自定义客户端工具或第三方应用程序的集群节点性能下降的机会。
- 6.12.0 版本提高了 Amazon EMR 的集群上日志管理进程守护程序的性能。因此，对于以高并发度运行步骤的 EMR 集群，性能下降的可能性较小。
- 在 Amazon EMR 6.12.0 版本中，日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 6.12.0 版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 6.12.0 版本支持 YARN Timeline Server 日志的日志轮换。这样可以最大限度地减少磁盘过度使用情况，特别是对于长时间运行的集群。

- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			酋）、加拿大（中部）、以色列（特拉维夫）
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.4.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
delta-standalone-connectors	0.6.0	Delta Connectors 提供不同的运行时，将 Delta Lake 与 Flink、Hive 和 Presto 等引擎集成。
emr-ddb	5.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.9.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.26.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.5.0	EMR S3 Select 连接器
emr-wal-cli	1.1.0	用于 emrwal 列表/删除的 cli。

组件	版本	描述
emrfs	2.57.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.17.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.17.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	3.3.3-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.17-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.17-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.17-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.4.17-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.17-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.17-amzn-0	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-5	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	3.1.3-amzn-5	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-5	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-5	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-5	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-5	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-5	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.13.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.13.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.13.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.13.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
iceberg	1.3.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.7.0	开源计算机视觉库。
phoenix-library	5.1.3	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.3	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.281-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.281-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.281-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	414-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	414-amzn-0	用于执行查询的各个部分的服务。
trino-client	414-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.4.0-amzn-0	Spark 命令行客户端。
spark-history-server	3.4.0-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.4.0-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.4.0-amzn-0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
spark-rapids	23.06.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-3	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-3	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.12.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-java-home	更改 Hadoop 的 KMS java 主页	Not available.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.

分类	描述	重新配置操作
delta-defaults	更改 Delta 的 delta-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-lakeformat ion	更改 Presto 的 lakeforma tion.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.

分类	描述	重新配置操作
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.

分类	描述	重新配置操作
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.12.0 的更改日志和发布说明

日期	事件	描述
2023-07-27	更新文档	更新 6.12 的 Java 选项并添加 Oozie 教程来更新 JVM
2023-07-21	部署完成	Amazon EMR 6.12.0 已全面部署到所有支持的区域
2023-07-21	文档发布	Amazon EMR 6.12.0 发布说明首次发布
2023-07-12	首次发布	Amazon EMR 6.12.0 首次面向部分商业区域部署

Amazon EMR 版本 6.11.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.11.1	emr-6.11.0	emr-6.10.1	emr-6.10.0
Amazon SDK for Java	1.12.446	1.12.446	1.12.397	1.12.397
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.2.0	2.2.0	2.2.0	2.2.0
Flink	1.16.0	1.16.0	1.16.0	1.16.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.15	2.4.15	2.4.15	2.4.15
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.13.0-amzn-0	0.13.0-amzn-0	0.12.2-amzn-0	0.12.2-amzn-0

	emr-6.11.1	emr-6.11.0	emr-6.10.1	emr-6.10.0
Hue	4.11.0	4.11.0	4.10.0	4.10.0
Iceberg	1.2.0-amzn-0	1.2.0-amzn-0	1.1.0-amzn-0	1.1.0-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.4.1	1.4.1	1.5.0	1.5.0
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.279	0.279	0.278	0.278
Spark	3.3.2	3.3.2	3.3.1	3.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.11.0	2.11.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	410	410	403	403
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.11.1 的信息。更改与 6.11.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

更改、增强功能和解决的问题

- 由于锁争用，如果在尝试停用节点的同时添加或移除节点，则该节点可能会陷入死锁。结果，Hadoop 资源管理器（YARN）变得无响应，并会影响所有传入和当前正在运行的容器。
- 此版本包括一项更改，允许高可用性集群在重启后从故障状态中恢复。
- 此版本包含针对 Hue 和 HBase 的安全补丁。
- 此版本修复了在 Spark 上使用 Amazon EMR 运行工作负载的集群可能会静默收到包含 `contains`、`startsWith`、`endsWith` 和 `like` 错误结果的问题。当您在 Amazon EMR Hive3 Metastore 服务器（HMS）中使用包含元数据的分区字段的表达式时，就会出现此问题。
- 此版本修复了没有用户定义函数（UDF）时在 Glue 端的节流问题。
- 此版本修复了在 YARN 停用时，在日志推送器能够将容器日志推送到 S3 之前，节点日志聚合服务会删除容器日志的问题。
- 此版本修复了 Hadoop 启用节点标签时 FairShare 调度器指标的问题。
- 此版本修复了您在 `spark-defaults.conf` 中为 `spark.yarn.heterogeneousExecutors.enabled` 配置设置默认 `true` 值时影响 Spark 性能的问题。
- 此版本修复了 Reduce Task 无法读取随机数据的问题。该问题因内存损坏错误导致 Hive 查询失败。
- 此版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 此版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 此版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.2.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
delta-standalone-connectors	0.6.0	Delta Connectors 提供不同的运行时，将 Delta Lake 与 Flink、Hive 和 Presto 等引擎集成。
emr-ddb	5.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.8.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.25.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.4.0	EMR S3 Select 连接器
emr-wal-cli	1.1.0	用于 emrwal 列表/删除的 cli。

组件	版本	描述
emrfs	2.56.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.16.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.16.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-3.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-3.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-3.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-3.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-3.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	3.3.3-amzn-3.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-3.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-3.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-3.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-3.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-3.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.15-amzn-1.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.15-amzn-1.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.15-amzn-1.1	HBase 命令行客户端。
hbase-rest-server	2.4.15-amzn-1.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.15-amzn-1.1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.15-amzn-1.1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-4.1	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	3.1.3-amzn-4.1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-4.1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-4.1	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-4.1	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-4.1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-4.1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.13.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.13.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.13.0-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.13.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
iceberg	1.2.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.279-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.279-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.279-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	410-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	410-amzn-0	用于执行查询的各个部分的服务。
trino-client	410-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.2-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.3.2-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.2-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.2-amzn-0.1	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
spark-rapids	23.02.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-2.1	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-2.1	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.11.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-def aults.conf 文件中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-lakeformat ion	更改 Presto 的 lakeforma tion.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.

分类	描述	重新配置操作
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.11.1 的更改日志和发布说明

日期	事件	描述
2023-08-30	更新发行说明	在发行说明中添加了几个与控制面板相关的修复
2023-08-21	文档发布	Amazon EMR 6.11.1 发布说明首次发布
2023-08-16	部署完成	Amazon EMR 6.11.1 已全面部署到所有支持的区域
2023-08-04	首次发布	Amazon EMR 6.11.1 首次面向部分商业区域部署

Amazon EMR 版本 6.11.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本

- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.11.0	emr-6.10.1	emr-6.10.0	emr-6.9.1
Amazon SDK for Java	1.12.446	1.12.397	1.12.397	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.2.0	2.2.0	2.2.0	2.1.0
Flink	1.16.0	1.16.0	1.16.0	1.15.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.15	2.4.15	2.4.15	2.4.13
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.13.0-amzn-0	0.12.2-amzn-0	0.12.2-amzn-0	0.12.1-amzn-0
Hue	4.11.0	4.10.0	4.10.0	4.10.0
Iceberg	1.2.0-amzn-0	1.1.0-amzn-0	1.1.0-amzn-0	0.14.1-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.4.1	1.5.0	1.5.0	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1

	emr-6.11.0	emr-6.10.1	emr-6.10.0	emr-6.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.279	0.278	0.278	0.276
Spark	3.3.2	3.3.1	3.3.1	3.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.11.0	2.10.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	410	403	403	398
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.11.0 的信息。更改与 6.10.0 有关。有关发布时间表的信息，请参阅更改日志。

新特征

- Amazon EMR 6.11.0 支持 Apache Spark 3.3.2-amzn-0、Apache Spark RAPIDS 23.02.0-amzn-0、CUDA 11.8.0、Apache Hudi 0.13.0-amzn-0、Apache Iceberg 1.2.0-amzn-0、Trino 410-amzn-0 和 PrestoDB 0.279-amzn-0。

更改、增强功能和解决的问题

- 在 Amazon EMR 6.11.0 中，DynamoDB 连接器已升级到 5.0.0 版。5.0.0 版本使用 Amazon SDK for Java 2.x。之前的版本使用的是 Amazon SDK for Java 1.x。由于此次升级，我们强烈建议您在将 DynamoDB 连接器与 Amazon EMR 6.11 配合使用之前，先测试您的代码。
- 当 Amazon EMR 6.11.0 的 DynamoDB 连接器调用 DynamoDB 服务时，它会使用您为 `dynamodb.endpoint` 属性提供的区域值。我们建议您在使用 `dynamodb.endpoint` 时也配置 `dynamodb.region`，并且两个属性都以相同的 Amazon Web Services 区域 为目标。如果您使用 `dynamodb.endpoint` 但不配置 `dynamodb.region`，则适用于 Amazon EMR 6.11.0 的 DynamoDB 连接器将返回一个无效的区域异常，并尝试协调来自 Amazon EC2 实例元数据服务（IMDS）的 Amazon Web Services 区域 信息。如果连接器无法从 IMDS 检索区域，则默认为美国东部（弗吉尼亚州北部）(`us-east-1`)。以下错误是您未正确配置该 `dynamodb.region` 属性时可能会遇到的无效区域异常的示例：`error software.amazon.awssdk.services.dynamodb.model.DynamoDbException: Credential should be scoped to a valid region.` 有关受 Amazon SDK for Java 升级到 2.x 影响的类的更多信息，请参阅 Amazon EMR – DynamoDB 连接器的 GitHub 存储库中的 Upgrade Amazon SDK for Java from 1.x to 2.x (#175) 提交。
- 此版本修复了在执行列重命名操作后使用 Delta Lake 在 Amazon S3 中存储 Delta 表数据时列数据变为 NULL 的问题。有关 Delta Lake 中此实验性功能的更多信息，请参阅《Delta Lake User Guide》中的 Column rename operation。
- 6.11.0 版本修复了通过从具有多个主节点的集群中复制一个主节点来创建边缘节点时可能出现的问题。复制的边缘节点可能会导致缩减操作的延迟，或者导致主节点的内存使用率过高。有关如何创建边缘节点以与 EMR 集群通信的更多信息，请参阅 GitHub `aws-samples` 存储库中的 Edge Node Creator。
- 6.11.0 版本改进了 Amazon EMR 用于在重启后将 Amazon EBS 卷重新挂载到实例的自动化流程。
- 6.11.0 版本修复了导致 Amazon EMR 向 Amazon CloudWatch 发布的 Hadoop 指标间歇性出现差距的问题。
- 6.11.0 版本修复了 EMR 集群的一个问题，即由于磁盘过度使用而导致对包含集群节点排除列表的 YARN 配置文件的更新中断。不完整的更新阻碍了未来对集群的缩减操作。此版本可确保您的集群保持正常运行，并确保扩展操作按预期进行。
- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。
- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

在 Amazon EMR 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false` 以解决此问题。

虽然该修复解决了 YARN-9608 引入的问题，但由于启用了托管扩展的集群上的随机数据丢失，它可能会导致 Hive 作业失败。在此版本中，我们还通过设置 Hive `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-shuffle-data` 工作负载来降低这种风险。此配置在 Amazon EMR 版本 6.11.0 及更高版本中提供。

- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.2.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
delta-standalone-connectors	0.6.0	Delta Connectors 提供不同的运行时，将 Delta Lake 与 Flink、Hive 和 Presto 等引擎集成。
emr-ddb	5.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.8.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.25.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.4.0	EMR S3 Select 连接器
emr-wal-cli	1.1.0	用于 emrwal 列表/删除的 cli。

组件	版本	描述
emrfs	2.56.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.16.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.16.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-3	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	3.3.3-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.15-amzn-1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.15-amzn-1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.15-amzn-1	HBase 命令行客户端。
hbase-rest-server	2.4.15-amzn-1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.15-amzn-1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.15-amzn-1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	3.1.3-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-4	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-4	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.13.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.13.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.13.0-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.13.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
iceberg	1.2.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.279-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.279-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.279-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	410-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	410-amzn-0	用于执行查询的各个部分的服务。
trino-client	410-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.2-amzn-0	Spark 命令行客户端。
spark-history-server	3.3.2-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.2-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.2-amzn-0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
spark-rapids	23.02.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-2	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-2	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.11.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-def aults.conf 文件中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-lakeformat ion	更改 Presto 的 lakeforma tion.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.

分类	描述	重新配置操作
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.11.0 的更改日志和发布说明

日期	事件	描述
2023-08-21	更新	修复了 Hadoop 3.3.3 引入的问题。
2023-07-26	更新	新的操作系统版本标签 `2.0.20230612.0` 和 `2.0.20230628.0` 。
2023-06-09	部署完成	Amazon EMR 6.11.0 已全面部署到所有支持的区域
2023-06-09	文档发布	Amazon EMR 6.11.0 发布说明首次发布
2023-06-08	首次发布	Amazon EMR 6.11.0 首次面向部分商业区域部署

Amazon EMR 版本 6.10.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.10.1	emr-6.10.0	emr-6.9.1	emr-6.9.0
Amazon SDK for Java	1.12.397	1.12.397	1.12.170	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.2.0	2.2.0	2.1.0	2.1.0
Flink	1.16.0	1.16.0	1.15.2	1.15.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.15	2.4.15	2.4.13	2.4.13
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.3.3
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.12.2-amzn-0	0.12.2-amzn-0	0.12.1-amzn-0	0.12.1-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.10.0
Iceberg	1.1.0-amzn-0	1.1.0-amzn-0	0.14.1-amzn-0	0.14.1-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.6.0
JupyterHub	1.5.0	1.5.0	1.4.1	1.4.1

	emr-6.10.1	emr-6.10.0	emr-6.9.1	emr-6.9.0
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.278	0.278	0.276	0.276
Spark	3.3.1	3.3.1	3.3.0	3.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.11.0	2.10.0	2.10.0
Tez	0.10.2	0.10.2	0.10.2	0.10.2
Trino (PrestoSQL)	403	403	398	398
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.10.1 的信息。更改与 6.10.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

更改、增强功能和解决的问题

- 由于锁争用，如果在尝试停用节点的同时添加或移除节点，则该节点可能会陷入死锁。结果，Hadoop 资源管理器（YARN）变得无响应，并会影响所有传入和当前正在运行的容器。

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  在 Amazon EMR 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false` 以解决此问题。

  虽然该修复解决了 YARN-9608 引入的问题，但由于启用了托管扩展的集群上的随机数据丢失，它可能会导致 Hive 作业失败。在此版本中，我们还通过设置 Hive `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-shuffle-data` 工作负载来降低这种风险。此配置在 Amazon EMR 版本 6.11.0 及更高版本中提供。
- 使用实例组配置的集群中的主节点失效转移后，指标收集器不会向控制面板发送任何指标。
- 此版本包括一项更改，允许高可用性集群在重启后从故障状态中恢复。
- 此版本包含针对 Hue 和 HBase 的安全补丁。
- 此版本修复了在 Spark 上使用 Amazon EMR 运行工作负载的集群可能会静默收到包含 `contains`、`startsWith`、`endsWith` 和 `like` 错误结果的问题。当您在 Amazon EMR Hive3 Metastore 服务器（HMS）中使用包含元数据的分区字段的表达式时，就会出现此问题。
- 此版本修复了没有用户定义函数（UDF）时在 Glue 端的节流问题。
- 此版本修复了在 YARN 停用时，在日志推送器能够将容器日志推送到 S3 之前，节点日志聚合服务会删除容器日志的问题。
- 此版本修复了 Hadoop 启用节点标签时 FairShare 调度器指标的问题。
- 此版本修复了您在 `spark-defaults.conf` 中为 `spark.yarn.heterogeneousExecutors.enabled` 配置设置默认 `true` 值时影响 Spark 性能的问题。
- 此版本修复了 Reduce Task 无法读取随机数据的问题。该问题因内存损坏错误导致 Hive 查询失败。
- 此版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 此版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。

- 此版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 此版本修复了通过从具有多个主节点的集群中复制一个主节点来创建边缘节点时可能出现的问题。复制的边缘节点可能会导致缩减操作的延迟，或者导致主节点的内存使用率过高。有关如何创建边缘节点以与 EMR 集群通信的更多信息，请参阅 GitHub `aws-samples` 存储库中的 Edge Node Creator。
- 此版本改进了 Amazon EMR 用于在重启后将 Amazon EBS 卷重新挂载到实例的自动化流程。
- 此版本修复了导致 Amazon EMR 向 Amazon CloudWatch 发布的 Hadoop 指标间歇性出现差距的问题。
- 此版本修复了 EMR 集群的一个问题，即由于磁盘过度使用而导致对包含集群节点排除列表的 YARN 配置文件的更新中断。不完整的更新阻碍了未来对集群的缩减操作。此版本可确保您的集群保持正常运行，并确保扩展操作按预期进行。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.2.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	3.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.7.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.24.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.3.0	EMR S3 Select 连接器
emr-wal-cli	1.0.0	用于 emrwal 列表/删除的 cli。
emrfs	2.55.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.16.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.16.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-2.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-2.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-2.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-2.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-2.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-2.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-2.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-2.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-2.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-2.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-2.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	2.4.15-amzn-0.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.15-amzn-0.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.15-amzn-0.1	HBase 命令行客户端。
hbase-rest-server	2.4.15-amzn-0.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.15-amzn-0.1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.15-amzn-0.1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-3.1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-3.1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-3.1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-3.1	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-3.1	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-3.1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-3.1	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hudi	0.12.2-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.12.2-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.12.2-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.12.2-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	1.1.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.5.0	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	6.0.0-SNAPSHOT	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	6.0.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.278.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.278.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.278.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	403-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	403-amzn-0	用于执行查询的各个部分的服务。
trino-client	403-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目

组件	版本	描述
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.1-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.3.1-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.1-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.1-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.12.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-1.1	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-1.1	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.10.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-lakeformation	更改 Presto 的 lakeformation.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.

分类	描述	重新配置操作
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.

分类	描述	重新配置操作
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.10.1 的更改日志和发布说明

日期	事件	描述
2023-08-30	更新发行说明	在发行说明中添加了几个与控制面板相关的修复
2023-08-21	文档发布	Amazon EMR 6.10.1 发布说明首次发布
2023-08-16	部署完成	Amazon EMR 6.10.1 已全面部署到所有支持的区域
2023-08-04	首次发布	Amazon EMR 6.10.1 首次面向部分商业区域部署

Amazon EMR 版本 6.10.0

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.10.0	emr-6.9.1	emr-6.9.0	emr-6.8.1
Amazon SDK for Java	1.12.397	1.12.170	1.12.170	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.2.0	2.1.0	2.1.0	-
Flink	1.16.0	1.15.2	1.15.2	1.15.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-6.10.0	emr-6.9.1	emr-6.9.0	emr-6.8.1
HBase	2.4.15	2.4.13	2.4.13	2.4.12
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.3.3	3.2.1
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.12.2-amzn-0	0.12.1-amzn-0	0.12.1-amzn-0	0.11.1-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.10.0
Iceberg	1.1.0-amzn-0	0.14.1-amzn-0	0.14.1-amzn-0	0.14.0-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.6.0	2.1.0
JupyterHub	1.5.0	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.278	0.276	0.276	0.273
Spark	3.3.1	3.3.0	3.3.0	3.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.11.0	2.10.0	2.10.0	2.9.1
Tez	0.10.2	0.10.2	0.10.2	0.9.2

	emr-6.10.0	emr-6.9.1	emr-6.9.0	emr-6.8.1
Trino (PrestoSQL)	403	398	398	388
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.10.0 的信息。更改与 6.9.0 有关。有关发布时间表的信息，请参阅更改日志。

新特征

- Amazon EMR 6.10.0 支持 Apache Spark 3.3.1、Apache Spark RAPIDS 22.12.0、CUDA 11.8.0、Apache Hudi 0.12.2-amzn-0、Apache Iceberg 1.1.0-amzn-0、Trino 403 和 PrestoDB 0.278.1。
- Amazon EMR 6.10.0 包含原生 Trino-Hudi 连接器，可提供对 Hudi 表中数据的读取权限。您可以使用 `trino-cli --catalog hudi` 激活连接器，并使用 `trino-connector-hudi` 配置连接器以满足您的要求。与 Amazon EMR 的原生集成意味着您不再需要使用 `trino-connector-hive` 来查询 Hudi 表。有关新连接器支持的配置列表，请参阅 Trino 文档的 Hudi connector 页面。
- Amazon EMR 版本 6.10.0 及更高版本支持 Apache Zeppelin 与 Apache Flink 集成。参阅 在 Amazon EMR 中通过 Zeppelin 使用 Flink 作业 了解更多信息。

已知问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  要在 Amazon EMR 6.10.0 中解决此问题，您可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强功能和解决的问题

- Amazon EMR 6.10.0 消除了对适用于 Apache Spark 的 Amazon Redshift 集成的 `minimal-json.jar` 依赖，并自动将所需的 Spark-Redshift 相关 jar 添加到 Spark 的执行程序类路径中：`spark-redshift.jar`、`spark-avro.jar` 和 `RedshiftJDBC.jar`。
- 6.10.0 版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 6.10.0 版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。
- Amazon EMR 6.10.0 支持 EMRFS 用户映射的区域端点。
- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。
- 6.10.0 版本修复了当所有剩余的 Spark 执行程序都位于使用 YARN 资源管理器的停用主机上时，导致 Spark 作业停滞的问题。
- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.2.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.7.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.24.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.3.0	EMR S3 Select 连接器
emr-wal-cli	1.0.0	用于 emrwal 列表/删除的 cli。
emrfs	2.55.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
flink-client	1.16.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.16.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-2	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	3.3.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.3.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.15-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.15-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.15-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.4.15-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.15-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.15-amzn-0	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	3.1.3-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-3	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-3	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.12.2-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.12.2-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.12.2-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.12.2-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
iceberg	1.1.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.5.0	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.8.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	6.0.0-SNAPSHOT	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	6.0.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.278.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.278.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.278.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	403-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	403-amzn-0	用于执行查询的各个部分的服务。
trino-client	403-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.1-amzn-0	Spark 命令行客户端。
spark-history-server	3.3.1-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.1-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.1-amzn-0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
spark-rapids	22.12.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-1	tez YARN 应用程序和库。
tez-on-worker	0.10.2-amzn-1	用于 Worker 节点的 tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.10.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-def aults.conf 文件中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-lakeformat ion	更改 Presto 的 lakeforma tion.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hudi	更改 Trino 的 hudi.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-logback	更改 Ranger KMS 的 kms-logback.xml 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.

分类	描述	重新配置操作
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.10.0 的更改日志和发布说明

日期	事件	描述
2023-08-21	更新	添加了 Hadoop 3.3.3 引入的一个已知问题。
2023-07-26	更新	新的操作系统版本标签 `2.0.20230612.0` 和 `2.0.20230628.0` 。
2023-03-02	部署完成	Amazon EMR 6.10 已全面部署到所有支持的区域
2023-03-02	文档发布	Amazon EMR 6.10 发布说明首次发布
2023-02-27	首次发布	Amazon EMR 6.10 面向部分商业区域部署

Amazon EMR 版本 6.9.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.9.1	emr-6.9.0	emr-6.8.1	emr-6.8.0
Amazon SDK for Java	1.12.170	1.12.170	1.12.170	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.1.0	2.1.0	-	-
Flink	1.15.2	1.15.2	1.15.1	1.15.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.13	2.4.13	2.4.12	2.4.12
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.3.3	3.2.1	3.2.1
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.12.1-amzn-0	0.12.1-amzn-0	0.11.1-amzn-0	0.11.1-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.10.0
Iceberg	0.14.1-amzn-0	0.14.1-amzn-0	0.14.0-amzn-0	0.14.0-amzn-0
JupyterEnterpriseGateway	2.6.0	2.6.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1

	emr-6.9.1	emr-6.9.0	emr-6.8.1	emr-6.8.0
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.9.1
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.276	0.276	0.273	0.273
Spark	3.3.0	3.3.0	3.3.0	3.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.10.0	2.10.0	2.9.1	2.9.1
Tez	0.10.2	0.10.2	0.9.2	0.9.2
Trino (PrestoSQL)	398	398	388	388
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.1
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.10

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.9.1 的信息。更改与 6.9.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

更改、增强功能和解决的问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

在 Amazon EMR 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false` 以解决此问题。

虽然该修复解决了 YARN-9608 引入的问题，但由于启用了托管扩展的集群上的随机数据丢失，它可能会导致 Hive 作业失败。在此版本中，我们还通过设置 Hive `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-shuffle-data` 工作负载来降低这种风险。此配置在 Amazon EMR 版本 6.11.0 及更高版本中提供。

- 使用实例组配置的集群中的主节点失效转移后，指标收集器不会向控制面板发送任何指标。
- 此版本消除了在向指标收集器端点发出失败的 HTTP 请求时进行重试。
- 此版本包括一项更改，允许高可用性集群在重启后从故障状态中恢复。
- 此版本修复了用户创建的大型 UID 导致溢出异常的问题。
- 此版本修复了 Amazon EMR 重新配置过程中的超时问题。
- 此版本包含安全修复。
- 此版本修复了在 Spark 上使用 Amazon EMR 运行工作负载的集群可能会静默收到包含 `contains`、`startsWith`、`endsWith` 和 `like` 错误结果的问题。当您在 Amazon EMR Hive3 Metastore 服务器（HMS）中使用包含元数据的分区字段的表达式时，就会出现此问题。
- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。
- 当您使用 HDFS 作为暂存目录并启用了合并小文件且该表包含静态分区路径时，Hive 可能会丢失数据。
- 此版本修复了若在 ETL 作业结束时启用合并小文件（默认禁用）时 Hive 的性能问题。
- 此版本修复了没有用户定义函数（UDF）时在 Glue 端的节流问题。
- 此版本修复了在 YARN 停用时，在日志推送器能够将容器日志推送到 S3 之前，节点日志聚合服务会删除容器日志的问题。
- 此版本修复了使用 HBase 永久存储文件跟踪功能对压缩/存档文件的处理。
- 此版本修复了您在 `spark-defaults.conf` 中为 `spark.yarn.heterogeneousExecutors.enabled` 配置设置默认 `true` 值时影响 Spark 性能的问题。

- 此版本修复了 Reduce Task 无法读取随机数据的问题。该问题因内存损坏错误导致 Hive 查询失败。
- 此版本修复了在 HDFS NameNode（NN）服务在节点替换期间卡在安全模式下时导致节点置备器失败的问题。
- 此版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 此版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 此版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 此版本修复了通过从具有多个主节点的集群中复制一个主节点来创建边缘节点时可能出现的问题。复制的边缘节点可能会导致缩减操作的延迟，或者导致主节点的内存使用率过高。有关如何创建边缘节点以与 EMR 集群通信的更多信息，请参阅 GitHub `aws-samples` 存储库中的 Edge Node Creator。
- 此版本改进了 Amazon EMR 用于在重启后将 Amazon EBS 卷重新挂载到实例的自动化流程。
- 此版本修复了导致 Amazon EMR 向 Amazon CloudWatch 发布的 Hadoop 指标间歇性出现差距的问题。
- 此版本修复了 EMR 集群的一个问题，即由于磁盘过度使用而导致对包含集群节点排除列表的 YARN 配置文件的更新中断。不完整的更新阻碍了未来对集群的缩减操作。此版本可确保您的集群保持正常运行，并确保扩展操作按预期进行。
- 此版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 此版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.1.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.6.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.23.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.2.0	EMR S3 Select 连接器
emrfs	2.54.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.15.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
flink-jobmanager-config	1.15.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-0.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-0.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-0.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-0.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-0.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-0.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-0.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	3.3.3-amzn-0.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-0.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-0.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-0.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.13-amzn-0.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.13-amzn-0.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.13-amzn-0.1	HBase 命令行客户端。
hbase-rest-server	2.4.13-amzn-0.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.13-amzn-0.1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.13-amzn-0.1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-2.1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-2.1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	3.1.3-amzn-2.1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-2.1	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-2.1	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-2.1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-2.1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.12.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.12.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.12.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.12.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.14.1-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.7.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	6.0.0-SNAPSHOT	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	6.0.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.276-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.276-amzn-0	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.276-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	398-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	398-amzn-0	用于执行查询的各个部分的服务。
trino-client	398-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.0-amzn-1.1	Spark 命令行客户端。
spark-history-server	3.3.0-amzn-1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.0-amzn-1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.0-amzn-1.1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.08.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。

组件	版本	描述
tensorflow	2.10.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-0.1	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.9.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.

分类	描述	重新配置操作
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-lakeformation	更改 Presto 的 lakeformation.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.

分类	描述	重新配置操作
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.

分类	描述	重新配置操作
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.9.1 的更改日志和发布说明

日期	事件	描述
2023-08-30	更新发行说明	在发行说明中添加了几个与控制面板相关的修复
2023-08-21	文档发布	Amazon EMR 6.9.1 发布说明首次发布
2023-08-16	部署完成	Amazon EMR 6.9.1 已全面部署到所有支持的区域
2023-08-04	首次发布	Amazon EMR 6.9.1 首次面向部分商业区域部署

Amazon EMR 发行版 6.9.0

应用程序版本

此发行版支持以下应用程序：Delta、Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseG 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.9.0	emr-6.8.1	emr-6.8.0	emr-6.7.0
Amazon SDK for Java	1.12.170	1.12.170	1.12.170	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.15
Delta	2.1.0	-	-	-
Flink	1.15.2	1.15.1	1.15.1	1.14.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-6.9.0	emr-6.8.1	emr-6.8.0	emr-6.7.0
HBase	2.4.13	2.4.12	2.4.12	2.4.4
HCatalog	3.1.3	3.1.3	3.1.3	3.1.3
Hadoop	3.3.3	3.2.1	3.2.1	3.2.1
Hive	3.1.3	3.1.3	3.1.3	3.1.3
Hudi	0.12.1-amzn-0	0.11.1-amzn-0	0.11.1-amzn-0	0.11.0-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.10.0
Iceberg	0.14.1-amzn-0	0.14.0-amzn-0	0.14.0-amzn-0	0.13.1-amzn-0
JupyterEnterpriseGateway	2.6.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.9.1	1.8.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.276	0.273	0.273	0.272
Spark	3.3.0	3.3.0	3.3.0	3.2.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.10.0	2.9.1	2.9.1	2.4.1
Tez	0.10.2	0.9.2	0.9.2	0.9.2

	emr-6.9.0	emr-6.8.1	emr-6.8.0	emr-6.7.0
Trino (PrestoSQL)	398	388	388	378
Zeppelin	0.10.1	0.10.1	0.10.1	0.10.0
ZooKeeper	3.5.10	3.5.10	3.5.10	3.5.7

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.9.0 的信息。更改与 Amazon EMR 发行版 6.8.0 有关。有关发布时间表的信息，请参阅更改日志。

新功能

- Amazon EMR 发行版 6.9.0 支持 Apache Spark RAPIDS 22.08.0、Apache Hudi 0.12.1、Apache Iceberg 0.14.1、Trino 398 和 Tez 0.10.2。
- Amazon EMR 发行版 6.9.0 包括一个新的开源应用程序，Delta Lake 2.1.0。
- Amazon EMR 发行版 6.9.0 及更高版本包含适用于 Apache Spark 的 Amazon Redshift 集成。本地集成之前是一种开源工具，现在是 Spark 连接器，您可以将其用于构建 Apache Spark 应用程序，这些应用程序可在 Amazon Redshift 和 Amazon Redshift Serverless 中读取和写入数据。有关更多信息，请参阅将适用于 Apache Spark 的 Amazon Redshift 集成与 Amazon EMR 结合使用。
- Amazon EMR 发行版 6.9.0 增加了对在集群缩减期间将日志存档到 Amazon S3 的支持。之前，您只能在集群终止期间将日志文件存档到 Amazon S3。这项新功能可确保即使在节点终止后，集群上生成的日志文件仍保留在 Amazon S3 上。有关更多信息，请参阅配置集群日志记录和调试。
- 为了支持长时间运行的查询，Trino 现在包括容错执行机制。容错执行通过重试失败的查询或其组件任务来减少查询失败。有关更多信息，请参阅Trino 中的容错执行。
- 您可以在 Amazon EMR 上使用 Apache Flink 对 Apache Hive 表或任何 Flink 表源（例如 Iceberg、Kinesis 或 Kafka）的元数据进行统一的 `BATCH` 和 `STREAM` 处理。您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 指定 Amazon Glue 数据目录作为 Flink 的元存储。有关更多信息，请参阅在 Amazon EMR 中配置 Flink。
- 现在，您可以在 EC2 集群上的 Amazon EMR 上使用 Amazon SageMaker Studio，为 Apache Spark、Apache Hive 和 Presto 查询指定 Amazon Identity and Access Management（IAM）运行时角色和基于 Amazon Lake Formation 的访问控制。有关更多信息，请参阅为 Amazon EMR 步骤配置运行时角色。

已知问题

- 对于 Amazon EMR 发行版 6.9.0，Trino 不适用于为 Apache Ranger 启用的集群。如果您需要将 Trino 与 Ranger 结合使用，请联系 Amazon Web Services Support。
- 如果您使用适用于 Apache Spark 的 Amazon Redshift 集成，并且具有 Parquet 格式的时间、timetz、时间戳或 timestamptz（精度为微秒），连接器会将时间值舍入为最接近的毫秒值。解决方法是使用文本卸载格式 `unload_s3_format` 参数。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 从 Amazon SageMaker Studio 连接到 Amazon EMR 集群可能会间歇性失败，并显示 403 Forbidden（403 禁止访问）响应代码。如果在集群上设置 IAM 角色的时间超过 60 秒，就会发生此错误。解决方法是安装 Amazon EMR 补丁以启用重试，并将超时增加到至少 300 秒。启动集群时，按照以下步骤应用引导操作。

  1. 使用以下 Amazon S3 URI 下载引导脚本和 RPM 文件。

     ```
     s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/gcsc/replace-rpms.sh
     s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/gcsc/emr-secret-agent-1.18.0-SNAPSHOT20221121212949.noarch.rpm
     ```

  2. 将上一步中的文件上传到您自己的 Amazon S3 存储桶中。存储桶必须位于您计划启动集群的同一 Amazon Web Services 区域。
  3. 启动集群时，执行以下引导操作。将 *bootstrap_URI* 和 *RPM_URI* 替换为来自 Amazon S3 的相应 URI。

```
--bootstrap-actions "Path=bootstrap_URI,Args=[RPM_URI]"
```

- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，SecretAgent 和 RecordServer 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

- Apache Flink 提供 Native S3 FileSystem 和 Hadoop FileSystem 连接器，允许应用程序创建 FileSink 并将数据写入 Amazon S3。此 FileSink 失败并出现以下两个异常之一。

  ```
  java.lang.UnsupportedOperationException: Recoverable writers on Hadoop are only
   supported for HDFS
  ```

  ```
  Caused by: java.lang.NoSuchMethodError:
   org.apache.hadoop.io.retry.RetryPolicies.retryOtherThanRemoteAndSaslException(Lorg/
  apache/hadoop/io/retry/RetryPolicy;Ljava/util/Map;)Lorg/apache/hadoop/io/retry/
  RetryPolicy;
                                          at
   org.apache.hadoop.yarn.client.RMProxy.createRetryPolicy(RMProxy.java:302) ~[hadoop-
  yarn-common-3.3.3-amzn-0.jar:?]
  ```

  解决方法是安装 Amazon EMR 补丁，该补丁可以修复 Flink 中的上述问题。要在启动集群时应用引导操作，请完成以下步骤。

  1. 将 flink-rpm 下载到 Amazon S3 存储桶中。您的 RPM 路径是 s3://*DOC-EXAMPLE-BUCKET*/rpms/flink/。
  2. 使用以下 URI 从 Amazon S3 下载引导脚本和 RPM 文件。将 *regionName* 替换为您计划启动集群的 Amazon Web Services 区域。

```
s3://emr-data-access-control-regionName/customer-bootstrap-actions/gcsc/replace-rpms.sh
```

3. Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。在 Amazon EMR 6.8.0 和 6.9.0 中，无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

   在 Amazon EMR 6.10.0 中，有一个解决此问题的方法，可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强和解决的问题

- 对于 Amazon EMR 发行版 6.9.0 及更高版本，Amazon EMR 安装的所有使用 Log4j 库的组件都使用 Log4j 版本 2.17.1 或更高版本。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。Amazon EMR 发行版 6.9.0 修复了此问题。
- 在使用 Spark SQL 读取数据时，Amazon EMR 6.9.0 添加对基于 Lake Formation 的访问控制及 Apache Hudi 的有限支持。支持针对使用 Spark SQL 的 SELECT 查询，并且仅限于列级访问控制。有关更多信息，请参阅 Hudi 和 Lake Formation。
- 当您使用 Amazon EMR 6.9.0 创建启用了节点标签的 Hadoop 集群时，YARN 指标 API 会返回所有分区的聚合信息，而不是默认分区。有关更多信息，请参阅 YARN-11414。
- 在 Amazon EMR 6.9.0 版本中，我们已将 Trino 更新到使用 Java 17 的 398 版本。之前支持的 Amazon EMR 6.8.0 Trino 版本是在 Java 11 上运行的 Trino 388。有关此变更的更多信息，请参阅 Trino 博客上的 Trino updates to Java 17。
- 此版本修复了 Apache BigTop 和 EC2 集群启动序列上的 Amazon EMR 之间的时间序列不匹配的问题。当系统尝试同时执行两个或多个操作而不是按正确的顺序执行它们时，就会发生这种计时序列不匹配。因此，某些集群配置会遇到实例启动超时和较慢的集群启动时间。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (6.8.*1*) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 1 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
delta	2.1.0	Delta lake 是一种适用于超大型分析数据集的开放表格式。
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.6.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.23.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.2.0	EMR S3 Select 连接器
emrfs	2.54.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.15.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
flink-jobmanager-config	1.15.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.3.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.3.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.3.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.3.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.3.3-amzn-1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.3.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.3.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	3.3.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.3.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.3.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.3.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.13-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.13-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.13-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.4.13-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.13-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.13-amzn-0	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	3.1.3-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-2	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-2	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.12.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.12.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.12.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.12.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.14.1-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.7.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	6.0.0-SNAPSHOT	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	6.0.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.276-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.276-amzn-0	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.276-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	398-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	398-amzn-0	用于执行查询的各个部分的服务。
trino-client	398-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.0-amzn-1	Spark 命令行客户端。
spark-history-server	3.3.0-amzn-1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.0-amzn-1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.0-amzn-1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.08.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。

组件	版本	描述
tensorflow	2.10.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.10.2-amzn-0	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.9.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.

分类	描述	重新配置操作
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
delta-defaults	更改 Delta 的 delta-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j2	更改 Livy log4j2.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-lakeformation	更改 Presto 的 lakeformation.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-delta	更改 Trino 的 delta.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-exchange-manager	更改 Trino 的 exchange-manager.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.

分类	描述	重新配置操作
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.

分类	描述	重新配置操作
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.9.0 的更改日志和发布说明

日期	事件	描述
2023-08-30	更新发行说明	添加了对定时序列不匹配问题的修复
2023-08-21	更新发行说明	在 Hadoop 3.3.3 中添加了一个已知问题。
2023-07-26	更新	新的操作系统版本标签 `2.0.20230612.0` 和 `2.0.20230628.0` 。
2022-12-13	发布说明已更新	增加了 SageMaker 运行时功能和已知问题

日期	事件	描述
2022-11-29	发布说明和文档已更新	添加了适用于 Apache Spark 的 Amazon Redshift 集成功能
2022-11-23	发布说明已更新	Log4j 条目已删除
2022-11-18	部署完成	Amazon EMR 6.9 已全面部署到所有支持的区域
2022-11-18	文档发布	Amazon EMR 6.9 发布说明首次发布
2022-11-14	首次发布	Amazon EMR 6.9 面向部分商业区域部署

Amazon EMR 版本 6.8.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.8.1	emr-6.8.0	emr-6.7.0	emr-6.6.0
Amazon SDK for Java	1.12.170	1.12.170	1.12.170	1.12.170
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.15	2.12.10
Delta	-	-	-	-
Flink	1.15.1	1.15.1	1.14.2	1.14.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.12	2.4.12	2.4.4	2.4.4
HCatalog	3.1.3	3.1.3	3.1.3	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.3	3.1.3	3.1.3	3.1.2
Hudi	0.11.1-amzn-0	0.11.1-amzn-0	0.11.0-amzn-0	0.10.1-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.10.0
Iceberg	0.14.0-amzn-0	0.14.0-amzn-0	0.13.1-amzn-0	0.13.1
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.9.1	1.8.0	1.8.0
Mahout	-	-	-	-

	emr-6.8.1	emr-6.8.0	emr-6.7.0	emr-6.6.0
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.273	0.273	0.272	0.267
Spark	3.3.0	3.3.0	3.2.1	3.2.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.9.1	2.9.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	388	388	378	367
Zeppelin	0.10.1	0.10.1	0.10.0	0.10.0
ZooKeeper	3.5.10	3.5.10	3.5.7	3.5.7

发布说明

以下发布说明包括有关 Amazon EMR 版本 6.8.1 的信息。更改与 6.8.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

更改、增强功能和解决的问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  在 Amazon EMR 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false` 以解决此问题。

虽然该修复解决了 YARN-9608 引入的问题，但由于启用了托管扩展的集群上的随机数据丢失，它可能会导致 Hive 作业失败。在此版本中，我们还通过设置 Hive `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-shuffle-data` 工作负载来降低这种风险。此配置在 Amazon EMR 版本 6.11.0 及更高版本中提供。

- 使用实例组配置的集群中的主节点失效转移后，指标收集器不会向控制面板发送任何指标。
- 此版本消除了在向指标收集器端点发出失败的 HTTP 请求时进行重试。
- 此版本包括一项更改，允许高可用性集群在重启后从故障状态中恢复。
- 此版本修复了用户创建的大型 UID 导致溢出异常的问题。
- 此版本修复了 Amazon EMR 重新配置过程中的超时问题。
- 此版本可防止出现重新配置失败可能会中断其他不相关的进程的问题。
- 此版本包含安全修复。
- 此版本修复了在 Spark 上使用 Amazon EMR 运行工作负载的集群可能会静默收到包含 `contains`、`startsWith`、`endsWith` 和 `like` 错误结果的问题。当您在 Amazon EMR Hive3 Metastore 服务器（HMS）中使用包含元数据的分区字段的表达式时，就会出现此问题。
- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。
- 当您使用 HDFS 作为暂存目录并启用了合并小文件且该表包含静态分区路径时，Hive 可能会丢失数据。
- 此版本修复了若在 ETL 作业结束时启用合并小文件（默认禁用）时 Hive 的性能问题。
- 此版本修复了没有用户定义函数（UDF）时在 Glue 端的节流问题。
- 此版本修复了在 YARN 停用时，在日志推送器能够将容器日志推送到 S3 之前，节点日志聚合服务会删除容器日志的问题。
- 此版本修复了使用 HBase 永久存储文件跟踪功能对压缩/存档文件的处理。
- 此版本修复了您在 `spark-defaults.conf` 中为 `spark.yarn.heterogeneousExecutors.enabled` 配置设置默认 `true` 值时影响 Spark 性能的问题。
- 此版本修复了 Reduce Task 无法读取随机数据的问题。该问题因内存损坏错误导致 Hive 查询失败。

- 此版本修复了在 HDFS NameNode（NN）服务在节点替换期间卡在安全模式下时导致节点置备器失败的问题。
- 此版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 此版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 此版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 此版本修复了通过从具有多个主节点的集群中复制一个主节点来创建边缘节点时可能出现的问题。复制的边缘节点可能会导致缩减操作的延迟，或者导致主节点的内存使用率过高。有关如何创建边缘节点以与 EMR 集群通信的更多信息，请参阅 GitHub `aws-samples` 存储库中的 Edge Node Creator。
- 此版本改进了 Amazon EMR 用于在重启后将 Amazon EBS 卷重新挂载到实例的自动化流程。
- 此版本修复了导致 Amazon EMR 向 Amazon CloudWatch 发布的 Hadoop 指标间歇性出现差距的问题。
- 此版本修复了 EMR 集群的一个问题，即由于磁盘过度使用而导致对包含集群节点排除列表的 YARN 配置文件的更新中断。不完整的更新阻碍了未来对集群的缩减操作。此版本可确保您的集群保持正常运行，并确保扩展操作按预期进行。
- 此版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 此版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。
- 此版本增加了对在集群缩减期间将日志存档到 Amazon S3 的支持。之前，您只能在集群终止期间将日志文件存档到 Amazon S3。这项新功能可确保即使在节点终止后，集群上生成的日志文件仍保留在 Amazon S3 上。有关更多信息，请参阅配置集群日志记录和调试。
- 此版本修复了引导操作的 Amazon S3 URI 以端口号结尾时出现的问题，例如：`a.b.c.d:4345`。Amazon EMR 错误地解析了这些 URI，因此任何相关的引导操作都将失败。
- 此版本修复了 Apache BigTop 和 EC2 集群启动序列上的 Amazon EMR 之间的时间序列不匹配的问题。当系统尝试同时执行两个或多个操作而不是按正确的顺序执行它们时，就会发生这种计时序列不匹配。因此，某些集群配置会遇到实例启动超时和较慢的集群启动时间。

- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.22.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.53.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.15.1	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
flink-jobmanager-config	1.15.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-8.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-8.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-8.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-8.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-8.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-8.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-8.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	3.2.1-amzn-8.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-8.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-8.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-8.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.12-amzn-0.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.12-amzn-0.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.12-amzn-0.1	HBase 命令行客户端。
hbase-rest-server	2.4.12-amzn-0.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.12-amzn-0.1	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.12-amzn-0.1	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-1.1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-1.1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	3.1.3-amzn-1.1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-1.1	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-1.1	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-1.1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-1.1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.11.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.11.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.11.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.11.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.14.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.7.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.273.3-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.273.3-amzn-0	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.273.3-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	388-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	388-amzn-0	用于执行查询的各个部分的服务。
trino-client	388-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.0-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.3.0-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.0-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.0-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.06.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。

组件	版本	描述
tensorflow	2.9.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.8.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.

分类	描述	重新配置操作
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.

分类	描述	重新配置操作
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.

分类	描述	重新配置操作
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.

分类	描述	重新配置操作
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.8.1 的更改日志和发布说明

日期	事件	描述
2023-08-30	更新发行说明	在发行说明中添加了几个与控制面板相关的修复
2023-08-21	文档发布	Amazon EMR 6.8.1 发布说明首次发布
2023-08-16	部署完成	Amazon EMR 6.8.1 已全面部署到所有支持的区域
2023-08-04	首次发布	Amazon EMR 6.8.1 首次面向部分商业区域部署

Amazon EMR 发行版 6.8.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本

- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.8.0	emr-6.7.0	emr-6.6.0	emr-6.5.0
Amazon SDK for Java	1.12.170	1.12.170	1.12.170	1.12.31
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.15.1	1.14.2	1.14.2	1.14.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.12	2.4.4	2.4.4	2.4.4
HCatalog	3.1.3	3.1.3	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.3	3.1.3	3.1.2	3.1.2
Hudi	0.11.1-amzn-0	0.11.0-amzn-0	0.10.1-amzn-0	0.9.0-amzn-1
Hue	4.10.0	4.10.0	4.10.0	4.9.0
Iceberg	0.14.0-amzn-0	0.13.1-amzn-0	0.13.1	0.12.0
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.9.1	1.8.0	1.8.0	1.8.0

	emr-6.8.0	emr-6.7.0	emr-6.6.0	emr-6.5.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.273	0.272	0.267	0.261
Spark	3.3.0	3.2.1	3.2.0	3.1.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.9.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	388	378	367	360
Zeppelin	0.10.1	0.10.0	0.10.0	0.10.0
ZooKeeper	3.5.10	3.5.7	3.5.7	3.5.7

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.8.0 的信息。更改与 6.7.0 有关。

新功能

- Amazon EMR 步骤功能现支持 Livy 端点和 JDBC/ODBC 客户端。有关更多信息，请参阅为 Amazon EMR 步骤配置运行时角色。
- Amazon EMR 发行版 6.8.0 随附 Apache HBase 发行版 2.4.12。借助此 HBase 发行版，您可以对 HBase 表进行存档和删除。Amazon S3 存档过程会将所有表文件重命名为存档目录。这一过程成本高昂且时间较长。现在，您可以跳过存档过程，快速删除大型表。有关更多信息，请参阅使用 HBase shell。

已知问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。在 Amazon EMR 6.8.0 和 6.9.0 中，无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  在 Amazon EMR 6.10.0 中，有一个解决此问题的方法，可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强和解决的问题

- 当 Amazon EMR 发行版 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark Shell 读取 Apache Phoenix 表时，Amazon EMR 会生成 `NoSuchMethodError`。Amazon EMR 发行版 6.8.0 修复了此问题。
- Amazon EMR 发行版 6.8.0 随附 Apache Hudi 0.11.1；但是，Amazon EMR 6.8.0 集群也与 Hudi 0.12.0 中的开源 `hudi-spark3.3-bundle_2.12` 兼容。
- Amazon EMR 发行版 6.8.0 随附 Apache Spar 3.3.0。此 Spark 发行版使用 Apache Log4j 2 和 `log4j2.properties` 文件，在 Spark 进程中配置 Log4j。如果您在集群中使用 Spark 或使用自定义配置参数创建 EMR 集群，并且希望升级到 Amazon EMR 发行版 6.8.0，则必须迁移到新的 `spark-log4j2` 配置分类和 Apache Log4j 2 的密钥格式。有关更多信息，请参阅从 Apache Log4j 1.x 迁移到 Log4j 2.x。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）、

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 9 月 6 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

已知问题

- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将 `spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

有关发布时间表的更多信息，请参阅更改日志。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.2	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.7.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.22.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.53.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.15.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.15.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-8	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-8	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-8	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-8	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-8	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-8	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-8	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-8	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	3.2.1-amzn-8	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-8	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-8	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.12-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.12-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.12-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.4.12-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.12-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.12-amzn-0	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-1	Hive 命令行客户端。

组件	版本	描述
hive-hbase	3.1.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.11.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.11.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.11.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.11.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.14.0-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.9.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.7.0	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.273.3-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.273.3-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.273.3-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。

组件	版本	描述
trino-coordinator	388-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	388-amzn-0	用于执行查询的各个部分的服务。
trino-client	388-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.3.0-amzn-0	Spark 命令行客户端。
spark-history-server	3.3.0-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.3.0-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.3.0-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.06.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.9.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.8.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode,

分类	描述	重新配置操作
		Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.

分类	描述	重新配置操作
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.

分类	描述	重新配置操作
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhu b_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.js on 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.

分类	描述	重新配置操作
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.p roperties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properti es 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.pr operties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.p roperties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresq l.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j2	更改 Spark 的 log4j2.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 6.8.0 的更改日志和发布说明

日期	事件	描述
2023-08-21	更新	在 Hadoop 3.3.3 中添加了一个已知问题。
2023-07-26	更新	新的操作系统版本标签 `2.0.20230612.0` 和 `2.0.20230628.0` 。
2022-09-06	部署完成	Amazon EMR 6.8 已全面部署到所有支持的区域
2022-09-06	初次发布	Amazon EMR 6.8 发布说明首次发布
2022-08-31	首次发布	Amazon EMR 6.8 面向部分商业区域发布

Amazon EMR 发行版 6.7.0

- 应用程序版本
- 发布说明

- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.7.0	emr-6.6.0	emr-6.5.0	emr-6.4.0
Amazon SDK for Java	1.12.170	1.12.170	1.12.31	1.12.31
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.14.2	1.14.2	1.14.0	1.13.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.4	2.4.4	2.4.4	2.4.4
HCatalog	3.1.3	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1

	emr-6.7.0	emr-6.6.0	emr-6.5.0	emr-6.4.0
Hive	3.1.3	3.1.2	3.1.2	3.1.2
Hudi	0.11.0-amzn-0	0.10.1-amzn-0	0.9.0-amzn-1	0.8.0-amzn-0
Hue	4.10.0	4.10.0	4.9.0	4.9.0
Iceberg	0.13.1-amzn-0	0.13.1	0.12.0	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.8.0	1.8.0	1.8.0	1.8.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.1.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.272	0.267	0.261	0.254.1
Spark	3.2.1	3.2.0	3.1.2	3.1.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	378	367	360	359
Zeppelin	0.10.0	0.10.0	0.10.0	0.9.0

	emr-6.7.0	emr-6.6.0	emr-6.5.0	emr-6.4.0
ZooKeeper	3.5.7	3.5.7	3.5.7	3.5.7

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.7.0 的信息。更改与 6.6.0 有关。

首次发布日期：2022 年 7 月 15 日

新功能

- Amazon EMR 现在支持 Apache Spark 3.2.1、Apache Hive 3.1.3、HUDI 0.11、PrestoDB 0.272 和 Trino 0.378。
- 通过 EMR 步骤（Spark、Hive）支持 EC2 集群上的 Amazon EMR 基于 IAM 角色和 Lake Formation 的访问控制。
- 在启用 Apache Ranger 的集群上支持 Apache Spark 数据定义语句。现在，这包括支持 Trino 应用程序在启用 Apache Ranger 的集群上读取和写入 Apache Hive 元数据。有关更多信息，请参阅在 Amazon EMR 上使用 Trino 和 Apache Ranger 启用联合治理。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 10 月 7 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2022 719.0	4.14.287	2022 年 8 月 10 日	us-west-1 , eu-west-3 , eu-north-1 , ap-south-1 , me-south-1

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 606.1	4.14.281	2022 年 7 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将

`spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

```
--bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

```
--bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
```

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.6.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.22.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.52.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
flink-jobmanager-config	1.14.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-7	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-7	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-7	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-7	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-7	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-7	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-7	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	3.2.1-amzn-7	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-7	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-7	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-7	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.4-amzn-3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.4-amzn-3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.4-amzn-3	HBase 命令行客户端。
hbase-rest-server	2.4.4-amzn-3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.4-amzn-3	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.4-amzn-3	Apache HBase 集群的修复工具。
hcatalog-client	3.1.3-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.3-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	3.1.3-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.3-amzn-0	Hive 命令行客户端。
hive-hbase	3.1.3-amzn-0	Hive-hbase 客户端。
hive-metastore-server	3.1.3-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.3-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.11.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.11.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.11.0-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.11.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.13.1-amzn-0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.0.194	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.272-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.272-amzn-0	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.272-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	378-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	378-amzn-0	用于执行查询的各个部分的服务。
trino-client	378-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.2.1-amzn-0	Spark 命令行客户端。
spark-history-server	3.2.1-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.2.1-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.2.1-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.02.0-amzn-1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。

组件	版本	描述
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.7	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.7	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.7.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.

分类	描述	重新配置操作
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.

分类	描述	重新配置操作
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-iceberg	更改 Trino 的 iceberg.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.

分类	描述	重新配置操作
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.

分类	描述	重新配置操作
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.6.0

- 应用程序版本
- 发布说明

- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.6.0	emr-6.5.0	emr-6.4.0	emr-6.3.1
Amazon SDK for Java	1.12.170	1.12.31	1.12.31	1.11.977
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.14.2	1.14.0	1.13.1	1.12.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.4	2.4.4	2.4.4	2.2.6
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1

	emr-6.6.0	emr-6.5.0	emr-6.4.0	emr-6.3.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.10.1-amzn-0	0.9.0-amzn-1	0.8.0-amzn-0	0.7.0-amzn-0
Hue	4.10.0	4.9.0	4.9.0	4.9.0
Iceberg	0.13.1	0.12.0	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.2.2
Livy	0.7.1	0.7.1	0.7.1	0.7.0
MXNet	1.8.0	1.8.0	1.8.0	1.7.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.1.2	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.267	0.261	0.254.1	0.245.1
Spark	3.2.0	3.1.2	3.1.2	3.1.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	367	360	359	350
Zeppelin	0.10.0	0.10.0	0.9.0	0.9.0

	emr-6.6.0	emr-6.5.0	emr-6.4.0	emr-6.3.1
ZooKeeper	3.5.7	3.5.7	3.5.7	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.6.0 的信息。更改与 6.5.0 有关。

首次发布日期：2022 年 5 月 9 日

文档更新日期：2022 年 6 月 15 日

新功能

- Amazon EMR 6.6 现在支持 Apache Spark 3.2、Apache Spark RAPIDS 22.02、CUDA 11、Apache Hudi 0.10.1、Apache Iceberg 0.13、Trino 0.367 和 PrestoDB 0.267。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 10 月 7 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2022 805.0	4.14.287	2022 年 8 月 30 日	us-west-1

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 719.0	4.14.287	2022 年 8 月 10 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 426.0	4.14.281	2022 年 6 月 10 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 406.1	4.14.275	2022 年 5 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

- 在 Amazon EMR 6.6 及更高版本中，使用 log4J 1.x 和 log4J 2.x 的应用程序将分别升级为使用 log4J 1.2.17（或更高版本）和 log4J 2.17.1（或更高版本），并且不需要使用提供的引导操作来缓解 CVE 问题。
- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管横向缩减和 Spark 编程指南。

- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 `spark.yarn.heterogeneousExecutors.enabled` 配置参数。

更改、增强和解决的问题

- 对于使用 EMR 默认 AMI 选项且仅安装常用应用程序（如 Apache Hadoop、Apache Spark 和 Apache Hive）的集群，Amazon EMR 平均可将启动时间缩短 80 秒。

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将 `spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。
- 在 Trino 长时间运行的集群上，Amazon EMR 6.6.0 在 Trino jvm.config 中启用了垃圾回收日志记录参数，以便从垃圾回收日志中获取更好的见解。此更改会将许多垃圾回收日志附加到 launcher.log (/var/log/trino/launcher.log) 文件。如果您在 Amazon EMR 6.6.0 中运行 Trino 集群，由于附加的日志，可能会在集群运行几天后出现节点磁盘空间不足的情况。

  这一问题的解决办法是在为 Amazon EMR 6.6.0 创建或克隆集群时，将以下脚本作为引导操作运行以禁用 jvm.config 中的垃圾回收日志记录参数。

```
#!/bin/bash
  set -ex
  PRESTO_PUPPET_DIR='/var/aws/emr/bigtop-deploy/puppet/modules/trino'
  sudo bash -c "sed -i '/-Xlog/d' ${PRESTO_PUPPET_DIR}/templates/jvm.config"
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。

- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 / 字符 (U+002F)。例如，在 s3://bucket/table/p=a b 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：!"#$%&'()*+,-。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，SecretAgent 和 RecordServer 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.5.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.20.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.50.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.14.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-6	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-6	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	3.2.1-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.4-amzn-2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.4-amzn-2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.4-amzn-2	HBase 命令行客户端。
hbase-rest-server	2.4.4-amzn-2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.4-amzn-2	用于向 HBase 提供 Thrift 终端节点的服务。
hbase-operator-tools	2.4.4-amzn-2	Apache HBase 集群的修复工具。
hcatalog-client	3.1.2-amzn-7	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-7	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-7	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-7	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-7	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	3.1.2-amzn-7	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-7	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.10.1-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.10.1-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.10.1-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.10.1-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.13.1	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器

组件	版本	描述
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	11.0.194	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-connectors	5.1.2	Apache Phoenix-Connectors for Spark-3
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.267-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.267-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.267-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	367-amzn-0	用于在 trino-worker 之中接受查询并管理查询的服务。

组件	版本	描述
trino-worker	367-amzn-0	用于执行查询的各个部分的服务。
trino-client	367-amzn-0	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.2.0-amzn-0	Spark 命令行客户端。
spark-history-server	3.2.0-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.2.0-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.2.0-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	22.02.0-amzn-0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.7	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.7	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.6.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the

分类	描述	重新配置操作
		Hadoop YARN services ResourceManager, NodeManager, ProxyServ er, and TimelineServer. Ad ditionally restarts Hadoop KMS, Ranger KMS, HiveServe r2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServ er, and TimelineServer. Ad ditionally restarts HBaseRegi onserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore , Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.

分类	描述	重新配置操作
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.

分类	描述	重新配置操作
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-d efaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhu b_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.js on 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.

分类	描述	重新配置操作
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-iceberg	更改 Trino 的 iceberg.p roperties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properti es 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.pr operties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.p roperties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresq l.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.5.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.5.0	emr-6.4.0	emr-6.3.1	emr-6.3.0
Amazon SDK for Java	1.12.31	1.12.31	1.11.977	1.11.977
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.14.0	1.13.1	1.12.1	1.12.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.4	2.4.4	2.2.6	2.2.6
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.9.0-amzn-1	0.8.0-amzn-0	0.7.0-amzn-0	0.7.0-amzn-0
Hue	4.9.0	4.9.0	4.9.0	4.9.0
Iceberg	0.12.0	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.2.2	1.2.2
Livy	0.7.1	0.7.1	0.7.0	0.7.0
MXNet	1.8.0	1.8.0	1.7.0	1.7.0
Mahout	-	-	-	-

	emr-6.5.0	emr-6.4.0	emr-6.3.1	emr-6.3.0
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	5.1.2	5.1.2	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.261	0.254.1	0.245.1	0.245.1
Spark	3.1.2	3.1.2	3.1.1	3.1.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	360	359	350	350
Zeppelin	0.10.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.5.7	3.5.7	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.5.0 的信息。更改与 6.4.0 有关。

首次发布日期：2022 年 1 月 20 日

发布更新日期：2022 年 3 月 21 日

新功能

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管横向缩减和 Spark 编程指南。
- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 `spark.yarn.heterogeneousExecutors.enabled` 配置参数。

- 支持 Apache Iceberg 开放表格式，用于大型分析数据集。
- 支持 ranger-trino-plugin 2.0.1-amzn-1
- 支持 toree 0.5.0

更改、增强和解决的问题

- Amazon EMR 6.5 发行版现在支持 Apache Iceberg 0.12.0，并通过适用于 Apache Spark 的 Amazon EMR 运行时、适用于 Presto 的 Amazon EMR 运行时和适用于 Apache Hive 的 Amazon EMR 运行时提供了运行时改进。
- Apache Iceberg 是 Amazon S3 中适用于大型数据集的开放表格式，可提供快速的大型表查询性能、原子提交、并发写入和 SQL 兼容表演进等功能。借助 EMR 6.5，您可以将 Apache Spark 3.1.2 与 Iceberg 表格式结合使用。
- Apache Hudi 0.9 增加了对 Spark SQL DDL 和 DML 的支持。从而让您可以仅使用 SQL 语句创建 upsert Hudi 表。Apache Hudi 0.9 还包括查询端和写入器端的性能改进。
- 适用于 Apache Hive 的 Amazon EMR 运行时取消了暂存操作期间的重命名操作，从而提高了 Apache Hive 在 Amazon S3 上的性能，此外还提高了用于修复表的元数据仓检查（MSCK）命令的性能。

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 高可用性（HA）的 Hbase 捆绑集群无法使用默认卷大小和实例类型进行预置。此问题的变通解决方法是增加根卷大小。
- 要将 Spark 操作与 Apache Oozie 一起使用，必须将以下配置添加到 Oozie `workflow.xml` 文件中。否则，Oozie 启动的 Spark 执行器的类路径中将丢失几个诸如 Hadoop 和 EMRFS 之类的关键库。

```
<spark-opts>--conf spark.yarn.populateHadoopClasspath=true</spark-opts>
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。

- 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-notebook-env	1.4.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.19.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.48.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.14.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-5	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	3.2.1-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-5	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.4-amzn-1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.4-amzn-1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.4-amzn-1	HBase 命令行客户端。
hbase-rest-server	2.4.4-amzn-1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	2.4.4-amzn-1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-6	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-6	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-6	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-6	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-6	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-6	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-6	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.9.0-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.9.0-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.9.0-amzn-1	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.9.0-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。

组件	版本	描述
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
iceberg	0.12.0	Apache Iceberg 是一种适用于超大型分析数据集的开放表格式。
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.261-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.261-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.261-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
trino-coordinator	360	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	360	用于执行查询的各个部分的服务。
trino-client	360	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.1.2-amzn-1	Spark 命令行客户端。
spark-history-server	3.1.2-amzn-1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.1.2-amzn-1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.1.2-amzn-1	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
spark-rapids	0.4.1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.7	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.7	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.5.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.

分类	描述	重新配置操作
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defaults.conf 文件中的值。	Not available.
iceberg-defaults	更改 Iceberg 的 iceberg-defaults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.

分类	描述	重新配置操作
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.
trino-connector-blackhole	更改 Trino 的 blackhole.properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra.properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)

分类	描述	重新配置操作
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.

分类	描述	重新配置操作
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.

分类	描述	重新配置操作
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.4.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.4.0	emr-6.3.1	emr-6.3.0	emr-6.2.1
Amazon SDK for Java	1.12.31	1.11.977	1.11.977	1.11.880
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.13.1	1.12.1	1.12.1	1.11.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.4.4	2.2.6	2.2.6	2.2.6-amzn-0
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.8.0-amzn-0	0.7.0-amzn-0	0.7.0-amzn-0	0.6.0-amzn-1

	emr-6.4.0	emr-6.3.1	emr-6.3.0	emr-6.2.1
Hue	4.9.0	4.9.0	4.9.0	4.8.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.2.2	1.2.2	1.1.0
Livy	0.7.1	0.7.0	0.7.0	0.7.0
MXNet	1.8.0	1.7.0	1.7.0	1.7.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.1	5.2.0
Phoenix	5.1.2	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.254.1	0.245.1	0.245.1	0.238.3
Spark	3.1.2	3.1.1	3.1.1	3.0.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.3.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	359	350	350	343
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.5.7	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.4.0 的信息。更改与 6.3.0 有关。

首次发布日期：2021 年 9 月 20 日

发布更新日期：2022 年 3 月 21 日

支持的应用程序

- Amazon SDK for Java 1.12.31
- CloudWatch Sink 2.2.0
- DynamoDB 连接器 4.16.0
- EMRFS 2.47.0
- Amazon EMR Goodies 3.2.0
- Amazon EMR Kinesis 连接器 3.5.0
- Amazon EMR 记录服务器 2.1.0
- Amazon EMR Scripts 2.5.0
- Flink 1.13.1
- Ganglia 3.7.2
- Amazon Glue Hive Metastore Client 3.3.0
- Hadoop 3.2.1-amzn-4
- HBase 2.4.4-amzn-0
- HBase-operator-tools 1.1.0
- HCatalog 3.1.2-amzn-5
- Hive 3.1.2-amzn-5
- Hudi 0.8.0-amzn-0
- Hue 4.9.0
- Java JDK Corretto-8.302.08.1（内部 1.8.0_302-b08）
- JupyterHub 1.4.1
- Livy 0.7.1-incubating
- MXNet 1.8.0
- Oozie 5.2.1

- Phoenix 5.1.2
- Pig 0.17.0
- Presto 0.254.1-amzn-0
- Trino 359
- Apache Ranger KMS（多主节点透明加密）版本 2.0.0
- ranger-plugins 2.0.1-amzn-0
- ranger-s3-plugin 1.2.0
- SageMaker Spark SDK 1.4.1
- Scala 2.12.10（OpenJDK 64 位服务器 VM，Java 1.8.0_282）
- Spark 3.1.2-amzn-0
- spark-rapids 0.4.1
- Sqoop 1.4.7
- TensorFlow 2.4.1
- tez 0.9.2
- Zeppelin 0.9.0
- Zookeeper 3.5.7
- 连接器和驱动程序：DynamoDB 连接器 4.16.0

新特征

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管扩展 和 Spark 编程指南。
- 在 Apache Ranger 启用的 Amazon EMR 集群上，您可以使用 Apache Spark SQL 将数据插入到 Apache Hive 元数据存储表中或使用`INSERT INTO`、`INSERT OVERWRITE`和`ALTER TABLE`更新 Apache Hive 元数据存储表。将 ALTER TABLE 与 Spark SQL 结合使用时，分区位置必须是表位置的子目录。如果某个分区的分区位置与表位置不同，Amazon EMR 目前不支持将数据插入该分区。
- PrestoSQL 已重命名为 Trino。
- Hive：在获取 LIMIT 子句中提到的记录数目后，通过立即停止查询执行可加快使用 LIMIT 子句执行简单 SELECT 查询的速度。简单 SELECT 查询是没有 GROUP BY/ORDER BY 子句的查询或没有减速阶段的查询。例如，`SELECT * from <TABLE> WHERE <Condition> LIMIT <Number>`。

Hudi 并发控制

- Hudi 目前支持乐观并发控制 (OCC)，它可以与 UPSERT 和 INSERT 等写入操作一起利用，以允许从多个写入器更改为同一 Hudi 表。这是文件级 OCC，因此任何两个提交（或写入器）可以写入同一表内，前提是它们的更改不冲突。有关更多信息，请参阅 Hadi 并发性控制。
- Amazon EMR 集群安装了 Zookeeper，可以利用它作为 OCC 的锁提供商。为了更便捷地使用此功能，Amazon EMR 集群预先配置了以下属性：

```
hoodie.write.lock.provider=org.apache.hudi.client.transaction.lock.ZookeeperBasedLockProvider
hoodie.write.lock.zookeeper.url=<EMR Zookeeper URL>
hoodie.write.lock.zookeeper.port=<EMR Zookeeper Port>
hoodie.write.lock.zookeeper.base_path=/hudi
```

要启用 OCC，您需要使用 Hudi 任务选项或使用 Amazon EMR 配置 API 在集群级别配置以下属性：

```
hoodie.write.concurrency.mode=optimistic_concurrency_control
hoodie.cleaner.policy.failed.writes=LAZY (Performs cleaning of failed writes lazily
 instead of inline with every write)
hoodie.write.lock.zookeeper.lock_key=<Key to uniquely identify the Hudi table> (Table
 Name is a good option)
```

Hudi 监控：Amazon CloudWatch 集成，用于报告 Hudi 指标

- Amazon EMR 支持向 Amazon CloudWatch 发布 Hudi 指标。通过设置以下所需配置来启用：

```
hoodie.metrics.on=true
hoodie.metrics.reporter.type=CLOUDWATCH
```

- 以下是您可以更改的可选 Hudi 配置：

设置	描述	Value
hoodie.metrics.cloudwatch.report.period.seconds	向 Amazon CloudWatch 报告指标的频率（以秒为单位）	默认值为 60 秒，对于 Amazon CloudWatch 提供的默认一分钟分辨率而言是可行的

设置	描述	Value
hoodie.metrics.cloudwatch.metric.prefix	要添加到每个指标名称的前缀	默认值为空（无前缀）
hoodie.metrics.cloudwatch.namespace	以此为发布指标的 Amazon CloudWatch 命名空间	默认值为 Hudi
hoodie.metrics.cloudwatch.maxDatumsPerRequest	向 Amazon CloudWatch 发出的请求中要包含的最大基准数	默认值为 20（与 Amazon CloudWatch 默认值相同）

Amazon EMR Hudi 配置的支持和改进

- 客户目前可以利用 EMR 配置 API 和重新配置功能在集群级别配置 Hudi 配置。与 Spark 和 Hive 等其他应用程序一样，通过 /etc/hudi/CONF/hudi-defaults.conf 引入了基于文件的新配置支持。EMR 配置了几个默认值以改善用户体验：

  —hoodie.datasource.hive_sync.jdbcurl 已配置为集群 Hive 服务器 URL，无需指定。这在 Spark 集群模式下运行任务时十分有效，而您之前必须指定 Amazon EMR 主 IP。

  — HBase 特定的配置，这对于将 HBase 索引与 Hudi 一起使用非常有用。

  — Zookeeper 锁提供商的特定配置，如并发控制下所讨论的内容，这令乐观并发控制 (OCC) 的使用更加方便。
- 还引入了其他更改，以减少需要通过的配置数量，并在可能的情况下自动推断：

  — 该partitionBy 关键字可用于指定分区列。

  — 启用 Hive Sync 时，不再强制通过HIVE_TABLE_OPT_KEY, HIVE_PARTITION_FIELDS_OPT_KEY, HIVE_PARTITION_EXTRACTOR_CLASS_OPT_KEY。这些值可以根据 Hudi 表名称和分区字段推断出来。

  —KEYGENERATOR_CLASS_OPT_KEY不强制通过，可以从更简单的SimpleKeyGenerator和ComplexKeyGenerator情况下推断。

Hudi 注意事项

- Hudi 不支持在 Hive 中用于读取时合并 (MoR) 和 Bootstrap 表格中的矢量化执行。例如，当`hive.vectorized.execution.enabled`设置为 true 时，Hudi 实时表的`count(*)`失败。作为解决方法，您可以通过将`hive.vectorized.execution.enabled`设置为`false`禁用矢量化读入。
- 多写作器支持与 Hudi 引导启动功能不兼容。
- Flink Streamer 和 Flink SQL 是此发行版中的实验性功能。建议不要在生产部署中使用这些功能。

更改、增强功能和解决的问题

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

- 以前，在多主节点集群上手动重启资源管理器会导致 Zookeeper znode 文件中的 Amazon EMR 集群进程守护程序（如 Zookeeper）重新加载以前停用或丢失的所有节点。在某些情况下，这会导致超出默认限制。Amazon EMR 现在会从 Zookeeper 文件中删除已停用或丢失超过一小时的节点记录，并且内部限制也有所提高。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 配置集群以修复 Apache YARN 时间轴服务器 1 和 1.5 版的性能问题

  Apache YARN 时间轴服务器版本 1 和 1.5 可能会对非常活跃的大型 EMR 集群造成性能问题，尤其是 `yarn.resourcemanager.system-metrics-publisher.enabled=true`，这是 Amazon

EMR 中的默认设置。开源 YARN 时间轴服务器 v2 解决了与 YARN 时间轴服务器可扩展性相关的性能问题。

此问题的其他解决方法包括：

- 配置 yarn.资源管理器.系统指标-发布者.启用=false 在 yarn-site.xml 中。
- 如下所述，在创建群集时启用此问题的修复程序。

以下 Amazon EMR 发行版包含针对此 YARN 时间线服务器性能问题的修复。

EMR 5.30.2、5.31.1、5.32.1、5.33.1、5.34.x、6.0.1、6.1.1、6.2.1、6.3.1、6.4.x

要对上述任何指定的 Amazon EMR 版本启用修复程序，请使用 `aws emr create-cluster` 命令参数：`--configurations file://./configurations.json` 在传入的配置 JSON 文件中将这些属性设置为 `true`。或者使用重新配置控制台 UI 启用修复程序。

配置 .json 文件内容的示例：

```
[
{
"Classification": "yarn-site",
"Properties": {
"yarn.resourcemanager.system-metrics-publisher.timeline-server-v1.enable-batch":
 "true",
"yarn.resourcemanager.system-metrics-publisher.enabled": "true"
},
"Configurations": []
}
]
```

- 默认情况下禁用 WebHDFS 和 HTTFS 服务器。您可以使用 Hadoop 配置重新启用 WebHDFS，`dfs.webhdfs.enabled`。HTTPFS 服务器可以通过使用 `sudo systemctl start hadoop-httpfs` 启动。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

- Hive：为最后任务，通过启用 HDFS 上的 scratch 目录，从而提高写入查询性能。最终任务的临时数据可写入 HDFS 而不是 Amazon S3，性能可以得到提高，因为数据从 HDFS 移动到最终表位置 (Amazon S3) 而不是在 Amazon S3 设备之间移动。
- Hive：使用 Glue 元存储分区修剪，查询编译时间最多可缩短 2.5 倍。
- 默认情况下，当 Hive 将内置 UDF 传递到 Hive 元存储服务器时，由于 Glue 只支持有限的表达式运算，所以只会将这些内置 UDF 的子集传递到 Glue 元存储。如果您设置`hive.glue.partition.pruning.client=true`，则所有分区修剪发生在客户端。如果您设置`hive.glue.partition.pruning.server=true`，则所有分区修剪发生在服务器端。

已知问题

- Hue 查询在 Amazon EMR 6.4.0 中不起作用，因为默认情况下 Apache Hadoop HTTPFS 服务器处于禁用状态。要在 Amazon EMR 6.4.0 上使用 Hue，请使用 `sudo systemctl start hadoop-httpfs` 在 Amazon EMR 主节点上手动启动 HTTPFS 服务器，或者使用 Amazon EMR 步骤。
- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。
- 在 Amazon EMR 6.4.0 版本中，Phoenix 不支持 Phoenix 连接器组件。
- 要将 Spark 操作与 Apache Oozie 一起使用，必须将以下配置添加到 Oozie `workflow.xml` 文件中。否则，Oozie 启动的 Spark 执行器的类路径中将丢失几个诸如 Hadoop 和 EMRFS 之类的关键库。

```
<spark-opts>--conf spark.yarn.populateHadoopClasspath=true</spark-opts>
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.3.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.47.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.13.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.13.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	3.2.1-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.4.4-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.4.4-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.4.4-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.4.4-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.4.4-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-5	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	3.1.2-amzn-5	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-5	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-5	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-5	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-5	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-5	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.8.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.8.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-trino	0.8.0-amzn-0	用于运行 Trino 以及 Hudi 的捆绑库。
hudi-spark	0.8.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.1.2	服务器和客户端的 phoenix 库
phoenix-query-server	5.1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.254.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.254.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.254.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。

组件	版本	描述
trino-coordinator	359	用于在 trino-worker 之中接受查询并管理查询的服务。
trino-worker	359	用于执行查询的各个部分的服务。
trino-client	359	Trino 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Trino 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.1.2-amzn-0	Spark 命令行客户端。
spark-history-server	3.1.2-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.1.2-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.1.2-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	0.4.1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.5.7	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.5.7	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.4.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode,

分类	描述	重新配置操作
		Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.

分类	描述	重新配置操作
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.

分类	描述	重新配置操作
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhu b_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.js on 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.

分类	描述	重新配置操作
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)

分类	描述	重新配置操作
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
trino-log	更改 Trino 的 log.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-config	更改 Trino 的 config.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-password-authenticator	更改 Trino 的 password-authenticator.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-env	更改 Trino 的 trino-env.sh 文件中的值。	Restarts Trino-Server (for Trino)
trino-node	更改 Trino 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-blackhole	更改 Trino 的 blackhole .properties 文件中的值。	Not available.
trino-connector-cassandra	更改 Trino 的 cassandra .properties 文件中的值。	Not available.
trino-connector-hive	更改 Trino 的 hive.properties 文件中的值。	Restarts Trino-Server (for Trino)
trino-connector-jmx	更改 Trino 的 jmx.properties 文件中的值。	Not available.
trino-connector-kafka	更改 Trino 的 kafka.properties 文件中的值。	Not available.
trino-connector-localfile	更改 Trino 的 localfile.properties 文件中的值。	Not available.
trino-connector-memory	更改 Trino 的 memory.properties 文件中的值。	Not available.
trino-connector-mongodb	更改 Trino 的 mongodb.properties 文件中的值。	Not available.
trino-connector-mysql	更改 Trino 的 mysql.properties 文件中的值。	Not available.
trino-connector-postgresql	更改 Trino 的 postgresql.properties 文件中的值。	Not available.
trino-connector-raptor	更改 Trino 的 raptor.properties 文件中的值。	Not available.
trino-connector-redis	更改 Trino 的 redis.properties 文件中的值。	Not available.
trino-connector-redshift	更改 Trino 的 redshift.properties 文件中的值。	Not available.

分类	描述	重新配置操作
trino-connector-tpch	更改 Trino 的 tpch.properties 文件中的值。	Not available.
trino-connector-tpcds	更改 Trino 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site .xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-def aults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.

分类	描述	重新配置操作
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.3.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.3.1	emr-6.3.0	emr-6.2.1	emr-6.2.0
Amazon SDK for Java	1.11.977	1.11.977	1.11.880	1.11.880
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10

	emr-6.3.1	emr-6.3.0	emr-6.2.1	emr-6.2.0
Delta	-	-	-	-
Flink	1.12.1	1.12.1	1.11.2	1.11.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.6	2.2.6	2.2.6-amzn-0	2.2.6-amzn-0
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.7.0-amzn-0	0.7.0-amzn-0	0.6.0-amzn-1	0.6.0-amzn-1
Hue	4.9.0	4.9.0	4.8.0	4.8.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.2.2	1.2.2	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.7.0	1.7.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.1	5.2.0	5.2.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.245.1	0.245.1	0.238.3	0.238.3
Spark	3.1.1	3.1.1	3.0.1	3.0.1

	emr-6.3.1	emr-6.3.0	emr-6.2.1	emr-6.2.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.3.1	2.3.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	350	350	343	343
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。

- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARNarn:aws:s3:::amazonlinux-2-repos-$region/*到策略（将$region替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，s3://bucket/table/p=a 是 s3://bucket/table/p=a b 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 / 字符 (U+002F)。例如，在 s3://bucket/table/p=a b 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：!"#$%&‘()*+,-。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.2.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.46.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.12.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.12.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-3.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-3.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-3.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-3.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-3.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-3.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-3.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-3.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-3.1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	3.2.1-amzn-3.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-3.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.6-amzn-1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.6-amzn-1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.6-amzn-1	HBase 命令行客户端。
hbase-rest-server	2.2.6-amzn-1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.6-amzn-1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-4	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-4	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	3.1.2-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.7.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.7.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-prestosql	0.7.0-amzn-0	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.7.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.2.2	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.245.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.245.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.245.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	350	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	350	用于执行查询的各个部分的服务。
prestosql-client	350	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统

组件	版本	描述
spark-client	3.1.1-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.1.1-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.1.1-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.1.1-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	0.4.1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.3.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
prestosql-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。	Not available.
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。	Not available.
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。	Not available.
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。	Not available.
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。	Not available.

分类	描述	重新配置操作
prestosql-connector-localfile	更改 PrestoSQL 的 localfile .properties 文件中的值。	Not available.
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。	Not available.
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。	Not available.
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。	Not available.
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。	Not available.
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。	Not available.
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。	Not available.
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。	Not available.
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。	Not available.
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site .xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.

分类	描述	重新配置操作
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.3.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.3.0	emr-6.2.1	emr-6.2.0	emr-6.1.1
Amazon SDK for Java	1.11.977	1.11.880	1.11.880	1.11.828
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.12.1	1.11.2	1.11.2	1.11.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.6	2.2.6-amzn-0	2.2.6-amzn-0	2.2.5
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2

	emr-6.3.0	emr-6.2.1	emr-6.2.0	emr-6.1.1
Hudi	0.7.0-amzn-0	0.6.0-amzn-1	0.6.0-amzn-1	0.5.2-incubating-amzn-2
Hue	4.9.0	4.8.0	4.8.0	4.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	-
JupyterHub	1.2.2	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.7.0	1.6.0
Mahout	-	-	-	-
Oozie	5.2.1	5.2.0	5.2.0	5.2.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.245.1	0.238.3	0.238.3	0.232
Spark	3.1.1	3.0.1	3.0.1	3.0.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.3.1	2.3.1	2.1.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	350	343	343	338
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.3.0 的信息。更改与 6.2.0 有关。

首次发布日期：2021 年 5 月 12 日

上次更新日期：2021 年 8 月 9 日

支持的应用程序

- Amazon SDK for Java 1.11.977
- CloudWatch Sink 2.1.0
- DynamoDB 连接器 4.16.0
- EMRFS 2.46.0
- Amazon EMR Goodies 3.2.0
- Amazon EMR Kinesis 连接器 3.5.0 版
- Amazon EMR 记录服务器版本 2.0.0
- Amazon EMR Scripts 2.5.0
- Flink 1.12.1
- Ganglia 3.7.2
- Amazon Glue Hive Metastore Client 3.2.0
- Hadoop 3.2.1-amzn-3
- HBase 2.2.6-amzn-1
- HBase-operator-tools 1.0.0
- HCatalog 3.1.2-amzn-0
- Hive 3.1.2-amzn-4
- Hudi 0.7.0-amzn-0
- Hue 4.9.0
- Java JDK Corretto-8.282.08.1（内部1.8.0_282-b08）
- JupyterHub 1.2.0
- Livy 0.7.0-incubating
- MXNet 1.7.0
- Oozie 5.2.1

- Phoenix 5.0.0
- Pig 0.17.0
- Presto 0.245.1-amzn-0
- PrestoSQL 350
- Apache Ranger KMS（多主节点透明加密）版本 2.0.0
- ranger-plugins 2.0.1-amzn-0
- ranger-s3-plugin 1.1.0
- SageMaker Spark SDK 1.4.1
- Scala 2.12.10（OpenJDK 64 位服务器 VM，Java 1.8.0_282）
- Spark 3.1.1-amzn-0
- spark-rapids 0.4.1
- Sqoop 1.4.7
- TensorFlow 2.4.1
- tez 0.9.2
- Zeppelin 0.9.0
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.16.0

新特征

- Amazon EMR 支持 Amazon S3 接入点，这是 Amazon S3 的一项功能，可让您轻松管理共享数据湖的访问。使用 Amazon S3 接入点别名，您可以在 Amazon EMR 上大规模简化数据访问。您可以将 Amazon S3 接入点与所有版本的 Amazon EMR 一起使用，在 Amazon EMR 可用的所有 Amazon 区域无需额外费用。要了解有关 Amazon S3 访问点和访问点别名的详细信息，请参阅《Amazon S3 用户指南》中的为接入点使用存储桶式别名。
- 新的 `DescribeReleaseLabel` 和 `ListReleaseLabel` API 参数提供 Amazon EMR 发行版标注详细信息。您可以以编程方式列出运行 API 请求的区域中提供的版本，并列出特定 Amazon EMR 发行版标注的可用应用程序。发行版标签参数还列出了支持指定应用程序（如 Spark）的 Amazon EMR 发行版。以编程方式启动 Amazon EMR 集群时会用到此信息。例如，您可以使用 `ListReleaseLabel` 结果中的最新发行版启动集群。有关更多信息，请参阅《Amazon EMR API 参考》中的 DescribeReleaseLabel 和 ListReleaseLabels。
- 借助 Amazon EMR 6.3.0，您可以启动与 Apache Ranger 在本地集成的集群。Apache Ranger 是一个开源框架，可跨 Hadoop 平台启用、监控和管理全面的数据安全。有关更多信息，请参阅 Apache

Ranger。通过本机集成，您可以自带 Apache Ranger，在 Amazon EMR 上强制实施精细数据访问控制。请参阅《Amazon EMR 管理指南》中的将 Amazon EMR 与 Apache Ranger 集成。

- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。
- 实例元数据服务（IMDS）V2 支持状态：对于 Amazon EMR 6.2 或更高版本，Amazon EMR 组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。如果您在早于 Amazon EMR 6.x 的发行版中禁用 IMDSv1，则会导致集群启动失败。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Spark SQL UI 说明了如何在 Spark 3.1 中将默认模式从 `extended` 更改为 `formatted`。Amazon EMR 已将其恢复为 `extended`，以在 Spark SQL UI 中包含逻辑计划信息。可以通过将 `spark.sql.ui.explainMode` 设置为 `formatted` 进行还原。

- 以下提交是从 Spark 主分支逆向移植的。
  - [SPARK-34752][BUILD] Bump Jetty 升级到 9.4.37，解决了 CVE-2020-27223 问题。
  - [SPARK-34534] 修复了使用 FetchShuffleBlocks 获取数据块时的数据块 ID 顺序。
  - [SPARK-34681] [SQL] 修复了在非同等条件下构建左侧时完整的外部混乱哈希连接的错误。
  - [SPARK-34497] [SQL] 修复了内置 JDBC 连接提供程序，以恢复 JVM 安全上下文更改。
- 为了提高与 Nvidia Spark RAPID 插件的互操作性，添加了解决以下问题的解决方法：在禁用自适应查询执行的情况下，使用 Nvidia Spark RAPID 时阻止动态分区修剪触发，请参阅 RAPIDS 问题 #1378 和 RAPIDS 问题 ##1386。有关新配置的详细信息`spark.sql.optimizer.dynamicPartitionPruning.enforceBroadcastReuse`，请参阅 RAPIDS 问题 ##1386。
- 文件输出提交程序默认算法已在开源 Spark 3.1 中将 v2 算法更改为 v1 算法。有关更多信息，请参阅 Amazon EMR 优化 Spark 性能 – 动态分区修剪。
- Amazon EMR 恢复为 v2 算法（早于 Amazon EMR 6.x 的发行版中使用默认算法），以防止性能下降。要恢复开源 Spark 3.1 行为，请将 `spark.hadoop.mapreduce.fileoutputcommitter.algorithm.version` 设置为 1。开源 Spark 进行了此更改，因为文件输出提交程序算法 v2 中的任务提交不是原子操作，在某些情况下可能会导致输出数据正确性问题。不过，算法 v1 中的任务提交也不是原子操作。在某些情况下，任务提交会包括在重命名之前执行的删除。这可能会导致出现无提示的数据正确性问题。
- 修复了早期 Amazon EMR 发行版中的托管扩展问题，并对托管扩展进行了改进，从而显著降低了应用程序故障率。
- 在每个新集群上已安装了Amazon Java SDK Bundle。这是一个包含所有服务 SDK 及其依赖项的单个 jar，而不是单个组件 jar。有关更多信息，请参阅 Java SDK Bundled Dependency。

已知问题

- 对于 Amazon EMR 6.3.0 和 6.2.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 `sudo systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。
- 启用 Amazon Glue 数据目录时，使用 Spark 访问具有空字符串位置 URI 的 Amazon Glue 数据库可能会失败。这种情况发生在早期的 Amazon EMR 发行版中，但 SPARK-31709（https://issues.apache.org/jira/browse/SPARK-31709）使其可应用于更多案例。例如，当在位置 URI 为空

字符串的默认 Amazon Glue 数据库内创建表时，`spark.sql("CREATE TABLE mytest (key string) location '/table_path';")` 会失败并显示消息“Cannot create a Path from an empty string（无法从空字符串创建路径）”。要解决此问题，请手动设置 Amazon Glue 数据库的位置 URI，然后使用 Spark 在这些数据库中创建表。

- 在 Amazon EMR 6.3.0 中，PrestoSQL 已从版本 343 升级到版本 350。开源中有两个与安全相关的更改与此版本更改相关。未定义表、架构或会话属性规则时，基于文件的目录访问控制已从 deny 更改为 `allow`。此外，基于文件的系统访问控制已更改为支持目录规则未定义的文件。在这种情况下，允许完全访问目录。

  有关更多信息，请参阅发行版 344（2020 年 10 月 9 日）。
- 请注意，所有人都可读取 Hadoop 用户目录（/home/hadoop）。它具有 Unix 755（drwxr-xr-x）目录权限，允许 Hive 等框架进行读取访问。您可以将文件放入 /home/hadoop 及其子目录中，但请注意这些目录的权限，做好对敏感信息的保护。
- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```
  2. 重新启动 InstanceController

```
$ sudo systemctl daemon-reload

$ sudo systemctl restart instance-controller
```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- **Important**

  运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- 要将 Spark 操作与 Apache Oozie 一起使用，必须将以下配置添加到 Oozie `workflow.xml` 文件中。否则，Oozie 启动的 Spark 执行器的类路径中将丢失几个诸如 Hadoop 和 EMRFS 之类的关键库。

```
<spark-opts>--conf spark.yarn.populateHadoopClasspath=true</spark-opts>
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	3.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.2.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.1.0	EMR S3 Select 连接器
emrfs	2.46.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.12.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.12.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	3.2.1-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-3	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.6-amzn-1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	2.2.6-amzn-1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.6-amzn-1	HBase 命令行客户端。
hbase-rest-server	2.2.6-amzn-1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.6-amzn-1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-4	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-4	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.7.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.7.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。

组件	版本	描述
hudi-prestosql	0.7.0-amzn-0	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.7.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.2.2	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.245.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.245.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.245.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	350	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	350	用于执行查询的各个部分的服务。
prestosql-client	350	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.1.1-amzn-0	Spark 命令行客户端。
spark-history-server	3.1.1-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	3.1.1-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.1.1-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	0.4.1	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.3.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Restarts Flink history server.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS services Namenode, Datanode, and ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.

分类	描述	重新配置操作
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.

分类	描述	重新配置操作
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.

分类	描述	重新配置操作
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.

分类	描述	重新配置操作
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)

分类	描述	重新配置操作
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
prestosql-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)

分类	描述	重新配置操作
prestosql-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。	Not available.
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。	Not available.
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。	Not available.
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。	Not available.
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。	Not available.
prestosql-connector-localfile	更改 PrestoSQL 的 localfile.properties 文件中的值。	Not available.
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。	Not available.
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。	Not available.
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。	Not available.
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。	Not available.
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。	Not available.
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。	Not available.
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。	Not available.
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.

分类	描述	重新配置操作
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie and HiveServer2.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.

分类	描述	重新配置操作
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.2.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.2.1	emr-6.2.0	emr-6.1.1	emr-6.1.0
Amazon SDK for Java	1.11.880	1.11.880	1.11.828	1.11.828
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.11.2	1.11.2	1.11.0	1.11.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.6-amzn-0	2.2.6-amzn-0	2.2.5	2.2.5
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.6.0-amzn-1	0.6.0-amzn-1	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2
Hue	4.8.0	4.8.0	4.7.1	4.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	-	-

	emr-6.2.1	emr-6.2.0	emr-6.1.1	emr-6.1.0
JupyterHub	1.1.0	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.6.0	1.6.0
Mahout	-	-	-	-
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.238.3	0.238.3	0.232	0.232
Spark	3.0.1	3.0.1	3.0.0	3.0.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.3.1	2.3.1	2.1.0	2.1.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	343	343	338	338
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。

- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 / 字符 (U+002F)。例如，在 s3://bucket/table/p=a b 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：!"#$%&'()*+,-。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-notebook-env	1.0.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.16.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.0.0	EMR S3 Select 连接器
emrfs	2.44.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.11.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-2.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-2.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-2.1	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	3.2.1-amzn-2.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-2.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-2.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-2.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-2.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-2.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-2.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-2.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.6-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.6-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.6-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.2.6-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	2.2.6-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-3	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-3	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.6.0-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-prestosql	0.6.0-amzn-1	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.6.0-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。

组件	版本	描述
hue-server	4.8.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.4.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.238.3-amzn-1	用于在 presto-worker 之中接受查询并管理查询的服务。

组件	版本	描述
presto-worker	0.238.3-amzn-1	用于执行查询的各个部分的服务。
presto-client	0.238.3-amzn-1	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	343	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	343	用于执行查询的各个部分的服务。
prestosql-client	343	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.0.1-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.0.1-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.0.1-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.0.1-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
spark-rapids	0.2.0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。

组件	版本	描述
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.3.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-preview1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.2.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.

分类	描述	重新配置操作
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Not available.
flink-log4j	更改 Flink log4j.properties 设置。	Not available.
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Not available.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
prestosql-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。	Not available.
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。	Not available.
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。	Not available.
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。	Not available.
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。	Not available.

分类	描述	重新配置操作
prestosql-connector-localfile	更改 PrestoSQL 的 localfile .properties 文件中的值。	Not available.
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。	Not available.
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。	Not available.
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。	Not available.
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。	Not available.
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。	Not available.
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。	Not available.
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。	Not available.
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。	Not available.
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site .xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.

分类	描述	重新配置操作
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.2.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.2.0	emr-6.1.1	emr-6.1.0	emr-6.0.1
Amazon SDK for Java	1.11.880	1.11.828	1.11.828	1.11.711
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.12.10
Delta	-	-	-	-
Flink	1.11.2	1.11.0	1.11.0	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.6-amzn-0	2.2.5	2.2.5	2.2.3
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2

	emr-6.2.0	emr-6.1.1	emr-6.1.0	emr-6.0.1
Hudi	0.6.0-amzn-1	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2	0.5.0-incubating-amzn-1
Hue	4.8.0	4.7.1	4.7.1	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	-	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.0.0
Livy	0.7.0	0.7.0	0.7.0	0.6.0
MXNet	1.7.0	1.6.0	1.6.0	1.5.1
Mahout	-	-	-	-
Oozie	5.2.0	5.2.0	5.2.0	5.1.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	0.17.0	-
Presto	0.238.3	0.232	0.232	0.230
Spark	3.0.1	3.0.0	3.0.0	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	-
TensorFlow	2.3.1	2.1.0	2.1.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	343	338	338	-
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.2.0 的信息。更改与 6.1.0 有关。

首次发布日期：2020 年 12 月 9 日

上次更新日期：2021 年 10 月 4 日

支持的应用程序

- Amazon SDK for Java 1.11.828
- emr-record-server 1.7.0
- Flink 1.11.2
- Ganglia 3.7.2
- Hadoop 3.2.1-amzn-1
- HBase 2.2.6-amzn-0
- HBase-operator-tools 1.0.0
- HCatalog 3.1.2-amzn-0
- Hive 3.1.2-amzn-3
- Hudi 0.6.0-amzn-1
- Hue 4.8.0
- JupyterHub 1.1.0
- Livy 0.7.0
- MXNet 1.7.0
- Oozie 5.2.0
- Phoenix 5.0.0
- Pig 0.17.0
- Presto 0.238.3-amzn-1
- PrestoSQL 343
- Spark 3.0.1-amzn-0
- spark-rapids 0.2.0
- TensorFlow 2.3.1
- Zeppelin 0.9.0-preview1

- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.16.0

新特征

- HBase：删除了提交阶段的重命名，添加了持久性 HFile 跟踪。请参阅《Amazon EMR 版本指南》中的持久性 HFile 跟踪。
- HBase：已逆向移植创建在压缩时强制缓存数据块的配置。
- PrestoDB：改进了动态分区修剪。基于规则的连接重新排序对未分区数据运行。
- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。
- 实例元数据服务（IMDS）V2 支持状态：对于 Amazon EMR 6.2 或更高版本，Amazon EMR 组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。如果您在早于 Amazon EMR 6.x 的发行版中禁用 IMDSv1，则会导致集群启动失败。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。

- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Spark：改进了 Spark 运行时的性能。

已知问题

- Amazon EMR 6.2 对 EMR 6.2.0 中的 /etc/cron.d/libinstance-controller-java 文件设置了错误权限。当文件的权限应为 644 (-rw-r--r--) 时，它们为 645 (-rw-r--r-x)。因此，Amazon EMR 6.2 版本不记录实例状态日志，并且 /emr/instance-log 目录为空。此问题已在 Amazon EMR 6.3.0 及更高版本中得到修复。

  要解决此问题，请在集群启动时将以下脚本作为引导操作运行。

  ```
  #!/bin/bash
  sudo chmod 644 /etc/cron.d/libinstance-controller-java
  ```

- 对于 Amazon EMR 6.2.0 和 6.3.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 `sudo systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。
- Amazon EMR 6.2.0 中存在一个问题：httpd 持续失败，导致 Ganglia 不可用。您会收到“cannot connect to the server（无法连接到服务器）”错误。要修复已在运行期间出现此问题的集群，请使用 SSH 连接到集群主节点并将行 `Listen 80` 添加到位于 `/etc/httpd/conf/httpd.conf` 的文件 `httpd.conf` 中。此问题已在 Amazon EMR 6.3.0 中得到修复。
- 使用安全配置时，HTTPD 在 EMR 6.2.0 集群会上失败。因此，Ganglia Web 应用程序用户界面不可用。要访问 Ganglia Web 应用程序用户界面，请将 `Listen 80` 添加到集群主节点上的 `/etc/httpd/conf/httpd.conf` 文件中。有关连接集群的更多信息，请参阅使用 SSH 连接到主节点。

  使用安全配置时，EMR Notebooks 也无法建立与 EMR 6.2.0 集群的连接。笔记本将无法列出内核和提交 Spark 任务。我们建议您改为将 EMR Notebooks 与其它版本的 Amazon EMR 结合使用。
- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon

EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
```

```
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- **Important**

  Amazon EMR 6.1.0 和 6.2.0 包含可能严重影响所有 Hudi 插入、更新插入和删除操作的性能问题。如果您计划将 Hudi 与 Amazon EMR 6.1.0 或 6.2.0 结合使用，请联系 Amazon Support，获取 Hudi RPM 补丁。

- **Important**

  运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- Amazon EMR 6.2.0 Maven 构件尚未发布。它们将随 Amazon EMR 未来版本一起发布。
- 使用 HBase 存储文件系统表的持久性 HFile 跟踪不支持 HBase 区域复制功能。有关 HBase 区域复制的更多信息，请参阅时间表一致的高可用读取。
- Amazon EMR 6.x 和 EMR 5.x Hive 分桶版本差异

  EMR 5.x 使用 OOS Apache Hive 2，而 EMR 6.x 使用 OOS Apache Hive 3。开源 Hive2 使用分桶版本 1，而开源 Hive3 使用分桶版本 2。Hive 2（EMR 5.x）和 Hive 3（EMR 6.x）之间的这一分桶版本差异将导致 Hive 分桶哈希函数不同。请参见以下示例。

  下表分别是在 EMR 6.x 和 EMR 5.x 中创建的示例。

```
-- Using following LOCATION in EMR 6.x
CREATE TABLE test_bucketing (id INT, desc STRING)
PARTITIONED BY (day STRING)
```

```
CLUSTERED BY(id) INTO 128 BUCKETS
LOCATION 's3://your-own-s3-bucket/emr-6-bucketing/';

-- Using following LOCATION in EMR 5.x
LOCATION 's3://your-own-s3-bucket/emr-5-bucketing/';
```

在 EMR 6.x 和 EMR 5.x 中插入相同的数据。

```
INSERT INTO test_bucketing PARTITION (day='01') VALUES(66, 'some_data');
INSERT INTO test_bucketing PARTITION (day='01') VALUES(200, 'some_data');
```

检查 S3 位置，显示分桶文件名不同，这是因为 EMR 6.x（Hive 3）和 EMR 5.x（Hive 2）之间的哈希函数不同。

```
[hadoop@ip-10-0-0-122 ~]$ aws s3 ls s3://your-own-s3-bucket/emr-6-bucketing/day=01/
2020-10-21 20:35:16         13 000025_0
2020-10-21 20:35:22         14 000121_0
[hadoop@ip-10-0-0-122 ~]$ aws s3 ls s3://your-own-s3-bucket/emr-5-bucketing/day=01/
2020-10-21 20:32:07         13 000066_0
2020-10-21 20:32:51         14 000072_0
```

您还可以通过以下方式查看版本之间的差异：在 EMR 6.x 的 Hive CLI 中运行以下命令。请注意，它将返回分桶版本 2。

```
hive> DESCRIBE FORMATTED test_bucketing;
...
Table Parameters:
    bucketing_version       2
...
```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.0.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.16.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.0.0	EMR S3 Select 连接器
emrfs	2.44.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.11.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-2	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-2	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	3.2.1-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.6-amzn-0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.6-amzn-0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.6-amzn-0	HBase 命令行客户端。
hbase-rest-server	2.2.6-amzn-0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.6-amzn-0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-3	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-3	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	3.1.2-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.6.0-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-prestosql	0.6.0-amzn-1	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.6.0-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.8.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.4.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.238.3-amzn-1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.238.3-amzn-1	用于执行查询的各个部分的服务。
presto-client	0.238.3-amzn-1	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	343	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	343	用于执行查询的各个部分的服务。
prestosql-client	343	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统

组件	版本	描述
spark-client	3.0.1-amzn-0	Spark 命令行客户端。
spark-history-server	3.0.1-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.0.1-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.0.1-amzn-0	YARN 从属项所需的 Apache Spark 库。
spark-rapids	0.2.0	加速 Apache Spark 和 GPU 的 Nvidia Spark RAPIDS。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.3.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-preview1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-6.2.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Not available.
flink-log4j	更改 Flink log4j.properties 设置。	Not available.
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Not available.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Secondary Namenode, Datanode, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	This classification should not be reconfigured.
hdfs-env	更改 HDFS 环境中的值。	Restarts Hadoop HDFS ZKFC.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat server.
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。	Sets configurations to launch Hive LLAP service.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Not available.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2, HiveMetastore, and Hive HCatalog-Server. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.

分类	描述	重新配置操作
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-blackhole	更改 Presto 的 blackhole .properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoDB)
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
prestosql-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。	Not available.
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。	Not available.
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。	Not available.
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。	Restarts Presto-Server (for PrestoSQL)
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。	Not available.
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。	Not available.

分类	描述	重新配置操作
prestosql-connector-localfile	更改 PrestoSQL 的 localfile .properties 文件中的值。	Not available.
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。	Not available.
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。	Not available.
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。	Not available.
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。	Not available.
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。	Not available.
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。	Not available.
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。	Not available.
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。	Not available.
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site .xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.

分类	描述	重新配置操作
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restart Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 6.1.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Oozie、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Amazon SDK for Java	1.11.828	1.11.828	1.11.711	1.11.711
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.11.12
Delta	-	-	-	-
Flink	1.11.0	1.11.0	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.5	2.2.5	2.2.3	2.2.3
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Hudi	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2	0.5.0-incubating-amzn-1	0.5.0-incubating-amzn-1
Hue	4.7.1	4.7.1	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.0.0	1.0.0
Livy	0.7.0	0.7.0	0.6.0	0.6.0
MXNet	1.6.0	1.6.0	1.5.1	1.5.1
Mahout	-	-	-	-
Oozie	5.2.0	5.2.0	5.1.0	5.1.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	-	-
Presto	0.232	0.232	0.230	0.230
Spark	3.0.0	3.0.0	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	-	-
TensorFlow	2.1.0	2.1.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	338	338	-	-
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.3.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.0.0	EMR S3 Select 连接器
emrfs	2.42.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-1.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-1.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-1.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-1.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-1.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-1.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-1.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-1.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-1.1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	3.2.1-amzn-1.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-1.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.5	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.5	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.5	HBase 命令行客户端。
hbase-rest-server	2.2.5	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.5	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-2	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	3.1.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.2-incubating-amzn-2	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.2-incubating-amzn-2	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-prestosql	0.5.2-incubating-amzn-2	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.5.2-incubating-amzn-2	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.7.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.6.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.3.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.232	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.232	用于执行查询的各个部分的服务。
presto-client	0.232	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	338	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	338	用于执行查询的各个部分的服务。
prestosql-client	338	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统

组件	版本	描述
spark-client	3.0.0-amzn-0.1	Spark 命令行客户端。
spark-history-server	3.0.0-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.0.0-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.0.0-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.1.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-preview1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅 配置应用程序。

emr-6.1.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-env	更改 HDFS 环境中的值。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。

分类	描述
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
prestosql-log	更改 Presto 的 log.properties 文件中的值。
prestosql-config	更改 Presto 的 config.properties 文件中的值。
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。
prestosql-connector-localfile	更改 PrestoSQL 的 localfile.properties 文件中的值。
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。

分类	描述
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 6.1.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Oozie、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Amazon SDK for Java	1.11.828	1.11.828	1.11.711	1.11.711
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.11.12
Delta	-	-	-	-
Flink	1.11.0	1.11.0	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.5	2.2.5	2.2.3	2.2.3
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Hudi	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2	0.5.0-incubating-amzn-1	0.5.0-incubating-amzn-1
Hue	4.7.1	4.7.1	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.0.0	1.0.0
Livy	0.7.0	0.7.0	0.6.0	0.6.0
MXNet	1.6.0	1.6.0	1.5.1	1.5.1
Mahout	-	-	-	-
Oozie	5.2.0	5.2.0	5.1.0	5.1.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	-	-
Presto	0.232	0.232	0.230	0.230
Spark	3.0.0	3.0.0	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	-	-
TensorFlow	2.1.0	2.1.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	338	338	-	-
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.1.0 的信息。更改与 6.0.0 有关。

首次发布日期：2020 年 9 月 4 日

上次更新日期：2020 年 10 月 15 日

支持的应用程序

- Amazon SDK for Java 1.11.828
- Flink 1.11.0
- Ganglia 3.7.2
- Hadoop 3.2.1-amzn-1
- HBase 2.2.5
- HBase-operator-tools 1.0.0
- HCatalog 3.1.2-amzn-0
- Hive 3.1.2-amzn-1
- Hudi 0.5.2-incubating
- Hue 4.7.1
- JupyterHub 1.1.0
- Livy 0.7.0
- MXNet 1.6.0
- Oozie 5.2.0
- Phoenix 5.0.0
- Presto 0.232
- PrestoSQL 338
- Spark 3.0.0-amzn-0
- TensorFlow 2.1.0
- Zeppelin 0.9.0-preview1
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.14.0

新特征

- 从 Amazon EMR 5.30.0 和 Amazon EMR 6.1.0 开始，支持 ARM 实例类型。
- 从 Amazon EMR 6.1.0 和 5.30.0.开始，支持 M6g 通用型实例类型。有关更多信息，请参阅《Amazon EMR 管理指南》中的支持的实例类型。
- 从 Amazon EMR 5.23.0 开始支持 EC2 置放群组功能，该功能可作为多主节点集群选项。目前，置放群组功能仅支持主节点类型，并会将 SPREAD 策略应用于这些主节点。SPREAD 策略将一小组实例放置在单独的基础硬件上，以防止发生硬件故障时出现多个主节点丢失的问题。有关更多信息，请参阅《Amazon EMR 管理指南》中的 EMR 与 EC2 置放群组的集成。
- 托管扩展 – 使用 Amazon EMR 版本 6.1.0 时，您可以启用 Amazon EMR 托管式自动扩缩功能，以根据工作负载自动增加或减少集群中实例或单位的数量。Amazon EMR 会持续评估集群指标，以便做出扩展决策，从而优化集群的成本和速度。Amazon EMR 5.30.0 及更高版本（但 6.0.0 除外）也提供了托管扩展。有关更多信息，请参阅《Amazon EMR 管理指南》中的扩缩集群资源。
- EMR 6.1.0 支持 PrestoSQL 338。有关更多信息，请参阅 Presto。
  - 仅在 EMR 6.1.0 及更高版本上支持 PrestoSQL，而 EMR 6.0.0 或 EMR 5.x 则不支持。
  - 可以继续使用应用程序名称 `Presto` 在集群上安装 PrestoDB。要在集群上安装 PrestoSQL，请使用应用程序名称 `PrestoSQL`。
  - 您可以安装 PrestoDB 或 PrestoSQL，但不能在同一个集群上同时安装两者。如果在尝试创建集群时同时指定了 PrestoDB 和 PrestoSQL，则会发生验证错误，而且集群创建请求失败。
  - 单主节点集群和多主节点集群均支持 PrestoSQL。在多主节点集群上，需要外部 Hive 元存储才能运行 PrestoSQL 或 PrestoDB。请参阅 Supported applications in an EMR cluster with multiple primary nodes。
- 支持在 Apache Hadoop 和 Apache Spark 上使用 Docker 对 ECR 进行自动身份验证：Spark 用户可以使用 Docker Hub 中的 Docker 镜像和 Amazon Elastic Container Registry（Amazon ECR）来定义环境和库依赖项。

  配置 Docker 和使用 Amazon EMR 6.x 通过 Docker 运行 Spark 应用程序。
- EMR 支持 Apache Hive ACID 事务：Amazon EMR 6.1.0 增加了对 Hive ACID 事务的支持，使其符合数据库的 ACID 属性。借助此功能，您可以使用 Amazon Simple Storage Service（Amazon S3）中的数据在 Hive 托管表中运行`INSERT`, `UPDATE`, `DELETE`,和`MERGE`操作。这是流式提取、数据重述、使用 MERGE 批量更新等使用案例的一项关键功能，并会缓慢更改维度。有关包括配置示例和使用案例在内的更多信息，请参阅 Amazon EMR 支持 Apache Hive ACID 事务。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- EMR 6.0.0 上不支持 Apache Fink，但集成了 Flink 1.11.0 的 EMR 6.1.0 可以支持 Apache Fink。这是首个正式支持 Hadoop 3 的 Fink 版本。请参阅 Apache Flink 1.11.0 发布公告。
- 默认 EMR 6.1.0 捆绑包中已经删除了 Ganglia。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定

设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
```

```
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- > **Important**
  >
  > Amazon EMR 6.1.0 和 6.2.0 包含可能严重影响所有 Hudi 插入、更新插入和删除操作的性能问题。如果您计划将 Hudi 与 Amazon EMR 6.1.0 或 6.2.0 结合使用，请联系 Amazon Support，获取 Hudi RPM 补丁。

- 如果使用 `spark.driver.extraJavaOptions` 和 `spark.executor.extraJavaOptions` 设置自定义垃圾回收配置，将会因为垃圾回收配置冲突导致 EMR 6.1 驱动程序/执行程序启动失败。使用 EMR 发行版 6.1.0 时，您应该使用属性 `spark.driver.defaultJavaOptions` 和 `spark.executor.defaultJavaOptions` 为驱动程序和执行程序指定自定义 Spark 垃圾回收配置。如要了解更多信息，请参阅 Apache Spark 运行时环境和在 Amazon EMR 6.1.0 上配置 Spark 垃圾回收。
- 在 Oozie 中使用 Pig（以及在 Hue 中，因为 Hue 使用 Oozie 操作来运行 Pig 脚本）会生成一个错误，即无法加载 native-lzo 库。此错误消息是信息性的，不会阻止 Pig 运行。
- Hudi 并发支持：目前 Hudi 不支持并发写入单个 Hudi 表。此外，Hudi 会回滚处于运行状态的写入器所做的所有更改后再允许新写入器启动。并发写入可能会干扰此机制并引入竞争条件，这会导致数据损坏。您应确保作为数据处理工作流程的一部分，任何时候都只有一个 Hudi 写入器对 Hudi 表进行操作。Hudi 支持多个并发读取器对同一 Hudi 表进行操作。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- Amazon EMR 6.1.0 中存在一个问题，会影响运行 Presto 的集群。在较长时间（天）后，集群可能会引发错误，例如“su: failed to execute /bin/bash: Resource temporarily unavailable”或“shell request failed on channel 0”。此问题是由内部 Amazon EMR 进程（InstanceController）产生过多的轻量级进程（LWP）导致的，这最终会导致 Hadoop 用户超出其 nproc 限制。这可以阻止用户打开其它进程。此问题的解决方案是：升级到 EMR 6.2.0。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.3.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.1.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	2.0.0	EMR S3 Select 连接器
emrfs	2.42.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-hdfs-journalnode	3.2.1-amzn-1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.5	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.5	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.5	HBase 命令行客户端。
hbase-rest-server	2.2.5	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.5	用于向 HBase 提供 Thrift 终端节点的服务。

组件	版本	描述
hcatalog-client	3.1.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-2	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.2-incubating-amzn-2	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.2-incubating-amzn-2	用于运行 Presto 以及 Hudl 的捆绑库。
hudi-prestosql	0.5.2-incubating-amzn-2	用于运行 PrestoSQL 以及 Hudi 的捆绑库。
hudi-spark	0.5.2-incubating-amzn-2	用于运行 Spark 以及 Hudi 的捆绑库。
hue-server	4.7.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.6.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.3.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.232	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.232	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.232	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
prestosql-coordinator	338	用于在 prestosql-worker 之中接受查询并管理查询执行的服务。
prestosql-worker	338	用于执行查询的各个部分的服务。
prestosql-client	338	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	2.0.0	Apache Ranger 密钥管理系统
spark-client	3.0.0-amzn-0	Spark 命令行客户端。
spark-history-server	3.0.0-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	3.0.0-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	3.0.0-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.1.0	适用于高性能数值计算的 TensorFlow 开源软件库。

组件	版本	描述
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-preview1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-6.1.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-env	更改 HDFS 环境中的值。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。

分类	描述
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
prestosql-log	更改 Presto 的 log.properties 文件中的值。
prestosql-config	更改 Presto 的 config.properties 文件中的值。
prestosql-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
prestosql-env	更改 Presto 的 presto-env.sh 文件中的值。
prestosql-node	更改 PrestoSQL 的 node.properties 文件中的值。
prestosql-connector-blackhole	更改 PrestoSQL 的 blackhole.properties 文件中的值。

分类	描述
prestosql-connector-cassandra	更改 PrestoSQL 的 cassandra.properties 文件中的值。
prestosql-connector-hive	更改 PrestoSQL 的 hive.properties 文件中的值。
prestosql-connector-jmx	更改 PrestoSQL 的 jmx.properties 文件中的值。
prestosql-connector-kafka	更改 PrestoSQL 的 kafka.properties 文件中的值。
prestosql-connector-localfile	更改 PrestoSQL 的 localfile.properties 文件中的值。
prestosql-connector-memory	更改 PrestoSQL 的 memory.properties 文件中的值。
prestosql-connector-mongodb	更改 PrestoSQL 的 mongodb.properties 文件中的值。
prestosql-connector-mysql	更改 PrestoSQL 的 mysql.properties 文件中的值。
prestosql-connector-postgresql	更改 PrestoSQL 的 postgresql.properties 文件中的值。
prestosql-connector-raptor	更改 PrestoSQL 的 raptor.properties 文件中的值。
prestosql-connector-redis	更改 PrestoSQL 的 redis.properties 文件中的值。
prestosql-connector-redshift	更改 PrestoSQL 的 redshift.properties 文件中的值。
prestosql-connector-tpch	更改 PrestoSQL 的 tpch.properties 文件中的值。

分类	描述
prestosql-connector-tpcds	更改 PrestoSQL 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 6.0.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Oozie、Phoe
Presto、Spark、TensorFlow、Tez、Zeppelin 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Amazon SDK for Java	1.11.828	1.11.828	1.11.711	1.11.711
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.11.12
Delta	-	-	-	-
Flink	1.11.0	1.11.0	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	2.2.5	2.2.5	2.2.3	2.2.3
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2	0.5.0-incubating-amzn-1	0.5.0-incubating-amzn-1
Hue	4.7.1	4.7.1	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.0.0	1.0.0
Livy	0.7.0	0.7.0	0.6.0	0.6.0
MXNet	1.6.0	1.6.0	1.5.1	1.5.1
Mahout	-	-	-	-

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Oozie	5.2.0	5.2.0	5.1.0	5.1.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	-	-
Presto	0.232	0.232	0.230	0.230
Spark	3.0.0	3.0.0	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	-	-
TensorFlow	2.1.0	2.1.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	338	338	-	-
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。

- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.6	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.5.0	EMR S3 Select 连接器
emrfs	2.39.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-0.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-0.1	用于存储数据块的 HDFS 节点级服务。

组件	版本	描述
hadoop-hdfs-library	3.2.1-amzn-0.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-0.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-0.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-0.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-0.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-0.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-0.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-0.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-0.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	2.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.3	HBase 命令行客户端。
hbase-rest-server	2.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	2.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-0	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.0-incubating-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.0-incubating-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器

组件	版本	描述
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.230	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.230	用于执行查询的各个部分的服务。
presto-client	0.230	Presto 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Presto 服务器）上。

组件	版本	描述
r	3.4.3	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-SNAPSHOT	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-6.0.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-env	更改 HDFS 环境中的值。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。

分类	描述
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。

分类	描述
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 6.0.0

应用程序版本

此发行版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Oozie、Phoe
Presto、Spark、TensorFlow、Tez、Zeppelin 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Amazon SDK for Java	1.11.828	1.11.828	1.11.711	1.11.711
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	2.12.10	2.12.10	2.11.12
Delta	-	-	-	-
Flink	1.11.0	1.11.0	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
HBase	2.2.5	2.2.5	2.2.3	2.2.3
HCatalog	3.1.2	3.1.2	3.1.2	3.1.2
Hadoop	3.2.1	3.2.1	3.2.1	3.2.1
Hive	3.1.2	3.1.2	3.1.2	3.1.2
Hudi	0.5.2-incubating-amzn-2	0.5.2-incubating-amzn-2	0.5.0-incubating-amzn-1	0.5.0-incubating-amzn-1
Hue	4.7.1	4.7.1	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.0.0	1.0.0
Livy	0.7.0	0.7.0	0.6.0	0.6.0
MXNet	1.6.0	1.6.0	1.5.1	1.5.1
Mahout	-	-	-	-
Oozie	5.2.0	5.2.0	5.1.0	5.1.0
Phoenix	5.0.0	5.0.0	5.0.0	5.0.0
Pig	0.17.0	0.17.0	-	-
Presto	0.232	0.232	0.230	0.230
Spark	3.0.0	3.0.0	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	-	-
TensorFlow	2.1.0	2.1.0	1.14.0	1.14.0

	emr-6.1.1	emr-6.1.0	emr-6.0.1	emr-6.0.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	338	338	-	-
Zeppelin	0.9.0	0.9.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 6.0.0 的信息。

首次发布日期：2020 年 3 月 10 日

支持的应用程序

- Amazon SDK for Java 1.11.711
- Ganglia 3.7.2
- Hadoop 3.2.1
- HBase 2.2.3
- HCatalog 3.1.2
- Hive 3.1.2
- Hudi 0.5.0-incubating
- Hue 4.4.0
- JupyterHub 1.0.0
- Livy 0.6.0
- MXNet 1.5.1
- Oozie 5.1.0
- Phoenix 5.0.0
- Presto 0.230
- Spark 2.4.4

- TensorFlow 1.14.0
- Zeppelin 0.9.0-SNAPSHOT
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.14.0

> **Note**
>
> Flink、Sqoop、Pig 和 Mahout 在 Amazon EMR 6.0.0 中不可用。

新特征

- YARN Docker 运行时支持 - YARN 应用程序（例如 Spark 作业）现在可以在 Docker 容器的上下文中运行。这可让您轻松定义 Docker 镜像中的依赖项，而无需在 Amazon EMR 集群上安装自定义库。有关更多信息，请参阅配置 Docker 集成和使用 Amazon EMR 6.0.0 通过 Docker 运行 Spark 应用程序。
- Hive LLAP 支持 - Hive 现在支持 LLAP 执行模式以提高查询性能。有关更多信息，请参阅使用 Hive LLAP。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。

- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Amazon Linux
  - Amazon Linux 2 是 EMR 6.x 发布版本系列的操作系统。
  - 使用 `systemd` 进行服务管理，而 Amazon Linux 1 中使用的是 `upstart`。
- Java 开发工具包 (JDK)
  - Coretto JDK 8 是 EMR 6.x 版本系列的默认 JDK。
- Scala
  - Scala 2.12 与 Apache Spark 和 Apache Livy 一起使用。
- Python 3
  - Python 3 现在是 EMR 中的默认 Python 版本。
- YARN 节点标注
  - 从 Amazon EMR 6.x 发行版系列开始，默认情况下禁用 YARN 节点标注功能。默认情况下，应用程序主进程可以在核心节点和任务节点上运行。您可以通过配置以下属性来启用 YARN 节点标注功能：`yarn.node-labels.enabled` 和 `yarn.node-labels.am.default-node-label-expression`。有关更多信息，请参阅了解主节点、核心节点和任务节点。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低

ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
```

```
sudo systemctl daemon-reload
```

- Spark 交互式 shell（包括 PySpark、SparkR 和 spark-shell）不支持将 Docker 与其它库一起使用。
- 要在 Amazon EMR 6.0.0 中使用 Python 3，您必须在 `yarn.nodemanager.env-whitelist` 中添加 `PATH`。
- 使用 Amazon Glue 数据目录作为 Hive 的元存储时，不支持 Live Long and Process（LLAP）功能。
- 将 Amazon EMR 6.0.0 与 Spark 和 Docker 集成使用时，您需要使用同一实例类型和相同数量的 EBS 卷配置集群中的实例，以避免在使用 Docker 运行时提交 Spark 任务时出现故障。
- 在 Amazon EMR 6.0.0 中，HBASE-24286 问题会影响 HBase on Amazon S3 存储模式。使用现有 S3 数据创建集群时，无法初始化 HBase 主服务器。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 `hadoop` 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 `hadoop` 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.6	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	3.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.5.0	EMR S3 Select 连接器
emrfs	2.39.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	3.2.1-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	3.2.1-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	3.2.1-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	3.2.1-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	3.2.1-amzn-0	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	3.2.1-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	3.2.1-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	3.2.1-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	3.2.1-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	3.2.1-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	3.2.1-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	2.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	2.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	2.2.3	HBase 命令行客户端。
hbase-rest-server	2.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	2.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	3.1.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	3.1.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	3.1.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	3.1.2-amzn-0	Hive 命令行客户端。
hive-hbase	3.1.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	3.1.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	3.1.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.0-incubating-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。

组件	版本	描述
hudi-presto	0.5.0-incubating-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MariaDB 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	5.0.0-HBase-2.0	服务器和客户端的 phoenix 库
phoenix-query-server	5.0.0-HBase-2.0	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.230	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.230	用于执行查询的各个部分的服务。
presto-client	0.230	Presto 命令行客户端，安装在 HA 集群的备用主服务器（未启动 Presto 服务器）上。
r	3.4.3	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.41+	Apache HTTP 服务器。
zeppelin-server	0.9.0-SNAPSHOT	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-6.0.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-env	更改 HDFS 环境中的值。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive	适用于 Apache Hive 的 Amazon EMR 辅助设置。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。

分类	描述
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。

分类	描述
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

分类	描述
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 5.x 发行版

本部分内容涵盖每个 Amazon EMR 5.x 发行版中可用的应用程序版本、发布说明、组件版本和配置分类。

启动集群时，有多个 Amazon EMR 发行版可供选择。这允许您测试和使用满足您解决方案兼容性需求的应用程序版本。可使用发行版标注指定版本。发行版标注的格式是 `emr-x.x.x`。For example, `emr-6.14.0`.

从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

有关每个 Amazon EMR 5.x 发行版本中应用程序版本的综合表格，请参阅Amazon EMR 5.x 发行版中的应用程序版本。

主题

- Amazon EMR 发行版 5.30.1
- Amazon EMR 发行版 5.30.0
- Amazon EMR 发行版 5.29.0
- Amazon EMR 发行版 5.28.1
- Amazon EMR 发行版 5.28.0
- Amazon EMR 发行版 5.27.1
- Amazon EMR 发行版 5.27.0
- Amazon EMR 发行版 5.26.0
- Amazon EMR 发行版 5.25.0
- Amazon EMR 发行版 5.24.1
- Amazon EMR 发行版 5.24.0
- Amazon EMR 发行版 5.23.1
- Amazon EMR 发行版 5.23.0
- Amazon EMR 发行版 5.22.0
- Amazon EMR 发行版 5.21.2
- Amazon EMR 发行版 5.21.1
- Amazon EMR 发行版 5.21.0
- Amazon EMR 发行版 5.20.1
- Amazon EMR 发行版 5.20.0
- Amazon EMR 发行版 5.19.1
- Amazon EMR 发行版 5.19.0
- Amazon EMR 发行版 5.18.1
- Amazon EMR 发行版 5.18.0
- Amazon EMR 发行版 5.17.2
- Amazon EMR 发行版 5.17.1
- Amazon EMR 发行版 5.17.0
- Amazon EMR 发行版 5.16.1
- Amazon EMR 发行版 5.16.0
- Amazon EMR 发行版 5.15.1
- Amazon EMR 发行版 5.15.0

- Amazon EMR 发行版 5.14.2
- Amazon EMR 发行版 5.14.1
- Amazon EMR 发行版 5.14.0
- Amazon EMR 发行版 5.13.1
- Amazon EMR 发行版 5.13.0
- Amazon EMR 发行版 5.12.3
- Amazon EMR 发行版 5.12.2
- Amazon EMR 发行版 5.12.1
- Amazon EMR 发行版 5.12.0
- Amazon EMR 发行版 5.11.4
- Amazon EMR 发行版 5.11.3
- Amazon EMR 发行版 5.11.2
- Amazon EMR 发行版 5.11.1
- Amazon EMR 发行版 5.11.0
- Amazon EMR 发行版 5.10.1
- Amazon EMR 发行版 5.10.0
- Amazon EMR 发行版 5.9.1
- Amazon EMR 发行版 5.9.0
- Amazon EMR 发行版 5.8.3
- Amazon EMR 发行版 5.8.2
- Amazon EMR 发行版 5.8.1
- Amazon EMR 发行版 5.8.0
- Amazon EMR 发行版 5.7.1
- Amazon EMR 发行版 5.7.0
- Amazon EMR 发行版 5.6.1
- Amazon EMR 发行版 5.6.0
- Amazon EMR 发行版 5.5.4
- Amazon EMR 发行版 5.5.3
- Amazon EMR 发行版 5.5.2
- Amazon EMR 发行版 5.5.1

- Amazon EMR 发行版 5.5.0
- Amazon EMR 发行版 5.4.1
- Amazon EMR 发行版 5.4.0
- Amazon EMR 发行版 5.3.2
- Amazon EMR 发行版 5.3.1
- Amazon EMR 发行版 5.3.0
- Amazon EMR 发行版 5.2.3
- Amazon EMR 发行版 5.2.2
- Amazon EMR 发行版 5.2.1
- Amazon EMR 发行版 5.2.0
- Amazon EMR 发行版 5.1.1
- Amazon EMR 发行版 5.1.0
- Amazon EMR 发行版 5.0.3
- Amazon EMR 发行版 5.0.0

Amazon EMR 5.x 发行版中的应用程序版本

有关列出每个 Amazon EMR 5.x 发行版中可用的应用程序版本的综合表格，请在浏览器打开 Amazon EMR 5.x 发行版中的应用程序版本。

Amazon EMR 版本 5.36.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.36.1	emr-5.36.0	emr-5.35.0	emr-5.34.0
Amazon SDK for Java	1.12.206	1.12.206	1.12.159	1.11.970
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.15	2.12.10	不可用
Delta	-	-	-	-
Flink	1.14.2	1.14.2	1.14.2	1.13.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.9	2.3.9	2.3.9	2.3.8
Hadoop	2.10.1	2.10.1	2.10.1	2.10.1
Hive	2.3.9	2.3.9	2.3.9	2.3.8
Hudi	0.10.1-amzn-1	0.10.1-amzn-1	0.9.0-amzn-2	0.9.0-amzn-0
Hue	4.10.0	4.10.0	4.10.0	4.9.0
Iceberg	-	-	-	-

	emr-5.36.1	emr-5.36.0	emr-5.35.0	emr-5.34.0
JupyterEnterpriseGateway	2.6.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.4.1
Livy	0.7.1	0.7.1	0.7.1	0.7.1
MXNet	1.8.0	1.8.0	1.8.0	1.8.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.1	5.2.1	5.2.1	5.2.1
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.267	0.267	0.266	0.261
Spark	2.4.8	2.4.8	2.4.8	2.4.8
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.10.0	0.10.0	0.10.0	0.10.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 版本 5.36.1 的信息。更改与 5.36.0 有关。有关发布时间表的信息，请参阅更改日志。

更改、增强功能和解决的问题

- Amazon EMR 版本 5.36.1 增加了对在集群缩减期间将日志存档到 Amazon S3 的支持。在之前的 5.x 版本中，您只能在集群终止期间将日志文件存档到 Amazon S3。这项改进可确保即使在节点终止后，集群上生成的日志文件仍保留在 Amazon S3 上。有关更多信息，请参阅配置集群日志记录和调试。
- 5.36.1 版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 5.36.1 版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。
- 5.36.1 版本修复了主节点上的 Amazon EMR 进程守护程序会维护集群中已终止实例的过时元数据的问题。维护陈旧的数据可能会导致集群上的 CPU 和内存使用量无限增长，并最终导致集群故障。
- 对于使用多个主节点启动的集群，5.36.1 版本修复了其中一个主节点上的 Amazon EC2 硬件故障可能导致第二个主节点出现故障并导致集群不稳定的问题。
- 对于配置了传输中加密的集群，托管扩展现在支持 Spark shuffle 数据感知。Spark shuffle 数据是 Spark 跨分区重新分配以执行特定操作的数据。在缩减期间，托管扩展会忽略带有随机数据的实例。这样可以防止任务的重新尝试和重新计算，这些都会给价格和性能带来高昂的代价。有关随机排序操作的更多信息，请参阅 Spark 编程指南。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-

component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.16.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.5.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.21.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.7.0	EMR S3 Select 连接器
emrfs	2.51.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.14.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.10.1-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.9-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.9-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.9-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.9-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.9-amzn-2	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.9-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.9-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.10.1-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.10.1-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.10.1-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.13.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	11.0.194	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.267-amzn-1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.267-amzn-1	用于执行查询的各个部分的服务。
presto-client	0.267-amzn-1	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.8-amzn-2	Spark 命令行客户端。
spark-history-server	2.4.8-amzn-2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.4.8-amzn-2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.8-amzn-2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.36.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
hudi-defaults	更改 Hudi 的 hudi-defa ults.conf 文件中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhu b_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.js on 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.

分类	描述	重新配置操作
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.

分类	描述	重新配置操作
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server.
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.

分类	描述	重新配置操作
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-def aults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.p roperties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

更改日志

发行版 5.36.1 的更改日志和发布说明

日期	事件	描述
2023-07-26	更新	新的操作系统版本标签 `2.0.20230612.0` 和 `2.0.20230628.0` 。

日期	事件	描述
2023-05-25	部署完成	Amazon EMR 5.36.1 已全面部署到所有支持的区域
2023-05-09	文档发布	Amazon EMR 5.36.1 发布说明首次发布
2023-05-04	首次发布	Amazon EMR 5.36.1 首次面向部分商业区域部署

Amazon EMR 发行版 5.36.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.36.0	emr-5.35.0	emr-5.34.0	emr-5.33.1
Amazon SDK for Java	1.12.206	1.12.159	1.11.970	1.11.970
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.15	2.12.10	不可用	不可用
Delta	-	-	-	-
Flink	1.14.2	1.14.2	1.13.1	1.12.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.9	2.3.9	2.3.8	2.3.7
Hadoop	2.10.1	2.10.1	2.10.1	2.10.1
Hive	2.3.9	2.3.9	2.3.8	2.3.7
Hudi	0.10.1-amzn-1	0.9.0-amzn-2	0.9.0-amzn-0	0.7.0-amzn-1
Hue	4.10.0	4.10.0	4.9.0	4.9.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.4.1	1.2.2
Livy	0.7.1	0.7.1	0.7.1	0.7.0
MXNet	1.8.0	1.8.0	1.8.0	1.7.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.36.0	emr-5.35.0	emr-5.34.0	emr-5.33.1
Oozie	5.2.1	5.2.1	5.2.1	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.267	0.266	0.261	0.245.1
Spark	2.4.8	2.4.8	2.4.8	2.4.7
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.10.0	0.10.0	0.10.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.36.0 的信息。更改与 5.35.0 有关。

首次发布日期：2022 年 6 月 15 日

新功能

- Amazon EMR 版本 5.36.0 在启用 Apache Ranger 的集群上，通过 Apache Spark 增加了对数据定义语言（DDL）的支持。这样，您就能够使用 Apache Ranger 管理操作的访问权限，例如创建、更改和删除 Amazon EMR 集群中的数据库和表。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (6.8.*1*) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 10 月 7 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2022 719.0	4.14.287	2022 年 8 月 10 日	美国西部（北加利福尼亚）、欧洲地区（巴黎）、欧洲地区（斯德哥尔摩）、欧洲地区（法兰克福）、亚太地区（孟买）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 426.0	4.14.281	2022 年 6 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

更改、增强和解决的问题

- Amazon EMR 5.36.0 升级现在支持：aws-sdk 1.12.206、Hadoop 2.10.1-amzn-4、Hive 2.3.9-amzn-2、Hudi 0.10.1-amzn-1、Spark 2.4.8-amzn-2、Presto 0.267-amzn-1、Amazon Glue 连接器 1.18.0、EMRFS 2.51.0。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.16.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.5.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.21.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.7.0	EMR S3 Select 连接器
emrfs	2.51.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.14.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.10.1-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.9-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.9-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.9-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.9-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.9-amzn-2	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.9-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.9-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.10.1-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.10.1-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.10.1-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.13.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	11.0.194	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.267-amzn-1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.267-amzn-1	用于执行查询的各个部分的服务。
presto-client	0.267-amzn-1	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.8-amzn-2	Spark 命令行客户端。
spark-history-server	2.4.8-amzn-2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.4.8-amzn-2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.8-amzn-2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.36.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server.

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.

分类	描述	重新配置操作
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-def aults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.p roperties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.35.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、Iceberg、JupyterEnterpriseGateway 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.35.0	emr-5.34.0	emr-5.33.1	emr-5.33.0
Amazon SDK for Java	1.12.159	1.11.970	1.11.970	1.11.970
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	2.12.10	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.14.2	1.13.1	1.12.1	1.12.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.9	2.3.8	2.3.7	2.3.7
Hadoop	2.10.1	2.10.1	2.10.1	2.10.1
Hive	2.3.9	2.3.8	2.3.7	2.3.7
Hudi	0.9.0-amzn-2	0.9.0-amzn-0	0.7.0-amzn-1	0.7.0-amzn-1

	emr-5.35.0	emr-5.34.0	emr-5.33.1	emr-5.33.0
Hue	4.10.0	4.9.0	4.9.0	4.9.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.4.1	1.2.2	1.2.2
Livy	0.7.1	0.7.1	0.7.0	0.7.0
MXNet	1.8.0	1.8.0	1.7.0	1.7.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.1	5.2.1	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.266	0.261	0.245.1	0.245.1
Spark	2.4.8	2.4.8	2.4.7	2.4.7
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.4.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.10.0	0.10.0	0.9.0	0.9.0
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

这是 Amazon EMR 发行版 5.35.0 的发布说明。

以下发布说明包括有关 Amazon EMR 发行版 5.35.0 的信息。更改与 5.34.0 有关。

首次发布日期: 2022 年 3 月 30 日

新功能

- 使用 log4J 1.x 和 log4J 2.x 的 Amazon EMR 发行版 5.35 应用程序将分别升级为使用 log4J 1.2.17（或更高版本）和 log4J 2.17.1（或更高版本），并且不需要使用引导操作来缓解之前发行版中的 CVE 问题。请参阅 缓解 CVE-2021-44228 的方法。

更改、增强和解决的问题

Flink 更改

更改类型	描述
升级	• 将 Flink 版本更新到 1.14.2。 • log4j 升级到 2.17.1。

Hadoop 更改

更改类型	描述
自 EMR 5.34.0 以来的 Hadoop 开源逆向移植	• YARN-10438: 处理 ClientRMService#getContainerReport() 中的空 containerId • YARN-7266: 时间轴服务器事件处理程序线程已锁定 • YARN-10438：如果 RollingLevelDb 文件损坏或丢失，ATS 1.5 将无法开启 • HADOOP-13500: 同步配置属性对象的迭代 • YARN-10651: CapacityScheduler 由于 AbstractYarnScheduler.updateNodeResource() 中的 NPE 崩溃

更改类型	描述
	• HDFS-12221: 替换 XmlEditsVisitor 中的 xerces • HDFS-16410: OfflineEditsXmlLoader 中不安全的 Xml 解析
Hadoop 更改和修复	• KMS 和 HttpFS 中使用的 Tomcat 升级到 8.5.75 • 在 FileSystemOptimizedCommitterV2 中，成功标记被写入创建提交程序时定义的 commitJob 输出路径。由于 commitJob 和任务级别输出路径可能不同，因此更正路径以使用清单文件中定义的路径。对于 Hive 任务，这会导致在执行动态分区或 UNION ALL 等操作时正确写入成功标记。

Hive 更改

更改类型	描述
Hive 升级到开源发行版 2.3.9，包括这些 JIRA 修复	• HIVE-17155: HiveConf.java 中的 findConfFile() 存在一些配置路径问题 • HIVE-24797: 在解析 Avro 架构时禁用验证原定设置值 • HIVE-21563: 通过禁用 registerAllFunctionsOnce 提升 Table#getEmptyTable 性能 • HIVE-18147: 测试可能失败，显示 java.net.BindException: 地址已在使用中 • HIVE-24608: 切换回 Hive 2.3.x HMS 客户端中的 get_table • HIVE-21200: 向量化 - 日期列显示 java.lang.UnsupportedOperationException for parquet • HIVE-19228: 删除 commons-httpclient 3.x 使用

更改类型	描述
自 EMR 5.34.0 以来的 Hive 开源逆向移植	• HIVE-19990: 在联接条件下使用时间间隔文本查询失败 • HIVE-25824: 将 branch-2.3 升级到 log4j 2.17.0 • TEZ-4062: 推测性尝试计划应在任务完成时中止 • TEZ-4108: 在推测性执行竞争条件期间出现 NullPointerException • TEZ-3918: 设置项 tez.task.log.level 无效
Hive 升级和修复	• 将 Log4j 版本升级到 2.17.1 • 将 ORC 版本升级到 1.4.3 • 修复了由于 ShuffleScheduler 中的惩罚线程导致的死锁
新特征	• 添加了在 AM 日志中打印 Hive 查询的功能 默认情况下，将禁用该功能。标记/配置 : `tez.am.emr.print.hive.query.in.log` 。状态（原定设置）: FALSE。

Oozie 更改

更改类型	描述
自 EMR 5.34.0 以来的 Oozie 开源逆向移植	• OOZIE-3652: 当发生 NoSuchFileException 时，Oozie 启动器应重试目录列表

Pig 更改

更改类型	描述
升级	• log4j 升级到 1.2.17。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.15.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.5.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.20.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.7.0	EMR S3 Select 连接器
emrfs	2.49.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.14.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.14.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.10.1-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-3	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.1-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.9-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.9-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.9-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.9-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.9-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.9-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.9-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.9.0-amzn-2	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.9.0-amzn-2	用于运行 Spark 以及 Hudi 的捆绑库。

组件	版本	描述
hudi-presto	0.9.0-amzn-2	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.10.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.13.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.266-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.266-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.266-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.8-amzn-1	Spark 命令行客户端。
spark-history-server	2.4.8-amzn-1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.8-amzn-1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.8-amzn-1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。

组件	版本	描述
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.35.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode,

分类	描述	重新配置操作
		SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.

分类	描述	重新配置操作
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.

分类	描述	重新配置操作
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-log4j2	更改 Hive 的 hive-log4 j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server.
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.34.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.34.0	emr-5.33.1	emr-5.33.0	emr-5.32.1
Amazon SDK for Java	1.11.970	1.11.970	1.11.970	1.11.890
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.13.1	1.12.1	1.12.1	1.11.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.8	2.3.7	2.3.7	2.3.7
Hadoop	2.10.1	2.10.1	2.10.1	2.10.1
Hive	2.3.8	2.3.7	2.3.7	2.3.7
Hudi	0.9.0-amzn-0	0.7.0-amzn-1	0.7.0-amzn-1	0.6.0-amzn-0
Hue	4.9.0	4.9.0	4.9.0	4.8.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.4.1	1.2.2	1.2.2	1.1.0
Livy	0.7.1	0.7.0	0.7.0	0.7.0
MXNet	1.8.0	1.7.0	1.7.0	1.7.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.34.0	emr-5.33.1	emr-5.33.0	emr-5.32.1
Oozie	5.2.1	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.261	0.245.1	0.245.1	0.240.1
Spark	2.4.8	2.4.7	2.4.7	2.4.7
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.4.1	2.3.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.10.0	0.9.0	0.9.0	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.34.0 的信息。更改与 5.33.1 有关。

首次发布日期：2022 年 1 月 20 日

发布更新日期：2022 年 3 月 21 日

新功能

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管扩展 和 Spark 编程指南。
- [Hudi] 简化了 Hudi 配置的改进。预设情况下禁用乐观并发控制。

更改、增强和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 以前，在多主节点集群上手动重启资源管理器会导致 Zookeeper znode 文件中的 Amazon EMR 集群进程守护程序（如 Zookeeper）重新加载以前停用或丢失的所有节点。在某些情况下，这会导致超出默认限制。Amazon EMR 现在会从 Zookeeper 文件中删除已停用或丢失超过一小时的节点记录，并且内部限制也有所提高。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- Zeppelin 已升级到版本 0.10.0。
- Livy 修复 - 已升级到 0.7.1
- Spark 性能提升 - 当 EMR 5.34.0 中的某些 Spark 配置值被覆盖时禁用异构执行器。
- 默认情况下禁用 WebHDFS 和 HTTFS 服务器。您可以使用 Hadoop 配置重新启用 WebHDFS，`dfs.webhdfs.enabled`。HTTPFS 服务器可以通过使用 `sudo systemctl start hadoop-httpfs` 启动。

已知问题

- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。

- Hue 查询在 Amazon EMR 6.4.0 中不起作用，因为默认情况下 Apache Hadoop HTTPFS 服务器处于禁用状态。要在 Amazon EMR 6.4.0 上使用 Hue，请使用 `sudo systemctl start hadoop-httpfs` 在 Amazon EMR 主节点上手动启动 HTTPFS 服务器，或者使用 Amazon EMR 步骤。
- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	emrfs	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.4.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.7.0	EMR S3 Select 连接器
emrfs	2.48.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.13.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.13.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-2	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.1-amzn-2	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.10.1-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.8-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.8-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.8-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.8-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.8-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.8-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.8-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.9.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.9.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.9.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.4.1	Jupyter notebook 的多用户服务器
livy-server	0.7.1-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.13.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.8.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.1	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.1	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.261-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.261-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.261-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.8-amzn-0	Spark 命令行客户端。
spark-history-server	2.4.8-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.8-amzn-0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.4.8-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.10.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.34.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.

分类	描述	重新配置操作
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.prop erties 中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server.
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.33.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.33.1	emr-5.33.0	emr-5.32.1	emr-5.32.0
Amazon SDK for Java	1.11.970	1.11.970	1.11.890	1.11.890
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.12.1	1.12.1	1.11.2	1.11.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.7	2.3.7	2.3.7
Hadoop	2.10.1	2.10.1	2.10.1	2.10.1
Hive	2.3.7	2.3.7	2.3.7	2.3.7
Hudi	0.7.0-amzn-1	0.7.0-amzn-1	0.6.0-amzn-0	0.6.0-amzn-0
Hue	4.9.0	4.9.0	4.8.0	4.8.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	2.1.0
JupyterHub	1.2.2	1.2.2	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.7.0	1.7.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.33.1	emr-5.33.0	emr-5.32.1	emr-5.32.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.245.1	0.245.1	0.240.1	0.240.1
Spark	2.4.7	2.4.7	2.4.7	2.4.7
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.4.1	2.3.1	2.3.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.9.0	0.9.0	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.33.0/5.33.1 的信息。更改与 5.32.0 有关。

首次发布日期：2021 年 4 月 19 日

上次更新日期：2021 年 8 月 9 日

升级

- 已将 Amazon Glue 连接器升级到版本 1.15.0
- 已将 Amazon SDK for Java 升级到版本 1.11.970
- 已将 EMRFS 升级到版本 2.46.0
- 已将 EMR Goodies 升级到版本 2.14.0
- 已将 EMR 记录服务器升级到版本 1.9.0

- 已将 EMR S3 Dist CP 升级到版本 2.18.0
- 已将 EMR Secret Agent 升级到版本 1.8.0
- 已将 Flink 升级到版本 1.12.1
- 已将 Hadoop 升级到版本 2.10.1-amzn-1
- 已将 Hive 升级到版本 2.3.7-amzn-4
- 已将 Hudi 升级到版本 0.7.0
- 已将 Hue 升级到版本 4.9.0
- 已将 OpenCV 升级到版本 4.5.0
- 已将 Presto 升级到版本 0.245.1-amzn-0
- 已将 R 升级到版本 4.0.2
- 已将 Spark 升级到版本 2.4.7-amzn-1
- 已将 TensorFlow 升级到版本 2.4.1
- 已将 Zeppelin 升级到版本 0.9.0

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。

- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 配置集群以修复 Apache YARN 时间轴服务器 1 和 1.5 版的性能问题

  Apache YARN 时间轴服务器版本 1 和 1.5 可能会对非常活跃的大型 EMR 集群造成性能问题，尤其是 `yarn.resourcemanager.system-metrics-publisher.enabled=true`，这是 Amazon EMR 中的默认设置。开源 YARN 时间轴服务器 v2 解决了与 YARN 时间轴服务器可扩展性相关的性能问题。

  此问题的其他解决方法包括：

  - 配置 yarn.资源管理器.系统指标-发布者.启用=false 在 yarn-site.xml 中。
  - 如下所述，在创建群集时启用此问题的修复程序。

  以下 Amazon EMR 发行版包含针对此 YARN 时间线服务器性能问题的修复。

  EMR 5.30.2、5.31.1、5.32.1、5.33.1、5.34.x、6.0.1、6.1.1、6.2.1、6.3.1、6.4.x

  要对上述任何指定的 Amazon EMR 版本启用修复程序，请使用 `aws emr create-cluster` 命令参数：`--configurations file://./configurations.json` 在传入的配置 JSON 文件中将这些属性设置为 `true`。或者使用重新配置控制台 UI 启用修复程序。

  配置 .json 文件内容的示例：

```
[
{
"Classification": "yarn-site",
"Properties": {
"yarn.resourcemanager.system-metrics-publisher.timeline-server-v1.enable-batch":
 "true",
"yarn.resourcemanager.system-metrics-publisher.enabled": "true"
},
"Configurations": []
}
]
```

- 现在，从 Hive 元存储中获取分区位置进行 Spark 插入查询时，Spark 运行时的速度更快。
- 升级了组件版本。有关组件版本的列表，请参阅本指南中的关于 Amazon EMR 发行版。

- 在每个新集群上已安装了Amazon Java SDK Bundle。这是一个包含所有服务 SDK 及其依赖项的单个 jar，而不是单个组件 jar。有关更多信息，请参阅 Java SDK Bundled Dependency。
- 修复了早期 Amazon EMR 发行版中的托管扩展问题，并对托管扩展进行了改进，从而显著降低了应用程序故障率。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

新特征

- Amazon EMR 支持 Amazon S3 接入点，这是 Amazon S3 的一项功能，可让您轻松管理共享数据湖的访问。使用 Amazon S3 接入点别名，您可以在 Amazon EMR 上大规模简化数据访问。您可以将 Amazon S3 接入点与所有版本的 Amazon EMR 一起使用，在 Amazon EMR 可用的所有 Amazon 区域无需额外费用。要了解有关 Amazon S3 访问点和访问点别名的详细信息，请参阅《Amazon S3 用户指南》中的为接入点使用存储桶式别名。
- Amazon EMR-5.33 支持新的 Amazon EC2 实例类型：c5a、c5ad、c6gn、c6gd、m6gd、d3、d3en、m5zn、r5b、r6gd。请参阅支持的实例类型。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 /etc/systemd/system/instance-controller.service，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- 对于 Amazon EMR 6.3.0 和 6.2.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 sudo

`systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。

- > **Important**
  >
  > 运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- GovCloud 区域中目前不支持使用控制台创建指定 Amazon Ranger 集成选项的安全配置。可以使用 CLI 完成安全配置。请参阅《Amazon EMR 管理指南》中的创建 EMR 安全配置。
- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	emrfs	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.2.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器
emrfs	2.46.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.12.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.12.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-1.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-1.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-1.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-1.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-1.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-1.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-1.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-1.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.10.1-amzn-1.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-1.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-1.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.7-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.7-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-4	Hive 命令行客户端。
hive-hbase	2.3.7-amzn-4	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.7-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.7-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.7.0-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.7.0-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.7.0-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.2.2	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.245.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.245.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.245.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.7-amzn-1.1	Spark 命令行客户端。
spark-history-server	2.4.7-amzn-1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.4.7-amzn-1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.7-amzn-1.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.9.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.33.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server.

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.

分类	描述	重新配置操作
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-def aults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.p roperties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.33.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.33.0	emr-5.32.1	emr-5.32.0	emr-5.31.1
Amazon SDK for Java	1.11.970	1.11.890	1.11.890	1.11.852
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.12.1	1.11.2	1.11.2	1.11.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.7	2.3.7	2.3.7
Hadoop	2.10.1	2.10.1	2.10.1	2.10.0
Hive	2.3.7	2.3.7	2.3.7	2.3.7
Hudi	0.7.0-amzn-1	0.6.0-amzn-0	0.6.0-amzn-0	0.6.0-amzn-0

	emr-5.33.0	emr-5.32.1	emr-5.32.0	emr-5.31.1
Hue	4.9.0	4.8.0	4.8.0	4.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	2.1.0	-
JupyterHub	1.2.2	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.7.0	1.6.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.245.1	0.240.1	0.240.1	0.238.3
Spark	2.4.7	2.4.7	2.4.7	2.4.6
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.4.1	2.3.1	2.3.1	2.1.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.9.0	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	emrfs	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.2.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.18.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器

组件	版本	描述
emrfs	2.46.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.12.1	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.12.1	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.10.1-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.1-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.7-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.7-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-4	Hive 命令行客户端。
hive-hbase	2.3.7-amzn-4	Hive-hbase 客户端。
hive-metastore-server	2.3.7-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.7-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.7.0-amzn-1	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.7.0-amzn-1	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.7.0-amzn-1	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.9.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.2.2	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.5.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.245.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.245.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.245.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。

组件	版本	描述
r	4.0.2	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.7-amzn-1	Spark 命令行客户端。
spark-history-server	2.4.7-amzn-1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.7-amzn-1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.7-amzn-1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.4.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.9.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.33.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-session	为 Kubernetes/Yarn 会话更改 Flink log4j-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.prop erties 中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server.
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zeppelin-site	更改 zeppelin-site.xml 中的配置设置。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.32.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.32.1	emr-5.32.0	emr-5.31.1	emr-5.31.0
Amazon SDK for Java	1.11.890	1.11.890	1.11.852	1.11.852
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.11.2	1.11.2	1.11.0	1.11.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.7	2.3.7	2.3.7
Hadoop	2.10.1	2.10.1	2.10.0	2.10.0
Hive	2.3.7	2.3.7	2.3.7	2.3.7
Hudi	0.6.0-amzn-0	0.6.0-amzn-0	0.6.0-amzn-0	0.6.0-amzn-0
Hue	4.8.0	4.8.0	4.7.1	4.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	2.1.0	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.7.0	1.6.0	1.6.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.32.1	emr-5.32.0	emr-5.31.1	emr-5.31.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.240.1	0.240.1	0.238.3	0.238.3
Spark	2.4.7	2.4.7	2.4.6	2.4.6
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.3.1	2.3.1	2.1.0	2.1.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。

- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.1.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.17.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器
emrfs	2.45.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
flink-client	1.11.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.11.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-0.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-0.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-0.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-0.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-0.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-0.1	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.10.1-amzn-0.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-0.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.1-amzn-0.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.1-amzn-0.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-0.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.7-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	2.3.7-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-3	Hive 命令行客户端。
hive-hbase	2.3.7-amzn-3	Hive-hbase 客户端。
hive-metastore-server	2.3.7-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.7-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.6.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.6.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.8.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。

组件	版本	描述
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68+	MySQL 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.240.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.240.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.240.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统

组件	版本	描述
spark-client	2.4.7-amzn-0.1	Spark 命令行客户端。
spark-history-server	2.4.7-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.7-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.7-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.3.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.32.1 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager,

分类	描述	重新配置操作
		NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.

分类	描述	重新配置操作
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。	Not available.

分类	描述	重新配置操作
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。	Restarts Hive HCatalog Server.
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.

分类	描述	重新配置操作
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.

分类	描述	重新配置操作
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log 4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-me trics2-hbase.properties 文件中的值。	Not available.
phoenix-hbase-site	更改 Phoenix 的 hbase-sit e.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.pro perties 文件中的值。	Restarts Phoenix-QueryServe r.
phoenix-metrics	更改 Phoenix 的 hadoop-me trics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.pr operties 文件中的值。	Restarts Presto-Server.

分类	描述	重新配置操作
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.

分类	描述	重新配置操作
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-def aults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.p roperties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.

分类	描述	重新配置操作
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.32.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterEnterpriseGateway、Jupyter 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.32.0	emr-5.31.1	emr-5.31.0	emr-5.30.2
Amazon SDK for Java	1.11.890	1.11.852	1.11.852	1.11.759
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.11.2	1.11.0	1.11.0	1.10.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.7	2.3.7	2.3.6
Hadoop	2.10.1	2.10.0	2.10.0	2.8.5
Hive	2.3.7	2.3.7	2.3.7	2.3.6
Hudi	0.6.0-amzn-0	0.6.0-amzn-0	0.6.0-amzn-0	0.5.2-incubating

	emr-5.32.0	emr-5.31.1	emr-5.31.0	emr-5.30.2
Hue	4.8.0	4.7.1	4.7.1	4.6.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	2.1.0	-	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.7.0	1.6.0	1.6.0	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.240.1	0.238.3	0.238.3	0.232
Spark	2.4.7	2.4.6	2.4.6	2.4.5
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.3.1	2.1.0	2.1.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.32.0 的信息。更改与 5.31.0 有关。

首次发布日期：2021 年 1 月 8 日

升级

- 已将 Amazon Glue 连接器升级到1.14.0
- 已将 Amazon SageMaker Spark SDK 升级到版本 1.4.1
- 已将 Amazon SDK for Java 升级到版本 1.11.890
- 已将 EMR DynamoDB 连接器升级到版本 4.16.0
- 已将 EMRFS 升级到版本 2.45.0
- 已将 EMR Log Analytics Metrics 升级到版本 1.18.0
- 已将 EMR MetricsAndEventsApiGateway 客户端升级到版本 1.5.0
- 已将 EMR 记录服务器升级到版本 1.8.0
- 已将 EMR S3 Dist CP 升级到版本 2.17.0
- 已将 EMR Secret Agent 升级到版本 1.7.0
- 已将 Flink 升级到版本 1.11.2
- 已将 Hadoop 升级到版本 2.10.1-amzn-0
- 已将 Hive 升级到版本 2.3.7-amzn-3
- 已将 Hue 升级到版本 4.8.0
- 已将 Mxnet 升级到版本 1.7.0
- 已将 OpenCV 升级到版本 4.4.0
- 已将 Presto 升级到版本 0.240.1-amzn-0
- 已将 Spark 升级到版本 2.4.7-amzn-0
- 已将 TensorFlow 升级到版本 2.3.1

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 升级了组件版本。
- 有关组件版本的列表，请参阅本指南中的关于 Amazon EMR 发行版。

新特征

- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 `spark.yarn.heterogeneousExecutors.enabled` 配置参数。
- 实例元数据服务（IMDS）V2 支持状态：Amazon EMR 5.23.1、5.27.1 和 5.32 或更高版本的组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。对于其它 5.x EMR 版本，禁用 IMDSv1 会导致集群启动失败。
- 从 Amazon EMR 5.32.0 开始，您可以启动与 Apache Ranger 在本地集成的集群。Apache Ranger 是一个开源框架，可跨 Hadoop 平台启用、监控和管理全面的数据安全。有关更多信息，请参阅 Apache Ranger。通过本机集成，您可以自带 Apache Ranger，在 Amazon EMR 上强制实施精细数据访问控制。请参阅《Amazon EMR 版本指南》中的 将 Amazon EMR 与 Apache Ranger 集成。

- Amazon EMR 发行版 5.32.0 支持 Amazon EMR on EKS。有关 EMR on EKS 入门的更多详细信息，请参阅什么是 Amazon EMR on EKS。
- Amazon EMR 发行版 5.32.0 版支持 Amazon EMR Studio（预览版）。有关 EMR Studio 入门的更多详细信息，请参阅 Amazon EMR Studio（预览版）。
- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。

已知问题

- 对于 Amazon EMR 6.3.0 和 6.2.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 `sudo systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。
- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

2. 重新启动 InstanceController

```
$ sudo systemctl daemon-reload

$ sudo systemctl restart instance-controller
```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- **Important**

  运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- GovCloud 区域中目前不支持使用控制台创建指定 Amazon Ranger 集成选项的安全配置。可以使用 CLI 完成安全配置。请参阅《Amazon EMR 管理指南》中的创建 EMR 安全配置。
- 在使用 Amazon EMR 5.31.0 或 5.32.0 的集群上启用了 AtRestEncryption 或 HDFS 加密时，Hive 查询会导致以下运行时系统异常。

```
TaskAttempt 3 failed, info=[Error: Error while running task ( failure ) :
 attempt_1604112648850_0001_1_01_000000_3:java.lang.RuntimeException:
 java.lang.RuntimeException: Hive Runtime Error while closing
 operators: java.io.IOException: java.util.ServiceConfigurationError:
 org.apache.hadoop.security.token.TokenIdentifier: Provider
 org.apache.hadoop.hbase.security.token.AuthenticationTokenIdentifier not found
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a  b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a  b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.16.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.1.0	适用于 EMR Notebooks（可提供 jupyter 企业网关）的 Conda env
emr-s3-dist-cp	2.17.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器
emrfs	2.45.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.2	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.11.2	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.1-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.1-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.1-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.1-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.1-amzn-0	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.1-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.1-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.1-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.1-amzn-0	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.10.1-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.1-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.7-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.7-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-3	Hive 命令行客户端。
hive-hbase	2.3.7-amzn-3	Hive-hbase 客户端。
hive-metastore-server	2.3.7-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.7-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.6.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.6.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.8.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.7.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.68	MySQL 数据库服务器。
nvidia-cuda	10.1.243	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.240.1-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.240.1-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.240.1-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.7-amzn-0	Spark 命令行客户端。
spark-history-server	2.4.7-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.7-amzn-0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.4.7-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.3.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

为处于运行状态的集群中的实例组指定配置时，将发生重新配置操作。Amazon EMR 仅为您修改的分类启动重新配置操作。有关更多信息，请参阅在正在运行的集群中重新配置实例组。

emr-5.32.0 分类

分类	描述	重新配置操作
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。	Restarts the ResourceManager service.

分类	描述	重新配置操作
container-executor	更改 Hadoop YARN 的 container-executor.cfg 文件中的值。	Not available.
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。	Not available.
core-site	更改 Hadoop 的 core-site.xml 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Ranger KMS, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
docker-conf	更改 docker 相关设置。	Not available.

分类	描述	重新配置操作
emrfs-site	更改 EMRFS 设置。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts HBaseRegionserver, HBaseMaster, HBaseThrift, HBaseRest, HiveServer2, Hive MetaStore, Hadoop Httpfs, and MapReduce-HistoryServer.
flink-conf	更改 flink-conf.yaml 设置。	Restarts Flink history server.
flink-log4j	更改 Flink log4j.properties 设置。	Restarts Flink history server.
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。	Not available.
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。	Restarts Flink history server.

分类	描述	重新配置操作
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts PhoenixQueryserver, HiveServer2, Hive MetaStore, and MapReduce-HistoryServer.
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Hadoop KMS, Hadoop Httpfs, and MapReduce-HistoryServer.
hadoop-ssl-server	更改 hadoop ssl 服务器配置	Not available.
hadoop-ssl-client	更改 hadoop ssl 客户端配置	Not available.
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。	Custom EMR specific property. Sets emrfs-site and hbase-site configs. See those for their associated restarts.

分类	描述	重新配置操作
hbase-env	更改 HBase 环境中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-log4j	更改 HBase 的 hbase-log 4j.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-metrics	更改 HBase 的 hadoop-me trics2-hbase.properties 文件中的值。	Restarts the HBase services RegionServer, HBaseMaster, ThriftServer, RestServer.
hbase-policy	更改 HBase 的 hbase-pol icy.xml 文件中的值。	Not available.
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。	Restarts the HBase services RegionServer, HBaseMast er, ThriftServer, RestServer. Additionally restarts Phoenix QueryServer.
hdfs-encryption-zones	配置 HDFS 加密区域。	Should not be reconfigured.
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。	Restarts the Hadoop HDFS services Namenode, SecondaryNamenode, Datanode, ZKFC, and Journalnode. Additionally restarts Hadoop Httpfs.
hcatalog-env	更改 HCatalog 的环境中的值。	Restarts Hive HCatalog Server.
hcatalog-server-jndi	更改 HCatalog 的 jndi.prop erties 中的值。	Restarts Hive HCatalog Server.

分类	描述	重新配置操作
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。	Restarts Hive HCatalog Server.
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。	Restarts Hive WebHCat Server.
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。	Restarts Hive WebHCat Server.
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。	Not available.
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。	Not available.
hive-env	更改 Hive 环境中的值。	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore.
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。	Restarts HiveServer2 and HiveMetastore.
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。	Not available.
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。	Not available.

分类	描述	重新配置操作
hive-site	更改 Hive 的 hive-site.xml 文件中的值	Restarts HiveServer2 and HiveMetastore. Runs Hive schemaTool CLI commands to verify hive-metastore. Also restarts Oozie and Zeppelin.
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值	Not available.
hue-ini	更改 Hue 的 ini 文件中的值	Restarts Hue. Also activates Hue config override CLI commands to pick up new configurations.
httpfs-env	更改 HTTPFS 环境中的值。	Restarts Hadoop Httpfs service.
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。	Restarts Hadoop Httpfs service.
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。	Not available.
hadoop-kms-env	更改 Hadoop KMS 环境中的值。	Restarts Hadoop-KMS service.
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j .properties 文件中的值。	Not available.
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。	Restarts Hadoop-KMS and Ranger-KMS service.
hudi-env	更改 Hudi 环境中的值。	Not available.

分类	描述	重新配置操作
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。	Not available.
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。	Not available.
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。	Not available.
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。	Not available.
livy-conf	更改 Livy 的 livy.conf 文件中的值。	Restarts Livy Server.
livy-env	更改 Livy 环境中的值。	Restarts Livy Server.
livy-log4j	更改 Livy log4j.properties 设置。	Restarts Livy Server.
mapred-env	更改 MapReduce 应用程序的环境中的值。	Restarts Hadoop MapReduce-HistoryServer.
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。	Restarts Hadoop MapReduce-HistoryServer.
oozie-env	更改 Oozie 的环境中的值。	Restarts Oozie.
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。	Restarts Oozie.
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。	Restarts Oozie.
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。	Not available.

分类	描述	重新配置操作
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。	Not available.
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。	Restarts Phoenix-QueryServer.
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。	Not available.
pig-env	更改 Pig 环境中的值。	Not available.
pig-properties	更改 Pig 的 pig.properties 文件中的值。	Restarts Oozie.
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。	Not available.
presto-log	更改 Presto 的 log.properties 文件中的值。	Restarts Presto-Server.
presto-config	更改 Presto 的 config.properties 文件中的值。	Restarts Presto-Server.
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。	Not available.
presto-env	更改 Presto 的 presto-env.sh 文件中的值。	Restarts Presto-Server.
presto-node	更改 Presto 的 node.properties 文件中的值。	Not available.
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-cassandra	更改 Presto 的 cassandra .properties 文件中的值。	Not available.
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。	Restarts Presto-Server.
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。	Not available.
presto-connector-kafka	更改 Presto 的 kafka.pro perties 文件中的值。	Not available.
presto-connector-localfile	更改 Presto 的 localfile .properties 文件中的值。	Not available.
presto-connector-memory	更改 Presto 的 memory.pr operties 文件中的值。	Not available.
presto-connector-mongodb	更改 Presto 的 mongodb.p roperties 文件中的值。	Not available.
presto-connector-mysql	更改 Presto 的 mysql.pro perties 文件中的值。	Not available.
presto-connector-postgresql	更改 Presto 的 postgresq l.properties 文件中的值。	Not available.
presto-connector-raptor	更改 Presto 的 raptor.pr operties 文件中的值。	Not available.
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。	Not available.
presto-connector-redshift	更改 Presto 的 redshift. properties 文件中的值。	Not available.
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。	Not available.

分类	描述	重新配置操作
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。	Not available.
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。	Restarts Ranger KMS Server.
ranger-kms-env	更改 Ranger KMS 环境中的值。	Restarts Ranger KMS Server.
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。	Not available.
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。	Not available.
recordserver-env	更改 EMR RecordServer 环境中的值。	Restarts EMR record server.
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。	Restarts EMR record server.
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。	Restarts EMR record server.
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。	This property modifies spark-defaults. See actions there.
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-env	更改 Spark 环境中的值。	Restarts Spark history server and Spark thrift server.

分类	描述	重新配置操作
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值	Not available.
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。	Restarts Spark history server and Spark thrift server.
sqoop-env	更改 Sqoop 的环境中的值。	Not available.
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。	Not available.
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。	Not available.
tez-site	更改 Tez 的 tez-site.xml 文件中的值。	Restarts Oozie.
yarn-env	更改 YARN 环境中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts MapReduce-HistoryServer.
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。	Restarts the Hadoop YARN services ResourceManager, NodeManager, ProxyServer, and TimelineServer. Additionally restarts Livy Server and MapReduce-HistoryServer.
zeppelin-env	更改 Zeppelin 环境中的值。	Restarts Zeppelin.

分类	描述	重新配置操作
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。	Restarts Zookeeper server.
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。	Restarts Zookeeper server.

Amazon EMR 发行版 5.31.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahout 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.31.1	emr-5.31.0	emr-5.30.2	emr-5.30.1
Amazon SDK for Java	1.11.852	1.11.852	1.11.759	1.11.759

	emr-5.31.1	emr-5.31.0	emr-5.30.2	emr-5.30.1
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.11.0	1.11.0	1.10.0	1.10.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.7	2.3.6	2.3.6
Hadoop	2.10.0	2.10.0	2.8.5	2.8.5
Hive	2.3.7	2.3.7	2.3.6	2.3.6
Hudi	0.6.0-amzn-0	0.6.0-amzn-0	0.5.2-incubating	0.5.2-incubating
Hue	4.7.1	4.7.1	4.6.0	4.6.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.6.0	1.6.0	1.5.1	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0

	emr-5.31.1	emr-5.31.0	emr-5.30.2	emr-5.30.1
Presto	0.238.3	0.238.3	0.232	0.232
Spark	2.4.6	2.4.6	2.4.5	2.4.5
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.1.0	2.1.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。

- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.15.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.15.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器
emrfs	2.43.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.0	Apache Flink 命令行客户端脚本和应用程序。
flink-jobmanager-config	1.11.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.0-amzn-0.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.0-amzn-0.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.0-amzn-0.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.0-amzn-0.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.0-amzn-0.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.0-amzn-0.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.0-amzn-0.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.10.0-amzn-0.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.0-amzn-0.1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.10.0-amzn-0.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.0-amzn-0.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.7-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.7-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.7-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.7-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.7-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.6.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.6.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.7.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.6.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.3.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.238.3-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.238.3-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.238.3-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.6-amzn-0.1	Spark 命令行客户端。
spark-history-server	2.4.6-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.6-amzn-0.1	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.4.6-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.1.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅 配置应用程序。

emr-5.31.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。

分类	描述
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。

分类	描述
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。

分类	描述
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.31.0

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahout 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.31.0	emr-5.30.2	emr-5.30.1	emr-5.30.0
Amazon SDK for Java	1.11.852	1.11.759	1.11.759	1.11.759
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.7
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.11.0	1.10.0	1.10.0	1.10.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.31.0	emr-5.30.2	emr-5.30.1	emr-5.30.0
HBase	1.4.13	1.4.13	1.4.13	1.4.13
HCatalog	2.3.7	2.3.6	2.3.6	2.3.6
Hadoop	2.10.0	2.8.5	2.8.5	2.8.5
Hive	2.3.7	2.3.6	2.3.6	2.3.6
Hudi	0.6.0-amzn-0	0.5.2-incubating	0.5.2-incubating	0.5.2-incubating
Hue	4.7.1	4.6.0	4.6.0	4.6.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.1.0
Livy	0.7.0	0.7.0	0.7.0	0.7.0
MXNet	1.6.0	1.5.1	1.5.1	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.2.0	5.2.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.238.3	0.232	0.232	0.232
Spark	2.4.6	2.4.5	2.4.5	2.4.5
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	2.1.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2

	emr-5.31.0	emr-5.30.2	emr-5.30.1	emr-5.30.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.31.0 的信息。更改与 5.30.1 有关。

首次发布日期：2020 年 10 月 9 日

上次更新日期：2020 年 10 月 15 日

升级

- 已将 Amazon Glue 连接器升级到版本 1.13.0
- 已将 Amazon SageMaker Spark SDK 升级到版本 1.4.0
- 已将 Amazon Kinesis 连接器升级到版本 3.5.9
- 已将 Amazon SDK for Java 升级到版本 1.11.852
- 已将 Bigtop-tomcat 升级到版本 8.5.56
- 已将 EMR FS 升级到版本 2.43.0
- 已将 EMR MetricsAndEventsApiGateway 客户端升级到版本 1.4.0
- 已将 EMR S3 Dist CP 升级到版本 2.15.0
- 已将 EMR S3 Select 升级到版本 1.6.0
- 已将 Flink 升级到版本 1.11.0
- 已将 Hadoop 升级到版本 2.10.0
- 已将 Hive 升级到版本 2.3.7
- 已将 Hudi 升级到版本 0.6.0
- 已将 Hue 升级到版本 4.7.1
- 已将 JupyterHub 升级到版本 1.1.0
- 已将 Mxnet 升级到版本 1.6.0

- 已将 OpenCV 升级到版本 4.3.0
- 已将 Presto 升级到版本 0.238.3
- 已将 TensorFlow 升级到版本 2.1.0

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Amazon EMR 5.31.0 及更高版本支持 Hive 列统计信息。
- 升级了组件版本。
- Amazon EMR 5.31.0 支持 EMRFS S3EC V2。在 S3 Java SDK 1.11.837 及更高版本中，引入了加密客户端版本 2（S3EC V2），并新增了各种安全增强功能。有关更多信息，请参阅下列内容：
  - S3 博客文章：更新至 Amazon S3 加密客户端。
  - Amazon SDK for Java 开发人员指南：将加密和解密客户端迁移到 V2。
  - EMR 管理指南：Amazon S3 客户端加密。

为保持向后兼容性，加密客户端 V1 在 SDK 中仍可用。

新特征

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

  使用引导操作 (BA) 设置 ulimit

  您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

  ```
  #!/bin/bash
  ```

```
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- 借助 Amazon EMR 5.31.0，您可以启动与 Lake Formation 集成的集群。该集成提供精细的列级数据筛选功能，用于筛选 Amazon Glue 数据目录中的数据库和表。它还支持从企业身份系统通过联合单点登录的方式登录 EMR Notebooks 或 Apache Zeppelin。有关更多信息，请参阅《Amazon EMR 管理指南》中的 将 Amazon EMR 与 Amazon Lake Formation 集成。

  Amazon EMR（集成 Lake Formation）目前已在 16 个Amazon区域推出：美国东部（俄亥俄和弗吉尼亚北部）、美国西部（加利福尼亚北部和俄勒冈）、亚太地区（孟买、首尔、新加坡、悉尼和东京）、加拿大（中部）、欧洲（法兰克福、爱尔兰、伦敦、巴黎和斯德哥尔摩）、南美洲（圣保罗）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 在使用 Amazon EMR 5.31.0 或 5.32.0 的集群上启用了 AtRestEncryption 或 HDFS 加密时，Hive 查询会导致以下运行时系统异常。

```
TaskAttempt 3 failed, info=[Error: Error while running task ( failure ) :
 attempt_1604112648850_0001_1_01_000000_3:java.lang.RuntimeException:
 java.lang.RuntimeException: Hive Runtime Error while closing
 operators: java.io.IOException: java.util.ServiceConfigurationError:
 org.apache.hadoop.security.token.TokenIdentifier: Provider
 org.apache.hadoop.hbase.security.token.AuthenticationTokenIdentifier not found
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.4.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.15.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.15.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.6.0	EMR S3 Select 连接器
emrfs	2.43.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.11.0	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
flink-jobmanager-config	1.11.0	为 Apache Flink JobManager 管理 EMR 节点上的资源。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.10.0-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.10.0-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.10.0-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.10.0-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.10.0-amzn-0	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.10.0-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.10.0-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.10.0-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.10.0-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.10.0-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.10.0-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.7-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.7-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.7-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.7-amzn-1	Hive 命令行客户端。

组件	版本	描述
hive-hbase	2.3.7-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.7-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.7-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.6.0-amzn-0	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-spark	0.6.0-amzn-0	用于运行 Spark 以及 Hudi 的捆绑库。
hudi-presto	0.6.0-amzn-0	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.7.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.6.0	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	4.3.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.238.3-amzn-0	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.238.3-amzn-0	用于执行查询的各个部分的服务。
presto-client	0.238.3-amzn-0	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.6-amzn-0	Spark 命令行客户端。
spark-history-server	2.4.6-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.4.6-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.6-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	2.1.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.31.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值

分类	描述
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。

分类	描述
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。

分类	描述
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。

分类	描述
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

分类	描述
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.30.2

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahou 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.30.2	emr-5.30.1	emr-5.30.0	emr-5.29.0
Amazon SDK for Java	1.11.759	1.11.759	1.11.759	1.11.682
Python	2.7、3.7	2.7、3.7	2.7、3.7	2.7、3.6

	emr-5.30.2	emr-5.30.1	emr-5.30.0	emr-5.29.0
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.10.0	1.10.0	1.10.0	1.9.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.13	1.4.10
HCatalog	2.3.6	2.3.6	2.3.6	2.3.6
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.6	2.3.6	2.3.6	2.3.6
Hudi	0.5.2-incubating	0.5.2-incubating	0.5.2-incubating	0.5.0-incubating
Hue	4.6.0	4.6.0	4.6.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.1.0	1.0.0
Livy	0.7.0	0.7.0	0.7.0	0.6.0
MXNet	1.5.1	1.5.1	1.5.1	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.2.0	5.1.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.232	0.232	0.232	0.227

	emr-5.30.2	emr-5.30.1	emr-5.30.0	emr-5.29.0
Spark	2.4.5	2.4.5	2.4.5	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

更改、增强和解决的问题

- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。

- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*-amzn-*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-`

component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.3.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.5.0	EMR S3 Select 连接器
emrfs	2.40.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.10.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-6.1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-6.1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-6.1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-6.1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-6.1	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-6.1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-6.1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-6.1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-6.1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-6.1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-6.1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.6-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.6-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.2-incubating	增量处理框架，以支持低延迟和高效率的数据管道。

组件	版本	描述
hudi-presto	0.5.2-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.6.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.232	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.232	用于执行查询的各个部分的服务。
presto-client	0.232	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.5-amzn-0.1	Spark 命令行客户端。
spark-history-server	2.4.5-amzn-0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.5-amzn-0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.5-amzn-0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。

组件	版本	描述
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.30.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。

分类	描述
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.30.1

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahou 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.30.1	emr-5.30.0	emr-5.29.0	emr-5.28.1
Amazon SDK for Java	1.11.759	1.11.759	1.11.682	1.11.659
Python	2.7、3.7	2.7、3.7	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.10.0	1.10.0	1.9.1	1.9.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.13	1.4.13	1.4.10	1.4.10
HCatalog	2.3.6	2.3.6	2.3.6	2.3.6
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5

	emr-5.30.1	emr-5.30.0	emr-5.29.0	emr-5.28.1
Hive	2.3.6	2.3.6	2.3.6	2.3.6
Hudi	0.5.2-incubating	0.5.2-incubating	0.5.0-incubating	0.5.0-incubating
Hue	4.6.0	4.6.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.1.0	1.0.0	1.0.0
Livy	0.7.0	0.7.0	0.6.0	0.6.0
MXNet	1.5.1	1.5.1	1.5.1	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.2.0	5.1.0	5.1.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.232	0.232	0.227	0.227
Spark	2.4.5	2.4.5	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2

	emr-5.30.1	emr-5.30.0	emr-5.29.0	emr-5.28.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.30.1 的信息。更改与 5.30.0 有关。

首次发布日期：2020 年 6 月 30 日

上次更新时间：2020 年 8 月 24 日

更改、增强功能和解决的问题

- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 修复了实例控制器进程生成无限量进程的问题。
- 修复了以下问题：Hue 无法运行 Hive 查询并显示“database is locked（数据库已锁定）”消息、阻止执行查询的问题。
- 修复了一个 Spark 问题，现在可以在 EMR 集群上同时运行更多任务。
- 修复了一个 Jupyter notebook 问题，该问题会导致 Jupyter 服务器中出现“too many files open error（打开过多文件错误）”。
- 修复了集群启动时间的问题。

新特征

- Amazon EMR 版本 6.x 和 EMR 版本 5.30.1 及更高版本提供了 Tez UI 和 YARN 时间线服务器持久性应用程序界面。无需通过 SSH 连接设置 Web 代理，访问永久性应用程序历史记录的一键式链接即可让您快速访问任务历史记录。活动和已终止集群的日志将在应用程序结束后保留 30 天。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看持久性应用程序用户界面。
- 可以使用 EMR Notebooks 执行 API 通过脚本或命令行来执行 EMR Notebooks。无需使用Amazon 控制台以编程方式控制 EMR Notebooks，即可启动、停止、列出和描述 EMR Notebooks 执行。借助参数化笔记本单元，您可以将不同的参数值传递给笔记本，而无需为每组新参数值创建笔记本副本。请参阅 EMR API 操作。有关示例代码，请参阅以编程方式执行 EMR Notebooks 的示例命令。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

  使用引导操作 (BA) 设置 ulimit

  您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

  ```
  #!/bin/bash
  for user in hadoop spark hive; do
  sudo tee /etc/security/limits.d/$user.conf << EOF
  $user - nofile 65536
  $user - nproc 65536
  ```

```
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- EMR Notebooks

  在 EMR 版本 5.30.1 上，默认情况下禁用在集群主节点上安装内核和其他 Python 库的功能。有关此功能的更多信息，请参阅在集群主节点上安装内核和 Python 库。

  要启动此功能，请执行以下操作：

  1. 确保附加到 EMR Notebooks 服务角色的权限策略允许执行以下操作：

     ```
     elasticmapreduce:ListSteps
     ```

     有关更多信息，请参阅 EMR Notebooks 的服务角色。

  2. 使用 Amazon CLI 在集群上运行一个设置 EMR Notebooks 的步骤，如以下示例所示。将 *us-east-1* 替换为您的集群所在的区域。有关更多信息，请参阅使用 Amazon CLI 向集群中添加步骤。

     ```
     aws emr add-steps  --cluster-id MyClusterID --steps
      Type=CUSTOM_JAR,Name=EMRNotebooksSetup,ActionOnFailure=CONTINUE,Jar=s3://us-
     east-1.elasticmapreduce/libs/script-runner/script-runner.jar,Args=["s3://
     awssupportdatasvcs.com/bootstrap-actions/EMRNotebooksSetup/emr-notebooks-setup.sh"]
     ```

- 托管扩展

  在未安装 Presto 的 5.30.0 和 5.30.1 的集群上进行托管扩展操作可能会导致应用程序故障或导致统一的实例组或实例集处于 ARRESTED 状态，尤其是在缩减操作之后快速执行扩展操作时。

  解决方法是即使您的任务不需要 Presto，也可以在使用 Amazon EMR 发行版 5.30.0 和 5.30.1 创建集群时，将 Presto 选为要安装的应用程序。

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.3.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.5.0	EMR S3 Select 连接器
emrfs	2.40.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.10.0	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-6	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.8.5-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.6-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-2	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.6-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.2-incubating	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.2-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.6.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.232	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.232	用于执行查询的各个部分的服务。
presto-client	0.232	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.5-amzn-0	Spark 命令行客户端。
spark-history-server	2.4.5-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.5-amzn-0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.4.5-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.30.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。

分类	描述
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。

分类	描述
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。

分类	描述
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.30.0

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahout 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.30.0	emr-5.29.0	emr-5.28.1	emr-5.28.0
Amazon SDK for Java	1.11.759	1.11.682	1.11.659	1.11.659
Python	2.7、3.7	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.10.0	1.9.1	1.9.0	1.9.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.30.0	emr-5.29.0	emr-5.28.1	emr-5.28.0
HBase	1.4.13	1.4.10	1.4.10	1.4.10
HCatalog	2.3.6	2.3.6	2.3.6	2.3.6
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.6	2.3.6	2.3.6	2.3.6
Hudi	0.5.2-incubating	0.5.0-incubating	0.5.0-incubating	0.5.0-incubating
Hue	4.6.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.1.0	1.0.0	1.0.0	1.0.0
Livy	0.7.0	0.6.0	0.6.0	0.6.0
MXNet	1.5.1	1.5.1	1.5.1	1.5.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.2.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.3
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.232	0.227	0.227	0.227
Spark	2.4.5	2.4.4	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2

	emr-5.30.0	emr-5.29.0	emr-5.28.1	emr-5.28.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.2
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.30.0 的信息。更改与 5.29.0 有关。

首次发布日期：2020 年 5 月 13 日

上次更新日期：2020 年 6 月 25 日

升级

- 已将 Amazon SDK for Java 升级到版本 1.11.759
- 已将 Amazon SageMaker Spark SDK 升级到版本 1.3.0
- 已将 EMR 记录服务器升级到版本 1.6.0
- 已将 Flink 升级到版本 1.10.0
- 已将 Ganglia 升级到版本 3.7.2
- 已将 HBase 升级到版本 1.4.13
- 已将 Hudi 升级到版本 0.5.2-incubating
- 已将 Hue 升级到版本 4.6.0
- 已将 JupyterHub 升级到版本 1.1.0
- 已将升级 Livy 到版本 0.7.0-incubating
- 已将 Oozie 升级到版本 5.2.0
- 已将 Presto 升级到版本 0.232
- 已将 Spark 升级到版本 2.4.5
- 升级的连接器和驱动程序：Amazon Glue Connector 1.12.0；Amazon Kinesis Connector 3.5.0；EMR DynamoDB Connector 4.14.0

新特征

- EMR Notebooks – 与使用 5.30.0 创建的 EMR 集群结合使用时，EMR Notebooks 内核在集群上运行。这可以提高笔记本的性能，并允许您安装和自定义内核。您还可以在集群主节点上安装 Python 库。有关更多信息，请参阅《EMR 管理指南》中的安装并使用内核和库。
- 托管扩展 – 使用 Amazon EMR 版本 5.30.0 及更高版本时，您可以启用 EMR 托管扩展，以根据工作负载自动增加或减少集群中实例或单位的数量。Amazon EMR 会持续评估集群指标，以便做出扩展决策，从而优化集群的成本和速度。有关更多信息，请参阅《Amazon EMR 管理指南》中的扩缩集群资源。
- 加密 Amazon S3 中存储的日志文件 – 使用 Amazon EMR 版本 5.30.0 及更高版本时，您可以使用 Amazon KMS 客户管理的密钥对 Amazon S3 中存储的日志文件进行加密。有关更多信息，请参阅《Amazon EMR 管理指南》中的加密存储在 Amazon S3 中的日志文件。
- Amazon Linux 2 支持 – 在 EMR 版本 5.30.0 及更高版本中，EMR 使用 Amazon Linux 2 操作系统。新的自定义 AMI（Amazon Machine Image）必须基于 Amazon Linux 2 AMI。有关更多信息，请参阅使用自定义 AMI。
- Presto 正常自动扩展 – 使用 5.30.0 的 EMR 集群可以设置自动扩展超时时段，以便 Presto 任务在其节点停用之前有时间完成运行。有关更多信息，请参阅使用采用 Graceful Decommission 的 Presto 自动扩展配置。
- 使用新的分配策略选项创建队列实例 – EMR 版本 5.12.1 及更高版本中提供了一个新的分配策略选项。它加快了集群预置、提高了 Spot 分配的准确性并减少了竞价型实例中断。需要更新非默认 EMR 服务角色。请查看配置实例集。
- sudo systemctl stop 和 sudo systemctl start 命令 – 在 EMR 版本 5.30.0 及更高版本（使用 Amazon Linux 2 操作系统）中，EMR 使用 `sudo systemctl stop` 和 `sudo systemctl start` 命令重新启动服务。有关更多信息，请参阅如何在 Amazon EMR 中重新启动服务？

更改、增强功能和解决的问题

- 默认情况下，EMR 版本 5.30.0 不安装 Ganglia。您可以在创建集群时明确选择 Ganglia 进行安装。
- Spark 性能优化。
- Presto 性能优化。
- Amazon EMR 版本 5.30.0 及更高版本默认使用 Python 3。
- 用于私有子网中服务访问的默认托管安全组已使用新规则进行更新。如果使用自定义安全组进行服务访问，则必须包含与默认托管安全组相同的规则。有关详细信息，请参阅适用于服务访问（私有子网）的 Amazon EMR 托管安全组。如果您对 Amazon EMR 使用自定义服务角色，则必须向

`ec2:describeSecurityGroups` 授予权限，以便 EMR 可以验证安全组是否已正确创建。如果您使用 `EMR_DefaultRole`，则此权限已包含在默认托管式策略中。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

  使用引导操作 (BA) 设置 ulimit

  您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

  ```
  #!/bin/bash
  ```

```
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- 托管扩展

  在未安装 Presto 的 5.30.0 和 5.30.1 的集群上进行托管扩展操作可能会导致应用程序故障或导致统一的实例组或实例集处于 ARRESTED 状态，尤其是在缩减操作之后快速执行扩展操作时。

  解决方法是即使您的任务不需要 Presto，也可以在使用 Amazon EMR 发行版 5.30.0 和 5.30.1 创建集群时，将 Presto 选为要安装的应用程序。

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- Hue 4.6.0 的默认数据库引擎是 SQLite，Hue 尝试使用外部数据库时，会引发问题。若要解决此问题，请在您的 `hue-ini` 配置分类中将 `engine` 设置为 `mysql`。Amazon EMR 版本 5.30.1 已修复这一问题。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.3.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.14.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.13.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.5.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-notebook-env	1.0.0	适用于 emr notebook 的 Conda env
emr-s3-dist-cp	emrfs	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.5.0	EMR S3 Select 连接器
emrfs	2.40.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.10.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-6	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.13	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.13	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.13	HBase 命令行客户端。
hbase-rest-server	1.4.13	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.13	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.6-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.6-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.2-incubating	增量处理框架，以支持低延迟和高效率的数据管道。

组件	版本	描述
hudi-presto	0.5.2-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.6.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.1.0	Jupyter notebook 的多用户服务器
livy-server	0.7.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mariadb-server	5.5.64	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.2.0	Oozie 命令行客户端。
oozie-server	5.2.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.232	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.232	用于执行查询的各个部分的服务。
presto-client	0.232	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.3	用于统计计算的 R 项目
ranger-kms-server	1.2.0	Apache Ranger 密钥管理系统
spark-client	2.4.5-amzn-0	Spark 命令行客户端。
spark-history-server	2.4.5-amzn-0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.5-amzn-0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.5-amzn-0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。

组件	版本	描述
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.30.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。

分类	描述
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
hudi-env	更改 Hudi 环境中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.29.0

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahout 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.29.0	emr-5.28.1	emr-5.28.0	emr-5.27.1
Amazon SDK for Java	1.11.682	1.11.659	1.11.659	1.11.615
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.9.1	1.9.0	1.9.0	1.8.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.10	1.4.10	1.4.10
HCatalog	2.3.6	2.3.6	2.3.6	2.3.5
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5

	emr-5.29.0	emr-5.28.1	emr-5.28.0	emr-5.27.1
Hive	2.3.6	2.3.6	2.3.6	2.3.5
Hudi	0.5.0-incubating	0.5.0-incubating	0.5.0-incubating	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.0.0	1.0.0	1.0.0	1.0.0
Livy	0.6.0	0.6.0	0.6.0	0.6.0
MXNet	1.5.1	1.5.1	1.5.1	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.3	4.14.3	4.14.3	4.14.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.227	0.227	0.227	0.224
Spark	2.4.4	2.4.4	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.2	0.8.1

	emr-5.29.0	emr-5.28.1	emr-5.28.0	emr-5.27.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.29.0 的信息。更改与 5.28.1 有关。

首次发布日期：2020 年 1 月 17 日

升级

- 已将 Amazon SDK for Java 升级到版本 1.11.682
- 已将 Hive 升级到版本 2.3.6
- 已将 Flink 升级到版本 1.9.1
- 已将 EMRFS 升级到版本 2.38.0
- 已将 EMR DynamoDB 连接器升级到版本 4.13.0

更改、增强功能和解决的问题

- Spark
  - Spark 性能优化。
- EMRFS
  - 将管理指南更新为 emrfs-site.xml 默认设置以实现了一致视图。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。

- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.6	Amazon SageMaker Spark 开发工具包
emr-ddb	4.13.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.12.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.13.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.4.0	EMR S3 Select 连接器
emrfs	2.38.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.9.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-hdfs-journalnode	2.8.5-amzn-5	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。

组件	版本	描述
hcatalog-client	2.3.6-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.6-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.0-incubating	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.0-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.227	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.227	用于执行查询的各个部分的服务。
presto-client	0.227	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。

组件	版本	描述
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.29.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。

分类	描述
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。

分类	描述
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.28.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahou 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.28.1	emr-5.28.0	emr-5.27.1	emr-5.27.0
Amazon SDK for Java	1.11.659	1.11.659	1.11.615	1.11.615

	emr-5.28.1	emr-5.28.0	emr-5.27.1	emr-5.27.0
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.9.0	1.9.0	1.8.1	1.8.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.10	1.4.10	1.4.10
HCatalog	2.3.6	2.3.6	2.3.5	2.3.5
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.6	2.3.6	2.3.5	2.3.5
Hudi	0.5.0-incubating	0.5.0-incubating	-	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.0.0	1.0.0	1.0.0	1.0.0
Livy	0.6.0	0.6.0	0.6.0	0.6.0
MXNet	1.5.1	1.5.1	1.4.0	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.3	4.14.3	4.14.2	4.14.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0

	emr-5.28.1	emr-5.28.0	emr-5.27.1	emr-5.27.0
Presto	0.227	0.227	0.224	0.224
Spark	2.4.4	2.4.4	2.4.4	2.4.4
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.14.0
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.2	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.28.1 的信息。更改与 5.28.0 有关。

首次发布日期：2020 年 1 月 10 日

更改、增强功能和解决的问题

- Spark
  - 修复了 Spark 兼容性问题。
- CloudWatch 指标
  - 修复了在具有多个主节点的 EMR 集群上发布的 Amazon CloudWatch Metrics。
- 已禁用日志消息
  - 已禁用假日志消息“...using old version (<4.5.8) of Apache http client”（使用低于版本 4.5.8 的 Apache http 客户端）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.6	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.12.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.11.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.13.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.37.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.9.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.5-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-5	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.6-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.6-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.0-incubating	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.0-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.227	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.227	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.227	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.28.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。

分类	描述
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。

分类	描述
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。

分类	描述
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。

分类	描述
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.28.0

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hudi、Hue、JupyterHub、Livy、MXNet、Mahou 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.28.0	emr-5.27.1	emr-5.27.0	emr-5.26.0
Amazon SDK for Java	1.11.659	1.11.615	1.11.615	1.11.595
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.9.0	1.8.1	1.8.1	1.8.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.10	1.4.10	1.4.10
HCatalog	2.3.6	2.3.5	2.3.5	2.3.5
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.6	2.3.5	2.3.5	2.3.5
Hudi	0.5.0-incubating	-	-	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.0.0	1.0.0	1.0.0	0.9.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0
MXNet	1.5.1	1.4.0	1.4.0	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.28.0	emr-5.27.1	emr-5.27.0	emr-5.26.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.3	4.14.2	4.14.2	4.14.2
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.227	0.224	0.224	0.220
Spark	2.4.4	2.4.4	2.4.4	2.4.3
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.14.0	1.13.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.2	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.28.0 的信息。更改与 5.27.0 有关。

首次发布日期：2019 年 11 月 12 日

升级

- 已将 Flink 升级到版本 1.9.0
- 已将 Hive 升级到版本 2.3.6
- 已将 MXNet 升级到版本 1.5.1
- 已将 Phoenix 升级到版本 4.14.3
- 已将 Presto 升级到版本 0.227
- 已将 Zeppelin 升级到版本 0.8.2

新特征

- 创建集群时，Amazon EMR 现在可以安装 Apache Hudi。有关更多信息，请参阅Hudi。
- （2019 年 11 月 25 日）您现在可以选择并行运行多个步骤以提高集群利用率并节省成本。您还可以取消待处理和正在运行的步骤。有关更多信息，请参阅使用 Amazon CLI 和控制台执行步骤。
- （2019 年 12 月 3 日）您现在可以在 Amazon Outposts 上创建和运行 EMR 集群。Amazon Outposts 启用本地设施中的本地Amazon服务、基础设施和操作模型。在 Amazon Outposts 环境中，您可以使用与Amazon云中相同的 Amazon API、工具和基础设施。有关更多信息，请参阅 EMR clusters on Amazon Outposts。
- （2020 年 3 月 11 日）从 Amazon EMR 版本 5.28.0 开始，您可以在 Amazon Local Zones 子网上创建和运行 Amazon EMR 集群，作为支持的Amazon区域的逻辑扩展。本地区域使得 Amazon EMR 功能和Amazon服务的子集（如计算和存储服务）在位置上与用户更近，从而为本地运行的应用程序提供非常低的延迟访问。有关可用的 Local Zones 列表，请参阅 Amazon Local Zones。有关访问可用Amazon Local Zones 的信息，请参阅区域、可用区和 Local Zones。

  Local Zones 目前不支持 Amazon EMR Notebooks，也不支持使用接口 VPC 终端节点（Amazon PrivateLink）直接连接到 Amazon EMR。

更改、增强功能和解决的问题

- 扩展了对高可用性集群的应用程序支持
  - 有关更多信息，请参阅 Amazon EMR Management Guide 中的 Supported applications in an EMR cluster with Multiple Primary Nodes。
- Spark
  - 性能优化
- Hive
  - 性能优化
- Presto
  - 性能优化

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.6	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.12.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.11.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.13.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.37.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.9.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.5-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-5	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.6-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.6-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.6-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.6-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.6-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.6-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.6-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hudi	0.5.0-incubating	增量处理框架，以支持低延迟和高效率的数据管道。
hudi-presto	0.5.0-incubating	用于运行 Presto 以及 Hudl 的捆绑库。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.5.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.3-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.3-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.227	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.227	用于执行查询的各个部分的服务。

组件	版本	描述
presto-client	0.227	Presto 命令行客户端，安装在 HA 集群的备用主节点（未启动 Presto 服务器）上。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.28.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。

分类	描述
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。

分类	描述
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。

分类	描述
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。

分类	描述
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.27.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.27.1	emr-5.27.0	emr-5.26.0	emr-5.25.0
Amazon SDK for Java	1.11.615	1.11.615	1.11.595	1.11.566
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.1	1.8.1	1.8.0	1.8.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.10	1.4.10	1.4.9
HCatalog	2.3.5	2.3.5	2.3.5	2.3.5
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.5	2.3.5	2.3.5	2.3.5
Hudi	-	-	-	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.0.0	1.0.0	0.9.6	0.9.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0
MXNet	1.4.0	1.4.0	1.4.0	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.27.1	emr-5.27.0	emr-5.26.0	emr-5.25.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.2	4.14.2	4.14.2	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.224	0.224	0.220	0.220
Spark	2.4.4	2.4.4	2.4.3	2.4.3
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.14.0	1.13.1	1.13.1
Tez	0.9.2	0.9.2	0.9.2	0.9.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.14

发布说明

这是补丁版本。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

实例元数据服务（IMDS）V2 支持状态：Amazon EMR 5.23.1、5.27.1 和 5.32 或更高版本的组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。对于其它 5.x EMR 版本，禁用 IMDSv1 会导致集群启动失败。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.4	Amazon SageMaker Spark 开发工具包
emr-ddb	4.12.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.11.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.13.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.36.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.5-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.5-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.5-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.5-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.5-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.5-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.5-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.2-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.2-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.224	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.224	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.27.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.27.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.27.0	emr-5.26.0	emr-5.25.0	emr-5.24.1
Amazon SDK for Java	1.11.615	1.11.595	1.11.566	1.11.546
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.1	1.8.0	1.8.0	1.8.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.10	1.4.9	1.4.9
HCatalog	2.3.5	2.3.5	2.3.5	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.5	2.3.5	2.3.5	2.3.4
Hudi	-	-	-	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	1.0.0	0.9.6	0.9.6	0.9.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0

	emr-5.27.0	emr-5.26.0	emr-5.25.0	emr-5.24.1
MXNet	1.4.0	1.4.0	1.4.0	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.2	4.14.2	4.14.1	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.224	0.220	0.220	0.219
Spark	2.4.4	2.4.3	2.4.3	2.4.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.14.0	1.13.1	1.13.1	1.12.0
Tez	0.9.2	0.9.2	0.9.2	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.14	3.4.14	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.27.0 的信息。更改与 5.26.0 有关。

首次发布日期：2019 年 9 月 23 日

升级

- Amazon SDK for Java 1.11.615
- Flink 1.8.1
- JupyterHub 1.0.0

- Spark 2.4.4
- Tensorflow 1.14.0
- 连接器和驱动程序：
  - DynamoDB 连接器 4.12.0

新特征

- （2019 年 10 月 24 日）所有 Amazon EMR 版本均在 EMR Notebooks 中提供以下新功能。
  - 您可以将 Git 存储库与 EMR Notebooks 关联，以将笔记本存储在版本控制的环境中。您可以通过远程 Git 存储库与同行共享代码，并重复使用现有的 Jupyter notebook。有关更多信息，请参阅《Amazon EMR 管理指南》中的将 Git 存储库与 Amazon EMR Notebooks 关联。
  - nbdime 实用工具现在可在 EMR Notebooks 中使用，简化笔记本比较和合并。
  - EMR Notebooks 现在支持 JupyterLab。JupyterLab 是一个基于 Web 的交互式开发环境，与 Jupyter notebook 完全兼容。现在，您可以选择在 JupyterLab 或 Jupyter notebook 编辑器中打开笔记本。
- （2019 年 10 月 30 日）借助 Amazon EMR 5.25.0 版及更高版本，您可以从控制台中的集群 Summary (摘要) 页面或 Application history (应用程序历史记录) 选项卡连接到 Spark 历史记录服务器 UI。您可以快速访问 Spark 历史记录服务器 UI，来查看应用程序指标并访问活动集群和终止集群的相关日志文件，而无需通过 SSH 连接设置 Web 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群外访问持久性应用程序用户界面。

更改、增强功能和解决的问题

- 具有多个主节点的 Amazon EMR 集群
  - 您可以在具有多个主节点的 Amazon EMR 集群上安装和运行 Flink。有关更多信息，请参阅支持的应用程序和功能。
  - 您可以在具有多个主节点的 Amazon EMR 集群上配置 HDFS 透明加密。有关更多信息，请参阅 HDFS Transparent Encryption on EMR clusters with Multiple Primary Nodes。
  - 现在，您可以修改在具有多个主节点的 Amazon EMR 集群上运行的应用程序的配置。有关更多信息，请参阅为运行的集群中的实例组提供配置。
- Amazon EMR-DynamoDB 连接器
  - Amazon EMR-DynamoDB 连接器现在支持以下 DynamoDB 数据类型：布尔值、列表、映射、项目、空值。有关更多信息，请参阅设置 Hive 表以运行 Hive 命令。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.4	Amazon SageMaker Spark 开发工具包
emr-ddb	4.12.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.11.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.13.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.36.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.5-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.5-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.5-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.5-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.5-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.5-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.5-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
jupyterhub	1.0.0	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.2-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.2-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.224	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.224	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.4	Spark 命令行客户端。
spark-history-server	2.4.4	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.4	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.4	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.14.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.27.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。

分类	描述
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。

分类	描述
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。

分类	描述
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。

分类	描述
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.26.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.26.0	emr-5.25.0	emr-5.24.1	emr-5.24.0
Amazon SDK for Java	1.11.595	1.11.566	1.11.546	1.11.546
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.0	1.8.0	1.8.0	1.8.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.10	1.4.9	1.4.9	1.4.9
HCatalog	2.3.5	2.3.5	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.5	2.3.5	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.4.0	4.4.0	4.4.0	4.4.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.6	0.9.6	0.9.6	0.9.6
Livy	0.6.0	0.6.0	0.6.0	0.6.0
MXNet	1.4.0	1.4.0	1.4.0	1.4.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.26.0	emr-5.25.0	emr-5.24.1	emr-5.24.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.2	4.14.1	4.14.1	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.220	0.220	0.219	0.219
Spark	2.4.3	2.4.3	2.4.2	2.4.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.13.1	1.13.1	1.12.0	1.12.0
Tez	0.9.2	0.9.2	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.14	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.26.0 的信息。更改与 5.25.0 有关。

首次发布日期：2019 年 8 月 8 日

上次更新日期：2019 年 8 月 19 日

升级

- Amazon SDK for Java 1.11.595
- HBase 1.4.10
- Phoenix 4.14.2
- 连接器和驱动程序：
  - DynamoDB 连接器 4.11.0

- MariaDB 连接器 2.4.2
- Amazon Redshift JDBC 驱动程序 1.2.32.1056

新特征

- （测试版）借助 Amazon EMR 5.26.0，您可以启动与 Lake Formation 集成的集群。此集成提供了对 Amazon Glue 数据目录中的数据库和表的精细列级别访问。它还支持从企业身份系统通过联合单点登录的方式登录 EMR Notebooks 或 Apache Zeppelin。有关更多信息，请参阅将 Amazon EMR 与 Amazon Lake Formation 集成（测试版）。
- （2019 年 8 月 19 日）所有支持安全组的 Amazon EMR 发行版现在均可提供 Amazon EMR 阻止公有访问功能。可在账户上设置阻止公有访问，适用于所有Amazon区域。如果与集群关联的任何安全组具有一个允许某端口上来自 IPv4 0.0.0.0/0 或 IPv6 ::/0（公有访问）的入站流量的规则，阻止公有访问将阻止集群启动，除非将该端口指定为例外。默认情况下，端口 22 是一个例外。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Amazon EMR 阻止公有访问。

更改、增强功能和解决的问题

- EMR Notebooks
  - 在 EMR 5.26.0 及更高版本中，EMR Notebooks 除了默认的 Python 库外，还支持笔记本范围的 Python 库。无需重新创建集群或重新将笔记本附加到集群，您即可从笔记本编辑器中安装笔记本范围的库。笔记本范围的库是在 Python 虚拟环境中创建的，因此适用于当前笔记本会话。这使得您可以隔离笔记本依赖项。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用笔记本范围的库。
- EMRFS
  - 您可以通过以下方式启用 ETag 验证功能（测试版）：将 `fs.s3.consistent.metadata.etag.verification.enabled` 设置为 `true`。启用后，EMRFS 使用 Amazon S3 ETag 验证所读取的对象是否为最新可用版本。此功能对更新后读取使用案例很有帮助，在这些案例中，将覆盖 Amazon S3 上的文件但保留相同名称。此 ETag 验证功能当前不可用于 S3 Select。有关更多信息，请参阅配置统一视图。
- Spark
  - 现在，默认情况下启用以下优化：动态分区修剪、DISTINCT before INTERSECT、改进了 JPIN（后跟 DISTINCT 查询）的 SQL 计划统计数据推理、展平标量子查询、优化的连接重排序和 Bloom 筛选条件连接。有关更多信息，请参阅优化 Spark 性能。
  - 改进了排序合并连接的整个阶段代码生成。

  - 改进了查询片段和子查询重用。
  - 改进了 Spark 启动时的预分配执行程序。
  - 连接的较小侧包含广播提示时，不再应用 Bloom 筛选条件连接。
- Tez
  - 已解决 Tez 中存在的问题。Tez UI 现可用于具有多个主节点的 Amazon EMR 集群。

已知问题

- 改进的“排序合并连接的整个阶段代码生成”功能在启用后会增加内存压力。此优化可提高性能，但如果 `spark.yarn.executor.memoryOverheadFactor` 未调整，不能提供足够的内存，则会导致任务重试或失败。要禁用此功能，请将 `spark.sql.sortMergeJoinExec.extendedCodegen.enabled` 设置为 false。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.4	Amazon SageMaker Spark 开发工具包
emr-ddb	4.11.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.10.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.12.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.35.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.0	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.10	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.10	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.10	HBase 命令行客户端。
hbase-rest-server	1.4.10	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.10	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.5-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.5-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.5-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.5-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.5-amzn-0	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.5-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.5-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.6	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。

组件	版本	描述
phoenix-library	4.14.2-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.2-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.220	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.220	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.3	Spark 命令行客户端。
spark-history-server	2.4.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.13.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.26.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。

分类	描述
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。

分类	描述
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.25.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Ooz和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.25.0	emr-5.24.1	emr-5.24.0	emr-5.23.1
Amazon SDK for Java	1.11.566	1.11.546	1.11.546	1.11.519
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.0	1.8.0	1.8.0	1.7.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.9	1.4.9	1.4.9	1.4.9
HCatalog	2.3.5	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.5	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.4.0	4.4.0	4.4.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.6	0.9.6	0.9.6	0.9.4
Livy	0.6.0	0.6.0	0.6.0	0.5.0
MXNet	1.4.0	1.4.0	1.4.0	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.25.0	emr-5.24.1	emr-5.24.0	emr-5.23.1
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.220	0.219	0.219	0.215
Spark	2.4.3	2.4.2	2.4.2	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.13.1	1.12.0	1.12.0	1.12.0
Tez	0.9.2	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.14	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.25.0 的信息。更改与 5.24.1 有关。

首次发布日期：2019 年 7 月 17 日

上次更新日期：2019 年 10 月 30 日

Amazon EMR 5.25.0

升级

- Amazon SDK for Java 1.11.566
- Hive 2.3.5
- Presto 0.220
- Spark 2.4.3

- TensorFlow 1.13.1
- Tez 0.9.2
- Zookeeper 3.4.14

新特征

- （2019 年 10 月 30 日）从 Amazon EMR 版本 5.25.0 开始，您可以从控制台中的集群 Summary (摘要) 页面或 Application history (应用程序历史记录) 选项卡连接到 Spark 历史记录服务器 UI。您可以快速访问 Spark 历史记录服务器 UI，来查看应用程序指标并访问活动集群和终止集群的相关日志文件，而无需通过 SSH 连接设置 Web 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群外访问持久性应用程序用户界面。

更改、增强功能和解决的问题

- Spark
  - 通过使用 Bloom 筛选条件预筛选输入，提高了某些连接的性能。默认情况下，优化处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.bloomFilterJoin.enabled` 设置为 `true`。
  - 改进了按字符串类型列分组的性能。
  - 改进了未安装 HBase 的集群 R4 实例类型的默认 Spark 执行程序内存和内核配置。
  - 解决了动态分区修剪功能之前存在的一个问题，即修剪的表必须位于联接的左侧。
  - 改进了 DISTINCT before INTERSECT 优化，以应用于涉及别名的其它情况。
  - 改进了 JOIN（后跟 DISTING 查询）的 SQL 计划统计数据推理。默认情况下，该改进处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.statsImprovements.enabled` 设置为 `true`。此优化是“Distinct before Intersect”功能所需的，将 `spark.sql.optimizer.distinctBeforeIntersect.enabled` 设置为 `true` 时将自动启用。
  - 根据表格大小和筛选条件优化了联接顺序。默认情况下，该优化处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.optimizer.sizeBasedJoinReorder.enabled` 设置为 `true`。

  有关更多信息，请参阅优化 Spark 性能。
- EMRFS
  - 现在，EMRFS 设置 `fs.s3.buckets.create.enabled` 默认处于禁用状态。通过测试，我们发现禁用此设置可提高性能并可防止意外创建 S3 存储桶。如果您的应用程序需使用此功能，则可

以通过以下方式启用：将 `emrfs-site` 配置分类中的 `fs.s3.buckets.create.enabled` 设置为 `true`。有关更多信息，请参阅在创建集群时提供配置。

- 安全配置中的本地磁盘加密和 S3 加密改进（2019 年 8 月 5 日）
  - 在安全配置设置中将 Amazon S3 加密设置与本地磁盘加密设置分开。
  - 发行版 5.24.0 及更高版本中添加了一个选项，可启用 EBS 加密。选择此选项后，除了存储卷之外，还会加密根设备卷。之前的版本需要使用自定义 AMI 来加密根设备卷。
  - 有关更多信息，请参阅《Amazon EMR 管理指南》中的加密选项。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.4	Amazon SageMaker Spark 开发工具包
emr-ddb	4.10.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.9.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.34.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.5-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.5-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.5-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.5-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.5-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.5-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.5-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.6	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.220	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.220	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.3	Spark 命令行客户端。
spark-history-server	2.4.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.13.1	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.2	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.14	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.14	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.25.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。

分类	描述
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。

分类	描述
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
recordserver-env	更改 EMR RecordServer 环境中的值。
recordserver-conf	更改 EMR RecordServer erver.properties 文件中的值。
recordserver-log4j	更改 EMR RecordServer log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.24.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Ooz 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.24.1	emr-5.24.0	emr-5.23.1	emr-5.23.0
Amazon SDK for Java	1.11.546	1.11.546	1.11.519	1.11.519
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.0	1.8.0	1.7.1	1.7.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.9	1.4.9	1.4.9	1.4.9
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.4.0	4.4.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.6	0.9.6	0.9.4	0.9.4
Livy	0.6.0	0.6.0	0.5.0	0.5.0
MXNet	1.4.0	1.4.0	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.24.1	emr-5.24.0	emr-5.23.1	emr-5.23.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.219	0.219	0.215	0.215
Spark	2.4.2	2.4.2	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.24.1 的信息。更改与 5.24.0 有关。

首次发布日期：2019 年 6 月 26 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.9.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.8.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.33.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.6	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.219	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.219	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.2	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.4.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.24.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。

分类	描述
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.24.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.24.0	emr-5.23.1	emr-5.23.0	emr-5.22.0
Amazon SDK for Java	1.11.546	1.11.519	1.11.519	1.11.510
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.8.0	1.7.1	1.7.1	1.7.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.9	1.4.9	1.4.9	1.4.9
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-

	emr-5.24.0	emr-5.23.1	emr-5.23.0	emr-5.22.0
Hue	4.4.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.6	0.9.4	0.9.4	0.9.4
Livy	0.6.0	0.5.0	0.5.0	0.5.0
MXNet	1.4.0	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.1.0	5.1.0
Phoenix	4.14.1	4.14.1	4.14.1	4.14.1
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.219	0.215	0.215	0.215
Spark	2.4.2	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.1
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.24.0 的信息。更改与 5.23.0 有关。

首次发布日期：2019 年 6 月 11 日

上次更新时间：2019 年 8 月 5 日

升级

- Flink 1.8.0
- Hue 4.4.0
- JupyterHub 0.9.6
- Livy 0.6.0
- MxNet 1.4.0
- Presto 0.219
- Spark 2.4.2
- Amazon SDK for Java 1.11.546
- 连接器和驱动程序：
  - DynamoDB 连接器 4.9.0
  - MariaDB 连接器 2.4.1
  - Amazon Redshift JDBC 驱动程序 1.2.27.1051

更改、增强功能和解决的问题

- Spark
  - 添加了对动态修剪分区的优化。默认情况下禁用优化。要启用该优化，请将 Spark 参数 `spark.sql.dynamicPartitionPruning.enabled` 设置为 `true`。
  - 改进了 `INTERSECT` 查询的性能。默认情况下禁用此优化。要启用该优化，请将 Spark 参数 `spark.sql.optimizer.distinctBeforeIntersect.enabled` 设置为 `true`。
  - 添加了对展平标量子查询的优化，可使用相同关系进行聚合。默认情况下禁用优化。要启用该优化，请将 Spark 参数 `spark.sql.optimizer.flattenScalarSubqueriesWithAggregates.enabled` 设置为 `true`。
  - 改进了整个阶段代码生成。

有关更多信息，请参阅优化 Spark 性能。

- 安全配置中的本地磁盘加密和 S3 加密改进（2019 年 8 月 5 日）
  - 在安全配置设置中将 Amazon S3 加密设置与本地磁盘加密设置分开。
  - 添加了一个启用 EBS 加密的选项。选择此选项后，除了存储卷之外，还会加密根设备卷。之前的版本需要使用自定义 AMI 来加密根设备卷。
  - 有关更多信息，请参阅《Amazon EMR 管理指南》中的加密选项。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.9.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.8.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.3.0	EMR S3 Select 连接器
emrfs	2.33.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.8.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-4	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-4	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.5-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.4-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.4.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.6	Jupyter notebook 的多用户服务器
livy-server	0.6.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.4.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.219	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.219	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.2	Spark 命令行客户端。
spark-history-server	2.4.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.24.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。

分类	描述
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。

分类	描述
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.23.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie、Phoenix、Pig、Presto、Spark、Sqoop、TensorFlow、Tez、Zeppelin 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.23.1	emr-5.23.0	emr-5.22.0	emr-5.21.2
Amazon SDK for Java	1.11.519	1.11.519	1.11.510	1.11.479
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.1	1.7.1	1.7.1	1.7.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.23.1	emr-5.23.0	emr-5.22.0	emr-5.21.2
HBase	1.4.9	1.4.9	1.4.9	1.4.8
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.3.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.1.0	5.0.0
Phoenix	4.14.1	4.14.1	4.14.1	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.215	0.215	0.215
Spark	2.4.0	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1

	emr-5.23.1	emr-5.23.0	emr-5.22.0	emr-5.21.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.1	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

这是补丁版本。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

实例元数据服务（IMDS）V2 支持状态：Amazon EMR 5.23.1、5.27.1 和 5.32 或更高版本的组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。对于其它 5.x EMR 版本，禁用 IMDSv1 会导致集群启动失败。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.8.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.7.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.32.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-3	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.8.5-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-3	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器

组件	版本	描述
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.23.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。

分类	描述
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.23.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.23.0	emr-5.22.0	emr-5.21.2	emr-5.21.1
Amazon SDK for Java	1.11.519	1.11.510	1.11.479	1.11.479
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.1	1.7.1	1.7.0	1.7.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.9	1.4.9	1.4.8	1.4.8
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-

	emr-5.23.0	emr-5.22.0	emr-5.21.2	emr-5.21.1
Hue	4.3.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.1.0	5.0.0	5.0.0
Phoenix	4.14.1	4.14.1	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.215	0.215	0.215
Spark	2.4.0	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.1	0.8.1	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.23.0 的信息。更改与 5.22.0 有关。

首次发布日期：2019 年 4 月 1 日

上次更新时间：2019 年 4 月 30 日

升级

- Amazon SDK for Java 1.11.519

新特征

- （2019 年 4 月 30 日）使用 Amazon EMR 5.23.0 及更高版本，您可以启动具有三个主节点的集群，以支持应用程序（如 YARN Resource Manager、HDFS NameNode、Spark、Hive 和 Ganglia）的高可用性。使用此功能，主节点不再发生潜在的单点故障。如果其中一个主节点出现故障，Amazon EMR 会自动故障转移到备用主节点，并将出现故障的主节点替换为具有相同配置和引导操作的新主节点。有关更多信息，请参阅计划和配置主节点。

已知问题

- Tez UI（已在 Amazon EMR 发行版 5.26.0 中修复）

  Tez UI 不能在具有多个主节点的 EMR 集群上运行。
- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

    1. 使用 SSH 连接到主节点命令行。
    2. 使用文本编辑器打开 `hue.ini` 文件。例如：

```
sudo vim /etc/hue/conf/hue.ini
```

3. 搜索术语 appblacklist，并将该行修改为以下内容：

```
appblacklist = search
```

4. 保存更改并重新启动 Hue，如以下示例所示：

```
sudo stop hue; sudo start hue
```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.8.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.7.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.32.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-hdfs-journalnode	2.8.5-amzn-3	用于管理 HA 集群上的 Hadoop 文件系统日志的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.23.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.22.0

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.22.0	emr-5.21.2	emr-5.21.1	emr-5.21.0
Amazon SDK for Java	1.11.510	1.11.479	1.11.479	1.11.479
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.1	1.7.0	1.7.0	1.7.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.9	1.4.8	1.4.8	1.4.8
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4

	emr-5.22.0	emr-5.21.2	emr-5.21.1	emr-5.21.0
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.3.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.1.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.1	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.215	0.215	0.215
Spark	2.4.0	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-

	emr-5.22.0	emr-5.21.2	emr-5.21.1	emr-5.21.0
Zeppelin	0.8.1	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.22.0 的信息。更改与 5.21.0 有关。

> **Important**
>
> 从 Amazon EMR 发行版 5.22.0 开始，Amazon EMR 专门使用 Amazon 签名版本 4 来对 Amazon S3 的请求进行身份验证。除非发布说明指出需专门使用签名版本 4，否则早期 Amazon EMR 发行版在某些情况下使用 Amazon 签名版本 2。有关更多信息，请参阅《Amazon Simple Storage Service 开发人员指南》中的对请求进行身份验证（Amazon签名版本 4）和对请求进行身份验证（Amazon签名版本 2）。

首次发布日期：2019 年 3 月 20 日

升级

- Flink 1.7.1
- HBase 1.4.9
- Oozie 5.1.0
- Phoenix 4.14.1
- Zeppelin 0.8.1
- 连接器和驱动程序：
  - DynamoDB 连接器 4.8.0
  - MariaDB 连接器 2.2.6
  - Amazon Redshift JDBC 驱动程序 1.2.20.1043

新特征

- 修改了仅限 EBS 存储的 EC2 实例类型的默认 EBS 配置。在使用 Amazon EMR 发行版 5.22.0 及更高版本创建集群时，默认 EBS 存储量根据实例大小而增加。此外，我们将增加的存储拆分到多个卷，从而提高了 IOPS 性能。如果要使用不同的 EBS 实例存储配置，您可以在创建 EMR 集群或将节点添加到现有集群时指定该配置。有关每个实例类型默认分配的存储容量和卷数的更多信息，请参阅《Amazon EMR 管理指南》中的实例的默认 EBS 存储。

更改、增强功能和解决的问题

- Spark
  - 在 YARN 上引入了一个新的配置属性 `spark.yarn.executor.memoryOverheadFactor`。此属性的值是一个缩放系数，它将内存开销值设置为执行程序内存的百分比，最小为 384 MB。如果内存开销设置为使用 `spark.yarn.executor.memoryOverhead`，则此属性不发挥任何作用。默认值为 `0.1875`，表示 18.75%。与 Spark 内部设置的 10% 的默认值相比，Amazon EMR 的此默认值在 YARN 容器中为执行器内存开销预留了更多空间。根据经验，Amazon EMR 默认值 18.75% 表明 TPC-DS 基准测试中与内存相关的故障较少。
  - 为了改进性能，已逆向移植 SPARK-26316。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

1. 使用 SSH 连接到主节点命令行。
2. 使用文本编辑器打开 hue.ini 文件。例如：

   ```
   sudo vim /etc/hue/conf/hue.ini
   ```

3. 搜索术语 appblacklist，并将该行修改为以下内容：

   ```
   appblacklist = search
   ```

4. 保存更改并重新启动 Hue，如以下示例所示：

   ```
   sudo stop hue; sudo start hue
   ```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.8.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.6.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.31.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.9	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.9	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.9	HBase 命令行客户端。
hbase-rest-server	1.4.9	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.9	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.1.0	Oozie 命令行客户端。
oozie-server	5.1.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.1-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.1-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.22.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.21.2

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.21.2	emr-5.21.1	emr-5.21.0	emr-5.20.1
Amazon SDK for Java	1.11.479	1.11.479	1.11.479	1.11.461
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.0	1.7.0	1.7.0	1.6.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.8	1.4.8	1.4.8	1.4.8
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4

	emr-5.21.2	emr-5.21.1	emr-5.21.0	emr-5.20.1
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.3.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.215	0.215	0.214
Spark	2.4.0	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-

	emr-5.21.2	emr-5.21.1	emr-5.21.0	emr-5.20.1
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.30.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-1	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.8.5-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.8	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.8	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.8	HBase 命令行客户端。
hbase-rest-server	1.4.8	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.8	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.4-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.21.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。

分类	描述
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。

分类	描述
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.21.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Ooz 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.21.1	emr-5.21.0	emr-5.20.1	emr-5.20.0
Amazon SDK for Java	1.11.479	1.11.479	1.11.461	1.11.461
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.6
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.0	1.7.0	1.6.2	1.6.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.8	1.4.8	1.4.8	1.4.8
HCatalog	2.3.4	2.3.4	2.3.4	2.3.4
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.4
Hudi	-	-	-	-
Hue	4.3.0	4.3.0	4.3.0	4.3.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.1
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.21.1	emr-5.21.0	emr-5.20.1	emr-5.20.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.215	0.214	0.214
Spark	2.4.0	2.4.0	2.4.0	2.4.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.12.0
Tez	0.9.1	0.9.1	0.9.1	0.9.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.21.1 的信息。更改与 5.21.0 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.30.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.5-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.8	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.8	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.8	HBase 命令行客户端。
hbase-rest-server	1.4.8	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.4.8	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器

组件	版本	描述
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.21.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。

分类	描述
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.21.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.21.0	emr-5.20.1	emr-5.20.0	emr-5.19.1
Amazon SDK for Java	1.11.479	1.11.461	1.11.461	1.11.433
Python	2.7、3.6	2.7、3.6	2.7、3.6	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.7.0	1.6.2	1.6.2	1.6.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.8	1.4.8	1.4.8	1.4.7
HCatalog	2.3.4	2.3.4	2.3.4	2.3.3
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.4	2.3.3
Hudi	-	-	-	-

	emr-5.21.0	emr-5.20.1	emr-5.20.0	emr-5.19.1
Hue	4.3.0	4.3.0	4.3.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.1	1.3.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.215	0.214	0.214	0.212
Spark	2.4.0	2.4.0	2.4.0	2.3.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.12.0	1.11.0
Tez	0.9.1	0.9.1	0.9.1	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.21.0 的信息。更改与 5.20.0 有关。

首次发布日期：2019 年 2 月 18 日

上次更新时间：2019 年 4 月 3 日

升级

- Flink 1.7.0
- Presto 0.215
- Amazon SDK for Java 1.11.479

新特征

- （2019 年 4 月 3 日）对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

更改、增强功能和解决的问题

- Zeppelin
  - 已逆向移植 ZEPPELIN-3878。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

要防止显示 Solr 错误消息:

1. 使用 SSH 连接到主节点命令行。
2. 使用文本编辑器打开 hue.ini 文件。例如：

   ```
   sudo vim /etc/hue/conf/hue.ini
   ```

3. 搜索术语 appblacklist，并将该行修改为以下内容：

   ```
   appblacklist = search
   ```

4. 保存更改并重新启动 Hue，如以下示例所示：

   ```
   sudo stop hue; sudo start hue
   ```

- Tez
  - 此问题已在 Amazon EMR 5.22.0 中得到修复。

    通过 http://*MasterDNS*:8080/tez-ui 连接到 Tez UI 时（通过 SSH 连接到集群主节点），显示错误“Adapter operation failed - Timeline server (ATS) is out of reach。Either it is down, or CORS is not enabled”，或任务不正常地显示为“N/A”。

    这是由于 Tez UI 使用 localhost（而没有使用主节点的主机名称）向 YARN 时间线服务器发出请求所致。解决方法：将脚本作为引导操作或步骤运行。脚本更新 Tez configs.env 文件中的主机名。有关更多信息以及脚本的位置信息，请参阅引导说明。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.11.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.30.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.7.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.8.5-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.8	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.8	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.8	HBase 命令行客户端。
hbase-rest-server	1.4.8	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.8	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.4-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.215	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.215	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.21.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。

分类	描述
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。

分类	描述
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.20.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Ooz 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.20.1	emr-5.20.0	emr-5.19.1	emr-5.19.0
Amazon SDK for Java	1.11.461	1.11.461	1.11.433	1.11.433
Python	2.7、3.6	2.7、3.6	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.6.2	1.6.2	1.6.1	1.6.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.8	1.4.8	1.4.7	1.4.7
HCatalog	2.3.4	2.3.4	2.3.3	2.3.3
Hadoop	2.8.5	2.8.5	2.8.5	2.8.5
Hive	2.3.4	2.3.4	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.3.0	4.3.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.9.4
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.1	1.3.0	1.3.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.20.1	emr-5.20.0	emr-5.19.1	emr-5.19.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.214	0.214	0.212	0.212
Spark	2.4.0	2.4.0	2.3.2	2.3.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.12.0	1.11.0	1.11.0
Tez	0.9.1	0.9.1	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.13

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器
emrfs	2.29.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.8	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.8	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.8	HBase 命令行客户端。
hbase-rest-server	1.4.8	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.8	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.4-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.214	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.214	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.20.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.20.0

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.20.0	emr-5.19.1	emr-5.19.0	emr-5.18.1
Amazon SDK for Java	1.11.461	1.11.433	1.11.433	1.11.393
Python	2.7、3.6	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.6.2	1.6.1	1.6.1	1.6.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.8	1.4.7	1.4.7	1.4.7
HCatalog	2.3.4	2.3.3	2.3.3	2.3.3

	emr-5.20.0	emr-5.19.1	emr-5.19.0	emr-5.18.1
Hadoop	2.8.5	2.8.5	2.8.5	2.8.4
Hive	2.3.4	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.3.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.9.4	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.1	1.3.0	1.3.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.214	0.212	0.212	0.210
Spark	2.4.0	2.3.2	2.3.2	2.3.2
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.12.0	1.11.0	1.11.0	1.9.0
Tez	0.9.1	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-

	emr-5.20.0	emr-5.19.1	emr-5.19.0	emr-5.18.1
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.13	3.4.12

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.20.0 的信息。更改与 5.19.0 有关。

首次发布日期：2018 年 12 月 18 日

上次更新时间：2019 年 1 月 22 日

升级

- Flink 1.6.2
- HBase 1.4.8
- Hive 2.3.4
- Hue 4.3.0
- MXNet 1.3.1
- Presto 0.214
- Spark 2.4.0
- TensorFlow 1.12.0
- Tez 0.9.1
- Amazon SDK for Java 1.11.461

新特征

- （2019 年 1 月 22 日）Amazon EMR 中的 Kerberos 已经得到改进，现在可支持对来自外部 KDC 的委托人进行身份验证。这集中了委托人管理，因为多个集群可以共享单个外部 KDC。此外，外部 KDC 可与 Active Directory 域建立跨领域信任关系。这使得所有集群可以从 Active Directory 对委托人进行身份验证。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Kerberos 身份验证。

更改、增强功能和解决的问题

- Amazon EMR 的默认 Amazon Linux AMI
  - Python 3 软件包已从 Python 3.4 升级到 3.6。
- 经 EMRFS S3 优化的提交程序
  - 现在，已默认启用经 EMRFS S3 优化的提交程序，从而改进写入性能。有关更多信息，请参阅使用经 EMRFS S3 优化的提交程序。
- Hive
  - 已逆向移植 HIVE-16686。
- 集成 Spark 和 Hive 的 Glue
  - 在 EMR 5.20.0 或更高版本中，当使用 Amazon Glue 数据目录作为元存储时，会自动为 Spark 和 Hive 启用并行分区修剪。此更改通过并行执行多个请求来检索分区，显著缩短查询计划时间。可同时执行的分段总数介于 1 到 10 之间。默认值为 5，这是建议的设置。您可以通过以下方式更改该值：指定 `hive-site` 配置分类中的属性 `aws.glue.partition.num.segments`。如果发生节流，则可以通过将值更改为 1 来关闭此功能。有关更多信息，请参阅 Amazon Glue 分段结构。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

    1. 使用 SSH 连接到主节点命令行。
    2. 使用文本编辑器打开 `hue.ini` 文件。例如：

       ```
       sudo vim /etc/hue/conf/hue.ini
       ```

    3. 搜索术语 `appblacklist`，并将该行修改为以下内容：

```
appblacklist = search
```

4. 保存更改并重新启动 Hue，如以下示例所示：

```
sudo stop hue; sudo start hue
```

- Tez
  - 此问题已在 Amazon EMR 5.22.0 中得到修复。

    通过 http://*MasterDNS*:8080/tez-ui 连接到 Tez UI 时（通过 SSH 连接到集群主节点），显示错误“Adapter operation failed - Timeline server (ATS) is out of reach。Either it is down, or CORS is not enabled”，或任务不正常地显示为“N/A”。

    这是由于 Tez UI 使用 `localhost`（而没有使用主节点的主机名称）向 YARN 时间线服务器发出请求所致。解决方法：将脚本作为引导操作或步骤运行。脚本更新 Tez `configs.env` 文件中的主机名。有关更多信息以及脚本的位置信息，请参阅引导说明。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 `hadoop` 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 `hadoop` 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.2.0	EMR S3 Select 连接器

组件	版本	描述
emrfs	2.29.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.8.5-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.8	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.8	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.8	HBase 命令行客户端。
hbase-rest-server	1.4.8	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.8	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.4-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.4-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.4-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.4-amzn-0	Hive 命令行客户端。

组件	版本	描述
hive-hbase	2.3.4-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.4-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.4-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.3.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.1	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。

组件	版本	描述
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.214	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.214	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.4.0	Spark 命令行客户端。
spark-history-server	2.4.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.4.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.4.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.12.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.9.1	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.20.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。

分类	描述
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。

分类	描述
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。

分类	描述
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。

分类	描述
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.19.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.19.1	emr-5.19.0	emr-5.18.1	emr-5.18.0
Amazon SDK for Java	1.11.433	1.11.433	1.11.393	1.11.393
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用

	emr-5.19.1	emr-5.19.0	emr-5.18.1	emr-5.18.0
Delta	-	-	-	-
Flink	1.6.1	1.6.1	1.6.0	1.6.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.7	1.4.7	1.4.7	1.4.7
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.5	2.8.5	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.9.4	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.0	1.3.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.212	0.212	0.210	0.210
Spark	2.3.2	2.3.2	2.3.2	2.3.2

	emr-5.19.1	emr-5.19.0	emr-5.18.1	emr-5.18.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.11.0	1.11.0	1.9.0	1.9.0
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.8.0	0.8.0
ZooKeeper	3.4.13	3.4.13	3.4.12	3.4.12

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.1.0	EMR S3 Select 连接器
emrfs	2.28.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.8.5-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.5-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.7	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.7	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.7	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.4.7	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.7	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.212	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.212	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.2	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.3.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.19.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。

分类	描述
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.19.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.19.0	emr-5.18.1	emr-5.18.0	emr-5.17.2
Amazon SDK for Java	1.11.433	1.11.393	1.11.393	1.11.336
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.6.1	1.6.0	1.6.0	1.5.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.7	1.4.7	1.4.7	1.4.6
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.5	2.8.4	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-

	emr-5.19.0	emr-5.18.1	emr-5.18.0	emr-5.17.2
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.9.4	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.3.0	1.2.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.212	0.210	0.210	0.206
Spark	2.3.2	2.3.2	2.3.2	2.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.11.0	1.9.0	1.9.0	1.9.0
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.8.0	0.7.3
ZooKeeper	3.4.13	3.4.12	3.4.12	3.4.12

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.19.0 的信息。更改与 5.18.0 有关。

首次发布日期：2018 年 11 月 7 日

上次更新时间：2018 年 11 月 19 日

升级

- Hadoop 2.8.5
- Flink 1.6.1
- JupyterHub 0.9.4
- MXNet 1.3.0
- Presto 0.212
- TensorFlow 1.11.0
- Zookeeper 3.4.13
- Amazon SDK for Java 1.11.433

新特征

- （2018 年 11 月 19 日）EMR Notebooks 是基于 Jupyter notebook 的托管环境。它支持适用于 PySpark、Spark SQL、Spark R 和 Scala 的 Spark magic 内核。EMR Notebooks 可在使用 Amazon EMR 发行版 5.18.0 及更高版本创建的集群上使用。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 EMR Notebooks。
- 使用 Spark 和 EMRFS 编写 Parquet 文件时，可以使用经 EMRFS S3 优化的提交程序。此提交程序改进了写入性能。有关更多信息，请参阅使用经 EMRFS S3 优化的提交程序。

更改、增强功能和解决的问题

- YARN
  - 修改了限制应用程序主进程在核心节点上运行的逻辑。此功能现在可使用 `yarn-site` 和 `capacity-scheduler` 配置分类中的 YARN 节点标注功能和属性。有关信息，请参阅。https://docs.amazonaws.cn/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html#emr-plan-spot-YARN.
- Amazon EMR 的默认 Amazon Linux AMI

- 默认情况下，不再安装 ruby18、php56 和 gcc48。如果需要，可以使用 yum 安装它们。
- 默认情况下，不再安装 aws-sdk ruby gem。如果需要，可以使用 `gem install aws-sdk` 进行安装。此外，还可以安装特定组件。例如，`gem install aws-sdk-s3`。

已知问题

- EMR Notebooks – 在某些情况下，打开多个笔记本编辑器时，笔记本编辑器可能无法连接到集群。如果发生这种情况，请清除浏览器 Cookie，然后重新打开笔记本编辑器。
- CloudWatch ContainerPending 指标和自动伸缩 – （已在 5.20.0 中修复）Amazon EMR 可能会发出一个 `ContainerPending` 负值。如果在自动伸缩规则中使用 `ContainerPending`，自动伸缩的行为方式可能会不符合预期。请避免在自动伸缩中使用 `ContainerPending`。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.2.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.7.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.5.1	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.1.0	EMR S3 Select 连接器
emrfs	2.28.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.5-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.5-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.5-amzn-0	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.8.5-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.5-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.5-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.5-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.5-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.5-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.5-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.7	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.7	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.7	HBase 命令行客户端。
hbase-rest-server	1.4.7	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.7	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.3-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.9.4	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.3.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.212	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.212	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.2	Spark 命令行客户端。
spark-history-server	2.3.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.2	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.11.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.13	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.13	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.19.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。

分类	描述
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。

分类	描述
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-memory	更改 Presto 的 memory.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
presto-connector-tpcds	更改 Presto 的 tpcds.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.18.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Ooz 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.18.1	emr-5.18.0	emr-5.17.2	emr-5.17.1
Amazon SDK for Java	1.11.393	1.11.393	1.11.336	1.11.336
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.6.0	1.6.0	1.5.2	1.5.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.7	1.4.7	1.4.6	1.4.6
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.2.0	1.2.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.18.1	emr-5.18.0	emr-5.17.2	emr-5.17.1
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.210	0.210	0.206	0.206
Spark	2.3.2	2.3.2	2.3.1	2.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.9.0	1.9.0	1.9.0	1.9.0
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.8.0	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.3	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.1.0	EMR S3 Select 连接器
emrfs	2.27.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.4-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.7	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.7	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.7	HBase 命令行客户端。
hbase-rest-server	1.4.7	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.7	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.210	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.210	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.2	Spark 命令行客户端。
spark-history-server	2.3.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.9.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.18.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.18.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.18.0	emr-5.17.2	emr-5.17.1	emr-5.17.0
Amazon SDK for Java	1.11.393	1.11.336	1.11.336	1.11.336
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.6.0	1.5.2	1.5.2	1.5.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.7	1.4.6	1.4.6	1.4.6
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-

	emr-5.18.0	emr-5.17.2	emr-5.17.1	emr-5.17.0
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.2.0	1.2.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.210	0.206	0.206	0.206
Spark	2.3.2	2.3.1	2.3.1	2.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.9.0	1.9.0	1.9.0	1.9.0
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.8.0	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.18.0 的信息。更改与 5.17.0 有关。

首次发布日期：2018 年 10 月 24 日

升级

- Flink 1.6.0
- HBase 1.4.7
- Presto 0.210
- Spark 2.3.2
- Zeppelin 0.8.0

新特征

- 您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发行版 5.18.0 开始）附带的准确版本的库和依赖项的任务代码。有关更多信息，请参阅使用 Amazon EMR 项目存储库检查依赖项。

更改、增强功能和解决的问题

- Hive
  - 添加了对 S3 Select 的支持。有关更多信息，请参阅将 S3 Select 与 Hive 结合使用以提高查询性能。
- Presto
  - 添加了对 S3 Select Pushdown 的支持。有关更多信息，请参阅使用 S3 Select Pushdown 搭配 Presto 提高性能。
- Spark
  - Spark 的默认 log4j 配置已更改为 Spark Streaming 任务每小时的滚动容器日志。这有助于防止删除长时间运行的 Spark Streaming 任务的日志。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.3	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.1.0	EMR S3 Select 连接器
emrfs	2.27.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.6.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.4-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.4.7	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.7	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.7	HBase 命令行客户端。
hbase-rest-server	1.4.7	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.7	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
nginx	1.12.1	nginx [引擎 x] 是 HTTP 和反向代理服务器
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.210	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.210	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.2	Spark 命令行客户端。
spark-history-server	2.3.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.9.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.8.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.18.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.17.2

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.17.2	emr-5.17.1	emr-5.17.0	emr-5.16.1
Amazon SDK for Java	1.11.336	1.11.336	1.11.336	1.11.336
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.5.2	1.5.2	1.5.2	1.5.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.6	1.4.6	1.4.6	1.4.4
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-

	emr-5.17.2	emr-5.17.1	emr-5.17.0	emr-5.16.1
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.2.0	1.2.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.206	0.206	0.206	0.203
Spark	2.3.1	2.3.1	2.3.1	2.3.1
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.9.0	1.9.0	1.9.0	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.3	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.0.0	EMR S3 Select 连接器
emrfs	2.26.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.5.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-1	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.8.4-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.4-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.6	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.6	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.6	HBase 命令行客户端。
hbase-rest-server	1.4.6	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.6	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	2.3.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.206	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.206	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.1	Spark 命令行客户端。
spark-history-server	2.3.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。

组件	版本	描述
tensorflow	1.9.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.17.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。

分类	描述
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。

分类	描述
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.17.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.17.1	emr-5.17.0	emr-5.16.1	emr-5.16.0
Amazon SDK for Java	1.11.336	1.11.336	1.11.336	1.11.336
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用

	emr-5.17.1	emr-5.17.0	emr-5.16.1	emr-5.16.0
Delta	-	-	-	-
Flink	1.5.2	1.5.2	1.5.0	1.5.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.6	1.4.6	1.4.4	1.4.4
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.4	2.8.4
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.5.0	0.5.0
MXNet	1.2.0	1.2.0	1.2.0	1.2.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.14.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.206	0.206	0.203	0.203
Spark	2.3.1	2.3.1	2.3.1	2.3.1

	emr-5.17.1	emr-5.17.0	emr-5.16.1	emr-5.16.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.9.0	1.9.0	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.17.1 的信息。更改与 5.17.0 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.3	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.0.0	EMR S3 Select 连接器
emrfs	2.26.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.5.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.8.4-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.4-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.6	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.6	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.6	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.4.6	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.6	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口

组件	版本	描述
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.206	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.206	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.1	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.3.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.9.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.17.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.17.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.17.0	emr-5.16.1	emr-5.16.0	emr-5.15.1
Amazon SDK for Java	1.11.336	1.11.336	1.11.336	1.11.333
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.5.2	1.5.0	1.5.0	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.6	1.4.4	1.4.4	1.4.4
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.4	2.8.3
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1

	emr-5.17.0	emr-5.16.1	emr-5.16.0	emr-5.15.1
Livy	0.5.0	0.5.0	0.5.0	0.4.0
MXNet	1.2.0	1.2.0	1.2.0	1.1.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.14.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.206	0.203	0.203	0.194
Spark	2.3.1	2.3.1	2.3.1	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	1.9.0	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.17.0 的信息。更改与 5.16.0 有关。

首次发布日期：2018 年 8 月 30 日

升级

- Flink 1.5.2
- HBase 1.4.6

- Presto 0.206

新特征

- 添加了对 Tensorflow 的支持。有关更多信息，请参阅TensorFlow。

更改、增强功能和解决的问题

- JupyterHub
  - Amazon S3 中添加了对笔记本持久性的支持。有关更多信息，请参阅在 Amazon S3 中配置笔记本的持久性。
- Spark
  - 添加了对 S3 Select 的支持。有关更多信息，请参阅将 S3 Select 与 Spark 结合使用以提高查询性能。
- 解决了 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中 Cloudwatch 指标和自动伸缩功能中存在的问题。

已知问题

- 创建使用 Kerberos 的集群时，如果安装了 Livy，Livy 将失败，并显示未启用简单身份验证的错误。重新启动 Livy 服务器可解决此问题。解决方法是在集群创建过程中添加一个在主节点上运行 `sudo restart livy-server` 的步骤。
- 如果您使用基于 Amazon Linux AMI（创建日期为 2018-08-11）的自定义 Amazon Linux AMI，则 Oozie 服务器无法启动。如果您使用 Oozie，请根据具有不同创建日期的 Amazon Linux AMI ID 创建自定义 AMI。您可以使用以下 Amazon CLI 命令返回所有 2018.03 版本的 HVM Amazon Linux AMI 的镜像 ID 列表以及发布日期，以便您可以根据需要选择合适的 Amazon Linux AMI。将 MyRegion 替换为您的区域标识符，如 us-west-2。

```
aws ec2 --region MyRegion describe-images --owner amazon --query 'Images[?
Name!=`null`]|[?starts_with(Name, `amzn-ami-hvm-2018.03`) == `true`].
[CreationDate,ImageId,Name]' --output text | sort -rk1
```

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.3	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.5.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emr-s3-select	1.0.0	EMR S3 Select 连接器
emrfs	2.26.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.5.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.4-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.4-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.6	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.6	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.6	HBase 命令行客户端。
hbase-rest-server	1.4.6	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.6	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.206	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.206	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.1	Spark 命令行客户端。
spark-history-server	2.3.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tensorflow	1.9.0	适用于高性能数值计算的 TensorFlow 开源软件库。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.17.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置

分类	描述
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-s3-conf	配置 Jupyter notebook S3 持久性。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.16.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.16.1	emr-5.16.0	emr-5.15.1	emr-5.15.0
Amazon SDK for Java	1.11.336	1.11.336	1.11.333	1.11.333
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.5.0	1.5.0	1.4.2	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.4	1.4.4	1.4.4	1.4.4
HCatalog	2.3.3	2.3.3	2.3.3	2.3.3
Hadoop	2.8.4	2.8.4	2.8.3	2.8.3
Hive	2.3.3	2.3.3	2.3.3	2.3.3
Hudi	-	-	-	-

	emr-5.16.1	emr-5.16.0	emr-5.15.1	emr-5.15.0
Hue	4.2.0	4.2.0	4.2.0	4.2.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.5.0	0.4.0	0.4.0
MXNet	1.2.0	1.2.0	1.1.0	1.1.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	5.0.0
Phoenix	4.14.0	4.14.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.203	0.203	0.194	0.194
Spark	2.3.1	2.3.1	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.12

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.25.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.5.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.8.4-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.4	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.4	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.4	HBase 命令行客户端。
hbase-rest-server	1.4.4	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.4	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-1	Hive 命令行客户端。

组件	版本	描述
hive-hbase	2.3.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.203	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.203	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.1	Spark 命令行客户端。
spark-history-server	2.3.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。

组件	版本	描述
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.16.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.16.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.16.0	emr-5.15.1	emr-5.15.0	emr-5.14.2
Amazon SDK for Java	1.11.336	1.11.333	1.11.333	1.11.297
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.5.0	1.4.2	1.4.2	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.4	1.4.4	1.4.4	1.4.2
HCatalog	2.3.3	2.3.3	2.3.3	2.3.2
Hadoop	2.8.4	2.8.3	2.8.3	2.8.3
Hive	2.3.3	2.3.3	2.3.3	2.3.2
Hudi	-	-	-	-

	emr-5.16.0	emr-5.15.1	emr-5.15.0	emr-5.14.2
Hue	4.2.0	4.2.0	4.2.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.5.0	0.4.0	0.4.0	0.4.0
MXNet	1.2.0	1.1.0	1.1.0	1.1.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	5.0.0	4.3.0
Phoenix	4.14.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.203	0.194	0.194	0.194
Spark	2.3.1	2.3.0	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.12	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.16.0 的信息。更改与 5.15.0 有关。

首次发布日期：2018 年 7 月 19 日

升级

- Hadoop 2.8.4
- Flink 1.5.0
- Livy 0.5.0
- MXNet 1.2.0
- Phoenix 4.14.0
- Presto 0.203
- Spark 2.3.1
- Amazon SDK for Java 1.11.336
- CUDA 9.2
- Redshift JDBC 驱动程序 1.2.15.1025

更改、增强功能和解决的问题

- HBase
  - 已逆向移植 HBASE-20723。
- Presto
  - 更改了配置，可支持 LDAP 身份验证。有关更多信息，请参阅为 Presto on Amazon EMR 使用 LDAP 身份验证。
- Spark
  - Apache Spark 版本 2.3.1（从 Amazon EMR 发行版 5.16.0 开始提供）解决了 CVE-2018-8024 和 CVE-2018-1334 问题。建议您将 Spark 的早期版本迁移到 Spark 2.3.1 版本或更高版本。

已知问题

- 此发行版不支持 c1.medium 或 m1.small 实例类型。使用这些实例类型的集群将无法启动。解决方法：指定其它实例类型或使用其它发行版。

- 创建使用 Kerberos 的集群时，如果安装了 Livy，Livy 将失败，并显示未启用简单身份验证的错误。重新启动 Livy 服务器可解决此问题。解决方法是在集群创建过程中添加一个在主节点上运行 `sudo restart livy-server` 的步骤。
- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.6.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emrfs	2.25.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.5.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.4-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.4-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.4-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.4-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.4-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.4-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.8.4-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.4-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.4-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.4-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.4	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.4	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.4	HBase 命令行客户端。
hbase-rest-server	1.4.4	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.4	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-1	Hive 命令行客户端。

组件	版本	描述
hive-hbase	2.3.3-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.5.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.2.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.2.88	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.14.0-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.14.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.203	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.203	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.1	Spark 命令行客户端。
spark-history-server	2.3.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。

组件	版本	描述
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.16.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-password-authenticator	更改 Presto 的 password-authenticator.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.15.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.15.1	emr-5.15.0	emr-5.14.2	emr-5.14.1
Amazon SDK for Java	1.11.333	1.11.333	1.11.297	1.11.297
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.2	1.4.2	1.4.2	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.4	1.4.4	1.4.2	1.4.2
HCatalog	2.3.3	2.3.3	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.3	2.3.3	2.3.2	2.3.2
Hudi	-	-	-	-

	emr-5.15.1	emr-5.15.0	emr-5.14.2	emr-5.14.1
Hue	4.2.0	4.2.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.1.0	1.1.0	1.1.0	1.1.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	5.0.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.194	0.194
Spark	2.3.0	2.3.0	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.12	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.24.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.8.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.4	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.4	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.4	HBase 命令行客户端。
hbase-rest-server	1.4.4	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.4	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-0	Hive 命令行客户端。

组件	版本	描述
hive-hbase	2.3.3-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.3-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.1.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。

组件	版本	描述
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.15.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。

分类	描述
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。

分类	描述
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.15.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.15.0	emr-5.14.2	emr-5.14.1	emr-5.14.0
Amazon SDK for Java	1.11.333	1.11.297	1.11.297	1.11.297
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.2	1.4.2	1.4.2	1.4.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.4	1.4.2	1.4.2	1.4.2
HCatalog	2.3.3	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.3	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-

	emr-5.15.0	emr-5.14.2	emr-5.14.1	emr-5.14.0
Hue	4.2.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	0.8.1
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.1.0	1.1.0	1.1.0	1.1.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	5.0.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.194	0.194
Spark	2.3.0	2.3.0	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.7
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.12	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.15.0 的信息。更改与 5.14.0 有关。

首次发布日期：2018 年 6 月 21 日

升级

- 已将 HBase 升级到 1.4.4
- 已将 Hive 升级到 2.3.3
- 已将 Hue 升级到 4.2.0
- 已将 Oozie 升级到 5.0.0
- 已将 Zookeeper 升级到 3.4.12
- 已将 Amazon SDK 升级到 1.11.333

更改、增强功能和解决的问题

- Hive
  - 已逆向移植 HIVE-18069。
- Hue
  - 更新了 Hue，启用 Kerberos 后可以使用 Livy 正确地进行身份验证。现在，在 Amazon EMR 中使用 Kerberos 时，支持 Livy。
- JupyterHub
  - 更新了 JupyterHub，因此 Amazon EMR 默认安装 LDAP 客户端库。
  - 修复了生成自签名凭证的脚本中的错误。

已知问题

- 此发行版不支持 c1.medium 或 m1.small 实例类型。使用这些实例类型的集群将无法启动。解决方法：指定其它实例类型或使用其它发行版。
- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.24.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.4	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.4	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.4	HBase 命令行客户端。
hbase-rest-server	1.4.4	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.4	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.3-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.3-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.3-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.3-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.3-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.3-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.3-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.2.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.1.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	5.0.0	Oozie 命令行客户端。
oozie-server	5.0.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.12	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.12	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.15.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.14.2

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.14.2	emr-5.14.1	emr-5.14.0	emr-5.13.1
Amazon SDK for Java	1.11.297	1.11.297	1.11.297	1.11.297
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.2	1.4.2	1.4.2	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.2	1.4.2	1.4.2	1.4.2
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	0.8.1	-

	emr-5.14.2	emr-5.14.1	emr-5.14.0	emr-5.13.1
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.1.0	1.1.0	1.1.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.194	0.194
Spark	2.3.0	2.3.0	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.7	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.23.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.2	HBase 命令行客户端。
hbase-rest-server	1.4.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.1.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.14.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.14.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.14.1	emr-5.14.0	emr-5.13.1	emr-5.13.0
Amazon SDK for Java	1.11.297	1.11.297	1.11.297	1.11.297
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.2	1.4.2	1.4.0	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.2	1.4.2	1.4.2	1.4.2
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	0.8.1	-	-

	emr-5.14.1	emr-5.14.0	emr-5.13.1	emr-5.13.0
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.1.0	1.1.0	1.0.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.194	0.194
Spark	2.3.0	2.3.0	2.3.0	2.3.0
Sqoop	1.4.7	1.4.7	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.14.1 的信息。更改与 5.14.0 有关。

首次发布日期：2018 年 10 月 17 日

更新了 Amazon EMR 的默认 AMI，解决了潜在的安全漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.23.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.2	HBase 命令行客户端。
hbase-rest-server	1.4.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.1.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.14.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.14.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、JupyterHub、Livy、MXNet、Mahout、Oozie 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.14.0	emr-5.13.1	emr-5.13.0	emr-5.12.3
Amazon SDK for Java	1.11.297	1.11.297	1.11.297	1.11.267
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.2	1.4.0	1.4.0	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.2	1.4.2	1.4.2	1.4.0
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	0.8.1	-	-	-

	emr-5.14.0	emr-5.13.1	emr-5.13.0	emr-5.12.3
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.1.0	1.0.0	1.0.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.194	0.188
Spark	2.3.0	2.3.0	2.3.0	2.2.1
Sqoop	1.4.7	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.14.0 的信息。更改与 5.13.0 有关。

首次发布日期：2018 年 6 月 4 日

升级

- 已将 Apache Flink 升级到 1.4.2
- 已将 Apache MXnet 升级到 1.1.0

- 已将 Apache Sqoop 升级到 1.4.7

新特征

- 添加了对 JupyterHub 的支持。有关更多信息，请参阅JupyterHub。

更改、增强功能和解决的问题

- EMRFS
  - 更新了对 Amazon S3 的 userAgent 字符串请求，更新为包含调用委托人的用户和组信息。这可以与 Amazon CloudTrail 日志结合使用，来获取更全面的请求跟踪。
- HBase
  - 提供了 HBASE-20447，它解决了可能导致缓存问题的问题，特别是拆分区域。
- MXnet
  - 新增了 OpenCV 库。
- Spark
  - Spark 使用 EMRFS 将 Parquet 文件写入 Amazon S3 位置时，FileOutputCommitter 算法已更新为使用版本 2 而非版本 1。这将减少重命名的数量，从而提高应用程序性能。此更改不会影响：
    - Spark 以外的应用程序。
    - 写入其它文件系统的应用程序，例如 HDFS（仍然使用 FileOutputCommitter 版本 1）。
    - 使用其它输出格式（如文本或 csv）的应用程序（已使用 EMRFS 直接写入）。

已知问题

- JupyterHub
  - 不支持在创建集群时使用配置分类设置 JupyterHub 和单个 Jupyter notebook。手动编辑每个用户的 jupyterhub_config.py 文件和 jupyter_notebook_config.py 文件。有关更多信息，请参阅配置 JupyterHub。
  - JupyterHub 无法在私有子网内的集群上启动，并显示消息 `Error: ENOENT: no such file or directory, open '/etc/jupyter/conf/server.crt'` 。这由生成自签名凭证的脚本中的错误所致。使用以下解决方法生成自签名凭证。在连接到主节点时执行所有命令。

    1. 将凭证生成脚本从容器复制到主节点：

    ```
    sudo docker cp jupyterhub:/tmp/gen_self_signed_cert.sh ./
    ```

2. 使用文本编辑器更改第 23 行，将公有主机名更改为本地主机名，如下所示：

```
local hostname=$(curl -s $EC2_METADATA_SERVICE_URI/local-hostname)
```

3. 运行脚本，生成自签名凭证：

```
sudo bash ./gen_self_signed_cert.sh
```

4. 将脚本生成的凭证文件移至 /etc/jupyter/conf/ 目录：

```
sudo mv /tmp/server.crt /tmp/server.key /etc/jupyter/conf/
```

您可以对 jupyter.log 文件执行 tail，来验证 JupyterHub 是否重新启动并返回 200 响应代码。例如：

```
tail -f /var/log/jupyter/jupyter.log
```

该命令应返回与以下示例类似的响应：

```
# [I 2018-06-14 18:56:51.356 JupyterHub app:1581] JupyterHub is now running at
 https://:9443/
# 19:01:51.359 - info: [ConfigProxy] 200 GET /api/routes
```

- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.23.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.2	HBase 命令行客户端。
hbase-rest-server	1.4.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.4.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
jupyterhub	0.8.1	Jupyter notebook 的多用户服务器
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。

组件	版本	描述
mxnet	1.1.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
opencv	3.4.0	开源计算机视觉库。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.7	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.14.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
container-log4j	更改 Hadoop YARN 的 container-log4j.properties 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。

分类	描述
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。

分类	描述
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
jupyter-notebook-conf	更改 Jupyter notebook 的 jupyter_notebook_config.py 文件中的值。
jupyter-hub-conf	更改 JupyterHubs 的 jupyterhub_config.py 文件中的值。
jupyter-sparkmagic-conf	更改 Sparkmagic 的 config.json 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。

分类	描述
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。

分类	描述
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.13.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.13.1	emr-5.13.0	emr-5.12.3	emr-5.12.2
Amazon SDK for Java	1.11.297	1.11.297	1.11.267	1.11.267
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用

	emr-5.13.1	emr-5.13.0	emr-5.12.3	emr-5.12.2
Delta	-	-	-	-
Flink	1.4.0	1.4.0	1.4.0	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.2	1.4.2	1.4.0	1.4.0
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	1.0.0	1.0.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.194	0.188	0.188
Spark	2.3.0	2.3.0	2.2.1	2.2.1

	emr-5.13.1	emr-5.13.0	emr-5.12.3	emr-5.12.2
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.22.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.2	HBase 命令行客户端。
hbase-rest-server	1.4.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.4.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.13.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。

分类	描述
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。

分类	描述
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.13.0

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.13.0	emr-5.12.3	emr-5.12.2	emr-5.12.1
Amazon SDK for Java	1.11.297	1.11.267	1.11.267	1.11.267
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.0	1.4.0	1.4.0	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.2	1.4.0	1.4.0	1.4.0
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3

	emr-5.13.0	emr-5.12.3	emr-5.12.2	emr-5.12.1
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	1.0.0	1.0.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.194	0.188	0.188	0.188
Spark	2.3.0	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3

	emr-5.13.0	emr-5.12.3	emr-5.12.2	emr-5.12.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.13.0 的信息。更改与 5.12.0 有关。

升级

- 已将 Spark 升级到 2.3.0
- 已将 HBase 升级到 1.4.2
- 已将 Presto 升级到 0.194
- 已将 Amazon SDK for Java 升级到 1.11.297

更改、增强功能和解决的问题

- Hive
  - 已逆向移植 HIVE-15436。增强了 Hive API 功能，仅返回视图。

已知问题

- MXNet 目前暂无 OpenCV 库。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.10.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.22.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.2	HBase 命令行客户端。
hbase-rest-server	1.4.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.4.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-2	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-2	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-2	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-2	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-2	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.194	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.194	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
r	3.4.1	用于统计计算的 R 项目
spark-client	2.3.0	Spark 命令行客户端。
spark-history-server	2.3.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.3.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.3.0	YARN 从属项所需的 Apache Spark 库。

组件	版本	描述
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.13.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。

分类	描述
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。

分类	描述
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.12.3

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.12.3	emr-5.12.2	emr-5.12.1	emr-5.12.0
Amazon SDK for Java	1.11.267	1.11.267	1.11.267	1.11.267
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.0	1.4.0	1.4.0	1.4.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.0	1.4.0	1.4.0	1.4.0
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.8.3

	emr-5.12.3	emr-5.12.2	emr-5.12.1	emr-5.12.0
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.1.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	1.0.0	1.0.0	1.0.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.13.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.188	0.188	0.188	0.188
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3

	emr-5.12.3	emr-5.12.2	emr-5.12.1	emr-5.12.0
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.9.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.21.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.0	HBase 命令行客户端。
hbase-rest-server	1.4.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.2-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.188	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.188	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.12.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.12.2

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.12.2	emr-5.12.1	emr-5.12.0	emr-5.11.4
Amazon SDK for Java	1.11.267	1.11.267	1.11.267	1.11.238
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.0	1.4.0	1.4.0	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.0	1.4.0	1.4.0	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.8.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.1.0	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-

	emr-5.12.2	emr-5.12.1	emr-5.12.0	emr-5.11.4
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	1.0.0	1.0.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.13.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.188	0.188	0.188	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.12.2 的信息。更改与 5.12.1 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.9.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.21.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.0	HBase 命令行客户端。
hbase-rest-server	1.4.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.2-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.188	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.188	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.12.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值

分类	描述
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.12.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.12.1	emr-5.12.0	emr-5.11.4	emr-5.11.3
Amazon SDK for Java	1.11.267	1.11.267	1.11.238	1.11.238
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.0	1.4.0	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.12.1	emr-5.12.0	emr-5.11.4	emr-5.11.3
HBase	1.4.0	1.4.0	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.8.3	2.7.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.1.0	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	1.0.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.13.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.188	0.188	0.187	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.12.1	emr-5.12.0	emr-5.11.4	emr-5.11.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.12.1 的信息。更改与 5.12.0 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.9.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.21.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。

组件	版本	描述
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.0	HBase 命令行客户端。
hbase-rest-server	1.4.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.0	用于向 HBase 提供 Thrift 终端节点的服务。

组件	版本	描述
hcatalog-client	2.3.2-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包

组件	版本	描述
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.188	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.188	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.12.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.12.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.12.0	emr-5.11.4	emr-5.11.3	emr-5.11.2
Amazon SDK for Java	1.11.267	1.11.238	1.11.238	1.11.238
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.4.0	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.4.0	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.8.3	2.7.3	2.7.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.1.0	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-

	emr-5.12.0	emr-5.11.4	emr-5.11.3	emr-5.11.2
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	1.0.0	0.12.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.13.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.188	0.187	0.187	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.12.0 的信息。更改与 5.11.1 有关。

升级

- Amazon SDK for Java1.11.238 升级到1.11.267。有关更多信息，请参阅 GitHub 上的 Amazon SDK for Java 更改日志。
- Hadoop 2.7.3 升级到 2.8.3。有关更多信息，请参阅 Apache Hadoop 发行版。

- Fink 1.3.2 升级到 1.4.0。有关详细信息，请参阅 Apache Flink 1.4.0 版本公告。
- HBase 1.3.1 升级到 1.4.0。有关详细信息，请参阅 HBase 版本公告。
- Hue 4.0.1 升级到 4.1.0。有关更多信息，请参阅发布说明。
- MxNet 0.12.0 升级到 1.0.0。有关更多信息，请参阅 GitHub 上的 MXNet 更改日志。
- Presto 0.187 升级到 0.188。有关更多信息，请参阅发布说明。

更改、增强功能和解决的问题

- Hadoop
  - `yarn.resourcemanager.decommissioning.timeout` 属性已更改为 `yarn.resourcemanager.nodemanager-graceful-decommission-timeout-secs`。您可以使用此属性自定义集群缩减。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群缩减。
  - Hadoop CLI 向 `cp`（复制）命令添加了 `-d` 选项，可指定直接复制。可以使用它来避免创建中间 `._COPYING_` 文件，这加快了在 Amazon S3 之间复制数据的速度。有关更多信息，请参阅 HADOOP-12384。
- Pig
  - 添加了 `pig-env` 配置分类，这简化了 Pig 环境属性的配置。有关更多信息，请参阅配置应用程序。
- Presto
  - 新增 `presto-connector-redshift` 配置分类，您可以将其用于配置 Presto `redshift.properties` 配置文件中的值。有关更多信息，请参阅 Presto 文档中 Redshift 连接器以及 配置应用程序。
  - 已添加对 EMRFS 的 Presto 支持，且已设为默认配置。Amazon EMR 早期发行版使用 PrestoS3FileSystem，它是唯一选项。有关更多信息，请参阅EMRFS 和 PrestoS3FileSystem 配置。

> **Note**
>
> 如果您使用 Amazon EMR 版本 5.12.0 查询 Amazon S3 中的底层数据，则可能会出现 Presto 错误。这是因为 Presto 无法从 `emrfs-site.xml` 提取配置分类值。解决方法是在 `usr/lib/presto/plugin/hive-hadoop2/` 下创建一个 `emrfs` 子目录，并在 `usr/lib/presto/plugin/hive-hadoop2/emrfs` 中创建一个指向现有 `/usr/share/aws/emr/emrfs/conf/emrfs-site.xml` 文件的符号链接。然后重新

启动 presto-server 进程（首先执行 `sudo presto-server stop`，然后执行 `sudo presto-server start`）。

- Spark
  - 已逆向移植 SPARK-22036：BigDecimal 乘法运算有时会返回空值。

已知问题

- MXNet 不包含 OpenCV 库。
- SparkR 不适用于使用自定义 AMI 创建的集群，因为默认情况下不会在集群节点上安装 R。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*-amzn-*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0.1	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.9.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.21.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.4.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.8.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.8.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.8.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.8.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.8.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.8.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.8.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.8.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.8.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.8.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.4.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.4.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.4.0	HBase 命令行客户端。
hbase-rest-server	1.4.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.4.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-1	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.2-amzn-1	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-1	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-1	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-1	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.1.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	1.0.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.1.85	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.13.0-HBase-1.4	服务器和客户端的 phoenix 库
phoenix-query-server	4.13.0-HBase-1.4	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.188	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.188	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.12.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-env	更改 Pig 环境中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-redshift	更改 Presto 的 redshift.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.11.4

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.11.4	emr-5.11.3	emr-5.11.2	emr-5.11.1
Amazon SDK for Java	1.11.238	1.11.238	1.11.238	1.11.238
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-

	emr-5.11.4	emr-5.11.3	emr-5.11.2	emr-5.11.1
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.187	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.8.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.11.4 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值

分类	描述
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.11.3

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.11.3	emr-5.11.2	emr-5.11.1	emr-5.11.0
Amazon SDK for Java	1.11.238	1.11.238	1.11.238	1.11.238
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.11.3	emr-5.11.2	emr-5.11.1	emr-5.11.0
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.2
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.2
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.187	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.11.3	emr-5.11.2	emr-5.11.1	emr-5.11.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.11.3 的信息。更改与 5.11.2 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0	Amazon SageMaker Spark 开发工具包

组件	版本	描述
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.8.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-6	用于存储数据块的 HDFS 节点级服务。

组件	版本	描述
hadoop-hdfs-library	2.7.3-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。

组件	版本	描述
hcatalog-client	2.3.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包

组件	版本	描述
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.11.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。

分类	描述
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。

分类	描述
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.11.2

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.11.2	emr-5.11.1	emr-5.11.0	emr-5.10.1
Amazon SDK for Java	1.11.238	1.11.238	1.11.238	1.11.221
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.2	2.3.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.2	2.3.2	2.3.2	2.3.1
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.11.2	emr-5.11.1	emr-5.11.0	emr-5.10.1
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.187	0.187
Spark	2.2.1	2.2.1	2.2.1	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.11.2 的信息。更改与 5.11.1 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.8.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.11.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。

分类	描述
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。

分类	描述
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。

分类	描述
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.11.1

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.11.1	emr-5.11.0	emr-5.10.1	emr-5.10.0
Amazon SDK for Java	1.11.238	1.11.238	1.11.221	1.11.221
Python	2.7、3.4	2.7、3.4	2.7、3.4	2.7、3.4
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.2	2.3.1	2.3.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3

	emr-5.11.1	emr-5.11.0	emr-5.10.1	emr-5.10.0
Hive	2.3.2	2.3.2	2.3.1	2.3.1
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	0.12.0	0.12.0
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.187	0.187
Spark	2.2.1	2.2.1	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.3

	emr-5.11.1	emr-5.11.0	emr-5.10.1	emr-5.10.0
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.11.1 的信息。更改与 Amazon EMR 5.8.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.8.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-6	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.11.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。

分类	描述
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。

分类	描述
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.11.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.11.0	emr-5.10.1	emr-5.10.0	emr-5.9.1
Amazon SDK for Java	1.11.238	1.11.221	1.11.221	1.11.183
Python	2.7、3.4	2.7、3.4	2.7、3.4	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.2	2.3.1	2.3.1	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.2	2.3.1	2.3.1	2.3.0
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	0.12.0	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.11.0	emr-5.10.1	emr-5.10.0	emr-5.9.1
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.187	0.184
Spark	2.2.1	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.3	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.11.0 的信息。更改与 5.10.0 有关。

升级

- Hive 2.3.2
- Spark 2.2.1
- SDK for Java1.11.238

新特征

- Spark

- 增加了 `spark.decommissioning.timeout.threshold` 设置，这将改进使用竞价型实例时的 Spark 停用行为。有关更多信息，请参阅配置节点停用行为。
- 向 Spark 添加了 `aws-sagemaker-spark-sdk` 组件，此组件将安装 Amazon SageMaker Spark 和用于 Spark 与 Amazon SageMaker 集成的关联依赖项。您可以使用 Amazon SageMaker Spark 构造使用 Amazon SageMaker 阶段的 Spark 机器学习 (ML) 管道。有关更多信息，请参阅 GitHub 上的 SageMaker Spark 自述文件和《Amazon SageMaker 开发人员指南》https://docs.amazonaws.cn/sagemaker/latest/dg/apache-spark.html中的将 Apache Spark 与 Amazon SageMaker 结合使用。

已知问题

- MXNet 不包含 OpenCV 库。
- 默认情况下，Hive 2.3.2 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
aws-sagemaker-spark-sdk	1.0	Amazon SageMaker Spark 开发工具包
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.8.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-6	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.7.3-amzn-6	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-6	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-6	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-6	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-6	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-6	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-6	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-6	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-6	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.2-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.2-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.2-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.2-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.2-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.2-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.2-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.1	Spark 命令行客户端。
spark-history-server	2.2.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。

组件	版本	描述
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.11.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。

分类	描述
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。

分类	描述
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。

分类	描述
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。

分类	描述
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.10.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.10.1	emr-5.10.0	emr-5.9.1	emr-5.9.0
Amazon SDK for Java	1.11.221	1.11.221	1.11.183	1.11.183
Python	2.7、3.4	2.7、3.4	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.2
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.1	2.3.1	2.3.0	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.1	2.3.1	2.3.0	2.3.0
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	4.0.1
Iceberg	-	-	-	-

	emr-5.10.1	emr-5.10.0	emr-5.9.1	emr-5.9.0
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	0.4.0
MXNet	0.12.0	0.12.0	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.17.0
Presto	0.187	0.187	0.184	0.184
Spark	2.2.0	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.3	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.7.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-5	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.3.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.10.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值

分类	描述
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.10.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、MXNet、Mahout、Oozie、Phoenix、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.10.0	emr-5.9.1	emr-5.9.0	emr-5.8.3
Amazon SDK for Java	1.11.221	1.11.183	1.11.183	1.11.160
Python	2.7、3.4	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.2	1.3.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.10.0	emr-5.9.1	emr-5.9.0	emr-5.8.3
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.1	2.3.0	2.3.0	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.1	2.3.0	2.3.0	2.3.0
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	4.0.1	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	0.4.0	-
MXNet	0.12.0	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.17.0	0.16.0
Presto	0.187	0.184	0.184	0.170
Spark	2.2.0	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.10.0	emr-5.9.1	emr-5.9.0	emr-5.8.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.3	0.7.2	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.10.0 的信息。更改与 Amazon EMR 5.9.0 发行版有关。

升级

- Amazon SDK for Java 1.11.221
- Hive 2.3.1
- Presto 0.187

新特征

- 添加了对 Kerberos 身份验证的支持。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Kerberos 身份验证。
- 添加了对适用于 EMRFS 的 IAM 角色的支持。有关更多信息，请参阅《Amazon EMR 管理指南》中的为处理 EMRFS 对 Amazon S3 的请求配置 IAM 角色。
- 添加了对基于 GPU 的 P2 和 P3 实例类型的支持。有关更多信息，请参阅 Amazon EC2 P2 实例和 Amazon EC2 P3 实例。NVIDIA 驱动程序 384.81 和 CUDA 驱动程序 9.0.176 默认安装在这些实例类型上。
- 添加了对 Apache MXNet 的支持。

更改、增强功能和解决的问题

- Presto
  - 添加了对使用 Amazon Glue 数据目录作为默认 Hive 元存储的支持。有关更多信息，请参阅将 Presto 与 Amazon Glue 数据目录结合使用。

  - 添加了对地理空间函数的支持。
  - 为联接添加了溢出到磁盘支持。
  - 添加了对 Redshift 连接器的支持。
- Spark
  - 已逆向移植 SPARK-20640，这使随机注册的 rpc 超时值和重试次数值可使用 `spark.shuffle.registration.timeout` 和 `spark.shuffle.registration.maxAttempts` 属性进行配置。
  - 已逆向移植 SPARK-21549，这更正了在将自定义 OutputFormat 写入非 HDFS 位置时出现的错误。
- 已逆向移植 Hadoop 13270
- 从基本 Amazon EMR AMI 中删除了 Numpy、Scipy 和 Matplotlib 库。如果您的应用程序需要这些库，应用程序存储库中提供了它们，因此您可以通过引导操作使用 `yum install` 在所有节点上安装它们。
- Amazon EMR 基本 AMI 不再包含应用程序 RPM 软件包，因此集群节点上不再存在 RPM 软件包。自定义 AMI 和 Amazon EMR 基本 AMI 现在引用 Amazon S3 中的 RPM 软件包存储库。
- 因为 Amazon EC2 中引入了按秒计费，默认的 Scale down behavior (缩减行为) 现在为 Terminate at task completion (在任务完成时终止) 而非 Terminate at instance hour (在实例小时边界终止)。有关更多信息，请参阅配置集群缩减。

已知问题

- MXNet 不包含 OpenCV 库。
- 默认情况下，Hive 2.3.1 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.5.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.7.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.20.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-5	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-5	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-5	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-5	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-5	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-5	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-5	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-5	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-5	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-5	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mxnet	0.12.0	用于深度学习的灵活的、可扩展且高效的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
nvidia-cuda	9.0.176	Nvidia 驱动程序和 Cuda 工具包
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.187	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.187	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.3	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.10.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。

分类	描述
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。

分类	描述
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。

分类	描述
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.9.1

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、Mahout、Oozie、Phoenix、Pig、Presto、Spark、Sqoop、Tez、Zeppelin 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.9.1	emr-5.9.0	emr-5.8.3	emr-5.8.2
Amazon SDK for Java	1.11.183	1.11.183	1.11.160	1.11.160
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.2	1.3.1	1.3.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.0	2.3.0	2.3.0	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3

	emr-5.9.1	emr-5.9.0	emr-5.8.3	emr-5.8.2
Hive	2.3.0	2.3.0	2.3.0	2.3.0
Hudi	-	-	-	-
Hue	4.0.1	4.0.1	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	0.4.0	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.17.0	0.16.0	0.16.0
Presto	0.184	0.184	0.170	0.170
Spark	2.2.0	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.2

	emr-5.9.1	emr-5.9.0	emr-5.8.3	emr-5.8.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.7.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.19.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.184	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.184	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.9.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。

分类	描述
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。

分类	描述
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。

分类	描述
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.9.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Livy、Mahout、Oozie、Phoenix、Pig、Presto、Spark、Sqoop、Tez、Zeppelin 和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.9.0	emr-5.8.3	emr-5.8.2	emr-5.8.1
Amazon SDK for Java	1.11.183	1.11.160	1.11.160	1.11.160
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.2	1.3.1	1.3.1	1.3.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.0	2.3.0	2.3.0	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.0	2.3.0	2.3.0	2.3.0
Hudi	-	-	-	-
Hue	4.0.1	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	0.4.0	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0

	emr-5.9.0	emr-5.8.3	emr-5.8.2	emr-5.8.1
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.17.0	0.16.0	0.16.0	0.16.0
Presto	0.184	0.170	0.170	0.170
Spark	2.2.0	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.9.0 的信息。更改与 Amazon EMR 5.8.0 发行版有关。

发布日期：2017 年 10 月 5 日

最近功能更新时间：2017 年 10 月 12 日

升级

- Amazon SDK for Java 1.11.183 版
- Flink 1.3.2
- Hue 4.0.1
- Pig 0.17.0
- Presto 0.184

新特征

- 添加了 Livy 支持 (0.4.0-incubating 版)。有关更多信息，请参阅Apache Livy。
- 添加了对 Hue Notebook for Spark 的支持。
- 添加了对 i3 系列 Amazon EC2 实例的支持 (2017 年 10 月 12 日)。

更改、增强功能和解决的问题

- Spark
  - 添加了一组新功能，有助于确保 Spark 能够更为正常地处理因手动调整大小或自动扩展策略请求导致的节点终止。有关更多信息，请参阅配置节点停用行为。
  - 使用 SSL 取代 3DES 为数据块传输服务提供 in-transit 加密，可在使用带 AES-NI 的 Amazon EC2 实例类型时增强性能。
  - 已逆向移植 SPARK-21494。
- Zeppelin
  - 已逆向移植 ZEPPELIN-2377。
- HBase
  - 添加了补丁 HBASE-18533，因此可以使用 `hbase-site` 配置分类为 HBase BucketCache 配置使用其它值。
- Hue
  - 添加了对 Hue 中 Hive 查询编辑器的 Amazon Glue 数据目录支持。
  - 默认情况下，Hue 中的超级用户可以访问允许 Amazon EMR IAM 角色访问的所有文件。新建用户不会自动拥有对 Amazon S3 filebrowser 的访问权限，并且必须为其组启用 `filebrowser.s3_access` 权限。
- 解决造成使用 Amazon Glue 数据目录创建的底层 JSON 数据不可访问的问题。

已知问题

- 当安装了所有应用程序且未更改默认 Amazon EBS 根卷大小时，集群启动会失败。作为解决方法，请从 Amazon CLI 使用 `aws emr create-cluster` 命令并指定一个更大的 `--ebs-root-volume-size` 参数。
- 默认情况下，Hive 2.3.0 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含

`hive.compute.query.using.stats=true` 的表并且将新文件上载到表 LOCATION，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.7.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.19.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
flink-client	1.3.2	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.7.3-amzn-4	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。

组件	版本	描述
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	4.0.1	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
livy-server	0.4.0-incubating	用于与 Apache Spark 交互的 REST 接口
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.184	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.184	用于执行查询的各个部分的服务。
pig-client	0.17.0	Pig 命令行客户端。

组件	版本	描述
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.9.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值

分类	描述
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
livy-conf	更改 Livy 的 livy.conf 文件中的值。
livy-env	更改 Livy 环境中的值。
livy-log4j	更改 Livy log4j.properties 设置。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。

分类	描述
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.8.3

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.8.3	emr-5.8.2	emr-5.8.1	emr-5.8.0
Amazon SDK for Java	1.11.160	1.11.160	1.11.160	1.11.160
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.1	1.3.1	1.3.1	1.3.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.8.3	emr-5.8.2	emr-5.8.1	emr-5.8.0
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.0	2.3.0	2.3.0	2.3.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.0	2.3.0	2.3.0	2.3.0
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.2.0	2.2.0	2.2.0	2.2.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.8.3	emr-5.8.2	emr-5.8.1	emr-5.8.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.6.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-3	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-1	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.8.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.8.2

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.8.2	emr-5.8.1	emr-5.8.0	emr-5.7.1
Amazon SDK for Java	1.11.160	1.11.160	1.11.160	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.1	1.3.1	1.3.1	1.3.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.8.2	emr-5.8.1	emr-5.8.0	emr-5.7.1
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.0	2.3.0	2.3.0	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.0	2.3.0	2.3.0	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.2.0	2.2.0	2.2.0	2.1.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.8.2	emr-5.8.1	emr-5.8.0	emr-5.7.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.8.2 的信息。更改与 5.8.1 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.6.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-3	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-1	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.8.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.8.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.8.1	emr-5.8.0	emr-5.7.1	emr-5.7.0
Amazon SDK for Java	1.11.160	1.11.160	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.1	1.3.1	1.3.0	1.3.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.8.1	emr-5.8.0	emr-5.7.1	emr-5.7.0
HBase	1.3.1	1.3.1	1.3.1	1.3.1
HCatalog	2.3.0	2.3.0	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.0	2.3.0	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.11.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.2.0	2.2.0	2.1.1	2.1.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.8.1	emr-5.8.0	emr-5.7.1	emr-5.7.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.2
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.8.1 的信息。更改与 Amazon EMR 5.8.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.6.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-3	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-1	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.8.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.8.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.8.0	emr-5.7.1	emr-5.7.0	emr-5.6.1
Amazon SDK for Java	1.11.160	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.1	1.3.0	1.3.0	1.2.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.8.0	emr-5.7.1	emr-5.7.0	emr-5.6.1
HBase	1.3.1	1.3.1	1.3.1	1.3.0
HCatalog	2.3.0	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.3.0	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.11.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.2.0	2.1.1	2.1.1	2.1.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.8.0	emr-5.7.1	emr-5.7.0	emr-5.6.1
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.2	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.8.0 的信息。更改与 Amazon EMR 5.7.0 发行版有关。

首次发布日期：2017 年 8 月 10 日

最近功能更新时间：2017 年 9 月 25 日

升级

- Amazon SDK 1.11.160
- Flink 1.3.1
- Hive 2.3.0。有关更多信息，请参阅 Apache Hive 站点上的发布说明。
- Spark 2.2.0。有关更多信息，请参阅 Apache Spark 站点上的发布说明。

新特征

- 添加了对查看应用程序历史记录的支持 (2017 年 9 月 25 日)。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。

更改、增强功能和解决的问题

- 与 Amazon Glue 数据目录集成
  - 为 Hive 和 Spark SQL 增加了使用 Amazon Glue 数据目录作为 Hive 元数据存储的功能。有关更多信息，请参阅将 Amazon Glue 数据目录用作 Hive 元存储。和使用 Amazon Glue 数据目录作为 Spark SQL 的元存储。

- 已向集群详细信息添加 Application history (应用程序历史记录)，这可让您查看 YARN 应用程序的历史数据以及 Spark 应用程序的其它详细信息。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。
- Oozie
  - 已逆向移植 OOZIE-2748。
- Hue
  - 已逆向移植 HUE-5859
- HBase
  - 添加了补丁，以使用 `getMasterInitializedTime` 通过 Java 管理扩展 (JMX) 公开 HBase 主服务器启动时间。
  - 添加了改进集群启动时间的补丁。

已知问题

- 当安装了所有应用程序且未更改默认 Amazon EBS 根卷大小时，集群启动会失败。作为解决方法，请从 Amazon CLI 使用 `aws emr create-cluster` 命令并指定一个更大的 `--ebs-root-volume-size` 参数。
- 默认情况下，Hive 2.3.0 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。
- Spark – 在使用 Spark 时，apppusher 进程守护程序存在文件处理程序泄漏问题，长时间运行的 Spark 任务在几个小时或几天后可能会出现此情况。要解决此问题，请连接到主节点并键入 `sudo /etc/init.d/apppusher stop`。这将停止 apppusher 进程守护程序，而 Amazon EMR 将自动重新启动它。
- 应用程序历史记录
  - 死 Spark 执行程序的历史数据不可用。
  - 应用程序历史记录对使用安全配置来启用传输中加密的集群不可用。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.4.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.4.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.4.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.6.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.3.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.3.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.3.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.3.0-amzn-0	Hive 命令行客户端。
hive-hbase	2.3.0-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.3.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.3.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-1	Pig 命令行客户端。
spark-client	2.2.0	Spark 命令行客户端。
spark-history-server	2.2.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.2.0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.2.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.8.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。

分类	描述
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。

分类	描述
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.7.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.7.1	emr-5.7.0	emr-5.6.1	emr-5.6.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.0	1.3.0	1.2.1	1.2.1
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.1	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-

	emr-5.7.1	emr-5.7.0	emr-5.6.1	emr-5.6.0
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.13.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.11.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.1	2.1.1	2.1.1	2.1.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.2	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.0	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.1	Spark 命令行客户端。
spark-history-server	2.1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.7.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。

分类	描述
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。

分类	描述
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。

分类	描述
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值

分类	描述
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.7.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.7.0	emr-5.6.1	emr-5.6.0	emr-5.5.4
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.3.0	1.2.1	1.2.1	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.1	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-

	emr-5.7.0	emr-5.6.1	emr-5.6.0	emr-5.5.4
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.13.0	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.11.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.1	2.1.1	2.1.1	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.2	0.7.1	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.7.0 的信息。更改与 Amazon EMR 5.6.0 发行版有关。

发布日期：2017 年 7 月 13 日

升级

- Flink 1.3.0
- Phoenix 4.11.0
- Zeppelin 0.7.2

新特征

- 添加了创建集群时指定自定义 Amazon Linux AMI 的功能。有关更多信息，请参阅使用自定义 AMI。

更改、增强功能和解决的问题

- HBase
  - 添加了配置 HBase 只读副本集群的功能。请参阅使用只读副本集群。
  - 多个错误修复和增强功能
- Presto – 添加了配置 `node.properties` 的功能。
- YARN – 添加了配置 `container-log4j.properties` 的功能
- Sqoop – 已逆向移植 SQOOP-2880，这将引入一个允许您设置 Sqoop 临时目录的参数。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.18.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.3.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.11.0-HBase-1.3	服务器和客户端的 phoenix 库
phoenix-query-server	4.11.0-HBase-1.3	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.1	Spark 命令行客户端。
spark-history-server	2.1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.7.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.6.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.6.1	emr-5.6.0	emr-5.5.4	emr-5.5.3
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.1	1.2.1	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.6.1	emr-5.6.0	emr-5.5.4	emr-5.5.3
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.13.0	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.1	2.1.1	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.6.1	emr-5.6.0	emr-5.5.4	emr-5.5.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。

组件	版本	描述
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.1	Spark 命令行客户端。
spark-history-server	2.1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.6.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.6.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.6.0	emr-5.5.4	emr-5.5.3	emr-5.5.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.1	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-5.6.0	emr-5.5.4	emr-5.5.3	emr-5.5.2
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.13.0	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.1	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4

	emr-5.6.0	emr-5.5.4	emr-5.5.3	emr-5.5.2
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.6.0 的信息。更改与 Amazon EMR 5.5.0 发行版有关。

发布日期：2017 年 6 月 5 日

升级

- Flink 1.2.1
- HBase 1.3.1
- Mahout 0.13.0。这是 Mahout 在 Amazon EMR 版本 5.0 及更高版本中支持 Spark 2.x 的第一个版本。
- Spark 2.1.1

更改、增强功能和解决的问题

- Presto
  - 添加了通过使用安全配置启用传输中加密，从而在 Presto 节点之间实现 SSL/TLS 安全通信的功能。有关更多信息，请参阅传输中的数据加密。
  - 已逆向移植 Presto 7661，它向 `VERBOSE` 语句添加了 `EXPLAIN ANALYZE` 选项，以报告有关查询计划的更详细、更低级别的统计数据。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.1	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.1	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.3.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.13.0	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.1	Spark 命令行客户端。
spark-history-server	2.1.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.6.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。

分类	描述
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。

分类	描述
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。

分类	描述
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-env	更改 Presto 的 presto-env.sh 文件中的值。
presto-node	更改 Presto 的 node.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。

分类	描述
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.5.4

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.5.4	emr-5.5.3	emr-5.5.2	emr-5.5.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.5.4	emr-5.5.3	emr-5.5.2	emr-5.5.1
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.16.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。

组件	版本	描述
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.5.4 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.5.3

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.5.3	emr-5.5.2	emr-5.5.1	emr-5.5.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.12.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.5.3	emr-5.5.2	emr-5.5.1	emr-5.5.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.170
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.1
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.10

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.5.3 的信息。更改与 5.5.2 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.16.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。

组件	版本	描述
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.5.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.5.2

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.5.2	emr-5.5.1	emr-5.5.0	emr-5.4.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.12.0	3.11.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.5.2	emr-5.5.1	emr-5.5.0	emr-5.4.1
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.170	0.166
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.1	0.7.0
ZooKeeper	3.4.10	3.4.10	3.4.10	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.5.2 的信息。更改与 5.5.1 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.16.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。

组件	版本	描述
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。

组件	版本	描述
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.5.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。

分类	描述
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.5.1

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.5.1	emr-5.5.0	emr-5.4.1	emr-5.4.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	1.2.0	1.2.0
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.3.0	1.3.0
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.12.0	3.11.0	3.11.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.5.1	emr-5.5.0	emr-5.4.1	emr-5.4.0
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.9.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.170	0.166	0.166
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.1	0.7.0	0.7.0
ZooKeeper	3.4.10	3.4.10	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.5.1 的信息。更改与 Amazon EMR 5.5.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.16.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.5.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.5.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.5.0	emr-5.4.1	emr-5.4.0	emr-5.3.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	1.2.0	flink-client
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.3.0	1.2.3
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.12.0	3.11.0	3.11.0	3.11.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-5.5.0	emr-5.4.1	emr-5.4.0	emr-5.3.2
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.9.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.170	0.166	0.166	0.157.1
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.1	0.7.0	0.7.0	0.6.2
ZooKeeper	3.4.10	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.5.0 的信息。更改与 Amazon EMR 5.4.0 发行版有关。

发布日期：2017 年 4 月 26 日

升级

- Hue 3.12
- Presto 0.170
- Zeppelin 0.7.1

- ZooKeeper 3.4.10

更改、增强功能和解决的问题

- Spark
  - 已将 Spark 补丁 (SPARK-20115) Fix DAGScheduler to recompute all the lost shuffle blocks when external shuffle service is unavailable 逆向移植到 2.1.0 版 Spark，此版本包含在本次发布中。
- Flink
  - Flink 现在使用 Scala 2.11 进行构建。如果您使用 Scala API 和库，我们建议您在项目中使用 Scala 2.11。
  - 解决了 `HADOOP_CONF_DIR` 和 `YARN_CONF_DIR` 默认值未正确设置，因此 `start-scala-shell.sh` 无法工作的问题。此外，还添加了使用 `env.hadoop.conf.dir` 或 `env.yarn.conf.dir` 配置类别中的 `/etc/flink/conf/flink-conf.yaml` 和 `flink-conf` 设置这些值的功能。
  - 推出了一个新的 EMR 特定的命令 `flink-scala-shell` 作为 `start-scala-shell.sh` 的包装程序。我们建议使用此命令而不是 `start-scala-shell`。新命令可简化执行。例如，`flink-scala-shell -n 2` 将使用任务并行度 2 启动 Flink Scala Shell。
  - 推出了一个新的 EMR 特定的命令 `flink-yarn-session` 作为 `yarn-session.sh` 的包装程序。我们建议使用此命令而不是 `yarn-session`。新命令可简化执行。例如，`flink-yarn-session -d -n 2` 将使用两个任务管理器在分离状态下启动长时间运行的 Flink 会话。
  - 解决了 (FLINK-6125) Commons httpclient is not shaded anymore in Flink 1.2 的问题。
- Presto
  - 添加了对 LDAP 身份验证的支持。将 LDAP 与 Presto on Amazon EMR 结合使用，需要您启用对 Presto 协调器的 HTTPS 访问（`config.properties` 中的 `http-server.https.enabled=true`）。有关配置详细信息，请参阅 Presto 文档中的 LDAP 身份验证。
  - 添加了对 `SHOW GRANTS` 的支持。
- Amazon EMR 基本 Linux AMI
  - Amazon EMR 发行版现在基于 Amazon Linux 2017.03。有关更多信息，请参阅 Amazon Linux AMI 2017.03 发布说明。
  - 从 Amazon EMR 基本 Linux 映像中删除了 Python 2.6。默认安装 Python 2.7 和 3.4。如果需要，您可以手动安装 Python 2.6。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.5.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.16.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	3.12.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.170	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.170	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.10	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.10	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.5.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.4.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.4.1	emr-5.4.0	emr-5.3.2	emr-5.3.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	1.2.0	flink-client	flink-client
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.3.0	1.2.3	1.2.3
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.11.0	3.11.0	3.11.0	3.11.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-5.4.1	emr-5.4.0	emr-5.3.2	emr-5.3.1
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.9.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.166	0.166	0.157.1	0.157.1
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.0	0.7.0	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.15.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	3.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.166	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.166	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。

组件	版本	描述
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.4.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。

分类	描述
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.4.0

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.4.0	emr-5.3.2	emr-5.3.1	emr-5.3.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.2.0	flink-client	flink-client	flink-client
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.3.0	1.2.3	1.2.3	1.2.3
HCatalog	2.1.1	2.1.1	2.1.1	2.1.1
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.1
Hudi	-	-	-	-
Hue	3.11.0	3.11.0	3.11.0	3.11.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-5.4.0	emr-5.3.2	emr-5.3.1	emr-5.3.0
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.3.0
Phoenix	4.9.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.166	0.157.1	0.157.1	0.157.1
Spark	2.1.0	2.1.0	2.1.0	2.1.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.7.0	0.6.2	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.4.0 的信息。更改与 Amazon EMR 5.3.0 发行版有关。

发布日期：2017 年 3 月 8 日

升级

- 已升级到 Flink 1.2.0
- 已升级到 Hbase 1.3.0
- 已升级到 Phoenix 4.9.0

> **Note**
>
> 如果您从早期版本的 Amazon EMR 升级到 Amazon EMR 发行版 5.4.0 或更高版本并使用二级索引，请升级本地索引，如 Apache Phoenix 文档中所述。Amazon EMR 将从 `hbase-site` 分类中删除所需配置，但索引需要重新填充。支持在线和离线升级索引。在线升级为默认值，这意味着，在从版本 4.8.0 或更高版本的 Phoenix 客户端初始化时重新填充索引。要指定离线升级，请在 `phoenix.client.localIndexUpgrade` 分类中将 `phoenix-site` 配置设置为 false，然后将 SSH 设置为主节点以运行 `psql [zookeeper] -1`。

- 已升级到 Presto 0.166
- 已升级到 Zeppelin 0.7.0

更改和增强功能

- 增加了对 r4 实例的支持。请参阅 Amazon EC2 实例类型。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.3.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.15.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.2.0	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.3.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.3.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.3.0	HBase 命令行客户端。
hbase-rest-server	1.3.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.3.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-hbase	2.1.1-amzn-0	Hive-hbase 客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server2	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.9.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.9.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.166	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.166	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.7.0	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.4.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。

分类	描述
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.3.2

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.3.2	emr-5.3.1	emr-5.3.0	emr-5.2.3
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	flink-client	flink-client	flink-client	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.1	2.1.1	2.1.1	2.1.0

	emr-5.3.2	emr-5.3.1	emr-5.3.0	emr-5.2.3
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.1	2.1.0
Hudi	-	-	-	-
Hue	3.11.0	3.11.0	3.11.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.3.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.157.1	0.157.1	0.157.1
Spark	2.1.0	2.1.0	2.1.0	2.0.2
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-

	emr-5.3.2	emr-5.3.1	emr-5.3.0	emr-5.2.3
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	emrfs	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	flink-client	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.3.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。

分类	描述
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.3.1

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.3.1	emr-5.3.0	emr-5.2.3	emr-5.2.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	flink-client	flink-client	1.1.3	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.1	2.1.1	2.1.0	2.1.0

	emr-5.3.1	emr-5.3.0	emr-5.2.3	emr-5.2.2
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.1	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.11.0	3.11.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.3.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.157.1	0.157.1	0.157.1
Spark	2.1.0	2.1.0	2.0.2	2.0.2
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-

	emr-5.3.1	emr-5.3.0	emr-5.2.3	emr-5.2.2
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.3.1 的信息。更改与 Amazon EMR 5.3.0 发行版有关。

发布日期：2017 年 2 月 7 日

进行了微小更改：逆向移植 Zeppelin 补丁，并更新了 Amazon EMR 的默认 AMI。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。

组件	版本	描述
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	emrfs	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	flink-client	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。

组件	版本	描述
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。

组件	版本	描述
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.3.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。

分类	描述
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。

分类	描述
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。

分类	描述
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值

分类	描述
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.3.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.3.0	emr-5.2.3	emr-5.2.2	emr-5.2.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	flink-client	1.1.3	1.1.3	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.1	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.1	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.11.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-

	emr-5.3.0	emr-5.2.3	emr-5.2.2	emr-5.2.1
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.3.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.157.1	0.157.1	0.157.1
Spark	2.1.0	2.0.2	2.0.2	2.0.2
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.3.0 的信息。更改与 Amazon EMR 5.2.1 发行版有关。

发布日期：2017 年 1 月 26 日

升级

- 已升级到 Hive 2.1.1
- 已升级到 Hue 3.11.0
- 已升级到 Spark 2.1.0
- 已升级到 Oozie 4.3.0
- 已升级到 Flink 1.1.4

更改和增强功能

- Hue 新增补丁可使您使用 `interpreters_shown_on_wheel` 设置配置解释器在笔记本选择轮盘上最先显示的内容，而不受 `hue.ini` 文件中排序的限制。
- 新增 `hive-parquet-logging` 配置分类，您可以将其用于配置 Hive `parquet-logging.properties` 文件中的值。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	emrfs	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	flink-client	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.1-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.1-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.1-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.1-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.1-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.1-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.11.0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.3.0	Oozie 命令行客户端。
oozie-server	4.3.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.1.0	Spark 命令行客户端。
spark-history-server	2.1.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.1.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.1.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.3.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-parquet-logging	更改 Hive parquet-logging.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。

分类	描述
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。

分类	描述
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。

分类	描述
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.2.3

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.2.3	emr-5.2.2	emr-5.2.1	emr-5.2.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	1.1.3	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0

	emr-5.2.3	emr-5.2.2	emr-5.2.1	emr-5.2.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.157.1	0.157.1	0.152.3
Spark	2.0.2	2.0.2	2.0.2	2.0.2
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-

	emr-5.2.3	emr-5.2.2	emr-5.2.1	emr-5.2.0
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2
ZooKeeper	3.4.9	3.4.9	3.4.9	3.4.8

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.13.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.2	Spark 命令行客户端。
spark-history-server	2.0.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.2.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.2.2

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.2.2	emr-5.2.1	emr-5.2.0	emr-5.1.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	1.1.3	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3

	emr-5.2.2	emr-5.2.1	emr-5.2.0	emr-5.1.1
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.157.1	0.152.3	0.152.3
Spark	2.0.2	2.0.2	2.0.2	2.0.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2

	emr-5.2.2	emr-5.2.1	emr-5.2.0	emr-5.1.1
ZooKeeper	3.4.9	3.4.9	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.2.2 的信息。更改与 Amazon EMR 5.2.1 发行版有关。

发布日期：2017 年 5 月 2 日

早期版本中已解决的已知问题

- 已逆向移植 SPARK-194459，解决了从包含 char/varchar 列的 ORC 表读取内容时可能失败的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.13.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。

组件	版本	描述
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.2	Spark 命令行客户端。
spark-history-server	2.0.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-5.2.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.2.1

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.2.1	emr-5.2.0	emr-5.1.1	emr-5.1.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	1.1.3	1.1.3
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.3
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3

	emr-5.2.1	emr-5.2.0	emr-5.1.1	emr-5.1.0
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.157.1	0.152.3	0.152.3	0.152.3
Spark	2.0.2	2.0.2	2.0.1	2.0.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.2

	emr-5.2.1	emr-5.2.0	emr-5.1.1	emr-5.1.0
ZooKeeper	3.4.9	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.2.1 的信息。更改与 Amazon EMR 5.2.0 发行版有关。

发布日期：2016 年 12 月 29 日

升级

- 已升级到 Presto 0.157.1。有关更多信息，请参阅 Presto 文档中的 Presto 发布说明。
- 已升级到 Zookeeper 3.4.9。有关更多信息，请参阅 Apache ZooKeeper 文档中的 ZooKeeper 发布说明。

更改和增强功能

- 在 Amazon EMR 4.8.3 及更高版本（但不包括 5.0.0、5.0.3 和 5.2.0 版）中添加了对 Amazon EC2 m4.16xlarge 实例类型的支持。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。
- 现在，Flink 和 YARN 配置路径的位置默认在 `/etc/default/flink` 中设置，您在运行 `flink` 或 `yarn-session.sh` 驱动程序脚本启动 Flink 任务时，无需设置环境变量 `FLINK_CONF_DIR` 和 `HADOOP_CONF_DIR`。
- 新增对 FlinkKinesisConsumer 类的支持。

早期版本中已解决的已知问题

- 修复了 Hadoop 中的一个问题，即 ReplicationMonitor 线程可能会因为在大型集群中复制和删除同一个文件导致的竞争而卡住很长时间。
- 修复了在作业状态未成功更新时 ControlledJob#toString 出现空指针异常 (NPE) 失败的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.13.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。

组件	版本	描述
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。

组件	版本	描述
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.2	Spark 命令行客户端。
spark-history-server	2.0.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。

组件	版本	描述
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.2.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。

分类	描述
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。

分类	描述
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。

分类	描述
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.2.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.2.0	emr-5.1.1	emr-5.1.0	emr-5.0.3
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	1.1.3	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.3	1.2.2
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.2.0	emr-5.1.1	emr-5.1.0	emr-5.0.3
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.152.3	0.152.3	0.152.3	0.152.3
Spark	2.0.2	2.0.1	2.0.1	2.0.1
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.2	0.6.1
ZooKeeper	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.2.0 的信息。更改与 Amazon EMR 5.1.0 发行版有关。

发布日期：2016 年 11 月 21 日

更改和增强功能

- 添加了适用于 HBase 的 Amazon S3 存储模式。
- 使您能够为 HBase rootdir 指定 Amazon S3 位置。有关更多信息，请参阅 Amazon S3 上的 HBase。

升级

- 已升级到 Spark 2.0.2

早期版本中已解决的已知问题

- 修复了限制为仅 EBS 实例类型上的 2 TB 的 /mnt 的问题。
- 修复了输出到相应的 .out 文件而不是常规 log4j 配置的 .log 文件 (每小时转动一次) 的 instance-controller 和 logpusher 日志的问题。.out 文件不会转动，因此这最终将填满 /emr 分区。此问题仅影响硬件虚拟机 (HVM) 实例类型。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.12.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.7.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。

组件	版本	描述
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.152.3	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.152.3	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.2	Spark 命令行客户端。

组件	版本	描述
spark-history-server	2.0.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.2.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase	适用于 Apache HBase 的 Amazon EMR 辅助设置。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。

分类	描述
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。

分类	描述
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.1.1

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.2	1.2.2
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.2
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.152.3	0.152.3	0.152.3	0.150
Spark	2.0.1	2.0.1	2.0.1	2.0.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.1	0.6.1
ZooKeeper	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.11.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.152.3	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.152.3	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.1	Spark 命令行客户端。
spark-history-server	2.0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.1.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。

分类	描述
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.1.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Flink、Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.2	1.2.2
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.2
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.152.3	0.152.3	0.152.3	0.150
Spark	2.0.1	2.0.1	2.0.1	2.0.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.1	0.6.1
ZooKeeper	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.1.0 的信息。更改与 Amazon EMR 5.0.3 发行版有关。

发布日期：2016 年 11 月 3 日

更改和增强功能

- 增加了对 Flink 1.1.3 的支持。
- Presto 已作为 Hue 的记事本部分中的选项添加。

升级

- 已升级到 HBase 1.2.3
- 已升级到 Zeppelin 0.6.2

早期版本中已解决的已知问题

- 修复了带 ORC 文件的 Amazon S3 上的 Tez 查询的性能低于早期 Amazon EMR 4.x 版本中的性能的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.11.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
flink-client	1.1.3	Apache Flink 命令行客户端脚本和应用程序。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.3	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.3	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.3	HBase 命令行客户端。
hbase-rest-server	1.2.3	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.3	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.152.3	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.152.3	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.1	Spark 命令行客户端。
spark-history-server	2.0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	2.0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.2	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.1.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
flink-conf	更改 flink-conf.yaml 设置。
flink-log4j	更改 Flink log4j.properties 设置。
flink-log4j-yarn-session	更改 Flink log4j-yarn-session.properties 设置。
flink-log4j-cli	更改 Flink log4j-cli.properties 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。

分类	描述
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.0.3

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、Spark、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.2	1.2.2
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.2
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.152.3	0.152.3	0.152.3	0.150
Spark	2.0.1	2.0.1	2.0.1	2.0.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.1	0.6.1
ZooKeeper	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 5.0.3 的信息。更改与 Amazon EMR 5.0.0 发行版有关。

发布日期：2016 年 10 月 24 日

升级

- 已升级到 Hadoop 2.7.3
- 已升级到 Presto 0.152.3，它包括对 Presto Web 界面的支持。可使用端口 8889 访问 Presto 协调器上的 Presto Web 界面。有关 Presto Web 界面的更多信息，请参阅 Presto 文档中的 Web 界面。
- 已升级到 Spark 2.0.1
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.10.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。

组件	版本	描述
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.152.3	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.152.3	用于执行查询的各个部分的服务。
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.1	Spark 命令行客户端。
spark-history-server	2.0.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.0.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。

分类	描述
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。

分类	描述
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。

分类	描述
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 5.0.0

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie、Phoenix、Pig、Presto、Spark、和 ZooKeeper。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	1.1.3	1.1.3	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.3	1.2.3	1.2.2	1.2.2
HCatalog	2.1.0	2.1.0	2.1.0	2.1.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.2

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
Hive	2.1.0	2.1.0	2.1.0	2.1.0
Hudi	-	-	-	-
Hue	3.10.0	3.10.0	3.10.0	3.10.0
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.16.0	0.16.0	0.16.0	0.16.0
Presto	0.152.3	0.152.3	0.152.3	0.150
Spark	2.0.1	2.0.1	2.0.1	2.0.0
Sqoop	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	0.6.2	0.6.2	0.6.1	0.6.1

	emr-5.1.1	emr-5.1.0	emr-5.0.3	emr-5.0.0
ZooKeeper	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

发布日期：2016 年 7 月 27 日

升级

- 已升级到 Hive 2.1
- 已升级到 Presto 0.150
- 已升级到 Spark 2.0
- 已升级到 Hue 3.10.0
- 已升级到 Pig 0.16.0
- 已升级到 Tez 0.8.4
- 已升级到 Zeppelin 0.6.1

更改和增强功能

- Amazon EMR 支持最新开源版本的 Hive（版本 2.1）和 Pig（版本 0.16.0）。如果您以前使用的是 Amazon EMR 上的 Hive 或 Pig，那么这可能会影响某些使用案例。有关更多信息，请参阅 Hive 和 Pig。
- Hive 和 Pig 的默认执行引擎现在是 Tez。要更改该设置，您可以在 `hive-site` 和 `pig-properties` 配置分类中分别编辑相应的值。
- 添加了增强型步骤调试功能，可让您查看步骤失败的根本原因 (如果服务可以确定原因)。有关更多信息，请参阅《Amazon EMR 管理指南》中的增强型步骤调试。
- 先前以“-Sandbox”结尾的应用程序不再拥有该后缀。这可能会中断您的自动化，例如，如果您使用脚本来启动具有这些应用程序的集群。下表显示了 Amazon EMR 4.7.2 与 Amazon EMR 5.0.0 中的应用程序名称。

  应用程序名称更改

Amazon EMR 4.7.2	Amazon EMR 5.0.0
Oozie-Sandbox	Oozie

Amazon EMR 4.7.2	Amazon EMR 5.0.0
Presto-Sandbox	Presto
Sqoop-Sandbox	Sqoop
Zeppelin-Sandbox	Zeppelin
ZooKeeper-Sandbox	ZooKeeper

- Spark 现在针对 Scala 2.11 进行编译。
- Java 8 现在是默认 JVM。所有应用程序均使用 Java 8 runtime 运行。对任何应用程序的字节代码目标都没有进行更改。大多数应用程序继续运行 Java 7。
- Zeppelin 现在包括身份验证功能。有关更多信息，请参阅 Zeppelin。
- 添加了对安全配置的支持，这使您可以更轻松地创建和应用加密选项。有关更多信息，请参阅数据加密。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.9.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-3	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.7.2-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	2.1.0-amzn-0	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	2.1.0-amzn-0	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	2.1.0-amzn-0	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	2.1.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	2.1.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	2.1.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.10.0-amzn-0	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.46	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.150	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.150	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.16.0-amzn-0	Pig 命令行客户端。
spark-client	2.0.0	Spark 命令行客户端。
spark-history-server	2.0.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	2.0.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	2.0.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.1-SNAPSHOT	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-5.0.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。

分类	描述
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j2	更改 HCatalog WebHCat 的 log4j2.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-beeline-log4j2	更改 Hive 的 beeline-log4j2.properties 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j2	更改 Hive 的 hive-exec-log4j2.properties 文件中的值。
hive-llap-daemon-log4j2	更改 Hive 的 llap-daemon-log4j2.properties 文件中的值。
hive-log4j2	更改 Hive 的 hive-log4j2.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。

分类	描述
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。

分类	描述
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-hive-site	更改 Spark 的 hive-site.xml 文件中的值
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。

分类	描述
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 4.x 发行版

本部分内容涵盖每个 Amazon EMR 4.x 发行版中可用的应用程序版本、发布说明、组件版本和配置分类。

启动集群时，有多个 Amazon EMR 发行版可供选择。这允许您测试和使用满足您解决方案兼容性需求的应用程序版本。可使用发行版标注指定版本。发行版标注的格式是 `emr-x.x.x`。For example, `emr-6.14.0`.

从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

有关每个 Amazon EMR 4.x 发行版本中应用程序版本的综合表格，请参阅Amazon EMR 4.x 发行版中的应用程序版本。

主题

- Amazon EMR 4.x 发行版中的应用程序版本
- 各 Amazon EMR 4.x 发行版之间的差异
- Amazon EMR 发行版 4.9.6
- Amazon EMR 发行版 4.9.5
- Amazon EMR 发行版 4.9.4
- Amazon EMR 发行版 4.9.3
- Amazon EMR 发行版 4.9.2

- Amazon EMR 发行版 4.9.1
- Amazon EMR 发行版 4.8.5
- Amazon EMR 发行版 4.8.4
- Amazon EMR 发行版 4.8.3
- Amazon EMR 发行版 4.8.2
- Amazon EMR 发行版 4.8.0
- Amazon EMR 发行版 4.7.4
- Amazon EMR 发行版 4.7.2
- Amazon EMR 发行版 4.7.1
- Amazon EMR 发行版 4.7.0
- Amazon EMR 发行版 4.6.0
- Amazon EMR 发行版 4.5.0
- Amazon EMR 发行版 4.4.0
- Amazon EMR 发行版 4.3.0
- Amazon EMR 发行版 4.2.0
- Amazon EMR 发行版 4.1.0
- Amazon EMR 发行版 4.0.0

Amazon EMR 4.x 发行版中的应用程序版本

有关列出每个 Amazon EMR 4.x 发行版中可用的应用程序版本的综合表格，请在浏览器打开 Amazon EMR 4.x 发行版中的应用程序版本。

各 Amazon EMR 4.x 发行版之间的差异

《Amazon EMR 管理指南》中有关 Amazon EMR 功能的文档指定开始提供某种功能的 Amazon EMR 发行版以及追溯至 4.0.0 的 Amazon EMR 功能之间的适用差异。

从 Amazon EMR 发行版 5.0.0 开始，一些应用程序已进行重大版本升级，安装或运行详细信息改变，另一些应用程序则从沙盒应用程序提升为本机应用程序。此节中的各个主题提供使用 Amazon EMR 4.x 发行版时特定于应用程序的明显差异。

主题

- 沙盒应用程序

- 使用 Amazon EMR 4.x 上的 Hive 的注意事项
- 使用 Amazon EMR 4.x 上的 Pig 的注意事项

沙盒应用程序

使用 Amazon EMR 4.x 发行版时，一些应用程序会被视为沙盒应用程序。沙盒应用程序是我们在初始 Amazon EMR 发行版时因需求提供的应用程序的早期版本。您可以使用控制台、Amazon CLI 或 API 让 Amazon EMR 安装沙盒应用程序（安装方式与本机应用程序的相同），但沙盒应用程序的支持和文档有限。沙盒应用程序在 Amazon EMR 发行版 5.0.0 及更高版本中变为完全受支持的本机应用程序。以下是 Amazon EMR 4.x 发行版中的沙盒应用程序。

- Oozie
- Presto
- Sqoop
- Zeppelin
- ZooKeeper

当您安装沙盒应用程序时，使用后缀 `-sandbox` 指示应用程序名称。例如，要安装沙盒版本的 *`Presto`*，请使用 `Presto-sandbox`。与完全受支持的应用程序相比，安装时间可能更长。此节中列出的每个应用程序的版本号对应于应用程序的社区版本。

Oozie（沙盒版本）

Oozie 从 Amazon EMR 发行版 4.1.0 开始作为沙盒应用程序提供。

默认情况下，不会使用沙盒版本安装 Oozie 示例。要安装以上示例，可使用 SSH 连接到主节点并运行 `install-oozie-examples`。

Oozie 沙盒版本信息

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.9.6	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
		httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.9.5	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.9.4	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.9.3	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.9.2	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.9.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.8.5	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.8.4	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.8.3	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.8.2	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.8.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.7.4	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.7.2	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.7.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.7.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.6.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.5.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.4.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Amazon EMR 发行版标签	Oozie 沙盒版本	随 Oozie 沙盒安装的组件
emr-4.3.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.2.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server
emr-4.1.0	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, oozie-client, oozie-server

Presto（沙盒版本）

Presto 从 Amazon EMR 发行版 4.1.0 开始作为沙盒应用程序提供。

Presto 沙盒版本信息

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.9.6	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.9.5	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.9.4	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server,

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
		presto-coordinator, presto-worker
emr-4.9.3	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.9.2	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.9.1	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.8.5	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.8.4	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.8.3	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.8.2	0.152.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.8.0	0.151	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.7.4	0.148	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.7.2	0.148	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.7.1	0.147	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-4.7.0	0.147	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.6.0	0.143	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker
emr-4.5.0	0.140	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker
emr-4.4.0	0.136	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 沙盒版本	随 Presto 沙盒安装的组件
emr-4.3.0	0.130	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker
emr-4.2.0	0.125	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker
emr-4.1.0	0.119	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, mysql-server, presto-coordinator, presto-worker

Sqoop（沙盒版本）

Sqoop 从 Amazon EMR 发行版 4.4.0 开始作为沙盒应用程序提供。

Sqoop 沙盒版本信息

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.9.6	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.9.5	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.9.4	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.9.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.9.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.9.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.8.5	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.8.4	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.8.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.8.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.8.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.7.4	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.7.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.7.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.7.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 沙盒版本	随 Sqoop 沙盒安装的组件
emr-4.6.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-4.5.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, sqoop-client
emr-4.4.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, sqoop-client

Zeppelin（沙盒版本）

Zeppelin 从 Amazon EMR 发行版 4.1.0 开始作为沙盒应用程序提供。

Zeppelin 沙盒版本信息

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.9.6	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.9.5	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.9.4	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
		on-yarn, spark-yarn-slave, zeppelin-server
emr-4.9.3	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.9.2	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.9.1	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.8.5	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.8.4	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.8.3	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.8.2	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.8.0	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.7.4	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.7.2	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.7.1	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.7.0	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.6.0	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.5.0	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.4.0	0.5.6	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.3.0	0.5.5	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-4.2.0	0.5.5	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 沙盒版本	随 Zeppelin 沙盒安装的组件
emr-4.1.0	0.6.0-SNAPSHOT	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Zookeeper（沙盒版本）

Zookeeper 从 Amazon EMR 发行版 4.6.0 开始作为沙盒应用程序提供。

ZooKeeper 沙盒版本信息

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
emr-4.9.6	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.9.5	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
		cemanager, zookeeper-client, zookeeper-server
emr-4.9.4	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.9.3	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.9.2	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
emr-4.9.1	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.8.5	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.8.4	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
emr-4.8.3	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.8.2	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.8.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
emr-4.7.4	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.7.2	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.7.1	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Zookeeper 沙盒版本	随 ZooKeeper 沙盒安装的组件
emr-4.7.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-4.6.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

使用 Amazon EMR 4.x 上的 Hive 的注意事项

本节介绍在使用 Amazon EMR 4.x 发行版上的 Hive 版本 1.0.0 时要注意的差异（与 Amazon EMR 5.x 发行版上的 Hive 2.x 相比）。

不支持 ACID 事务

使用 Amazon EMR 4.x 发行版时，4.x 发行版上的 Hive 不支持 Hive 数据存储在 Amazon S3 中的 ACID 事务。如果您尝试在 Amazon S3 中创建事务表，将出现异常。

对 Amazon S3 中的表的读写操作

Amazon EMR 4.x 发行版上的 Hive 可直接写入 Amazon S3，无需使用临时文件。这可以改善性能，但导致您无法在同一个 Hive 语句内读写 Amazon S3 中的同一个表。一种解决办法是在 HDFS 中创建并使用临时表。

以下示例显示了如何使用多 Hive 语句更新 Amazon S3 中的表。这些语句基于 Amazon S3 中一个名为 `my_s3_table` 的表在 HDFS 中创建一个名为 `tmp` 的临时表。之后将使用此临时表的内容更新 Amazon S3 中的表。

```
CREATE TEMPORARY TABLE tmp LIKE my_s3_table;
INSERT OVERWRITE TABLE tmp SELECT ....;
INSERT OVERWRITE TABLE my_s3_table SELECT * FROM tmp;
```

Log4j 与 Log4j 2 的对比

Amazon EMR 4.x 发行版上的 Hive 使用 Log4j。从版本 5.0.0 开始，默认情况下使用 Log4j 2。这些版本可能需要不同的日志记录配置。有关更多信息，请参阅 Apache Log4j 2。

MapReduce 是默认执行引擎

Amazon EMR 4.x 发行版上的 Hive 使用 MapReduce 作为默认执行引擎。从 Amazon EMR 版本 5.0.0 开始，Tez 为默认引擎，这将改善大多数工作流的性能。

Hive 授权

Amazon EMR 4.x 发行版上的 Hive 对于 HDFS 支持 Hive 授权，但对于 EMRFS 和 Amazon S3 不支持此授权。默认情况下，Amazon EMR 集群在禁用授权的状态下运行。

Amazon S3 中的 Hive 文件合并操作

如果 `hive.merge.mapfiles` 为 `true`，Amazon EMR 4.x 发行版上的 Hive 将在仅映射任务结束时合并小型文件。仅当任务的平均输出大小低于 `hive.merge.smallfiles.avgsize` 设置时，才会触发合并。如果最终输出路径位于 HDFS 中，那么 Amazon EMR Hive 的行为将完全相同。但是，如果输出路径位于 Amazon S3 中，将忽略 `hive.merge.smallfiles.avgsize` 参数。在那种情况下，如果 `hive.merge.mapfiles` 设置为 `true`，会始终触发合并任务。

使用 Amazon EMR 4.x 上的 Pig 的注意事项

Pig 版本 0.14.0 安装在使用 Amazon EMR 4.x 发行版创建的集群上。Pig 在 Amazon EMR 5.0.0 中已升级到版本 0.16.0。下面介绍了明显差异。

不同的默认执行引擎

Amazon EMR 4.x 发行版上的 Pig 版本 0.14.0 使用 MapReduce 作为默认执行引擎。Pig 0.16.0 及更高版本均使用 Apache Tez。您可在 `exectype=mapreduce` 配置分类中显式设置 `pig-properties` 以使用 MapReduce。

已删除的 Pig 用户定义函数（UDF）

从 Pig 0.16.0 开始，删除了在 Amazon EMR 4.x 发行版上的 Pig 中可用的自定义 UDF。大多数 UDF 具有您可替换的等效函数。下表列出了已删除的 UDF 和等效函数。有关更多信息，请参阅 Apache Pig 网站上的内置函数。

已删除的 UDF	等效函数
FORMAT_DT(dtformat, date)	GetHour(date)、GetMinute(date)、GetMonth(date)、GetSecond(date)、GetWeek(date)、GetYear(date)、GetDay(date)
EXTRACT(string, pattern)	REGEX_EXTRACT_ALL(string, pattern)
REPLACE(string, pattern, replacement)	REPLACE(string, pattern, replacement)
DATE_TIME()	ToDate()
DURATION(dt, dt2)	WeeksBetween(dt, dt2)、YearsBetween(dt, dt2)、SecondsBetween(dt, dt2)、MonthsBetween(dt, dt2)、MinutesBetween(dt, dt2)、HoursBetween(dt, dt2)
EXTRACT_DT(format, date)	GetHour(date)、GetMinute(date)、GetMonth(date)、GetSecond(date)、GetWeek(date)、GetYear(date)、GetDay(date)
OFFSET_DT(date, duration)	AddDuration(date, duration)、SubtractDuration(date, duration)
PERIOD(dt, dt2)	WeeksBetween(dt, dt2)、YearsBetween(dt, dt2)、SecondsBetween(dt, dt2)、MonthsBetween(dt, dt2)、MinutesBetween(dt, dt2)、HoursBetween(dt, dt2)

已删除的 UDF	等效函数
CAPITALIZE(string)	UCFIRST(string)
CONCAT_WITH()	CONCAT()
INDEX_OF()	INDEXOF()
LAST_INDEX_OF()	LAST_INDEXOF()
SPLIT_ON_REGEX()	STRSPLT()
UNCAPITALIZE()	LCFIRST()

以下 UDF 已被删除，没有等效函数：
FORMAT()、LOCAL_DATE()、LOCAL_TIME()、CENTER()、LEFT_PAD()、REPEAT()、REPLACE_ONCE

已停止使用 Grunt 命令

某些 Grunt 命令已从 Pig 0.16.0 开始停用。下表列出了 Pig 0.14.0 中的 Grunt 命令以及当前版本中的等效命令（如果适用）。

Pig 0.14.0 和等效的当前 Grunt 命令

Pig 0.14.0 Grunt 命令	0.16.0 及更高版本中的 Pig Grunt 命令
cat <non-hdfs-path>)	fs -cat <non-hdfs-path>;
cd <non-hdfs-path>;	无等效函数
ls <non-hdfs-path>;	fs -ls <non-hdfs-path>;
move <non-hdfs-path> <non-hdfs-path>;	fs -mv <non-hdfs-path> <non-hdfs-path>;
copy <non-hdfs-path> <non-hdfs-path>;	fs -cp <non-hdfs-path> <non-hdfs-path>;
copyToLocal <non-hdfs-path> <local-path>;	fs -copyToLocal <non-hdfs-path> <local-path>;
copyFromLocal <local-path> <non-hdfs-path>;	fs -copyFromLocal <local-path> <non-hdfs-path>;

Pig 0.14.0 Grunt 命令	0.16.0 及更高版本中的 Pig Grunt 命令
mkdir <non-hdfs-path>;	fs -mkdir <non-hdfs-path>;
rm <non-hdfs-path>;	fs -rm -r -skipTrash <non-hdfs-path>;
rmf <non-hdfs-path>;	fs -rm -r -skipTrash <non-hdfs-path>;

针对非 HDFS 主目录删除的功能

Amazon EMR 4.x 发行版上的 Pig 0.14.0 具有两种机制，以允许无主目录的 hadoop 用户之外的用户运行 Pig 脚本。第一种机制是自动后备，将初始工作目录设置为根目录 (如果主目录不存在)。第二种机制是 `pig.initial.fs.name` 属性，它允许您更改初始工作目录。

这两种机制从 Amazon EMR 版本 5.0.0 开始不可用，因此用户必须在 HDFS 上有一个主目录。这不适用于 hadoop 用户，因为在启动时会配置一个主目录。使用 Hadoop jar 步骤运行的脚本默认为由 Hadoop 用户运行，除非使用 `command-runner.jar` 显式指定了其它用户。

Amazon EMR 发行版 4.9.6

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.6	emr-4.9.5	emr-4.9.4	emr-4.9.3
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-4.9.6	emr-4.9.5	emr-4.9.4	emr-4.9.3
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。

组件	版本	描述
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。

组件	版本	描述
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.9.6 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。

分类	描述
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。

分类	描述
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.9.5

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.5	emr-4.9.4	emr-4.9.3	emr-4.9.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-4.9.5	emr-4.9.4	emr-4.9.3	emr-4.9.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-

	emr-4.9.5	emr-4.9.4	emr-4.9.3	emr-4.9.2
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.9.5 的信息。更改与 4.9.4 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- HBase
  - 此版本解决了潜在的安全漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-4.9.5 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.9.4

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.4	emr-4.9.3	emr-4.9.2	emr-4.9.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-4.9.4	emr-4.9.3	emr-4.9.2	emr-4.9.1
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.9.4 的信息。更改与 4.9.3 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。

组件	版本	描述
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。

组件	版本	描述
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.9.4 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。

分类	描述
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。

分类	描述
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.9.3

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.3	emr-4.9.2	emr-4.9.1	emr-4.8.5
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-4.9.3	emr-4.9.2	emr-4.9.1	emr-4.8.5
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-

	emr-4.9.3	emr-4.9.2	emr-4.9.1	emr-4.8.5
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.9.3 的信息。更改与 Amazon EMR 4.9.2 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.9.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。

分类	描述
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.9.2

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.2	emr-4.9.1	emr-4.8.5	emr-4.8.4
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-4.9.2	emr-4.9.1	emr-4.8.5	emr-4.8.4
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.9.2 的信息。更改与 Amazon EMR 4.9.1 发行版有关。

发布日期：2017 年 7 月 13 日

此版本略微进行了一些改动、错误修复和增强。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.3.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.17.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.9.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.9.1

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.9.1	emr-4.8.5	emr-4.8.4	emr-4.8.3
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-4.9.1	emr-4.8.5	emr-4.8.4	emr-4.8.3
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.157.1
Spark	1.6.3	1.6.3	1.6.3	1.6.3
Sqoop	-	-	-	-

	emr-4.9.1	emr-4.8.5	emr-4.8.4	emr-4.8.3
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.9

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.9.1 的信息。更改与 Amazon EMR 4.8.4 发行版有关。

发布日期：2017 年 4 月 10 日

早期版本中已解决的已知问题

- 已逆向移植 HIVE-9976 和 HIVE-10106
- 修复了 YARN 中的一个问题，即，大量节点（大于 2000 个）和容器（大于 5000 个）会导致内存不足错误，例如："Exception in thread main java.lang.OutOfMemoryError"。

更改和增强功能

- Amazon EMR 发行版现在基于 Amazon Linux 2017.03。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2017.03-release-notes/。
- 从 Amazon EMR 基本 Linux 映像中删除了 Python 2.6。如果需要，您可以手动安装 Python 2.6。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.3.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.15.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-9	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-9	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-9	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-9	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-9	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-9	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。

组件	版本	描述
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.9.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值

分类	描述
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。

分类	描述
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。

分类	描述
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.8.5

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.8.5	emr-4.8.4	emr-4.8.3	emr-4.8.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.3
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-

	emr-4.8.5	emr-4.8.4	emr-4.8.3	emr-4.8.2
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.157.1	0.152.3
Spark	1.6.3	1.6.3	1.6.3	1.6.2
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1

	emr-4.8.5	emr-4.8.4	emr-4.8.3	emr-4.8.2
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.9	3.4.8

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	emrfs	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-8	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-8	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	1.0.0-amzn-8	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-8	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-8	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-8	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.8.5 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。

分类	描述
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。

分类	描述
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。

分类	描述
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

分类	描述
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.8.4

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.8.4	emr-4.8.3	emr-4.8.2	emr-4.8.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用

	emr-4.8.4	emr-4.8.3	emr-4.8.2	emr-4.8.0
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.2
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.3	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0

	emr-4.8.4	emr-4.8.3	emr-4.8.2	emr-4.8.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.157.1	0.152.3	0.151
Spark	1.6.3	1.6.3	1.6.2	1.6.2
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.4
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.6.1
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.9	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.8.4 的信息。更改与 Amazon EMR 4.8.3 发行版有关。

发布日期：2017 年 2 月 7 日

此版本略微进行了一些改动、错误修复和增强。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或

aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	emrfs	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。

组件	版本	描述
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。

组件	版本	描述
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-8	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-8	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-8	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-8	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-8	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-8	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.54+	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。

组件	版本	描述
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.25+	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。

组件	版本	描述
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.8.4 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。

分类	描述
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.8.3

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.8.3	emr-4.8.2	emr-4.8.0	emr-4.7.4
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.2	1.2.1
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.3	2.7.2	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-

	emr-4.8.3	emr-4.8.2	emr-4.8.0	emr-4.7.4
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.157.1	0.152.3	0.151	0.148
Spark	1.6.3	1.6.2	1.6.2	1.6.2
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.4	0.8.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.6.1	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.9	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.8.3 的信息。更改与 Amazon EMR 4.8.2 发行版有关。

发布日期：2016 年 12 月 29 日

升级

- 已升级到 Presto 0.157.1。有关更多信息，请参阅 Presto 文档中的 Presto 发布说明。
- 已升级到 Spark 1.6.3。有关更多信息，请参阅 Apache Spark 文档中的 Spark 发布说明。
- 已升级到 ZooKeeper 3.4.9。有关更多信息，请参阅 Apache ZooKeeper 文档中的 ZooKeeper 发布说明。

更改和增强功能

- 在 Amazon EMR 4.8.3 及更高版本（但不包括 5.0.0、5.0.3 和 5.2.0 版）中添加了对 Amazon EC2 m4.16xlarge 实例类型的支持。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

早期版本中已解决的已知问题

- 修复了 Hadoop 中的一个问题，即 ReplicationMonitor 线程可能会因为在大型集群中复制和删除同一个文件导致的竞争而卡住很长时间。
- 修复了在作业状态未成功更新时 ControlledJob#toString 出现空指针异常 (NPE) 失败的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.2.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.13.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-1	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.3-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.3-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.3-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-1	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-8	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	1.0.0-amzn-8	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-8	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-8	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-8	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-8	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.157.1	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.157.1	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.3	Spark 命令行客户端。
spark-history-server	1.6.3	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.3	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.3	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.9	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.9	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.8.3 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。

分类	描述
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。

分类	描述
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。

分类	描述
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。

分类	描述
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.8.2

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.8.2	emr-4.8.0	emr-4.7.4	emr-4.7.2
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.2	1.2.2	1.2.1	1.2.1
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.3	2.7.2	2.7.2	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.2

	emr-4.8.2	emr-4.8.0	emr-4.7.4	emr-4.7.2
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.152.3	0.151	0.148	0.148
Spark	1.6.2	1.6.2	1.6.2	1.6.2
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.4	0.8.3	0.8.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.6.1	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.8.2 的信息。更改与 Amazon EMR 4.8.0 发行版有关。

发布日期：2016 年 10 月 24 日

升级

- 已升级到 Hadoop 2.7.3
- 已升级到 Presto 0.152.3，它包括对 Presto Web 界面的支持。可使用端口 8889 访问 Presto 协调器上的 Presto Web 界面。有关 Presto Web 界面的更多信息，请参阅 Presto 文档中的 Web 界面。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	4.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emrfs	2.10.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.3-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.3-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.3-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.3-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.3-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.3-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.3-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。

组件	版本	描述
hadoop-yarn-nodemanager	2.7.3-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resour cemanager	2.7.3-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.3-amzn-0	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-7	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-7	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-7	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-7	Hive 命令行客户端。

组件	版本	描述
hive-metastore-server	1.0.0-amzn-7	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-7	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.52	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.152.3	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.152.3	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.2	Spark 命令行客户端。

组件	版本	描述
spark-history-server	1.6.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.8.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。

分类	描述
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。

分类	描述
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。

分类	描述
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.8.0

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.8.0	emr-4.7.4	emr-4.7.2	emr-4.7.1
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-4.8.0	emr-4.7.4	emr-4.7.2	emr-4.7.1
HBase	1.2.2	1.2.1	1.2.1	1.2.1
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.2	2.7.2	2.7.2	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.2	0.12.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.151	0.148	0.148	0.147
Spark	1.6.2	1.6.2	1.6.2	1.6.1
Sqoop	-	-	-	-

	emr-4.8.0	emr-4.7.4	emr-4.7.2	emr-4.7.1
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.4	0.8.3	0.8.3	0.8.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.6.1	0.5.6	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.8.0 的信息。更改与 Amazon EMR 4.7.2 发行版有关。

发布日期：2016 年 9 月 7 日

升级

- 已升级到 HBase 1.2.2
- 已升级到 Presto-Sandbox 0.151
- 已升级到 Tez 0.8.4
- 已升级到 Zeppelin-Sandbox 0.6.1

更改和增强功能

- 修复了 YARN 中的一个问题，ApplicationMaster 将在其中尝试清除不再存在的容器，因为它们的实例已终止。
- 更正了 Oozie 示例中 Hive2 操作的 hive-server2 URL。

- 添加了对其它 Presto 目录的支持。
- 已逆向移植修补程序：HIVE-8948、HIVE-12679、HIVE-13405、PHOENIX-3116、HADOOP-12689
- 添加了对安全配置的支持，这使您可以更轻松地创建和应用加密选项。有关更多信息，请参阅数据加密。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.9.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-4	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-4	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-4	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-4	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-4	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.2-amzn-4	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-4	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-4	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.2-amzn-4	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-4	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.2	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.2	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.2	HBase 命令行客户端。
hbase-rest-server	1.2.2	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.2	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-7	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-7	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-7	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-7	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-7	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	1.0.0-amzn-7	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.51	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.151	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.151	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.2	Spark 命令行客户端。
spark-history-server	1.6.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。

组件	版本	描述
spark-on-yarn	1.6.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.4	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.6.1	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.8.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。

分类	描述
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。

分类	描述
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hiveserver2-site	更改 Hive Server2 的 hiveserver2-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-blackhole	更改 Presto 的 blackhole.properties 文件中的值。
presto-connector-cassandra	更改 Presto 的 cassandra.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
presto-connector-jmx	更改 Presto 的 jmx.properties 文件中的值。
presto-connector-kafka	更改 Presto 的 kafka.properties 文件中的值。
presto-connector-localfile	更改 Presto 的 localfile.properties 文件中的值。
presto-connector-mongodb	更改 Presto 的 mongodb.properties 文件中的值。
presto-connector-mysql	更改 Presto 的 mysql.properties 文件中的值。
presto-connector-postgresql	更改 Presto 的 postgresql.properties 文件中的值。

分类	描述
presto-connector-raptor	更改 Presto 的 raptor.properties 文件中的值。
presto-connector-redis	更改 Presto 的 redis.properties 文件中的值。
presto-connector-tpch	更改 Presto 的 tpch.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.7.4

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.7.4	emr-4.7.2	emr-4.7.1	emr-4.7.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.75
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2

	emr-4.7.4	emr-4.7.2	emr-4.7.1	emr-4.7.0
HBase	1.2.1	1.2.1	1.2.1	1.2.1
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.2	2.7.2	2.7.2	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.2	0.12.0	0.12.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	4.7.0
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.148	0.148	0.147	0.147
Spark	1.6.2	1.6.2	1.6.1	1.6.1
Sqoop	-	-	-	-

	emr-4.7.4	emr-4.7.2	emr-4.7.1	emr-4.7.0
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.3	0.8.3	0.8.3	0.8.3
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

这是补丁发行版本，用于将针对请求的Amazon签名版本 4 身份验证添加至 Amazon S3。所有应用程序和组件都与之前的 Amazon EMR 发行版相同。

> **Important**
>
> 在此发行版中，Amazon EMR 仅使用Amazon签名版本 4 来对针对 Amazon S3 的请求进行身份验证。有关更多信息，请参阅新功能。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.8.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。

组件	版本	描述
hadoop-hdfs-datanode	2.7.2-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.2-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.1	HBase 命令行客户端。
hbase-rest-server	1.2.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。

组件	版本	描述
hbase-thrift-server	1.2.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-6	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-6	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-6	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-6	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-6	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-6	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.46	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库

组件	版本	描述
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.148	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.148	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.2	Spark 命令行客户端。
spark-history-server	1.6.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.3	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。

组件	版本	描述
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.7.4 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。

分类	描述
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。

分类	描述
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。

分类	描述
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.7.2

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.7.2	emr-4.7.1	emr-4.7.0	emr-4.6.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.75	1.10.27
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.1	1.2.1	1.2.1	1.2.0
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.2	2.7.2	2.7.2	2.7.2
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-

	emr-4.7.2	emr-4.7.1	emr-4.7.0	emr-4.6.0
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.2	0.12.0	0.12.0	0.11.1
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	4.7.0	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.148	0.147	0.147	0.143
Spark	1.6.2	1.6.1	1.6.1	1.6.1
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.3	0.8.3	0.8.3	-
Trino (PrestoSQL)	-	-	-	-

	emr-4.7.2	emr-4.7.1	emr-4.7.0	emr-4.6.0
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	3.4.8	3.4.8

发布说明

以下发布说明包括有关 Amazon EMR 4.7.2 的信息。

发布日期：2016 年 7 月 15 日

特征

- 已升级到 Mahout 0.12.2
- 已升级到 Presto 0.148
- 已升级到 Spark 1.6.2
- 您现在可以使用 URI 作为参数来创建将与 EMRFS 配合使用的 AWSCredentialsProvider。有关更多信息，请参阅为 EMRFS 创建 AWSCredentialsProvider。
- EMRFS 现在允许用户使用 `emrfs-site.xml` 中的 `fs.s3.consistent.dynamodb.endpoint` 属性来为其一致视图元数据配置自定义 DynamoDB 终端节点。
- 在 `/usr/bin` 中添加了一个名为 `spark-example` 的脚本，它将包装 `/usr/lib/spark/spark/bin/run-example`，因此您可以直接运行示例。例如，要运行 Spark 分配的附带的 SparkPi 示例，可从命令行运行 `spark-example SparkPi 100` 或将 `command-runner.jar` 作为 API 中的一个步骤运行。

早期版本中已解决的已知问题

- 修复了 Oozie 在安装 Spark 后拥有的 `spark-assembly.jar` 未位于正确位置 (这导致使用 Oozie 启动 Spark 应用程序失败) 的问题。
- 修复了与 YARN 容器中基于 Spark Log4j 的登录有关的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.1.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.8.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。

组件	版本	描述
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-3	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-3	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-3	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-3	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-3	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.2-amzn-3	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-3	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-3	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-3	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-3	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。

组件	版本	描述
hbase-region-server	1.2.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.1	HBase 命令行客户端。
hbase-rest-server	1.2.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-6	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-6	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-6	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-6	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-6	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-6	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.2	用于机器学习的库。
mysql-server	5.5.46	MySQL 数据库服务器。

组件	版本	描述
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.148	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.148	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.2	Spark 命令行客户端。
spark-history-server	1.6.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.2	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.3	tez YARN 应用程序和库。
webserver	2.4.23	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.7.2 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hadoop-ssl-server	更改 hadoop ssl 服务器配置
hadoop-ssl-client	更改 hadoop ssl 客户端配置
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。

分类	描述
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。

分类	描述
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。

分类	描述
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.7.1

- 应用程序版本
- 发布说明
- 组件版本

- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.7.1	emr-4.7.0	emr-4.6.0	emr-4.5.0
Amazon SDK for Java	1.10.75	1.10.75	1.10.27	1.10.27
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.1	1.2.1	1.2.0	-
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.2	2.7.2	2.7.2	2.7.2

	emr-4.7.1	emr-4.7.0	emr-4.6.0	emr-4.5.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.12.0	0.12.0	0.11.1	0.11.1
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	4.7.0	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.147	0.147	0.143	0.140
Spark	1.6.1	1.6.1	1.6.1	1.6.1
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.3	0.8.3	-	-

	emr-4.7.1	emr-4.7.0	emr-4.6.0	emr-4.5.0
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	3.4.8	-

发布说明

以下发布说明包括有关 Amazon EMR 4.7.1 的信息。

发布日期：2016 年 6 月 10 日

早期版本中已解决的已知问题

- 修复了延长带有私有子网的 VPC 中启动的集群的启动时间的问题。此错误仅影响使用 Amazon EMR 4.7.0 发行版启动的集群。
- 修复了在 Amazon EMR 中错误处理针对使用 Amazon EMR 4.7.0 发行版启动的集群的文件列表的问题。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.7.1	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-2	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.2-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.2-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.1	HBase 命令行客户端。
hbase-rest-server	1.2.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-5	用于操作 hcatalog-server 的“hcat”命令行客户端。

组件	版本	描述
hcatalog-server	1.0.0-amzn-5	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-5	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-5	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-5	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-5	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.0	用于机器学习的库。
mysql-server	5.5.46	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器

组件	版本	描述
presto-coordinator	0.147	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.147	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.1	Spark 命令行客户端。
spark-history-server	1.6.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.3	tez YARN 应用程序和库。
webserver	2.4.18	Apache HTTP 服务器。
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.7.1 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。

分类	描述
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。

分类	描述
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.7.0

应用程序版本

此版本支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Phoenix、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Tez、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本

- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.7.0	emr-4.6.0	emr-4.5.0	emr-4.4.0
Amazon SDK for Java	1.10.75	1.10.27	1.10.27	1.10.27
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.1	1.2.0	-	-
HCatalog	1.0.0	1.0.0	1.0.0	1.0.0
Hadoop	2.7.2	2.7.2	2.7.2	2.7.1
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-

	emr-4.7.0	emr-4.6.0	emr-4.5.0	emr-4.4.0
MXNet	-	-	-	-
Mahout	0.12.0	0.11.1	0.11.1	0.11.1
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	4.7.0	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.147	0.143	0.140	0.136
Spark	1.6.1	1.6.1	1.6.1	1.6.0
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	1.4.6
TensorFlow	-	-	-	-
Tez	0.8.3	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.6	0.5.6
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	3.4.8	-	-

发布说明

> **Important**
>
> Amazon EMR 4.7.0 已弃用。请改用 Amazon EMR 4.7.1 或更高版本。

发布日期：2016 年 6 月 2 日

特征

- 已添加 Apache Phoenix 4.7.0
- 已添加 Apache Tez 0.8.3
- 已升级到 HBase 1.2.1
- 已升级到 Mahout 0.12.0
- 已升级到 Presto 0.147
- 已将 Amazon SDK for Java 升级到 1.10.75
- 已从 `mapreduce.cluster.local.dir` 中的 `mapred-site.xml` 属性中删除最终标志以允许用户以本地模式运行 Pig。
- 集群上可用的 Amazon Redshift JDBC 驱动程序

  Amazon Redshift JDBC 驱动程序现在包含在 `/usr/share/aws/redshift/jdbc` 中。`/usr/share/aws/redshift/jdbc/RedshiftJDBC41.jar` 是与 JDBC 4.1 兼容的驱动程序，`/usr/share/aws/redshift/jdbc/RedshiftJDBC4.jar` 是与 JDBC 4.0 兼容的 Amazon Redshift 驱动程序。有关更多信息，请参阅 Amazon Redshift 管理指南中的配置 JDBC 连接。
- Java 8

  OpenJDK 1.7 是用于所有应用程序 (Presto 除外) 的默认 JDK。但是，将同时安装 OpenJDK 1.7 和 1.8。有关如何为应用程序设置 `JAVA_HOME` 的信息，请参阅配置应用程序以使用 Java 8。

早期版本中已解决的已知问题

- 修复了一个内核问题，该问题已明显影响了 emr-4.6.0 中的 Amazon EMR 的吞吐量优化 HDD（st1）EBS 卷的性能。
- 修复了在不选择 Hadoop 作为应用程序的情况下指定任何 HDFS 加密区域时集群将失败的问题。

- 已将默认 HDFS 编写策略从 RoundRobin 更改为 AvailableSpaceVolumeChoosingPolicy。未通过 RoundRobin 配置正确利用某些卷，这将导致核心节点失败且 HDFS 不可靠。
- 修复了与 EMRFS CLI 有关的问题，此问题将在创建默认 DynamoDB 元数据表以获得一致视图时导致异常。
- 修复了在分段重命名和复制操作期间可能发生在 EMRFS 中的死锁问题。
- 修复了与 EMRFS 有关的问题，此问题导致 CopyPart 大小默认为 5 MB。默认值现已相应地设置为 128 MB。
- 修复了与 Zeppelin upstart 配置有关的问题，此问题可能会阻止您停止服务。
- 修复了与 Spark 和 Zeppelin 有关的问题，此问题会阻止您使用 s3a:// URI 方案，因为 /usr/lib/hadoop/hadoop-aws.jar 未在其各自的类路径中正确加载。
- 已逆向移植 HUE-2484。
- 已从 Hue 3.9.0 (不存在 JIRA) 逆向移植 commit 来修复与 HBase 浏览器示例有关的问题。
- 已逆向移植 HIVE-9073。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.2.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.4.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.7.1	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-2	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.7.2-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-2	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hadoop-yarn-timeline-server	2.7.2-amzn-2	用于检索 YARN 应用程序的当前信息和历史信息的服务。
hbase-hmaster	1.2.1	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.1	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.1	HBase 命令行客户端。
hbase-rest-server	1.2.1	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.1	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-5	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-5	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。

组件	版本	描述
hcatalog-webhcat-server	1.0.0-amzn-5	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-5	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-5	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-5	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-7	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.12.0	用于机器学习的库。
mysql-server	5.5.46	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
phoenix-library	4.7.0-HBase-1.2	服务器和客户端的 phoenix 库
phoenix-query-server	4.7.0-HBase-1.2	向 Avatica API 提供 JDBC 访问权限以及协议缓冲区和 JSON 格式访问权限的轻量级服务器
presto-coordinator	0.147	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.147	用于执行查询的各个部分的服务。

组件	版本	描述
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.1	Spark 命令行客户端。
spark-history-server	1.6.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
tez-on-yarn	0.8.3	tez YARN 应用程序和库。
webserver	2.4.18	Apache HTTP 服务器。
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.7.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。

分类	描述
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。

分类	描述
phoenix-hbase-metrics	更改 Phoenix 的 hadoop-metrics2-hbase.properties 文件中的值。
phoenix-hbase-site	更改 Phoenix 的 hbase-site.xml 文件中的值。
phoenix-log4j	更改 Phoenix 的 log4j.properties 文件中的值。
phoenix-metrics	更改 Phoenix 的 hadoop-metrics2-phoenix.properties 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
tez-site	更改 Tez 的 tez-site.xml 文件中的值。

分类	描述
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.6.0

应用程序版本

此发行版支持以下应用程序：Ganglia、HBase、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox、Zeppelin-Sandbox 和 ZooKeeper-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.6.0	emr-4.5.0	emr-4.4.0	emr-4.3.0
Amazon SDK for Java	1.10.27	1.10.27	1.10.27	1.10.27
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.7.2
HBase	1.2.0	-	-	-
HCatalog	1.0.0	1.0.0	1.0.0	-
Hadoop	2.7.2	2.7.2	2.7.1	2.7.1
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.1	0.11.1	0.11.1	0.11.0

	emr-4.6.0	emr-4.5.0	emr-4.4.0	emr-4.3.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.143	0.140	0.136	0.130
Spark	1.6.1	1.6.1	1.6.0	1.6.0
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	1.4.6	-
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.6	0.5.5
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	3.4.8	-	-	-

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.6.0 的信息。

- 已添加 HBase 1.2.0
- 已添加 Zookeeper-Sandbox 3.4.8
- 已升级到 Presto-Sandbox 0.143
- Amazon EMR 发行版现在基于 Amazon Linux 2016.03.0。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.03-release-notes/。
- 影响吞吐量优化 HDD（st1）EBS 卷类型的问题

  Linux 内核版本 4.2 及更高版本中的问题将显著影响 EMR 的吞吐量优化 HDD (st1) EBS 卷上的性能。此版本 (emr-4.6.0) 使用内核版本 4.4.5，因此会受到影响。因此，如果您打算使用 st1 EBS 卷，我们建议您不要使用 emr-4.6.0。您可将 emr-4.5.0 或早期 Amazon EMR 发行版与 st1 配合使用，而不会产生影响。此外，我们将随将来版本一起提供修复程序。
- Python 默认值

  现在，默认情况下已安装 Python 3.4，但 Python 2.7 将保留系统默认值。您可以使用引导操作将 Python 3.4 配置为系统默认值；也可以使用配置 API 将 PYSPARK_PYTHON 导出设置为 `/usr/bin/python3.4` 分类中的 `spark-env` 以便影响 PySpark 所使用的 Python 版本。
- Java 8

  OpenJDK 1.7 是用于所有应用程序 (Presto 除外) 的默认 JDK。但是，将同时安装 OpenJDK 1.7 和 1.8。有关如何为应用程序设置 `JAVA_HOME` 的信息，请参阅配置应用程序以使用 Java 8。

早期版本中已解决的已知问题

- 修复了应用程序预置有时会因生成的密码导致随机失败的问题。
- 之前，`mysqld` 已安装在所有节点上。现在，它仅安装在主实例上，而且仅在所选应用程序将 `mysql-server` 作为组件包含时安装。当前，以下应用程序包含 `mysql-server` 组件：HCatalog、Hive、Hue、Presto-Sandbox 和 Sqoop-Sandbox。
- 已将 `yarn.scheduler.maximum-allocation-vcores` 从默认值 32 更改为 80，这修复了 emr-4.4.0 中引入的一个问题，此问题主要在使用集群（其内核实例类型为具有高于 32 的 YARN 虚拟内核集的几个大型实例类型之一）中的 `maximizeResourceAllocation` 选项时与 Spark 时一起出现；也就是说，此问题影响了 c4.8xlarge、cc2.8xlarge、hs1.8xlarge、i2.8xlarge、m2.4xlarge、r3.8xlarge、d2.8xlarge 或 m4.10xlarge。
- s3-dist-cp 现在对所有 Amazon S3 提名使用 EMRFS，并且不再过渡到临时 HDFS 目录。
- 修复了与针对客户端加密分段上载的异常处理有关的问题。

- 添加了允许用户更改 Amazon S3 存储类的选项。默认情况下，此设置为 STANDARD。emrfs-site 配置分类设置为 fs.s3.storageClass，可能的值为 STANDARD、STANDARD_IA 和 REDUCED_REDUNDANCY。有关存储类的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的存储类。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.3.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.6.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.2-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-1	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.7.2-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hbase-hmaster	1.2.0	适用于负责协调区域和执行管理命令的 HBase 集群的服务。
hbase-region-server	1.2.0	用于服务于一个或多个 HBase 区域的服务。
hbase-client	1.2.0	HBase 命令行客户端。
hbase-rest-server	1.2.0	用于向 HBase 提供 RESTful HTTP 终端节点的服务。
hbase-thrift-server	1.2.0	用于向 HBase 提供 Thrift 终端节点的服务。
hcatalog-client	1.0.0-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-4	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。

组件	版本	描述
hue-server	3.7.1-amzn-6	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.11.1	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.143	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.143	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.1	Spark 命令行客户端。
spark-history-server	1.6.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
webserver	2.4	Apache HTTP 服务器。

组件	版本	描述
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。
zookeeper-server	3.4.8	用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。
zookeeper-client	3.4.8	ZooKeeper 命令行客户端。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.6.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hbase-env	更改 HBase 环境中的值。
hbase-log4j	更改 HBase 的 hbase-log4j.properties 文件中的值。
hbase-metrics	更改 HBase 的 hadoop-metrics2-hbaase.properties 文件中的值。

分类	描述
hbase-policy	更改 HBase 的 hbase-policy.xml 文件中的值。
hbase-site	更改 HBase 的 hbase-site.xml 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。

分类	描述
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。

分类	描述
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。
zookeeper-config	更改 ZooKeeper 的 zoo.cfg 文件中的值。
zookeeper-log4j	更改 ZooKeeper 的 log4j.properties 文件中的值。

Amazon EMR 发行版 4.5.0

应用程序版本

此发行版支持以下应用程序：Ganglia、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox 和 Zeppelin-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.5.0	emr-4.4.0	emr-4.3.0	emr-4.2.0
Amazon SDK for Java	1.10.27	1.10.27	1.10.27	1.10.27
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.7.2	3.6.0
HBase	-	-	-	-
HCatalog	1.0.0	1.0.0	-	-
Hadoop	2.7.2	2.7.1	2.7.1	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.1	0.11.1	0.11.0	0.11.0

	emr-4.5.0	emr-4.4.0	emr-4.3.0	emr-4.2.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.2.0
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.140	0.136	0.130	0.125
Spark	1.6.1	1.6.0	1.6.0	1.5.2
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	1.4.6	-	-
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.6	0.5.5	0.5.5
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.5.0 的信息。

发布日期：2016 年 4 月 4 日

特征

- 已升级到 Spark 1.6.1
- 已升级到 Hadoop 2.7.2
- 已升级到 Presto 0.140
- 添加了对 Amazon S3 服务器端加密的 Amazon KMS 支持。

早期版本中已解决的已知问题

- 修复了重启节点后无法启动 MySQL 和 Apache 服务器的问题。
- 修复了 IMPORT 未正确使用存储在 Amazon S3 中的非分区表的问题
- 修复了与 Presto 有关的问题，此问题导致在写入 Hive 表时要求暂存目录为 `/mnt/tmp` 而不是 `/tmp`。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。

组件	版本	描述
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.2.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.5.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.2-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.2-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.2-amzn-0	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.7.2-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.2-amzn-0	用于 HDFS 操作的 HTTP 终端节点。

组件	版本	描述
hadoop-kms-server	2.7.2-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.2-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.2-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.2-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hcatalog-client	1.0.0-amzn-4	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-4	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-4	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-4	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-4	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-4	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-5	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序

组件	版本	描述
mahout-client	0.11.1	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.140	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.140	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.1	Spark 命令行客户端。
spark-history-server	1.6.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.1	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。
webserver	2.4	Apache HTTP 服务器。
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-4.5.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。

分类	描述
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。

分类	描述
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

Amazon EMR 发行版 4.4.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此发行版支持以下应用程序：Ganglia、HCatalog、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark、Sqoop-Sandbox 和 Zeppelin-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.4.0	emr-4.3.0	emr-4.2.0	emr-4.1.0
Amazon SDK for Java	1.10.27	1.10.27	1.10.27	不可用
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.7.2	3.6.0	-
HBase	-	-	-	-
HCatalog	1.0.0	-	-	-
Hadoop	2.7.1	2.7.1	2.6.0	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-

	emr-4.4.0	emr-4.3.0	emr-4.2.0	emr-4.1.0
Hue	3.7.1	3.7.1	3.7.1	3.7.1
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.1	0.11.0	0.11.0	0.11.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.2.0	4.0.1
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.136	0.130	0.125	0.119
Spark	1.6.0	1.6.0	1.5.2	1.5.0
Sqoop	-	-	-	-
Sqoop-Sandbox	1.4.6	-	-	-
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-

	emr-4.4.0	emr-4.3.0	emr-4.2.0	emr-4.1.0
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.6	0.5.5	0.5.5	0.6.0-SNAPSHOT
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.4.0 的信息。

发布日期：2016 年 3 月 14 日

特征

- 已添加 HCatalog 1.0.0
- 已添加 Sqoop-Sandbox 1.4.6
- 已升级到 Presto 0.136
- 已升级到 Zeppelin 0.5.6
- 已升级到 Mahout 0.11.1
- 默认情况下已启用 `dynamicResourceAllocation`。
- 已添加针对此版本的所有配置分类的表。有关更多信息，请参阅配置应用程序中的“配置分类”表。

早期版本中已解决的已知问题

- 修复了 `maximizeResourceAllocation` 设置不为 YARN ApplicationMaster 进程守护程序预留足够内存的问题。
- 修复了遇到的与自定义 DNS 相关的问题。如果 `resolve.conf` 中的任何条目位于提供的自定义条目之前，则自定义条目不可解析。此行为受 VPC 中集群的影响，其中，默认 VPC 名称服务器已作为顶部条目插入 `resolve.conf` 中。
- 修复了默认 Python 已移至版本 2.7 且未为该版本安装 boto 的问题。

- 修复了 YARN 容器和 Spark 应用程序将生成唯一 Ganglia 轮询数据库 (rrd) 文件的问题，此问题会导致第一个附加到实例的磁盘填满。修复后，YARN 容器级别指标和 Spark 应用程序级别指标均已禁用。
- 修复了导致日志推送程序中删除所有空日志文件夹的问题。这会造成 Hive CLI 无法记录日志，因为日志推送程序已删除 `user` 下的 `/var/log/hive` 空文件夹。
- 修复了影响 Hive 导入的问题，此问题影响分区并导致导入期间出现错误。
- 修复了 EMRFS 和 s3-dist-cp 未正确处理包含句点的存储桶名称的问题。
- 更改了 EMRFS 中的行为，以便在启用版本控制的存储桶中，不会持续创建 `_$folder$` 标记文件，从而有助于提高启用版本控制的存储桶的性能。
- 更改了 EMRFS 中的行为，使它不会使用说明文件，已启用客户端加密的情况除外。如果您要在使用客户端加密时删除说明文件，可将 emrfs-site.xml 属性 `fs.s3.cse.cryptoStorageMode.deleteInstructionFiles.enabled` 设置为 true。
- 更改了 YARN 日志聚合以在聚合目标中将日志保留两天。默认目标为您的集群的 HDFS 存储。如果您要更改此持续时间，请在创建集群时使用 `yarn.log-aggregation.retain-seconds` 配置分类来更改 `yarn-site` 的值。与往常一样，您可以在创建集群时使用 `log-uri` 参数将应用程序日志保存到 Amazon S3。

已应用的修补程序

- HIVE-9655
- HIVE-9183
- HADOOP-12810

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*-amzn-*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.2.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.4.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.1-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.1-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.1-amzn-1	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.1-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.1-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.1-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.1-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.1-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.1-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hcatalog-client	1.0.0-amzn-3	用于操作 hcatalog-server 的“hcat”命令行客户端。
hcatalog-server	1.0.0-amzn-3	用于为分布式应用程序提供 HCatalog、表和存储管理层的服务。
hcatalog-webhcat-server	1.0.0-amzn-3	用于向 HCatalog 提供 REST 接口的 HTTP 终端节点。
hive-client	1.0.0-amzn-3	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-3	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	1.0.0-amzn-3	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-5	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.11.1	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.136	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.136	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.0	Spark 命令行客户端。
spark-history-server	1.6.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.0	YARN 从属项所需的 Apache Spark 库。
sqoop-client	1.4.6	Apache Sqoop 命令行客户端。

组件	版本	描述
webserver	2.4	Apache HTTP 服务器。
zeppelin-server	0.5.6-incubating	支持交互式数据分析的基于 Web 的笔记本电脑。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.4.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hcatalog-env	更改 HCatalog 的环境中的值。
hcatalog-server-jndi	更改 HCatalog 的 jndi.properties 中的值。
hcatalog-server-proto-hive-site	更改 HCatalog 的 proto-hive-site.xml 中的值。
hcatalog-webhcat-env	更改 HCatalog WebHCat 的环境中的值。

分类	描述
hcatalog-webhcat-log4j	更改 HCatalog WebHCat 的 log4j.properties 中的值。
hcatalog-webhcat-site	更改 HCatalog WebHCat 的 webhcat-site.xml 文件中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。

分类	描述
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
sqoop-env	更改 Sqoop 的环境中的值。
sqoop-oraoop-site	更改 Sqoop OraOop 的 oraoop-site.xml 文件中的值。
sqoop-site	更改 Sqoop 的 sqoop-site.xml 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

Amazon EMR 发行版 4.3.0

- 应用程序版本

- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Ganglia、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark 和 Zeppelin-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Amazon SDK for Java	1.10.27	1.10.27	不可用	不可用
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.6.0	-	-
HBase	-	-	-	-
HCatalog	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Hadoop	2.7.1	2.6.0	2.6.0	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	-
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.0	0.11.0	0.11.0	0.10.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.0.1	-
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.130	0.125	0.119	-
Spark	1.6.0	1.5.2	1.5.0	1.4.1
Sqoop	-	-	-	-
Sqoop-Sandbox	-	-	-	-
TensorFlow	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.5	0.5.5	0.6.0-SNAPSHOT	-
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.3.0 的信息。

发布日期：2016 年 1 月 19 日

特征

- 已升级到 Hadoop 2.7.1
- 已升级到 Spark 1.6.0
- 已将 Ganglia 升级到 3.7.2
- 已将 Presto 升级到 0.130
- 将 `spark.dynamicAllocation.enabled` 设置为 true 时，Amazon EMR 已对其做出一些更改；默认情况下，此项为 false。如果设置为 true，则会影响由 `maximizeResourceAllocation` 设置设定的默认值：
  - 若`spark.dynamicAllocation.enabled`设为 true，则 `spark.executor.instances` 将不被 `maximizeResourceAllocation` 设置。
  - 目前，`spark.driver.memory` 设置根据集群中的实例类型进行配置，与 `spark.executors.memory` 设置的方式类似。但是，由于 Spark 驱动应用程序可在主实例或核心实例之一上运行 (例如在 YARN 客户端和集群模式下分别进行)，`spark.driver.memory` 设置根据更小实例类型的实例类型，在两个实例组之间进行。

- 目前，spark.default.parallelism 设置为 YARN 容器可用的 CPU 内核数的两倍。在上一版本中，这是该值的一半。
- 为 Spark YARN 过程预留的内存开销计算精确性经过优化，从而使得 Spark 可用内存总量略有增加（即 spark.executor.memory）。

早期版本中已解决的已知问题

- 默认情况下，现已启用 YARN 日志聚合。
- 修复了在启用 YARN 日志聚合后日志未推送至集群的 Amazon S3 日志存储桶的问题。
- YARN 容器大小现跨所有节点类型具有新的最小值 32。
- 修复了与 Ganglia 有关的问题，此问题已导致大型集群中主节点上的磁盘 I/O 过多。
- 修复了在关闭集群时阻止应用程序日志推送至 Amazon S3 的问题。
- 修复了 EMRFS CLI 中导致某些命令失败的问题。
- 修复了与 Zeppelin 有关的问题，此问题已阻止依赖项在基础 SparkContext 中加载。
- 修复了因发出尝试添加实例的调整大小命令导致的问题。
- 修复了 Hive 中的问题，此问题导致 CREATE TABLE AS SELECT 对 Amazon S3 进行过多的列表调用。
- 修复了在安装 Hue、Oozie 和 Ganglia 时无法正常预置大型集群的问题。
- 修复了 s3-dist-cp 中的问题，此问题导致即使在失败并出现错误的情况下仍将返回零退出代码。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 emr 或 aws 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *CommunityVersion*-amzn-*EmrVersion* 的发行版标注。*EmrVersion* 从 0 开始。例如，假设已对名为 myapp-component 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 2.2-amzn-2。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.1.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.3.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
ganglia-monitor	3.7.2	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.7.2	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.7.1	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.7.1-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.7.1-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.7.1-amzn-0	HDFS 命令行客户端和库

组件	版本	描述
hadoop-hdfs-namenode	2.7.1-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.7.1-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.7.1-amzn-0	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.7.1-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.7.1-amzn-0	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.7.1-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hive-client	1.0.0-amzn-2	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-2	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-2	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-5	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.11.0	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。

组件	版本	描述
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.130	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.130	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.6.0	Spark 命令行客户端。
spark-history-server	1.6.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.6.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.6.0	YARN 从属项所需的 Apache Spark 库。
webserver	2.4	Apache HTTP 服务器。
zeppelin-server	0.5.5-incubating-amzn-1	支持交互式数据分析的基于 Web 的笔记本电脑。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.3.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。

分类	描述
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

Amazon EMR 发行版 4.2.0

应用程序版本

此版本支持以下应用程序：Ganglia、Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark 和 Zeppelin-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Amazon SDK for Java	1.10.27	1.10.27	不可用	不可用
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.6.0	-	-
HBase	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
HCatalog	-	-	-	-
Hadoop	2.7.1	2.6.0	2.6.0	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	-
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.0	0.11.0	0.11.0	0.10.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.0.1	-
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.130	0.125	0.119	-
Spark	1.6.0	1.5.2	1.5.0	1.4.1
Sqoop	-	-	-	-
Sqoop-Sandbox	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.5	0.5.5	0.6.0-SNAPSHOT	-
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

以下发布说明包括有关 Amazon EMR 发行版 4.2.0 的信息。

发布日期：2015 年 11 月 18 日

特征

- 已添加 Ganglia 支持
- 已升级到 Spark 1.5.2
- 已升级到 Presto 0.125
- 已将 Oozie 升级到 4.2.0
- 已将 Zeppelin 升级到 0.5.5
- 已将 Amazon SDK for Java 升级到 1.10.27

早期版本中已解决的已知问题

- 修复了与 EMRFS CLI 有关的问题，此问题发导致不使用默认元数据表名称。
- 修复了在 Amazon S3 中使用 ORC 支持的表时遇到的问题。

- 修复了遇到的 Python 版本在 Spark 配置中不匹配的问题。
- 修复了 YARN 节点状态因 VPC 中集群的 DNS 问题导致无法报告的问题。
- 修复了 YARN 停用节点时遇到的问题，该问题会导致应用程序挂起或无法计划新应用程序。
- 修复了集群终止且状态为 TIMED_OUT_STARTING 时遇到的问题。
- 修复了在其它内部版本中包含 EMRFS Scala 依赖项时遇到的问题。Scala 依赖项已被删除。

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.0.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.2.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。

组件	版本	描述
ganglia-monitor	3.6.0	适用于 Hadoop 生态系统应用程序的嵌入式 Ganglia 代理以及 Ganglia 监控代理。
ganglia-metadata-collector	3.6.0	用于从 Ganglia 监控代理中聚合指标的 Ganglia 元数据收集器。
ganglia-web	3.5.10	用于查看由 Ganglia 元数据收集器收集的指标的 Web 应用程序。
hadoop-client	2.6.0-amzn-2	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.6.0-amzn-2	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.6.0-amzn-2	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.6.0-amzn-2	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.6.0-amzn-2	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.6.0-amzn-2	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.6.0-amzn-2	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.6.0-amzn-2	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.6.0-amzn-2	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hive-client	1.0.0-amzn-1	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-5	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.11.0	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.2.0	Oozie 命令行客户端。
oozie-server	4.2.0	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.125	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.125	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.5.2	Spark 命令行客户端。

组件	版本	描述
spark-history-server	1.5.2	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.5.2	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.5.2	YARN 从属项所需的 Apache Spark 库。
webserver	2.4	Apache HTTP 服务器。
zeppelin-server	0.5.5-incubating-amzn-0	支持交互式数据分析的基于 Web 的笔记本电脑。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.2.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。

分类	描述
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。

分类	描述
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
presto-connector-hive	更改 Presto 的 hive.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
spark-metrics	更改 Spark 的 metrics.properties 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。
zeppelin-env	更改 Zeppelin 环境中的值。

Amazon EMR 发行版 4.1.0

应用程序版本

此版本支持以下应用程序：Hadoop、Hive、Hue、Mahout、Oozie-Sandbox、Pig、Presto-Sandbox、Spark 和 Zeppelin-Sandbox。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Amazon SDK for Java	1.10.27	1.10.27	不可用	不可用
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-
Flink	-	-	-	-
Ganglia	3.7.2	3.6.0	-	-
HBase	-	-	-	-
HCatalog	-	-	-	-
Hadoop	2.7.1	2.6.0	2.6.0	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	-
Iceberg	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.0	0.11.0	0.11.0	0.10.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.0.1	-
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.130	0.125	0.119	-
Spark	1.6.0	1.5.2	1.5.0	1.4.1
Sqoop	-	-	-	-
Sqoop-Sandbox	-	-	-	-
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.5	0.5.5	0.6.0-SNAPSHOT	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.1.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.0.0	针对 Amazon S3 优化的分布式复制应用程序。

组件	版本	描述
emrfs	2.1.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
hadoop-client	2.6.0-amzn-1	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.6.0-amzn-1	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-library	2.6.0-amzn-1	HDFS 命令行客户端和库
hadoop-hdfs-namenode	2.6.0-amzn-1	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.6.0-amzn-1	用于 HDFS 操作的 HTTP 终端节点。
hadoop-kms-server	2.6.0-amzn-1	基于 Hadoop 的 KeyProvider API 的加密密钥管理服务器。
hadoop-mapred	2.6.0-amzn-1	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.6.0-amzn-1	用于管理单个节点上的容器的 YARN 服务。
hadoop-yarn-resourcemanager	2.6.0-amzn-1	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hive-client	1.0.0-amzn-1	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-1	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。

组件	版本	描述
hive-server	1.0.0-amzn-1	用于将 Hive 查询作为 Web 请求接受的服务。
hue-server	3.7.1-amzn-4	用于使用 Hadoop 生态系统应用程序分析数据的 Web 应用程序
mahout-client	0.11.0	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
oozie-client	4.0.1	Oozie 命令行客户端。
oozie-server	4.0.1	用于接受 Oozie 工作流请求的服务。
presto-coordinator	0.119	用于在 presto-worker 之中接受查询并管理查询的服务。
presto-worker	0.119	用于执行查询的各个部分的服务。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.5.0	Spark 命令行客户端。
spark-history-server	1.5.0	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.5.0	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.5.0	YARN 从属项所需的 Apache Spark 库。
zeppelin-server	0.6.0-incubating-SNAPSHOT	支持交互式数据分析的基于 Web 的笔记本电脑。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅[配置应用程序]。

emr-4.1.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-encryption-zones	配置 HDFS 加密区域。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
hue-ini	更改 Hue 的 ini 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
hadoop-kms-acls	更改 Hadoop 的 kms-acls.xml 文件中的值。

分类	描述
hadoop-kms-env	更改 Hadoop KMS 环境中的值。
hadoop-kms-log4j	更改 Hadoop 的 kms-log4j.properties 文件中的值。
hadoop-kms-site	更改 Hadoop 的 kms-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
oozie-env	更改 Oozie 的环境中的值。
oozie-log4j	更改 Oozie 的 oozie-log4j.properties 文件中的值。
oozie-site	更改 Oozie 的 oozie-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
presto-log	更改 Presto 的 log.properties 文件中的值。
presto-config	更改 Presto 的 config.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。

分类	描述
zeppelin-env	更改 Zeppelin 环境中的值。

Amazon EMR 发行版 4.0.0

- 应用程序版本
- 发布说明
- 组件版本
- 配置分类

应用程序版本

此版本支持以下应用程序：Hadoop、Hive、Mahout、Pig 和 Spark。

下表列出了此版本的 Amazon EMR 中提供的应用程序版本以及前三个 Amazon EMR 发行版中的应用程序版本（若适用）。

有关每个发行版的 Amazon EMR 的应用程序版本的全面历史记录，请参见以下主题：

- Amazon EMR 6.x 发行版中的应用程序版本
- Amazon EMR 5.x 发行版中的应用程序版本
- Amazon EMR 4.x 发行版中的应用程序版本

应用程序版本信息

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Amazon SDK for Java	1.10.27	1.10.27	不可用	不可用
Python	不可用	不可用	不可用	不可用
Scala	不可用	不可用	不可用	不可用
Delta	-	-	-	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Flink	-	-	-	-
Ganglia	3.7.2	3.6.0	-	-
HBase	-	-	-	-
HCatalog	-	-	-	-
Hadoop	2.7.1	2.6.0	2.6.0	2.6.0
Hive	1.0.0	1.0.0	1.0.0	1.0.0
Hudi	-	-	-	-
Hue	3.7.1	3.7.1	3.7.1	-
Iceberg	-	-	-	-
JupyterEnterpriseGateway	-	-	-	-
JupyterHub	-	-	-	-
Livy	-	-	-	-
MXNet	-	-	-	-
Mahout	0.11.0	0.11.0	0.11.0	0.10.0
Oozie	-	-	-	-
Oozie-Sandbox	4.2.0	4.2.0	4.0.1	-
Phoenix	-	-	-	-
Pig	0.14.0	0.14.0	0.14.0	0.14.0
Presto	-	-	-	-
Presto-Sandbox	0.130	0.125	0.119	-

	emr-4.3.0	emr-4.2.0	emr-4.1.0	emr-4.0.0
Spark	1.6.0	1.5.2	1.5.0	1.4.1
Sqoop	-	-	-	-
Sqoop-Sandbox	-	-	-	-
TensorFlow	-	-	-	-
Tez	-	-	-	-
Trino (PrestoSQL)	-	-	-	-
Zeppelin	-	-	-	-
Zeppelin-Sandbox	0.5.5	0.5.5	0.6.0-SNAPSHOT	-
ZooKeeper	-	-	-	-
ZooKeeper-Sandbox	-	-	-	-

发布说明

组件版本

下面列出了 Amazon EMR 随此发行版一起安装的组件。一些组件作为大数据应用程序包的一部分安装。其它组件是 Amazon EMR 独有的，并且已为系统流程和功能安装这些组件。它们通常以 `emr` 或 `aws` 开头。最新的 Amazon EMR 发行版中的大数据应用程序包通常是在社区中找到的最新版本。我们会尽快在 Amazon EMR 中提供社区发行版。

Amazon EMR 中的某些组件与社区版本不同。这些组件具有以下形式的 *`CommunityVersion`*`-amzn-`*`EmrVersion`* 的发行版标注。*`EmrVersion`* 从 0 开始。例如，假设已对名为 `myapp-component` 的版本 2.2 的开源社区组件进行三次修改，以包含在不同的 Amazon EMR 发行版中，则其发行版将为 `2.2-amzn-2`。

组件	版本	描述
emr-ddb	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon DynamoDB 连接器。
emr-goodies	2.0.0	适用于 Hadoop 生态系统的方便易用的库。
emr-kinesis	3.0.0	适用于 Hadoop 生态系统应用程序的 Amazon Kinesis 连接器。
emr-s3-dist-cp	2.0.0	针对 Amazon S3 优化的分布式复制应用程序。
emrfs	2.0.0	适用于 Hadoop 生态系统应用程序的 Amazon S3 连接器。
hadoop-client	2.6.0-amzn-0	Hadoop 命令行客户端，如“hdfs”、“hadoop”或“yarn”。
hadoop-hdfs-datanode	2.6.0-amzn-0	用于存储数据块的 HDFS 节点级服务。
hadoop-hdfs-namenode	2.6.0-amzn-0	用于跟踪文件名和数据块位置的 HDFS 服务。
hadoop-httpfs-server	2.6.0-amzn-0	用于 HDFS 操作的 HTTP 终端节点。
hadoop-mapred	2.6.0-amzn-0	用于运行 MapReduce 应用程序的 MapReduce 执行引擎库。
hadoop-yarn-nodemanager	2.6.0-amzn-0	用于管理单个节点上的容器的 YARN 服务。

组件	版本	描述
hadoop-yarn-resourcemanager	2.6.0-amzn-0	用于分配和管理集群资源与分布式应用程序的 YARN 服务。
hive-client	1.0.0-amzn-0	Hive 命令行客户端。
hive-metastore-server	1.0.0-amzn-0	用于访问 Hive 元存储 (一个用于存储 Hadoop 操作中的 SQL 的元数据的语义存储库) 的服务。
hive-server	1.0.0-amzn-0	用于将 Hive 查询作为 Web 请求接受的服务。
mahout-client	0.10.0	用于机器学习的库。
mysql-server	5.5	MySQL 数据库服务器。
pig-client	0.14.0-amzn-0	Pig 命令行客户端。
spark-client	1.4.1	Spark 命令行客户端。
spark-history-server	1.4.1	用于查看完整的 Spark 应用程序的生命周期的已记录事件的 Web UI。
spark-on-yarn	1.4.1	适用于 YARN 的内存中执行引擎。
spark-yarn-slave	1.4.1	YARN 从属项所需的 Apache Spark 库。

配置分类

配置分类允许您自定义应用程序。这些通常与应用程序的配置 XML 文件（例如 `hive-site.xml`）相对应。有关更多信息，请参阅配置应用程序。

emr-4.0.0 分类

分类	描述
capacity-scheduler	更改 Hadoop 的 capacity-scheduler.xml 文件中的值。
core-site	更改 Hadoop 的 core-site.xml 文件中的值。
emrfs-site	更改 EMRFS 设置。
hadoop-env	更改适用于所有 Hadoop 组件的 Hadoop 环境中的值。
hadoop-log4j	更改 Hadoop 的 log4j.properties 文件中的值。
hdfs-site	更改 HDFS 的 hdfs-site.xml 中的值。
hive-env	更改 Hive 环境中的值。
hive-exec-log4j	更改 Hive 的 hive-exec-log4j.properties 文件中的值。
hive-log4j	更改 Hive 的 hive-log4j.properties 文件中的值。
hive-site	更改 Hive 的 hive-site.xml 文件中的值
httpfs-env	更改 HTTPFS 环境中的值。
httpfs-site	更改 Hadoop 的 httpfs-site.xml 文件中的值。
mapred-env	更改 MapReduce 应用程序的环境中的值。
mapred-site	更改 MapReduce 应用程序的 mapred-site.xml 文件中的值。
pig-properties	更改 Pig 的 pig.properties 文件中的值。
pig-log4j	更改 Pig 的 log4j.properties 文件中的值。
spark	适用于 Apache Spark 的 Amazon EMR 辅助设置。

分类	描述
spark-defaults	更改 Spark 的 spark-defaults.conf 文件中的值。
spark-env	更改 Spark 环境中的值。
spark-log4j	更改 Spark 的 log4j.properties 文件中的值。
yarn-env	更改 YARN 环境中的值。
yarn-site	更改 YARN 的 yarn-site.xml 文件中的值。

Amazon EMR 2.x 和 3.x AMI 版本

> **Note**
>
> Amazon 正在将所有 Amazon API 端点的 TLS 配置更新到最低 TLS 版本 1.2。Amazon EMR 3.10 及更低版本仅支持 TLS 1.0/1.1 连接。2023 年 12 月 4 日之后，您将无法使用 Amazon EMR 3.10 及更低版本创建集群。
> 如果您使用 Amazon EMR 3.10 或更低版本，我们建议您立即测试您的工作负载并将其迁移到最新的 Amazon EMR 版本。有关更多信息，请参阅 Amazon 安全博客。

Amazon EMR 2.x 和 3.x 发行版（称为 AMI 版本）是为出于兼容原因而需要它们的现有解决方案提供的。我们建议不要使用这些发行版创建新集群或新解决方案。它们缺少更新的发行版功能并且包含过时的应用程序包。

建议您使用最新的 Amazon EMR 发行版构建解决方案。

2.x 和 3.x 系列发行版与最新的 Amazon EMR 发行版之间的差异范围非常大。从您创建和配置集群的方式到集群上应用程序的端口和目录结构，均存在差异。

本节将尝试介绍 Amazon EMR 最重要的差异以及特定的应用程序配置和管理差异。介绍并不全面。如果您在 2.x 或 3.x 系列中创建并使用集群，可能遇到此节中未涉及的差异。

主题

- 使用 Amazon EMR 的早期 AMI 版本创建集群
- 使用 Amazon EMR 的早期 AMI 版本安装应用程序

- 使用 Amazon EMR 的早期 AMI 版本自定义集群和应用程序配置
- Amazon EMR 的早期 AMI 版本的 Hive 应用程序细节
- Amazon EMR 的早期 AMI 版本的 HBase 应用程序细节
- Amazon EMR 的早期 AMI 版本的 Pig 应用程序细节
- Amazon EMR 的早期 AMI 版本的 Spark 应用程序细节
- Amazon EMR 的早期 AMI 版本的 S3DistCp 实用程序差异

使用 Amazon EMR 的早期 AMI 版本创建集群

Amazon EMR 2.x 和 3.x 发行版是按 AMI 版本引用的。对于 Amazon EMR 版本 4.0.0 及更高版本，通过使用版本标签（如 `emr-5.11.0`）按发行版来引用版本。此更改在您使用 Amazon CLI 或以编程方式创建集群时最明显。

当您使用 Amazon CLI 通过 AMI 发行版创建集群时，请使用 `--ami-version` 选项（如 `--ami-version 3.11.0`）。Amazon EMR 4.0.0 及更高版本中推出的许多选项、功能和应用程序在您指定 `--ami-version` 时不可用。有关更多信息，请参阅《Amazon CLI 命令参考》中的 create-cluster。

以下示例 Amazon CLI 命令使用 AMI 版本启动集群。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.11.0 \
--applications Name=Hue Name=Hive Name=Pig \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-groups InstanceGroupType=MASTER,InstanceCount=1,\
InstanceType=m3.xlarge InstanceGroupType=CORE,InstanceCount=2,\
InstanceType=m3.xlarge --bootstrap-actions Path=s3://elasticmapreduce/bootstrap-
actions/configure-hadoop,\
Name="Configuring infinite JVM reuse",Args=["-m","mapred.job.reuse.jvm.num.tasks=-1"]
```

所有 Amazon EMR 发行版均以编程方式在 EMR API 中使用 `RunJobFlowRequest` 操作创建集群。以下示例 Java 代码使用 AMI 发行版 3.11.0 创建集群。

```
RunJobFlowRequest request = new RunJobFlowRequest()
   .withName("AmiVersion Cluster")
   .withAmiVersion("3.11.0")
   .withInstances(new JobFlowInstancesConfig()
    .withEc2KeyName("myKeyPair")
    .withInstanceCount(1)
    .withKeepJobFlowAliveWhenNoSteps(true)
    .withMasterInstanceType("m3.xlarge")
    .withSlaveInstanceType("m3.xlarge");
```

以下 RunJobFlowRequest 调用改用版本标签：

```
RunJobFlowRequest request = new RunJobFlowRequest()
   .withName("ReleaseLabel Cluster")
   .withReleaseLabel("emr-5.36.1")
   .withInstances(new JobFlowInstancesConfig()
    .withEc2KeyName("myKeyPair")
    .withInstanceCount(1)
    .withKeepJobFlowAliveWhenNoSteps(true)
    .withMasterInstanceType("m3.xlarge")
    .withSlaveInstanceType("m3.xlarge");
```

配置集群大小

当您的集群运行时，Hadoop 会决定处理数据所需的映射器和 Reducer 任务的数量。更大的集群应当有更多任务，以更好地利用资源并缩短处理时间。通常，EMR 集群在整个集群运行期间会保持相同的大小；当您创建集群时，需要设置任务数量。当您调整正在运行的集群大小时，您可以改变集群执行过程中的处理。因此，您可以在集群的生命周期内改变任务数量，而不是使用固定数量的任务。有两个配置选项可以帮助设置理想的任务数量：

- mapred.map.tasksperslot
- mapred.reduce.tasksperslot

您可以在 mapred-conf.xml 文件中设置这两个选项。当您向集群提交任务时，任务客户端会检查当前整个集群范围中可用的 map-reduce 槽位总数。然后，作业客户端会使用以下公式设置任务数量：

- mapred.map.tasks = mapred.map.tasksperslot * 集群中的映射槽位数
- mapred.reduce.tasks = mapred.reduce.tasksperslot * 集群中的 reduce 槽位

如果未配置任务数量，则作业客户端只会读取 tasksperslot 参数。通过引导操作 (对于全部集群) 或添加一个更改配置的步骤 (对每个作业单独执行)。您可以随时覆盖任务数量。

Amazon EMR 可以承受任务节点故障，即使任务节点不可用，也会继续执行集群。Amazon EMR 将自动预置额外的任务节点，以替换出现故障的节点。

对于每个集群步骤，您可以有不同数量的任务节点。您还可以向正在运行的集群添加步骤，以修改任务节点的数量。由于所有步骤在默认情况下都保证按顺序运行，您可以为任何步骤指定运行任务节点的数量。

使用 Amazon EMR 的早期 AMI 版本安装应用程序

当使用 AMI 版本时，可通过任何方式安装应用程序，包括使用 NewSupportedProductsRunJobFlow 操作的 参数，使用引导操作以及使用 Step 操作。

使用 Amazon EMR 的早期 AMI 版本自定义集群和应用程序配置

Amazon EMR 发行版 4.0.0 推出了使用配置分类的精简应用程序配置方法。有关更多信息，请参阅配置应用程序。当使用 AMI 版本时，使用引导操作以及您传递的参数配置应用程序。例如，configure-hadoop 和 configure-daemons 引导操作设置 Hadoop 和 YARN 特定环境属性（如 --namenode-heap-size）。在更新的版本中，这些属性是使用 hadoop-env 和 yarn-env 配置分类配置的。有关执行常见配置的引导操作，请参阅 Github 上的 emr-bootstrap-actions 存储库。

以下各表将引导操作映射到更新的 Amazon EMR 发行版中的配置分类。

Hadoop

受影响的应用程序文件名	AMI 版本引导操作	配置分类
core-site.xml	configure-hadoop -c	core-site
log4j.properties	configure-hadoop -l	hadoop-log4j
hdfs-site.xml	configure-hadoop -s	hdfs-site
不适用	不适用	hdfs-encryption-zones
mapred-site.xml	configure-hadoop -m	mapred-site
yarn-site.xml	configure-hadoop -y	yarn-site

受影响的应用程序文件名	AMI 版本引导操作	配置分类
httpfs-site.xml	configure-hadoop -t	httpfs-site
capacity-scheduler.xml	configure-hadoop -z	capacity-scheduler
yarn-env.sh	configure-daemons --resourcemanager-opts	yarn-env

Hive

受影响的应用程序文件名	AMI 版本引导操作	配置分类
hive-env.sh	不适用	hive-env
hive-site.xml	hive-script --install-hive-site ${MY_HIVE_SITE_FILE}	hive-site
hive-exec-log4j.properties	不适用	hive-exec-log4j
hive-log4j.properties	不适用	hive-log4j

EMRFS

受影响的应用程序文件名	AMI 版本引导操作	配置分类
emrfs-site.xml	configure-hadoop -e	emrfs-site
不适用	s3get -s s3://custom-provider.jar -d /usr/share/aws/emr/auxlib/	emrfs-site (带新设置 fs.s3.cse.encryptionMaterialsProvider.uri)

有关所有分类的列表，请参阅配置应用程序。

应用程序环境变量

当使用 AMI 版本时，将结合使用 hadoop-user-env.sh 脚本与 configure-daemons 引导操作来配置 Hadoop 环境。该脚本包括以下操作：

```
#!/bin/bash
export HADOOP_USER_CLASSPATH_FIRST=true;
echo "HADOOP_CLASSPATH=/path/to/my.jar" >> /home/hadoop/conf/hadoop-user-env.sh
```

在 Amazon EMR 版本 4.x 中，使用 hadoop-env 配置分类执行相同的操作，如以下示例中所示：

```
[
      {
         "Classification":"hadoop-env",
         "Properties":{

         },
         "Configurations":[
            {
               "Classification":"export",
               "Properties":{
                  "HADOOP_USER_CLASSPATH_FIRST":"true",
                  "HADOOP_CLASSPATH":"/path/to/my.jar"
               }
            }
         ]
      }
   ]
```

再举一例，使用 configure-daemons 并传递 --namenode-heap-size=2048 和 --namenode-opts=-XX:GCTimeRatio=19 等效于以下配置分类。

```
[
      {
         "Classification":"hadoop-env",
         "Properties":{

         },
         "Configurations":[
            {
```

```
            "Classification":"export",
            "Properties":{
               "HADOOP_DATANODE_HEAPSIZE":  "2048",
          "HADOOP_NAMENODE_OPTS":  "-XX:GCTimeRatio=19"
            }
          }
       ]
    }
 ]
```

其他应用程序环境变量不再在 /home/hadoop/.bashrc 中定义。相反，它们主要在 /etc/default 文件中基于每个组件或应用程序设置，例如 /etc/default/hadoop。/usr/bin/ 中由应用程序 RPM 安装的包装脚本还可以在涉及实际 bin 脚本之前设置其他环境变量。

服务端口

当使用 AMI 版本时，某些服务使用自定义端口。

端口设置中的更改

设置	AMI 版本 3.x	开源默认值
fs.default.name	hdfs://emrDeterminedIP:9000	默认值 (hdfs://*emrDeterminedIP* :8020)
dfs.datanode.address	0.0.0.0:9200	默认值 (0.0.0.0:50010)
dfs.datanode.http.address	0.0.0.0:9102	默认值 (0.0.0.0:50075)
dfs.datanode.https.address	0.0.0.0:9402	默认值 (0.0.0.0:50475)
dfs.datanode.ipc.address	0.0.0.0:9201	默认值 (0.0.0.0:50020)
dfs.http.address	0.0.0.0:9101	默认值 (0.0.0.0:50070)
dfs.https.address	0.0.0.0:9202	默认值 (0.0.0.0:50470)
dfs.secondary.http.address	0.0.0.0:9104	默认值 (0.0.0.0:50090)
yarn.nodemanager.address	0.0.0.0:9103	默认值 (${yarn.nodemanager.hostname}:0)

设置	AMI 版本 3.x	开源默认值
yarn.nodemanager.localizer.address	0.0.0.0:9033	默认值 (${yarn.nodemanager.hostname}:8040)
yarn.nodemanager.webapp.address	0.0.0.0:9035	默认值 (${yarn.nodemanager.hostname}:8042)
yarn.resourcemanager.address	*emrDeterminedIP* :9022	默认值 (${yarn.resourcemanager.hostname}:8032)
yarn.resourcemanager.admin.address	*emrDeterminedIP* :9025	默认值 (${yarn.resourcemanager.hostname}:8033)
yarn.resourcemanager.resource-tracker.address	*emrDeterminedIP* :9023	默认值 (${yarn.resourcemanager.hostname}:8031)
yarn.resourcemanager.scheduler.address	*emrDeterminedIP* :9024	默认值 (${yarn.resourcemanager.hostname}:8030)
yarn.resourcemanager.webapp.address	0.0.0.0:9026	默认值 (${yarn.resourcemanager.hostname}:8088)
yarn.web-proxy.address	*emrDeterminedIP* :9046	默认值 (无值)
yarn.resourcemanager.hostname	0.0.0.0 (默认值)	*emrDeterminedIP*

> **Note**
>
> *emrDeterminedIP* 是 Amazon EMR 生成的 IP 地址。

用户

当使用 AMI 版本时，用户 hadoop 将运行所有进程并拥有所有文件。在 Amazon EMR 版本 4.0.0 及更高版本中，用户位于应用程序和组件级别。

安装顺序、安装的构件和日志文件位置

当使用 AMI 版本时，应用程序构件及其配置目录安装在 /home/hadoop/*application* 目录中。例如，如果已安装 Hive，目录将是 /home/hadoop/hive。在 Amazon EMR 版本 4.0.0 及更高版本中，应用程序构件安装在 /usr/lib/*application* 目录中。当使用 AMI 版本时，日志文件位于不同的位置。下表列出了位置。

Amazon S3 上的日志位置的更改

守护进程或应用程序	目录位置
instance-state	node/*instance-id* /instance-state/
hadoop-hdfs-namenode	daemons/*instance-id* /hadoop-hadoop-namenode.log
hadoop-hdfs-datanode	daemons/*instance-id* /hadoop-hadoop-datanode.log
hadoop-yarn (ResourceManager)	daemons/*instance-id* /yarn-hadoop-resourcemanager
hadoop-yarn (代理服务器)	daemons/*instance-id* /yarn-hadoop-proxyserver
mapred-historyserver	daemons/*instance-id* /
httpfs	daemons/*instance-id* /httpfs.log
hive-server	node/*instance-id* /hive-server/hive-server.log
hive-metastore	node/*instance-id* /apps/hive.log
Hive CLI	node/*instance-id* /apps/hive.log
YARN 应用程序用户日志和容器日志	task-attempts/
Mahout	不适用
Pig	不适用

守护进程或应用程序	目录位置
spark-historyserver	不适用
mapreduce 任务历史记录文件	jobs/

Command Runner

当使用 AMI 版本时，许多脚本或程序 (如 `/home/hadoop/contrib/streaming/hadoop-streaming.jar`) 未位于 shell 登录路径环境上，因此您需要在使用 jar 文件 (如 command-runner.jar 或 script-runner.jar) 执行脚本时指定完整路径。`command-runner.jar` 位于 AMI 上，因此无需知道完整 URI，这与 `script-runner.jar` 的情况相同。

重复因子

重复因子使您可以配置启动 Hadoop JVM 的时间。您可以在执行每项任务时启动新的 Hadoop JVM，这将实现更好的任务隔离；也可以在各项任务之间共享 JVM，以降低框架开销。如果您处理的是许多小文件，合理的做法是多次重复使用 JVM，以摊销启动成本。然而，如果每项任务耗时较长或处理的数据量较大，您可以选择不重复使用 JVM，以确保释放出所有内存供后续任务使用。当使用 AMI 版本时，您可以使用 `configure-hadoop` 引导操作设置 `mapred.job.reuse.jvm.num.tasks` 属性来自定义重复因子。

以下示例演示如何设置无限 JVM 重复使用的 JVM 重复使用因子。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^) 。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.11.0 \
--applications Name=Hue Name=Hive Name=Pig \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-groups InstanceGroupType=MASTER,InstanceCount=1,InstanceType=m3.xlarge \
InstanceGroupType=CORE,InstanceCount=2,InstanceType=m3.xlarge \
--bootstrap-actions Path=s3://elasticmapreduce/bootstrap-actions/configure-hadoop,\
Name="Configuring infinite JVM reuse",Args=["-m","mapred.job.reuse.jvm.num.tasks=-1"]
```

Amazon EMR 的早期 AMI 版本的 Hive 应用程序细节

日志文件

使用 Amazon EMR AMI 版本 2.x 和 3.x 时，Hive 日志将保存到 `/mnt/var/log/apps/`。为了支持 Hive 的并行版本，您运行的这个 Hive 版本还会确定日志文件名称，如下表所示。

Hive 版本	日志文件名称
0.13.1	`hive.log` **Note** 从此版本开始，Amazon EMR 使用非版本化的文件名 `hive.log`。次要版本将与主要版本共享同一个日志位置。
0.11.0	`hive_0110.log` **Note** 0.11.0 的次要版本，如 Hive 0.11.0.1，和 Hive 0.11.0 一起共享相同的日志文件位置。
0.8.1	`hive_081.log` **Note** Hive 0.8.1 的次要版本，如 Hive 0.8.1.1，和 Hive 0.8.1 一起共享相同的日志文件位置。
0.7.1	`hive_07_1.log`

Hive 版本	日志文件名称
	Note Hive 0.7.1 的次要版本，如 Hive 0.7.1.3 和 Hive 0.7.1.4，和 Hive 0.7.1 一起共享相同的日志文件位置。
0.7	`hive_07.log`
0.5	`hive_05.log`
0.4	`hive.log`

拆分输入功能

要使用 0.13.1 之前的 Hive 版本（3.11.0 之前的 Amazon EMR AMI 版本）实现拆分输入功能，请使用以下命令：

```
hive> set hive.input.format=org.apache.hadoop.hive.ql.io.HiveCombineSplitsInputFormat;
hive> set mapred.min.split.size=100000000;
```

Hive 0.13.1 已弃用此功能。要在 Amazon EMR AMI 版本 3.11.0 中获得同样的拆分输入格式功能，可以使用以下命令：

```
set hive.hadoop.supports.splittable.combineinputformat=true;
```

Thrift 服务端口

Thrift 是一种 RPC 框架，用于定义紧凑型二进制序列化格式，以保存数据结构供后续分析使用。通常，Hive 会配置服务器在以下端口上运行。

Hive 版本	端口号
Hive 0.13.1	10000
Hive 0.11.0	10004

Hive 版本	端口号
Hive 0.8.1	10003
Hive 0.7.1	10002
Hive 0.7	10001
Hive 0.5	10000

有关 Thrift 服务的更多信息，请参阅 http://wiki.apache.org/thrift/。

使用 Hive 恢复分区

Amazon EMR 包含一条 Hive 查询语言语句，该语句可以从位于 Amazon S3 中的表数据恢复表的分区。以下示例对此进行了介绍。

```
CREATE EXTERNAL TABLE (json string) raw_impression
PARTITIONED BY (dt string)
LOCATION 's3://elastic-mapreduce/samples/hive-ads/tables/impressions';
ALTER TABLE logs RECOVER PARTITIONS;
```

分区目录和数据必须处于表定义所指定的位置，而且必须根据 Hive 惯例命名：如`dt=2009-01-01`。

> **Note**
>
> Hive 0.13.1 版之后使用 `msck repair table` 在本机支持此功能，因此不支持 `recover partitions`。有关更多信息，请参阅 https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL。

将 Hive 变量传递到脚本

要使用 Amazon CLI 将变量传递到 Hive 步骤，请键入以下命令，使用您的 EC2 密钥对的名称替换 *myKey*，使用您的存储桶名称替换 *mybucket*。在此示例中，SAMPLE 是 -d 开关后面的变量值。此变量在 Hive 脚本中的定义如下：`${SAMPLE}`。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.9 \
--applications Name=Hue Name=Hive Name=Pig \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-type m3.xlarge --instance-count 3 \
--steps Type=Hive,Name="Hive Program",ActionOnFailure=CONTINUE,\
Args=[-f,s3://elasticmapreduce/samples/hive-ads/libs/response-time-stats.q,-d,\
INPUT=s3://elasticmapreduce/samples/hive-ads/tables,-d,OUTPUT=s3://mybucket/hive-ads/
output/,\
-d,SAMPLE=s3://elasticmapreduce/samples/hive-ads/]
```

指定外部元存储位置

以下步骤介绍了如何覆盖 Hive 元数据仓位置的默认配置值和使用重新配置的元数据仓位置启动集群。

创建位于 EMR 集群外的元数据仓

1. 使用 Amazon RDS 创建 MySQL 或 Aurora 数据库。

   有关如何创建 Amazon RDS 数据库的信息，请参阅 Amazon RDS 入门。
2. 修改您的安全组，以允许在数据库与 ElasticMapReduce-Master 安全组之间建立 JDBC 连接。

   有关如何针对访问权限修改安全组的信息，请参阅《Amazon RDS 用户指南》中的 Amazon RDS 安全组。
3. 在 `hive-site.xml` 中设置 JDBC 配置值：

   a. 创建包含以下信息的 `hive-site.xml` 配置文件：

```
<configuration>
  <property>
    <name>javax.jdo.option.ConnectionURL</name>
    <value>jdbc:mariadb://hostname:3306/hive?createDatabaseIfNotExist=true</
value>
    <description>JDBC connect string for a JDBC metastore</description>
  </property>
```

```
  <property>
    <name>javax.jdo.option.ConnectionUserName</name>
    <value>hive</value>
    <description>Username to use against metastore database</description>
  </property>
  <property>
    <name>javax.jdo.option.ConnectionPassword</name>
    <value>password</value>
    <description>Password to use against metastore database</description>
  </property>
</configuration>
```

hostname 是运行数据库的 Amazon RDS 实例的 DNS 地址。*username* 和 *password* 是数据库的凭证。有关连接到 MySQL 和 Aurora 数据库实例的更多信息，请参阅《Amazon RDS 用户指南》中的连接到运行 MySQL 数据库引擎的数据库实例和连接到 Aurora 数据库集群。

JDBC 驱动程序由 Amazon EMR 进行安装。

> **Note**
>
> 值属性不应该包含任何空格或回车。所有内容应显示在一行中。

b. 将 `hive-site.xml` 文件保存到 Amazon S3 中的位置上，如 `s3://mybucket/hive-site.xml`。

4. 创建一个集群，以指定自定义 `hive-site.xml` 文件的 Amazon S3 位置。

以下示例命令演示执行此操作的 Amazon CLI 命令。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.10 \
--applications Name=Hue Name=Hive Name=Pig \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-type m3.xlarge --instance-count 3 \
--bootstrap-actions Name="Install Hive Site Configuration",\
Path="s3://region.elasticmapreduce/libs/hive/hive-script",\
```

```
Args=["--base-path","s3://elasticmapreduce/libs/hive","--install-hive-site",\
"--hive-site=s3://mybucket/hive-site.xml","--hive-versions","latest"]
```

使用 JDBC 连接到 Hive

要通过 JDBC 连接 Hive，您需要下载 JDBC 驱动程序并安装 SQL 客户端。以下示例说明如何使用 SQL Workbench/J 通过 JDBC 连接 Hive。

下载 JDBC 驱动程序

1. 下载并解压适用于您想访问的 Hive 版本的驱动程序。根据您在创建 Amazon EMR 集群时选择的 AMI，Hive 版本有所不同。

   - Hive 0.13.1 JDBC 驱动程序：https://amazon-odbc-jdbc-drivers.s3.amazonaws.com/public/AmazonHiveJDBC_1.0.4.1004.zip
   - Hive 0.11.0 JDBC 驱动程序：https://mvnrepository.com/artifact/org.apache.hive/hive-jdbc/0.11.0
   - Hive 0.8.1 JDBC 驱动程序：https://mvnrepository.com/artifact/org.apache.hive/hive-jdbc/0.8.1
2. 安装 SQL Workbench/J。有关更多信息，请参阅 SQL Workbench/J 用户手册中的安装并启动 SQL Workbench/J。
3. 创建到集群主节点的 SSH 隧道。连接端口因 Hive 版本而异。下表中提供了适用于 Linux ssh 用户的示例命令以及适用于 Windows 用户的 PuTTY 命令

   Linux SSH 命令

Hive 版本	命令
0.13.1	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10000:localhost:10000 hadoop@ master-public-dns-name`
0.11.0	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10004:localhost:10004 hadoop@ master-public-dns-name`
0.8.1	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10003:localhost:10003 hadoop@ master-public-dns-name`
0.7.1	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10002:localhost:10002 hadoop@ master-public-dns-name`

Hive 版本	命令
0.7	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10001:localhost:10001 hadoop@ master-public-dns-name`
0.5	`ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10000:localhost:10000 hadoop@ master-public-dns-name`

Windows PuTTY 隧道设置

Hive 版本	隧道设置
0.13.1	Source port (源端口)：10000 Destination (目标)：*master-public-dns-name* :10000
0.11.0	Source port (源端口)：10004 Destination (目标)：*master-public-dns-name* :10004
0.8.1	Source port (源端口)：10003 Destination (目标)：*master-public-dns-name* :10003

4. 将 JDBC 驱动程序添加到 SQL Workbench。

   a. 在 Select Connection Profile (选择连接配置文件) 对话框中，选择 Manage Drivers (管理驱动程序)。

   b. 选择 Create a new entry (创建新条目) (空白页) 图标。

   c. 在名称字段中，键入 **`Hive JDBC`**。

   d. 对于 Library (库)，请单击 Select the JAR file(s) (选择 JAR 文件) 图标。

   e. 选择如下表中所示的 JAR 文件。

Hive 驱动程序版本	要添加的 JAR 文件

0.13.1	`hive_metastore.jar`
	`hive_service.jar`
	`HiveJDBC3.jar`
	`libfb303-0.9.0.jar`

Hive 驱动程序版本	要添加的 JAR 文件
	libthrift-0.9.0.jar log4j-1.2.14.jar ql.jar slf4j-api-1.5.8.jar slf4j-log4j12-1.5.8.jar TCLIServiceClient.jar
0.11.0	hadoop-core-1.0.3.jar hive-exec-0.11.0.jar hive-jdbc-0.11.0.jar hive-metastore-0.11.0.jar hive-service-0.11.0.jar libfb303-0.9.0.jar commons-logging-1.0.4.jar slf4j-api-1.6.1.jar
0.8.1	hadoop-core-0.20.205.jar hive-exec-0.8.1.jar hive-jdbc-0.8.1.jar hive-metastore-0.8.1.jar hive-service-0.8.1.jar libfb303-0.7.0.jar libthrift-0.7.0.jar log4j-1.2.15.jar slf4j-api-1.6.1.jar slf4j-log4j12-1.6.1.jar
0.7.1	hadoop-0.20-core.jar hive-exec-0.7.1.jar hive-jdbc-0.7.1.jar hive-metastore-0.7.1.jar hive-service-0.7.1.jar libfb303.jar commons-logging-1.0.4.jar slf4j-api-1.6.1.jar slf4j-log4j12-1.6.1.jar

Hive 驱动程序版本	要添加的 JAR 文件
0.7	hadoop-0.20-core.jar hive-exec-0.7.0.jar hive-jdbc-0.7.0.jar hive-metastore-0.7.0.jar hive-service-0.7.0.jar libfb303.jar commons-logging-1.0.4.jar slf4j-api-1.5.6.jar slf4j-log4j12-1.5.6.jar
0.5	hadoop-0.20-core.jar hive-exec-0.5.0.jar hive-jdbc-0.5.0.jar hive-metastore-0.5.0.jar hive-service-0.5.0.jar libfb303.jar log4j-1.2.15.jar commons-logging-1.0.4.jar

f. 在 Please select one driver (请选择一个驱动程序) 对话框中，根据下表选择一个驱动程序并单击 OK (确定)。

Hive 版本	驱动程序类名
0.13.1	com.amazon.hive.jdbc3.HS2Driver
0.11.0	org.apache.hadoop.hive.jdbc.HiveDriver.jar
0.8.1	org.apache.hadoop.hive.jdbc.HiveDriver.jar
0.7.1	org.apache.hadoop.hive.jdbc.HiveDriver.jar

Hive 版本	驱动程序类名
0.7	`org.apache.hadoop.hive.jdbc.HiveDriver.jar`
0.5	`org.apache.hadoop.hive.jdbc.HiveDriver.jar`

5. 当您返回到 Select Connection Profile（选择连接配置文件）对话框时，验证 Driver（驱动程序）字段是否设置为 Hive JDBC，然后在 URL 字段中根据下表提供 JDBC 连接字符串。

Hive 版本	JDBC 连接字符串
0.13.1	`jdbc:hive2://localhost:10000/default`
0.11.0	`jdbc:hive://localhost:10004/default`
0.8.1	`jdbc:hive://localhost:10003/default`

如果集群使用 AMI 版本 3.3.1 或更高版本，则在 Select Connection Profile (选择连接配置文件) 对话框中，在 Username (用户名) 字段中键入 **hadoop**。

Amazon EMR 的早期 AMI 版本的 HBase 应用程序细节

支持的 HBase 版本

HBase 版本	AMI 版本	Amazon CLI 配置参数	HBase 版本详细信息

0.94.18	3.1.0 和更高版本	`--ami-version 3.1`
		
		`--ami-version 3.2`
		
		`--ami-version 3.3`
			• 错误修复和增强功能。

HBase 版本	AMI 版本	Amazon CLI 配置参数	HBase 版本详细信息
		`--applications Name=HBase`	

0.94.7	3.0-3.0.4	`--ami-version 3.0`
		
		`--applications Name=HBase`	

0.92	2.2 和更高版本	`--ami-version 2.2 or later`
		
		`--applications Name=HBase`	

HBase 集群前提条件

使用 Amazon EMR AMI 版本 2.x 和 3.x 创建的集群应满足 HBase 的以下要求。

- Amazon CLI（可选）– 要通过命令行与 HBase 进行交互，请下载并安装最新版本的 Amazon CLI。有关更多信息，请参阅《Amazon Command Line Interface 用户指南》中的安装 Amazon Command Line Interface。
- 至少两个实例（可选）– 集群的主节点运行 HBase 主服务器和 Zookeeper，任务节点运行 HBase 区域服务器。为获得最佳性能，HBase 集群至少应在两个 EC2 实例上运行，但在进行评估时，您可以在单一节点上运行 HBase。
- 长时间运行的集群 – HBase 只能在长时间运行的集群上运行。默认情况下，CLI 和 Amazon EMR 控制台会创建长时间运行的集群。
- Amazon EC2 密钥对集（推荐）– 要使用 Secure Shell（SSH）网络协议连接主节点并运行 HBase Shell 命令，必须在创建集群时使用 Amazon EC2 密钥对。
- The correct AMI and Hadoop versions (正确的 AMI 和 Hadoop 版本) — 目前，仅 Hadoop 20.205 或后续版本支持 HBase 群集。
- Ganglia(可选) - 要监控 HBase 性能指标，可以在创建群集时安装 Ganglia。
- 用于存储日志的 Amazon S3 存储桶（可选）– HBase 的日志可在主节点上使用。如果要将这些日志复制到 Amazon S3 中，请在创建集群时指定接收日志文件的 S3 存储桶。

创建带 HBase 的集群

下表列出了使用控制台通过 Amazon EMR AMI 发行版创建带 HBase 的集群时可用的选项。

Field	操作
Restore from backup (从备份中还原)	指定是否要使用存储在 Amazon S3 中的数据预加载 HBase 集群。
Backup location (备份位置)	指定 Amazon S3 中还原所用备份的存储位置的 URI。
Backup version (备份版本)	在 Backup Location (备份位置) 处指定要使用的备份的版本名称（可选）。如果将此字段留空，Amazon EMR 会使用 Backup Location (备份位置) 处的最新备份填充新 HBase 集群。
Schedule Regular Backups (安排定期备份)	指定是否要计划自动增量备份。第一次备份是完整备份，为以后的增量备份创建基线。
Consistent backup (一致性备份)	请指定备份是否应该保持一致。一致性备份指的是，在初始备份阶段为保持节点同步而暂停写入操作的备份。将任何因此而暂停的写入操作放入队列中，然后在同步完成时恢复操作。
Backup frequency (备份频率)	计划备份间隔的天/小时/分钟数。
Backup location (备份位置)	存储备份的 Amazon S3 URI。每个 HBase 集群的备份位置应该不同，确保差异备份保持正确。
Backup start time (备份启动时间)	指定进行首次备份的时间。您可以将此值设置为 now，使第一次备份在集群运行时立即开始，也可以输入 ISO 格式的日期和时间。例如，2012-06-15T20:00Z 会将开始时间设置为 2012 年 6 月 15 日下午 8 点 (UTC)。

以下示例 Amazon CLI 命令启动带 HBase 和其他应用程序的集群：

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^) 。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.3 \
               --applications Name=Hue Name=Hive Name=Pig Name=HBase \
               --use-default-roles --ec2-attributes KeyName=myKey \
               --instance-type c1.xlarge --instance-count 3 --termination-protected
```

在 Hive 和 HBase HBase 集群之间建立连接后 (如上一过程所示)。您可以通过在 Hive 中创建外部表访问存储在 HBase 集群上的数据。

从 Hive 提示符中运行时，以下示例创建了一个外部表，此表引用了存储在名为 `inputTable` 的 HBase 表上的数据。然后，您可以引用 Hive 语句中的 `inputTable`，查询和修改存储在 HBase 集群上的数据。

> **Note**
>
> 以下示例使用了 AMI 2.3.3 中的 protobuf-java-2.4.0a.jar，但是您应该修改此示例以匹配您的版本。要检查您有哪一版本的 Protocol Buffers JAR，请在 Hive 命令提示符处运行命令：`! ls /home/hadoop/lib;`。

```
add jar lib/emr-metrics-1.0.jar ;
               add jar lib/protobuf-java-2.4.0a.jar ;

               set hbase.zookeeper.quorum=ec2-107-21-163-157.compute-1.amazonaws.com ;

               create external table inputTable (key string, value string)
                    stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
                     with serdeproperties ("hbase.columns.mapping" = ":key,f1:col1")
                     tblproperties ("hbase.table.name" = "t1");

               select count(*) from inputTable ;
```

自定义 HBase 配置

尽管默认设置应当用于大多数应用程序，但是您可以灵活地修改 HBase 配置设置。为此，请运行两种引导操作脚本之一：

- configure-hbase-daemons – 配置主守护进程、regionserver 守护进程和 zookeeper 守护进程的属性。这些属性包括堆大小和 HBase 守护程序启动时传递至 Java 虚拟机 (JVM) 的选项。将这些属性设置为引导操作中的参数。此引导操作可修改 HBase 集群上的 /home/hadoop/conf/hbase-user-env.sh 配置文件。
- configure-hbase – 配置 HBase 站点特定的设置，例如 HBase 主服务器应绑定到的端口，以及客户端 CLI 客户端应重试操作的最大次数。您可以将这些属性逐一设置为引导操作中的参数，或者您可以指定 XML 配置文件在 Amazon S3 中的位置。此引导操作可修改 HBase 集群上的 /home/hadoop/conf/hbase-site.xml 配置文件。

> **Note**
>
> 与其他引导操作一样，这些脚本只有在创建集群时才可以运行，您不能使用它们更改目前运行的 HBase 集群的配置。

当您运行 configure-hbase 或 configure-hbase-daemons 引导操作时，您指定的值会覆盖默认值。任何未显式设置的值都会接受默认值。

使用这些引导操作配置 HBase 与在 Amazon EMR 中使用引导操作配置 Hadoop 设置和 Hadoop 守护进程属性类似。区别在于，HBase 没有每个进程的内存选项。而是使用 `--daemon-opts` 参数设置内存选项，使用要配置的守护程序名称代替 *daemon*。

配置 HBase 守护进程

Amazon EMR 提供了引导操作 `s3://region.elasticmapreduce/bootstrap-actions/configure-hbase-daemons`，您可以使用此操作更改 HBase 守护进程的配置，其中 *region (##)* 是您要在其中启动 HBase 集群的区域。

要使用 Amazon CLI 配置 HBase 守护程序，请在启动集群时添加 `configure-hbase-daemons` 引导操作，以配置一个或多个 HBase 守护程序。您可以设置以下属性。

属性	描述
hbase-master-opts	控制 JVM 如何运行主守护程序的选项。如果设置此选项，则它会覆盖默认的 HBASE_MASTER_OPTS 变量。
regionserver-opts	控制 JVM 如何运行区域服务器守护程序的选项。如果设置此选项，则它会覆盖默认的 HBASE_REGIONSERVER_OPTS 变量。
zookeeper-opts	控制 JVM 如何运行 zookeeper 守护程序的选项。如果设置此选项，则它会覆盖默认的 HBASE_ZOOKEEPER_OPTS 变量。

有关这些选项的更多信息，请参阅 HBase 文档中的 hbase-env.sh。

以下示例中显示的是为 zookeeper-opts 和 hbase-master-opts 配置值的引导操作。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.3 \
--applications Name=Hue Name=Hive Name=Pig Name=HBase \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-type c1.xlarge --instance-count 3 --termination-protected \
--bootstrap-actions Path=s3://elasticmapreduce/bootstrap-actions/configure-hbase-daemons,\
Args=["--hbase-zookeeper-opts=-Xmx1024m -XX:GCTimeRatio=19","--hbase-master-opts=-Xmx2048m","--hbase-regionserver-opts=-Xmx4096m"]
```

配置 HBase 站点设置

Amazon EMR 提供了引导操作 s3://elasticmapreduce/bootstrap-actions/configure-hbase，您可以使用此操作更改 HBase 的配置。您可以将配置值逐一设置为引导操作中的参数，或者您可以指定 XML 配置文件在 Amazon S3 中的位置。如果您只需要设置几个配置设置，那么逐一设置

配置值会很有用。如果您需要做出很多更改，或如果您要保存配置设置以便重新使用，那么非常适合使用 XML 文件进行设置。

> **Note**
>
> 您可以为 Amazon S3 存储桶名称加上区域前缀，如 `s3://region.elasticmapreduce/bootstrap-actions/configure-hbase`，其中的 *region (##)* 是要在其中启动 HBase 集群的区域。

此引导操作可修改 HBase 集群上的 `/home/hadoop/conf/hbase-site.xml` 配置文件。只有当 HBase 集群启动时，才可以运行引导操作。

有关可配置的 HBase 站点设置的更多信息，请参阅 HBase 文档中的默认配置。

设置当您启动 HBase 集群时的 `configure-hbase` 引导操作，并在 `hbase-site.xml` 中指定要更改的值。

使用 Amazon CLI 指定单个 HBase 站点设置

- 要更改 `hbase.hregion.max.filesize` 设置，请键入以下命令，并将 *myKey* 替换为您的 Amazon EC2 密钥对的名称。

  > **Note**
  >
  > 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.3 \
--applications Name=Hue Name=Hive Name=Pig Name=HBase \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-type c1.xlarge --instance-count 3 --termination-protected \
--bootstrap-actions Path=s3://elasticmapreduce/bootstrap-actions/configure-
hbase,Args=["-s","hbase.hregion.max.filesize=52428800"]
```

通过 Amazon CLI 使用 XML 文件指定 HBase 站点设置

1. 创建 hbase-site.xml 的自定义版本。您的自定义文件必须是有效的 XML。要减少出现错误的机会，请使用 Amazon EMR HBase 主节点上 /home/hadoop/conf/hbase-site.xml 处的 hbase-site.xml 默认副本开始，然后编辑该文件的副本，而不是从头创建全新文件。可以给您的新文件指定一个新名称，或保留 hbase-site.xml。
2. 将您的自定义 hbase-site.xml 文件上传到 Amazon S3 存储桶。应当设置权限，以便启动集群的Amazon账户可以访问此文件。如果启动集群的Amazon账户还拥有 Amazon S3 存储桶，它将具有访问权限。
3. 设置当您启动 HBase 集群时的 configure-hbase 引导操作，并在其中包含您的自定义 hbase-site.xml 文件所在的位置。以下示例将 HBase 站点配置值设置为文件 s3://mybucket/my-hbase-site.xml 中指定的值。键入以下命令，将 *myKey* 替换为您的 EC2 密钥对的名称，将 *mybucket* 替换为您的 Amazon S3 存储桶的名称。

   > **Note**
   >
   > 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

   ```
   aws emr create-cluster --name "Test cluster" --ami-version 3.3 \
           --applications Name=Hue Name=Hive Name=Pig Name=HBase \
           --use-default-roles --ec2-attributes KeyName=myKey \
           --instance-type c1.xlarge --instance-count 3 --termination-protected \
           --bootstrap-actions Path=s3://elasticmapreduce/bootstrap-actions/configure-hbase,Args=["--site-config-file","s3://mybucket/config.xml"]
   ```

   如果指定多个选项来自定义 HBase 操作，则必须在每个键值对的前面加上 -s 选项开关，如以下示例所示：

   ```
         --bootstrap-actions s3://elasticmapreduce/bootstrap-actions/configure-hbase,Args=["-s","zookeeper.session.timeout=60000"]
   ```

通过设置代理并打开 SSH 连接，您可以打开浏览器窗口，通过 http://*master-public-dns-name*:60010/master-status 查看 HBase UI，其中 *master-public-dns-name* 是 HBase 集群中主节点的公有 DNS 地址。

您可以使用 SSH 连接主节点，然后导航到 `mnt/var/log/hbase` 目录，从而查看当前的 HBase 日志。除非您在集群启动时启用了针对 Amazon S3 的日志记录，否则这些日志将在集群终止后不再可用。

备份和还原 HBase

Amazon EMR 可通过手动方式或按照计划自动将 HBase 数据备份到 Amazon S3 中。您可以执行完整备份和增量备份。拥有 HBase 数据的备份版本后，您就可以将该版本还原到 HBase 集群。您可以还原到当前正在运行的 HBase 集群，也可以启动预填充了备份数据的新集群。

备份过程中，HBase 会继续执行写入命令。虽然这样可确保集群在整个备份过程中都处于可用状态，但存在的风险是，正在备份的数据可能与并行执行的任何写入操作不一致。要了解可能出现的不一致情况，您必须考虑到，HBase 是在集群中的各节点处分配写入操作的。如果写入操作发生在轮询特定节点之后，则备份存档中不会包含该数据。甚至，您可能会发现，备份存档中可能未包含 HBase 集群的早期写入操作 (发送到已轮询过的节点)。而包含了后期写入操作 (发送到尚未轮询的节点)。

如果需要进行一致性备份，您必须在备份过程的初始部分中暂停对 HBase 的写入操作，保持节点同步。您可以通过在请求备份时指定 `--consistent` 参数完成此操作。指定此参数后，系统会将此期间的写入操作加入队列并在同步完成后立即执行这些操作。您还可以计划重复备份，一次备份过程中遗失的数据会在后续备份过程中得到备份，从而解决随时间推移而产生的任何不一致问题。

在备份 HBase 数据时，应该为每个集群指定不同的备份目录。完成此操作的一种简单方法是，将集群标识符用作备份目录指定路径的一部分。例如，`s3://mybucket/backups/j-3AEXXXXXX16F2`。这样可确保以后所有的增量备份都引用正确的 HBase 集群。

在准备删除不再需要的旧备份文件时，我们建议您首先对 HBase 数据进行完整备份。这样可确保保留所有数据并为以后的增量备份提供基线。完成完整备份后，您可以导航到备份位置并手动删除旧备份文件。

HBase 备份过程将 S3DistCp 用于复制操作，这在临时文件存储空间方面存在某些限制。

使用控制台备份和还原 HBase

借助控制台，您可以启动新集群并向其中填充先前 HBase 备份中的数据。您还可以借助控制台安排 HBase 数据的定期增量备份。使用 CLI 时还可以执行其他备份和还原功能，如将数据还原到已在运行的集群、执行手动备份和计划自动完整备份。

通过控制台使用已存档 HBase 数据填充新的集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 选择创建集群。
3. 在 Software Configuration (软件配置) 部分中，为 Additional Applications (其他应用程序) 选择 HBase 和 Configure and add (配置并添加)。
4. 在 Add Application (添加应用程序) 对话框中，选中 Restore From Backup (从备份中恢复)。
5. 对于 Backup Location (备份位置)，指定要加载到新 HBase 集群中的备份的位置。该位置应是 `s3://myawsbucket/backups/` 格式的 Amazon S3 URL。
6. 对于 Backup Version (备份版本)，您可以选择通过设置值来指定要加载的备份版本的名称。如果没有设置 Backup Version (备份版本) 的值，Amazon EMR 会加载指定位置中的最新备份。
7. 选择 Add (添加) 并继续使用所需的其他选项创建集群。

使用控制台计划 HBase 数据的自动备份

1. 在 Software Configuration (软件配置) 部分中，为 Additional Applications (其他应用程序) 选择 HBase 和 Configure and add (配置并添加)。
2. 选择 Schedule Regular Backups (安排定期备份)。
3. 请指定备份是否应该保持一致。一致性备份指的是，在初始备份阶段为保持节点同步而暂停写入操作的备份。将任何因此而暂停的写入操作放入队列中，然后在同步完成时恢复操作。
4. 通过在 Backup Frequency (备份频率) 中输入一个数字并选择 Days (天)、Hours (小时数) 或 Minutes (分钟数)来设置备份频率。首先运行的自动备份将是完整备份；在此之后，Amazon EMR 将根据您指定的日程安排保存增量备份。
5. 指定 Amazon S3 中应当存储备份的位置。每个 HBase 集群都应当备份至 Amazon S3 中的单独位置，以确保正确地计算增量备份。
6. 通过为 Backup Start Time (备份启动时间) 设置一个值，指定应当何时进行第一次备份。您可以将此值设置为 now，使第一次备份在集群运行时立即开始，也可以输入 ISO 格式的日期和时间。例如，2013-09-26T20:00Z 会将开始时间设置为 2013 年 9 月 26 日下午 8 点 (UTC)。
7. 选择 Add（添加）。
8. 根据需要继续使用其他选项创建集群。

使用 CloudWatch 监控 HBase

Amazon EMR 向 CloudWatch 报告三个指标，您可以用这些指标监控您的 HBase 备份。这些指标会免费推送到 CloudWatch，每五分钟一次。

指标	描述
HBaseBackupFailed	最后一次备份是否失败。默认设置为 0，如果上一次备份尝试失败，则更新为 1。仅为 HBase 集群报告此指标。 使用案例：监控 HBase 备份 单位：计数
HBaseMostRecentBackupDuration	完成上一次备份所需的时长。无论最后完成的备份成功或失败，都会设置此指标。进行备份的同时，此指标返回备份开始之后的分钟数。仅为 HBase 集群报告此指标。 使用案例：监控 HBase 备份 单位：分钟
HBaseTimeSinceLastSuccessfulBackup	在您的集群上，最后一次成功 HBase 备份开始之后经过的分钟数。仅为 HBase 集群报告此指标。 使用案例：监控 HBase 备份 单位：分钟

为 HBase 配置 Ganglia

使用 configure-hbase-for-ganglia 引导操作为 HBase 配置 Ganglia。此引导操作可配置 HBase 以向 Ganglia 发布指标。

在必须在启动集群时配置 HBase 和 Ganglia；无法将 Ganglia 报告添加到已在运行的集群。

Ganglia 还可以将日志文件存储在 /mnt/var/log/ganglia/rrds 处的服务器上。如果您配置了集群以将日志文件保存到 Amazon S3 存储桶，Ganglia 日志文件也会保存在那里。

要通过 Ganglia 为 HBase 启动集群，请使用 configure-hbase-for-ganglia 引导操作，如以下示例中所示。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Test cluster" --ami-version 3.3 \
--applications Name=Hue Name=Hive Name=Pig Name=HBase Name=Ganglia \
--use-default-roles --ec2-attributes KeyName=myKey \
--instance-type c1.xlarge --instance-count 3 --termination-protected \
--bootstrap-actions Path=s3://elasticmapreduce/bootstrap-actions/configure-hbase-for-ganglia
```

在配置了 Ganglia 的情况下启动集群后，您就可以使用主节点上运行的图形界面来访问 Ganglia 图形和报告。

Amazon EMR 的早期 AMI 版本的 Pig 应用程序细节

支持的 Pig 版本

您可以添加到集群的 Pig 版本取决于您所使用的 Amazon EMR AMI 的版本和 Hadoop 的版本。下表显示的是哪些 AMI 和 Hadoop 版本与哪些 Pig 版本兼容。我们建议使用最新版本的 Pig，以便利用各种性能增强和新的功能。

当您使用 API 安装 Pig 时，会使用默认版本，除非在通过调用 `--pig-versions`RunJobFlow 将 Pig 加载到群集的步骤中，将 指定为参数。

Pig 版本	AMI 版本	配置参数	Pig 版本详细信息

0.12.0
 	
发布说明
 	
文档
	3.1.0 和更高版本	`--ami-version 3.1`
		
		`--ami-version 3.2`
		
		`--ami-version 3.3`
			添加对以下各项的支持：
			
			• 无 JVM 实施对 UDF 进行流式处理
			• ASSERT 和 IN 运算符
			• CASE 表达式

Pig 版本	AMI 版本	配置参数	Pig 版本详细信息
			• AvroStorage 作为 Pig 内置函数。 • ParquetLoader 和 ParquetStorer 作为内置函数 • BigInteger 和 BigDecimal 类型
0.11.1.1 发布说明 文档	2.2 和更高版本	`--pig-versions 0.11.1.1` `--ami-version 2.2`	改进了在输入位于 Amazon S3 中的情况下，LOAD 命令针对 PigStorage 的性能。
0.11.1 发布说明 文档	2.2 和更高版本	`--pig-versions 0.11.1` `--ami-version 2.2`	增加了对 JDK 7、Hadoop 2、Groovy 用户定义的函数、SchemaTuple 优化、新运算符等支持。有关更多信息，请参阅 Pig 0.11.1 更改日志。
0.9.2.2 发布说明 文档	2.2 和更高版本	`--pig-versions 0.9.2.2` `--ami-version 2.2`	添加了对于 Hadoop 1.0.3 的支持。
0.9.2.1 发布说明 文档	2.2 和更高版本	`--pig-versions 0.9.2.1` `--ami-version 2.2`	增加了对 MapR 的支持。

Pig 版本	AMI 版本	配置参数	Pig 版本详细信息
0.9.2 发布说明 文档	2.2 和更高版本	`--pig-versions 0.9.2` `--ami-version 2.2`	包括多项性能改进和错误修复。有关 Pig 0.9.2 的更改的完整信息，请转到 Pig 0.9.2 更改日志。
0.9.1 发布说明 文档	2.0	`--pig-versions 0.9.1` `--ami-version 2.0`	
0.6 发布说明	1.0	`--pig-versions 0.6` `--ami-version 1.0`	
0.3 发布说明	1.0	`--pig-versions 0.3` `--ami-version 1.0`	

Pig 版本详细信息

Amazon EMR 支持可能应用了其他 Amazon EMR 补丁的某些 Pig 版本。您可以配置要在 Amazon EMR 集群上运行的 Pig 版本。有关此操作的详细信息，请参阅 Apache Pig。以下部分介绍了各种 Pig 版本以及应用到 Amazon EMR 上所加载版本的修补程序。

Pig 修补

本节介绍应用到 Amazon EMR 可用的 Pig 版本的自定义修补程序。

Pig 0.11.1.1 修补

Amazon EMR 版本的 Pig 0.11.1.1 是一个维护版，改进了在输入位于 Amazon S3 中的情况下 LOAD 命令针对 PigStorage 的性能。

Pig 0.11.1 补丁

Amazon EMR 版本的 Pig 0.11.1 包含 Apache Software Foundation 提供的所有更新以及从 Pig 0.9.2.2 版本开始累积的 Amazon EMR 补丁。但是，Pig 0.11.1 中没有特定于 Amazon EMR 的新补丁。

Pig 0.9.2 补丁

Apache Pig 0.9.2 是 Pig 的维护版。Amazon EMR 团队已将以下补丁应用到 Amazon EMR 版本的 Pig 0.9.2。

修补	描述
PIG-1429	将布尔数据类型以第一个类数据类型的形式添加到 Pig。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-1429。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.10
PIG-1824	支持 Jython UDF 中的导入模块。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-1824。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.10
PIG-2010	在分布式缓存上捆绑已注册的 JAR。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-2010。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.11
PIG-2456	添加 ~/.pigbootup 文件，用户可以在其中指定默认的 Pig 语句。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-2456。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.11

修补	描述
PIG-2623	支持使用 Amazon S3 路径注册 UDF。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-2623。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.10、0.11

Pig 0.9.1 补丁

Amazon EMR 团队已将以下补丁应用到 Amazon EMR 版本的 Pig 0.9.1。

修补	描述
支持 dfs 中的 JAR 文件和 Pig 脚本	添加对于 HDFS、Amazon S3 或者其他分布式文件系统中存储的运行脚本和已注册 JAR 文件的支持。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-1505。 状态：Committed (已提交) 已在以下 Apache Pig 版本中修复：0.8.0
支持 Pig 中的多个文件系统	添加对于 Pig 脚本的支持，以便从一个文件系统中读取数据，并写入另一个文件系统。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-1564。 状态：Not Committed (未提交) 已在以下 Apache Pig 版本中修复：不适用
添加 Piggybank 日期/时间和字符串 UDF	添加日期/时间和字符串 UDF，以支持自定义 Pig 脚本。有关更多信息，请转到 https://issues.apache.org/jira/browse/PIG-1565。 状态：Not Committed (未提交) 已在以下 Apache Pig 版本中修复：不适用

交互式的和批处理的 Pig 集群

Amazon EMR 可让您以两种模式运行 Pig 脚本：

- 交互式
- Batch

当您使用控制台或 Amazon CLI 启动长时间运行的集群时，您可以使用 ssh 以 Hadoop 用户身份连接到主节点，并使用 Grunt shell 以交互方式开发并运行您的 Pig 脚本。以交互方式使用 Pig，您就可以比批处理方式更轻松地修改 Pig 脚本。以交互模式成功修改 Pig 脚本之后，可以将脚本上传到 Amazon S3，并使用批处理模式在生产环境中运行该脚本。您还可以根据需要，在正在运行的集群中以交互方式提交 Pig 命令来分析和转换数据。

在批处理模式下，需要将 Pig 脚本上传到 Amazon S3，然后作为操作步骤将此工作提交到集群。Pig 步骤可提交到长时间运行的集群或临时集群。

Amazon EMR 的早期 AMI 版本的 Spark 应用程序细节

以交互方式或批处理模式使用 Spark

Amazon EMR 可让您以两种模式运行 Spark 应用程序：

- 交互式
- Batch

当您使用控制台或 Amazon CLI 启动长时间运行的集群时，可以使用 SSH 以 Hadoop 用户身份连接到主节点，并使用 Spark Shell 以交互方式开发并运行 Spark 应用程序。与批处理环境相比，以交互方式使用 Spark 能够让您更轻松地对 Spark 应用程序进行原型设计或测试。在交互模式下成功修改 Spark 应用程序后，您可以将该应用程序 JAR 或 Python 程序放到 Amazon S3 上集群主节点的本地文件系统上。然后，您可以将应用程序作为批处理工作流程提交。

在批处理模式中，将 Spark 脚本上传到 Amazon S3 或本地主节点文件系统，然后将此工作作为步骤提交到集群。Spark 步骤可提交到长时间运行的集群或暂时性集群。

创建安装了 Spark 的集群

使用控制台启动安装了 Spark 的集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 选择创建集群。
3. 对于 Software Configuration (软件配置)，请选择您需要的 AMI 发布版。
4. 对于 Applications to be installed (要安装的应用程序)，从列表中选择 Spark，然后选择 Configure and add (配置并添加)。
5. 添加参数以按需更改 Spark 配置。有关更多信息，请参阅配置 Spark。选择 Add（添加）。
6. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

以下示例演示如何使用 Java 创建带 Spark 的集群：

```
AmazonElasticMapReduceClient emr = new AmazonElasticMapReduceClient(credentials);
SupportedProductConfig sparkConfig = new SupportedProductConfig()
   .withName("Spark");

RunJobFlowRequest request = new RunJobFlowRequest()
   .withName("Spark Cluster")
   .withAmiVersion("3.11.0")
   .withNewSupportedProducts(sparkConfig)
   .withInstances(new JobFlowInstancesConfig()
    .withEc2KeyName("myKeyName")
    .withInstanceCount(1)
    .withKeepJobFlowAliveWhenNoSteps(true)
    .withMasterInstanceType("m3.xlarge")
    .withSlaveInstanceType("m3.xlarge")
   );
RunJobFlowResult result = emr.runJobFlow(request);
```

配置 Spark

您在创建集群时通过运行位于 Github 上的 awslabs/emr-bootstrap-actions/spark 存储库的引导操作配置 Spark。有关引导操作接受的参数，请参阅存储库中的 README。引导操作配置 `$SPARK_CONF_DIR/spark-defaults.conf` 文件中的属性。有关设置的更多信息，请参阅 Spark 文档中的 Spark 配置主题。您可以将以下 URL 中的“latest”替换为您要安装的 Spark 的版本号，例如，`2.2.0` http://spark.apache.org/docs/latest/configuration.html。

您也可以在每次提交应用程序时动态配置 Spark。使用 spark 配置文件提供了便于执行程序自动充分利用资源分配的设置。有关更多信息，请参阅覆盖 Spark 默认配置设置。

更改 Spark 默认设置

以下示例演示如何使用 Amazon CLI 创建 spark.executor.memory 设置为 2G 的集群。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --name "Spark cluster" --ami-version 3.11.0 \
--applications Name=Spark, Args=[-d,spark.executor.memory=2G] --ec2-attributes
 KeyName=myKey \
--instance-type m3.xlarge --instance-count 3 --use-default-roles
```

向 Spark 提交工作

要向集群提交工作，请使用步骤在 EMR 集群上运行 spark-submit 脚本。使用 addJobFlowStepsAmazonElasticMapReduceClient 中的 方法添加此步骤：

```
AWSCredentials credentials = new BasicAWSCredentials(accessKey, secretKey);
AmazonElasticMapReduceClient emr = new AmazonElasticMapReduceClient(credentials);
StepFactory stepFactory = new StepFactory();
AddJobFlowStepsRequest req = new AddJobFlowStepsRequest();
req.withJobFlowId("j-1K48XXXXXXHCB");

List<StepConfig> stepConfigs = new ArrayList<StepConfig>();

StepConfig sparkStep = new StepConfig()
 .withName("Spark Step")
 .withActionOnFailure("CONTINUE")
 .withHadoopJarStep(stepFactory.newScriptRunnerStep("/home/hadoop/spark/bin/spark-
submit","--class","org.apache.spark.examples.SparkPi","/home/hadoop/spark/lib/spark-
examples-1.3.1-hadoop2.4.0.jar","10"));

stepConfigs.add(sparkStep);
req.withSteps(stepConfigs);
AddJobFlowStepsResult result = emr.addJobFlowSteps(req);
```

覆盖 Spark 默认配置设置

建议您为不同的应用程序覆盖 Spark 默认配置值。您可以在提交应用程序时使用步骤完成此操作 (实质上是向 spark-submit 传递选项)。例如，您可能需要通过更改 spark.executor.memory 来更改为执行者进程分配的内存。您可以为 --executor-memory 开关提供与下类似的参数：

```
/home/hadoop/spark/bin/spark-submit --executor-memory 1g --class
 org.apache.spark.examples.SparkPi /home/hadoop/spark/lib/spark-examples*.jar 10
```

同样地，您也可以调节 --executor-cores 和 --driver-memory。在步骤中，您可以向步骤提供以下参数：

```
--executor-memory 1g --class org.apache.spark.examples.SparkPi /home/hadoop/spark/lib/
spark-examples*.jar 10
```

您还可以使用 --conf 选项调节没有内置开关的设置。有关可调节的其他设置的更多信息，请参阅 Apache Spark 文档中的动态加载 Spark 属性主题。

Amazon EMR 的早期 AMI 版本的 S3DistCp 实用程序差异

Amazon EMR 支持的 S3DistCp 版本

Amazon EMR AMI 发行版中支持以下 S3DistCp 版本。可直接在集群上找到 1.0.7 之后的 S3DistCp 版本。使用 /home/hadoop/lib 中的 JAR 以获得最新功能。

版本	描述	发行日期
1.0.8	添加 --appendToLastFile 、--requirePreviousManifest 和 --storageClass 选项。	2014 年 1 月 3 日
1.0.7	添加了 --s3ServerSideEncryption 选项。	2013 年 5 月 2 日
1.0.6	添加了 --s3Endpoint 选项。	2012 年 8 月 6 日
1.0.5	提高了指定要运行哪个 S3DistCp 版本的能力。	2012 年 6 月 27 日

版本	描述	发行日期
1.0.4	改进了 `--deleteOnSuccess` 选项。	2012 年 6 月 19 日
1.0.3	添加了对于 `--numberFiles` 和 `--startingIndex` 选项的支持。	2012 年 6 月 12 日
1.0.2	使用组时改进了文件命名。	2012 年 6 月 6 日
1.0.1	S3DistCp 的初始版本。	2012 年 1 月 19 日

向集群添加 S3DistCp 复制步骤

要向正在运行的集群添加 S3DistCp 复制步骤，请键入以下命令，将 *j-3GYXXXXXX9IOK* 替换为集群 ID，并将 *mybucket* 替换为 Amazon S3 存储桶名称。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr add-steps --cluster-id j-3GYXXXXXX9IOK \
--steps Type=CUSTOM_JAR,Name="S3DistCp step",Jar=/home/hadoop/lib/emr-s3distcp-1.0.jar,
\
Args=["--s3Endpoint,s3-eu-west-1.amazonaws.com",\
"--src,s3://mybucket/logs/j-3GYXXXXXX9IOJ/node/",\
"--dest,hdfs:///output",\
"--srcPattern,.*[a-zA-Z,]+"]
```

Example 将 Amazon CloudFront 日志加载到 HDFS

此示例通过向正在运行的集群添加步骤将 Amazon CloudFront 日志加载到 HDFS 中。在此过程中，压缩格式由 Gzip（CloudFront 默认格式）更改为 LZO。这很有用，因为使用 LZO 压缩的数据在解压缩时能拆分成多个映射，所以，与 Gzip 格式不同，您不必等到压缩完成。当您使用 Amazon EMR 分析数据时，这可以提供更好的性能。此示例还通过以下方式提高性能：使用在 `--groupBy` 选项中指定的

正则表达式，将给定小时内的所有日志组合成为单个文件。Amazon EMR 集群处理几个大型 LZO 压缩文件的效率比处理许多小型 Gzip 压缩文件的效率更高。要拆分 LZO 文件，您必须为这些文件编制索引并使用 hadoop-lzo 第三方库。

要将 Amazon CloudFront 日志加载到 HDFS 中，请键入以下命令，将 *j-3GYXXXXXX9IOK* 替换为集群 ID，并将 *mybucket* 替换为 Amazon S3 存储桶名称。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr add-steps --cluster-id j-3GYXXXXXX9IOK \
--steps Type=CUSTOM_JAR,Name="S3DistCp step",Jar=/home/hadoop/lib/emr-s3distcp-1.0.jar,
\
Args=["--src,s3://mybucket/cf","--dest,hdfs:///local",\
"--groupBy,.*XABCD12345678.([0-9]+-[0-9]+-[0-9]+-[0-9]+).*",\
"--targetSize,128",
"--outputCodec,lzo","--deleteOnSuccess"]
```

考虑上述示例在以下 CloudFront 日志文件上运行的情况。

```
s3://DOC-EXAMPLE-BUCKET1/cf/XABCD12345678.2012-02-23-01.HLUS3JKx.gz
s3://DOC-EXAMPLE-BUCKET1/cf/XABCD12345678.2012-02-23-01.I9CNAZrg.gz
s3://DOC-EXAMPLE-BUCKET1/cf/XABCD12345678.2012-02-23-02.YRRwERSA.gz
s3://DOC-EXAMPLE-BUCKET1/cf/XABCD12345678.2012-02-23-02.dshVLXFE.gz
s3://DOC-EXAMPLE-BUCKET1/cf/XABCD12345678.2012-02-23-02.LpLfuShd.gz
```

S3DistCp 复制、连接和压缩文件成为以下两份文件，其中，文件名由正则表达式的匹配来确定。

```
hdfs:///local/2012-02-23-01.lzo
hdfs:///local/2012-02-23-02.lzo
```

新增功能

本页介绍了 Amazon EMR 6.x 和 Amazon EMR 5.x 最新发行版中的更改和可用功能。Amazon EMR 发行版 6.14.0 页面和 Amazon EMR 发行版 5.36.1 页面上也提供了这些发布说明，以及应用程序版本、组件版本和每个发行版的可用配置分类。

订阅 RSS 源，通过 https://docs.amazonaws.cn/emr/latest/ReleaseGuide/amazon-emr-release-notes.rss 获取 Amazon EMR 发布说明，以便在新的 Amazon EMR 发行版可用时接收更新。

有关早期发行版的发布说明，请参阅 发布说明的 Amazon EMR 存档。

> **Note**
>
> Amazon EMR 发行版现在使用 Amazon 签名版本 4（SigV4）对发送到 Amazon S3 的请求进行身份验证。我们建议您使用支持 SigV4 的 Amazon EMR 发行版，这样您可以访问新的 S3 存储桶，避免工作负载中断。有关更多信息和支持 SigV4 的 Amazon EMR 发行版列表，请参阅 Amazon EMR 和 Amazon 签名版本 4。

缓解 CVE-2021-44228 的方法

> **Note**
>
> 对于 Amazon EMR 发行版 6.9.0 及更高版本，Amazon EMR 安装的所有使用 Log4j 库的组件都使用 Log4j 版本 2.17.1 或更高版本。

在 EC2 上运行的 Amazon EMR

CVE-2021-44228 中讨论的问题在处理来自不可信来源的输入时，与 2.0.0 到 2.14.1 之间的 Apache Log4j 核心版本相关。随 Amazon EMR 5.x 发行版（最高 5.34.0）和 EMR 6.x 发行版（最高 Amazon EMR 6.5.0）一起启动的 Amazon EMR 集群包括开源框架，例如 Apache Hive、Flink、HUDI、Presto 和 Trino，均使用这些版本的 Apache Log4j。但是，有许多客户使用安装在其 Amazon EMR 集群上的开源框架来处理和记录来自不可信来源的输入。

我们建议您按照下节所述，应用“适用于 Log4j CVE-2021-44228 的 Amazon EMR 引导操作解决方案”。此解决方案还解决了 CVE-2021-45046 问题。

> **Note**
>
> Amazon EMR 的引导操作脚本已于 2022 年 9 月 7 日更新，包括对 Oozie 的增量错误修复和改进。如果您使用 Oozie，则按照下节所述，应用更新后的 Amazon EMR 引导操作解决方案。

Amazon EMR on EKS

如果您使用默认配置的 Amazon EMR on EKS，则不会受到 CVE-2021-44228 中所述问题的影响，也不必应用 适用于 Log4j CVE-2021-44228 和 CVE-2021-45046 的 Amazon EMR 引导操作解决方案 部分中所述的解决方案。对于 Amazon EMR on EKS，适用于 Spark 的 Amazon EMR 运行时使用 Apache Log4j 版本 1.2.17。使用 Amazon EMR on EKS 时，不得将 `log4j.appender` 组件的默认设置更改为 `log`。

适用于 Log4j CVE-2021-44228 和 CVE-2021-45046 的 Amazon EMR 引导操作解决方案

此解决方案提供了必须在 Amazon EMR 集群上应用的 Amazon EMR 引导操作。对于每个 Amazon EMR 发行版，您都会在下面找到一个指向引导操作脚本的链接。要应用此引导操作，您应完成以下步骤：

1. 将与 Amazon EMR 发行版对应的脚本复制到 Amazon Web Services 账户 中的本地 S3 存储桶。请确保您使用的是 Amazon EMR 发行版特定的引导脚本。
2. 为 EMR 集群设置引导操作，以按照 EMR 文档中所述的说明运行复制到 S3 存储桶的脚本。如果您为 EMR 集群配置了其他引导操作，请确保将此脚本设置为要执行的第一个引导操作脚本。
3. 终止现有的 EMR 集群，然后使用引导操作脚本启动新集群。Amazon 建议您在测试环境中测试引导脚本并验证应用程序，然后再将其应用程序应用于生产环境。如果您没有为 EMR 次要版本（例如 6.3.0）使用最新版本，则必须使用最新版本（例如 6.3.1），然后应用上面讨论的解决方案。

CVE-2021-44228 和 CVE-2021-45046 - EMR 版本的引导脚本

Amazon EMR 发行版	脚本位置	脚本发布日期
6.5.0	`s3://elasticmapreduce/bootstrap-actions/`	2022 年 3 月 24 日

Amazon EMR 发行版	脚本位置	脚本发布日期
	log4j/patch-log4j-emr-6.5.0-v2.sh	
6.4.0	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-6.4.0-v2.sh	2022 年 3 月 24 日
6.3.1	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-6.3.1-v2.sh	2022 年 3 月 24 日
6.2.1	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-6.2.1-v2.sh	2022 年 3 月 24 日
6.1.1	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-6.1.1-v2.sh	2021 年 12 月 14 日
6.0.1	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-6.0.1-v2.sh	2021 年 12 月 14 日
5.34.0	s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.34.0-v2.sh	2021 年 12 月 12 日

Amazon EMR 发行版	脚本位置	脚本发布日期
5.33.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.33.1-v2.sh`	2021 年 12 月 12 日
5.32.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.32.1-v2.sh`	2021 年 12 月 13 日
5.31.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.31.1-v2.sh`	2021 年 12 月 13 日
5.30.2	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.30.2-v2.sh`	2021 年 12 月 14 日
5.29.0	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.29.0-v2.sh`	2021 年 12 月 14 日
5.28.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.28.1-v2.sh`	2021 年 12 月 15 日
5.27.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.27.1-v2.sh`	2021 年 12 月 15 日

Amazon EMR 发行版	脚本位置	脚本发布日期
5.26.0	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.26.0-v2.sh`	2021 年 12 月 15 日
5.25.0	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.25.0-v2.sh`	2021 年 12 月 15 日
5.24.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.24.1-v2.sh`	2021 年 12 月 15 日
5.23.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.23.1-v2.sh`	2021 年 12 月 15 日
5.22.0	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.22.0-v2.sh`	2021 年 12 月 15 日
5.21.2	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.21.2-v2.sh`	2021 年 12 月 15 日
5.20.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.20.1-v2.sh`	2021 年 12 月 15 日

Amazon EMR 发行版	脚本位置	脚本发布日期
5.19.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.19.1-v2.sh`	2021 年 12 月 15 日
5.18.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.18.1-v2.sh`	2021 年 12 月 15 日
5.17.2	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.17.2-v2.sh`	2021 年 12 月 15 日
5.16.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.16.1-v2.sh`	2021 年 12 月 15 日
5.15.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.15.1-v2.sh`	2021 年 12 月 15 日
5.14.2	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.14.2-v2.sh`	2021 年 12 月 15 日
5.13.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.13.1-v2.sh`	2021 年 12 月 15 日

Amazon EMR 发行版	脚本位置	脚本发布日期
5.12.3	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.12.3-v2.sh`	2021 年 12 月 15 日
5.11.4	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.11.4-v2.sh`	2021 年 12 月 15 日
5.10.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.10.1-v2.sh`	2021 年 12 月 15 日
5.9.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.9.1-v2.sh`	2021 年 12 月 15 日
5.8.3	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.8.3-v2.sh`	2021 年 12 月 15 日
5.7.1	`s3://elasticmapreduce/bootstrap-actions/log4j/patch-log4j-emr-5.7.1-v2.sh`	2021 年 12 月 15 日

EMR 发布版本	截至 2021 年 12 月的最新修订
6.3.0	6.3.1

EMR 发布版本	截至 2021 年 12 月的最新修订
6.2.0	6.2.1
6.1.0	6.1.1
6.0.0	6.0.1
5.33.0	5.33.1
5.32.0	5.32.1
5.31.0	5.31.1
5.30.0 或 5.30.1	5.30.2
5.28.0	5.28.1
5.27.0	5.27.1
5.24.0	5.24.1
5.23.0	5.23.1
5.21.0 或 5.21.1	5.21.2
5.20.0	5.20.1
5.19.0	5.19.1
5.18.0	5.18.1
5.17.0 或 5.17.1	5.17.2
5.16.0	5.16.1
5.15.0	5.15.1
5.14.0 或 5.14.1	5.14.2
5.13.0	5.13.1

EMR 发布版本	截至 2021 年 12 月的最新修订
5.12.0、5.12.1、5.12.2	5.12.3
5.11.0、5.11.1、5.11.2、5.11.3	5.11.4
5.9.0	5.9.1
5.8.0、5.8.1、5.8.2	5.8.3
5.7.0	5.7.1

常见问题

- EMR 5 之前的 EMR 版本是否受到 CVE-2021-44228 的影响？

  否。EMR 版本 5 之前的 EMR 版本使用 2.0 之前的 Log4j 版本。

- 此解决方案是否解决了 CVE-2021-45046 的问题？

  此解决方案还解决了 CVE-2021-45046 问题。

- 该解决方案是否处理了我在 EMR 集群上安装的自定义应用程序？

  引导脚本仅更新 EMR 安装的 JAR 文件。如果您通过使用自定义 Amazon Linux AMI 或通过任何其他机制，通过引导操作在您的 EMR 集群上安装自定义应用程序和 JAR 文件并将其作为提交至集群的步骤运行，请与您的应用程序提供商合作，以确定您的自定义应用程序是否受 CVE-2021- 44228 影响，并确定合适的解决方案。

- 我应该如何使用 EKS 上的 EMR 处理自定义 Docker 映像？

  如果您使用自定义 Docker 映像将自定义应用程序添加到 EKS 上的 Amazon EMR 或者使用自定义应用程序文件将任务提交到 EKS 上的 Amazon EMR，请与应用程序供应商合作，以确定您的自定义应用程序是否受到 CVE-2021-44228 的影响，并确定合适的解决方案。

- 引导脚本如何缓解 CVE-2021-44228 和 CVE-2021-45046 中描述的问题？

  引导脚本通过添加一组新的指令来更新 EMR 启动指令。这些新指令删除了 EMR 安装的所有开源框架通过 Log4j 使用的 JndiLookup 类文件。这遵循了 Apache 发布的用于解决 Log4j 问题的建议。

- 使用 Log4j 版本 2.17.1 或更高版本的 EMR 是否有更新？

不超过版本 5.34 的 EMR 5 发行版以及不超过版本 6.5 的 EMR 6 发行版使用不兼容最新版本 Log4j 的较旧版本开源框架。如果您继续使用这些发行版，我们建议您应用引导操作来缓解 CVE 中讨论的问题。在 EMR 5 发行版 5.34 和 EMR 6 发行版 6.5 以后，使用 log4J 1.x 和 log4J 2.x 的应用程序将分别升级为使用 log4J 1.2.17（或更高版本）和 log4J 2.17.1（或更高版本），并且不需要使用上面介绍的引导操作来缓解 CVE 问题。

- EMR 版本是否受到 CVE-2021-45105 的影响？

  由 Amazon EMR 安装的具有 EMR 默认配置的应用程序不受 CVE-2021-45105 的影响。在 Amazon EMR 安装的应用程序中，只有 Apache Hive 将 Apache Log4j 与上下文查找结合使用，而且它不会以允许处理不适当的输入数据的方式使用非默认模式布局。

- Amazon EMR 是否受以下任何 CVE 披露的影响？

  下表包含了与 Log4J 相关的 CVE 列表，并说明了每个 CVE 是否影响 Amazon EMR。此表中的信息仅适用于 Amazon EMR 使用原定设置配置安装应用程序的情形。

CVE	对 EMR 的影响	注意
CVE-2022-23302	否	Amazon EMR 不会设置 Log4j JMSSink
CVE-2022-23305	否	Amazon EMR 不会设置 Log4j JDBCAppender
CVE-2022-23307	否	Amazon EMR 不会设置 Log4j Chainsaw
CVE-2020-9493	否	Amazon EMR 不会设置 Log4j Chainsaw
CVE-2021-44832	否	Amazon EMR 不会使用 JNDI 连接字符串设置 Log4j JDBCAppender
CVE-2021-4104	否	Amazon EMR 不会使用 Log4j JMSAppender

CVE	对 EMR 的影响	注意
CVE-2020-9488	否	Amazon EMR 安装的应用程序不会使用 Log4j SMTPAppender
CVE-2019-17571	否	Amazon EMR 会阻止对集群的公有访问权限且不会启动 SocketServer
CVE-2019-17531	否	建议您升级到最新的 Amazon EMR 版本。Amazon EMR 5.33.0 及更高版本使用 jackson-databind 2.6.7.4 或更高版本，EMR 6.1.0 及更高版本使用 jackson-databind 2.10.0 或更高版本。这些版本的 jackson-databind 不受此 CVE 的影响。

Amazon EMR 6.14.0（6.x 系列的最新版本）

从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

以下发布说明包括有关 Amazon EMR 发行版 6.14.0 的信息。更改与 6.13.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.14.0 supports Apache Spark 3.4.1, Apache Spark RAPIDS 23.06.0-amzn-2, Flink 1.17.1, Iceberg 1.3.1, and Trino 422.
- Amazon EMR 托管式自动扩缩功能 现已在 `ap-southeast-3` 亚太地区（雅加达）区域开放，可用于您使用 Amazon EMR 6.14.0 及更高版本创建的集群。

更改、增强功能和解决的问题

- 6.14.0 发行版通过在 Amazon EC2 上运行的 Amazon EMR 来优化日志管理。因此，您可能会看到集群日志的存储成本略有降低。
- 6.14.0 发行版改进了扩展工作流，以满足 Amazon EBS 卷大小差异很大的不同核心实例需求。此改进仅适用于核心节点；任务节点的缩减操作不受影响。
- 6.14.0 发行版改进了 Amazon EMR 与 Apache Hadoop YARN ResourceManager and HDFS NameNode 等开源应用程序交互的方式。此改进降低了集群扩展导致操作延迟的风险，并减少了由于开源应用程序连接问题导致的启动故障。
- 6.14.0 发行版优化了集群启动时的应用程序安装。此改进缩短了某些 Amazon EMR 应用程序组合的集群启动时间。
- 6.14.0 发行版修复了在具有自定义域的 VPC 上运行的集群遇到核心节点或任务节点重启时，集群的缩减操作可能会停滞的问题。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 906.0	4.14.322	2023 年 9 月 11 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			买）、亚太地区（海得拉巴）、亚太地区（东京）、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

Amazon EMR 5.36.1（5.x 系列的最新版本）

从初始发布日期的第一个区域开始，新的 Amazon EMR 发行版将在几天内陆续在不同区域提供。在此期间，您所在区域可能无法提供最新发行版。

以下发布说明包括有关 Amazon EMR 版本 5.36.1 的信息。更改与 5.36.0 有关。有关发布时间表的信息，请参阅更改日志。

更改、增强功能和解决的问题

- Amazon EMR 版本 5.36.1 增加了对在集群缩减期间将日志存档到 Amazon S3 的支持。在之前的 5.x 版本中，您只能在集群终止期间将日志文件存档到 Amazon S3。这项改进可确保即使在节点终止后，集群上生成的日志文件仍保留在 Amazon S3 上。有关更多信息，请参阅配置集群日志记录和调试。
- 5.36.1 版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 5.36.1 版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。

- 5.36.1 版本修复了主节点上的 Amazon EMR 进程守护程序会维护集群中已终止实例的过时元数据的问题。维护陈旧的数据可能会导致集群上的 CPU 和内存使用量无限增长，并最终导致集群故障。
- 对于使用多个主节点启动的集群，5.36.1 版本修复了其中一个主节点上的 Amazon EC2 硬件故障可能导致第二个主节点出现故障并导致集群不稳定的问题。
- 对于配置了传输中加密的集群，托管扩展现在支持 Spark shuffle 数据感知。Spark shuffle 数据是 Spark 跨分区重新分配以执行特定操作的数据。在缩减期间，托管扩展会忽略带有随机数据的实例。这样可以防止任务的重新尝试和重新计算，这些都会给价格和性能带来高昂的代价。有关随机排序操作的更多信息，请参阅 Spark 编程指南。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

Amazon EMR 和 Amazon 签名版本 4

Amazon EMR 发行版现在使用 Amazon 签名版本 4（SigV4）对发送到 Amazon S3 的请求进行身份验证。2020 年 6 月 24 日之后在 Amazon S3 中创建的存储桶不支持由签名版本 2（SigV2）签名的请求。2020 年 6 月 24 日或之前创建的存储桶将继续支持 SigV2。建议您迁移到支持 SigV4 的 Amazon EMR 发行版，这样您可以访问新的 S3 存储桶，避免工作负载中断。

如果您使用的是 Amazon EMR 中包含的应用程序，例如 Apache Spark、Apache Hive 和 Presto，则无需更改应用程序代码即可使用 SigV4。如果您使用的是 Amazon EMR 中未包含的自定义应用程序，

则可能需要更新代码才能使用 SigV4。有关更多信息，请参阅《Amazon S3 用户指南》中的从签名版本 2 转向签名版本 4。

以下 Amazon EMR 发行版支持 SigV4：
emr-4.7.4、emr-4.8.5、emr-4.9.6、emr-4.10.1、emr-5.1.1、emr-5.2.3、emr-5.3.2、emr-5.4.1、emr-5.5.4、emr-5.21.2、and emr-5.22.0 及更高版本。

发布说明的 Amazon EMR 存档

下面提供了所有 Amazon EMR 发行版的发布说明。有关每个发行版的全面版本信息，请参阅 Amazon EMR 6.x 发行版、Amazon EMR 5.x 发行版 和 Amazon EMR 4.x 发行版。

订阅 RSS 源，通过 https://docs.amazonaws.cn/emr/latest/ReleaseGuide/amazon-emr-release-notes.rss 获取 Amazon EMR 发布说明，以便在新的 Amazon EMR 发行版可用时接收更新。

发行版 6.14.0

以下发布说明包括有关 Amazon EMR 发行版 6.14.0 的信息。更改与 6.13.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.14.0 supports Apache Spark 3.4.1, Apache Spark RAPIDS 23.06.0-amzn-2, Flink 1.17.1, Iceberg 1.3.1, and Trino 422.
- Amazon EMR 托管式自动扩缩功能 现已在 `ap-southeast-3` 亚太地区（雅加达）区域开放，可用于您使用 Amazon EMR 6.14.0 及更高版本创建的集群。

更改、增强功能和解决的问题

- 6.14.0 发行版通过在 Amazon EC2 上运行的 Amazon EMR 来优化日志管理。因此，您可能会看到集群日志的存储成本略有降低。
- 6.14.0 发行版改进了扩展工作流，以满足 Amazon EBS 卷大小差异很大的不同核心实例需求。此改进仅适用于核心节点；任务节点的缩减操作不受影响。
- 6.14.0 发行版改进了 Amazon EMR 与 Apache Hadoop YARN ResourceManager and HDFS NameNode 等开源应用程序交互的方式。此改进降低了集群扩展导致操作延迟的风险，并减少了由于开源应用程序连接问题导致的启动故障。
- 6.14.0 发行版优化了集群启动时的应用程序安装。此改进缩短了某些 Amazon EMR 应用程序组合的集群启动时间。

- 6.14.0 发行版修复了在具有自定义域的 VPC 上运行的集群遇到核心节点或任务节点重启时，集群的缩减操作可能会停滞的问题。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 906.0	4.14.322	2023 年 9 月 11 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			酋）、加拿大（中部）、以色列（特拉维夫）

发行版 6.13.0

以下发布说明包括有关 Amazon EMR 版本 6.13.0 的信息。更改与 6.12.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.13.0 supports Apache Spark 3.4.1, Apache Spark RAPIDS 23.06.0-amzn-1, CUDA Toolkit 11.8.0, and JupyterHub 1.5.0.

更改、增强功能和解决的问题

- 6.13.0 版本改进了 Amazon EMR 日志管理进程守护程序，以确保在发出集群终止命令时，所有日志都定期上传到 Amazon S3。这有助于更快地终止集群。
- 6.13.0 版本增强了 Amazon EMR 日志管理功能，确保所有日志文件一致而及时地上传到 Amazon S3。这尤其有利于长期运行的 EMR 集群。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 808.0	4.14.320	2023 年 8 月 24 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

版本 6.12.0

以下发布说明包括有关 Amazon EMR 版本 6.12.0 的信息。更改与 6.11.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

新特征

- Amazon EMR 6.12.0 supports Apache Spark 3.4.0, Apache Spark RAPIDS 23.06.0-amzn-0, CUDA 11.8.0, Apache Hudi 0.13.1-amzn-0, Apache Iceberg 1.3.0-amzn-0, Trino 414, and PrestoDB 0.281.
- Amazon EMR 发布 6.12.0 及更高版本支持 LDAP 通过 HiveServer2（HS2）、Trino、Presto 和 Hue 与 Apache Livy、Apache Live 集成。您还可以在使用 6.12.0 或更高版本的 EMR 集群上安装 Apache Spark 和 Apache Hadoop，并将它们配置为使用 LDAP。有关更多信息，请参阅使用 Active Directory 或 LDAP 服务器通过 Amazon EMR 进行身份验证。

更改、增强功能和解决的问题

- Amazon EMR 6.12.0 及更高版本为 Flink 提供 Java 11 运行时系统支持。有关更多信息，请参阅将 Flink 配置为使用 Java 11 运行。
- Amazon EMR 6.12.0 默认支持所有搭载 Amazon Corretto 8 的应用程序，但 Trino 除外。对于 Trino，Amazon EMR 从Amazon EMR 版本 6.9.0 开始默认支持 Amazon Corretto 17。Amazon EMR 还支持某些搭载 Amazon Corretto 11 和 17 的应用程序。下表列出了这些应用程序。如果要更改集群上的默认 JVM，请按照在集群上运行的每个应用程序的 配置应用程序来使用特定 Java 虚拟机 中的说明进行操作。一个集群只能使用一个 Java 运行时系统版本。Amazon EMR 不支持在同一集群的不同运行时系统版本上运行不同的节点或应用程序。

  虽然 Amazon EMR 在 Apache Spark、Apache Hadoop 和 Apache Hive 上同时支持 Amazon Corretto 11 和 17，但当您使用这些版本的 Corretto 时，某些工作负载的性能可能会下降。我们建议您在更改默认值之前先测试工作负载。

  Amazon EMR 6.12 中应用程序的默认 Java 版本

应用程序	Java/Amazon Corretto 版本（默认为粗体）
Delta	17、11、8
Flink	11、8
Ganglia	8
HBase	11、8
HCatalog	17、11、8

应用程序	Java/Amazon Corretto 版本（默认为粗体）
Hadoop	17、11、8
Hive	17、11、8
Hudi	17、11、8
Iceberg	17、11、8
Livy	17、11、8
Oozie	17、11、8
Phoenix	8
PrestoDB	8
Spark	17、11、8
Spark RAPIDS	17、11、8
Sqoop	8
Tez	17、11、8
Trino	17
Zeppelin	8
Pig	8
Zookeeper	8

- 6.12.0 版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 6.12.0 版本修复了一个问题，即当处于正常停用状态的核心节点在完全停用之前出于任何原因变得运行不正常时，集群的缩减操作可能会停滞不前。
- 6.12.0 版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。

- 6.12.0 版本通过提高记录实例状态变化的速度，提高了 Amazon EMR 运行状况监控服务的性能和效率。这一改进降低了运行多个自定义客户端工具或第三方应用程序的集群节点性能下降的机会。
- 6.12.0 版本提高了 Amazon EMR 的集群上日志管理进程守护程序的性能。因此，对于以高并发度运行步骤的 EMR 集群，性能下降的可能性较小。
- 在 Amazon EMR 6.12.0 版本中，日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 6.12.0 版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 6.12.0 版本支持 YARN Timeline Server 日志的日志轮换。这样可以最大限度地减少磁盘过度使用情况，特别是对于长时间运行的集群。
- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京）、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

版本 6.11.1

以下发布说明包括有关 Amazon EMR 版本 6.11.1 的信息。更改与 6.11.0 有关。有关发布时间表的更多信息，请参阅 更改日志。

更改、增强功能和解决的问题

- 由于锁争用，如果在尝试停用节点的同时添加或移除节点，则该节点可能会陷入死锁。结果，Hadoop 资源管理器（YARN）变得无响应，并会影响所有传入和当前正在运行的容器。
- 此版本包括一项更改，允许高可用性集群在重启后从故障状态中恢复。
- 此版本包含针对 Hue 和 HBase 的安全补丁。
- 此版本修复了在 Spark 上使用 Amazon EMR 运行工作负载的集群可能会静默收到包含 `contains`、`startsWith`、`endsWith` 和 `like` 错误结果的问题。当您在 Amazon EMR Hive3 Metastore 服务器（HMS）中使用包含元数据的分区字段的表达式时，就会出现此问题。
- 此版本修复了没有用户定义函数（UDF）时在 Glue 端的节流问题。
- 此版本修复了在 YARN 停用时，在日志推送器能够将容器日志推送到 S3 之前，节点日志聚合服务会删除容器日志的问题。
- 此版本修复了 Hadoop 启用节点标签时 FairShare 调度器指标的问题。
- 此版本修复了您在 `spark-defaults.conf` 中为 `spark.yarn.heterogeneousExecutors.enabled` 配置设置默认 `true` 值时影响 Spark 性能的问题。
- 此版本修复了 Reduce Task 无法读取随机数据的问题。该问题因内存损坏错误导致 Hive 查询失败。
- 此版本为运行 Presto 或 Trino 的 EMR 集群的集群扩展工作流程添加了新的重试机制。这一改进降低了由于单个调整大小操作失败而导致集群大小调整无限期停滞的风险。它还可以提高集群利用率，因为您的集群可以更快地向上和向下扩展。
- 此版本改进了集群缩减逻辑，因此您的集群不会尝试将核心节点缩减到低于集群 HDFS 复制因子设置的范围。这符合您的数据冗余要求，并减少了扩展操作可能停滞的机会。
- 日志管理进程守护程序已升级，可以识别本地实例存储中所有包含打开文件句柄的使用中的日志，以及相关的进程。此次升级可确保 Amazon EMR 在日志存档到 Amazon S3 后正确删除文件并回收存储空间。
- 此版本包括日志管理进程守护程序增强功能，可删除本地集群文件系统中空的、未使用的步骤目录。过多的空目录会降低 Amazon EMR 进程守护程序的性能并导致磁盘过度使用。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

版本 6.11.0

以下发布说明包括有关 Amazon EMR 版本 6.11.0 的信息。更改与 6.10.0 有关。有关发布时间表的信息，请参阅更改日志。

新特征

- Amazon EMR 6.11.0 支持 Apache Spark 3.3.2-amzn-0、Apache Spark RAPIDS 23.02.0-amzn-0、CUDA 11.8.0、Apache Hudi 0.13.0-amzn-0、Apache Iceberg 1.2.0-amzn-0、Trino 410-amzn-0 和 PrestoDB 0.279-amzn-0。

更改、增强功能和解决的问题

- 在 Amazon EMR 6.11.0 中，DynamoDB 连接器已升级到 5.0.0 版。5.0.0 版本使用 Amazon SDK for Java 2.x。之前的版本使用的是 Amazon SDK for Java 1.x。由于此次升级，我们强烈建议您在将 DynamoDB 连接器与 Amazon EMR 6.11 配合使用之前，先测试您的代码。
- 当 Amazon EMR 6.11.0 的 DynamoDB 连接器调用 DynamoDB 服务时，它会使用您为 `dynamodb.endpoint` 属性提供的区域值。我们建议您在使用 `dynamodb.endpoint` 时也配置 `dynamodb.region`，并且两个属性都以相同的 Amazon Web Services 区域 为目标。如果您使用 `dynamodb.endpoint` 但不配置 `dynamodb.region`，则适用于 Amazon EMR 6.11.0 的 DynamoDB 连接器将返回一个无效的区域异常，并尝试协调来自 Amazon EC2 实例元数据服务（IMDS）的 Amazon Web Services 区域 信息。如果连接器无法从 IMDS 检索区域，则默认为美国东部（弗吉尼亚州北部）(`us-east-1`)。以下错误是您未正确配置该 `dynamodb.region` 属性时可能会遇到的无效区域异常的示例：`error software.amazon.awssdk.services.dynamodb.model.DynamoDbException: Credential should be scoped to a valid region.` 有关受 Amazon SDK for Java 升级到 2.x 影响的类的更多信息，请参阅 Amazon EMR – DynamoDB 连接器的 GitHub 存储库中的 Upgrade Amazon SDK for Java from 1.x to 2.x (#175) 提交。
- 此版本修复了在执行列重命名操作后使用 Delta Lake 在 Amazon S3 中存储 Delta 表数据时列数据变为 NULL 的问题。有关 Delta Lake 中此实验性功能的更多信息，请参阅《Delta Lake User Guide》中的 Column rename operation。
- 6.11.0 版本修复了通过从具有多个主节点的集群中复制一个主节点来创建边缘节点时可能出现的问题。复制的边缘节点可能会导致缩减操作的延迟，或者导致主节点的内存使用率过高。有关如何创建边缘节点以与 EMR 集群通信的更多信息，请参阅 GitHub `aws-samples` 存储库中的 Edge Node Creator。
- 6.11.0 版本改进了 Amazon EMR 用于在重启后将 Amazon EBS 卷重新挂载到实例的自动化流程。
- 6.11.0 版本修复了导致 Amazon EMR 向 Amazon CloudWatch 发布的 Hadoop 指标间歇性出现差距的问题。
- 6.11.0 版本修复了 EMR 集群的一个问题，即由于磁盘过度使用而导致对包含集群节点排除列表的 YARN 配置文件的更新中断。不完整的更新阻碍了未来对集群的缩减操作。此版本可确保您的集群保持正常运行，并确保扩展操作按预期进行。

- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。
- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  在 Amazon EMR 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false` 以解决此问题。

  虽然该修复解决了 YARN-9608 引入的问题，但由于启用了托管扩展的集群上的随机数据丢失，它可能会导致 Hive 作业失败。在此版本中，我们还通过设置 Hive `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-shuffle-data` 工作负载来降低这种风险。此配置在 Amazon EMR 版本 6.11.0 及更高版本中提供。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

版本 6.10.0

以下发布说明包括有关 Amazon EMR 版本 6.10.0 的信息。更改与 6.9.0 有关。有关发布时间表的信息，请参阅更改日志。

新特征

- Amazon EMR 6.10.0 支持 Apache Spark 3.3.1、Apache Spark RAPIDS 22.12.0、CUDA 11.8.0、Apache Hudi 0.12.2-amzn-0、Apache Iceberg 1.1.0-amzn-0、Trino 403 和 PrestoDB 0.278.1。
- Amazon EMR 6.10.0 包含原生 Trino-Hudi 连接器，可提供对 Hudi 表中数据的读取权限。您可以使用 `trino-cli --catalog hudi` 激活连接器，并使用 `trino-connector-hudi` 配置连接器以满足您的要求。与 Amazon EMR 的原生集成意味着您不再需要使用 `trino-connector-hive` 来查询 Hudi 表。有关新连接器支持的配置列表，请参阅 Trino 文档的 Hudi connector 页面。
- Amazon EMR 版本 6.10.0 及更高版本支持 Apache Zeppelin 与 Apache Flink 集成。参阅 在 Amazon EMR 中通过 Zeppelin 使用 Flink 作业 了解更多信息。

已知问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  要在 Amazon EMR 6.10.0 中解决此问题，您可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强功能和解决的问题

- Amazon EMR 6.10.0 消除了对适用于 Apache Spark 的 Amazon Redshift 集成的 `minimal-json.jar` 依赖，并自动将所需的 Spark-Redshift 相关 jar 添加到 Spark 的执行程序类路径中：`spark-redshift.jar`、`spark-avro.jar` 和 `RedshiftJDBC.jar`。
- 6.10.0 版本改进了集群上日志管理进程守护程序，以监控 EMR 集群中的其他日志文件夹。这一改进最大限度地减少了磁盘过度使用情况。
- 6.10.0 版本在集群上日志管理进程守护程序停止后会自动重启该守护程序。这一改进降低了由于磁盘过度使用而导致节点出现运行状况不佳的风险。
- Amazon EMR 6.10.0 支持 EMRFS 用户映射的区域端点。
- Amazon EMR 6.10.0 及更高版本的默认根卷大小已增加到 15 GB。早期版本的默认根卷大小为 10 GB。

- 6.10.0 版本修复了当所有剩余的 Spark 执行程序都位于使用 YARN 资源管理器的停用主机上时，导致 Spark 作业停滞的问题。
- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（米兰）、欧洲（西班牙）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）

发行版 6.9.0

以下发布说明包括有关 Amazon EMR 发行版 6.9.0 的信息。更改与 Amazon EMR 发行版 6.8.0 有关。有关发布时间表的信息，请参阅更改日志。

新功能

- Amazon EMR 发行版 6.9.0 支持 Apache Spark RAPIDS 22.08.0、Apache Hudi 0.12.1、Apache Iceberg 0.14.1、Trino 398 和 Tez 0.10.2。
- Amazon EMR 发行版 6.9.0 包括一个新的开源应用程序，Delta Lake 2.1.0。
- Amazon EMR 发行版 6.9.0 及更高版本包含适用于 Apache Spark 的 Amazon Redshift 集成。本地集成之前是一种开源工具，现在是 Spark 连接器，您可以将其用于构建 Apache Spark 应用程序，这些应用程序可在 Amazon Redshift 和 Amazon Redshift Serverless 中读取和写入数据。有关更多信息，请参阅将适用于 Apache Spark 的 Amazon Redshift 集成与 Amazon EMR 结合使用。
- Amazon EMR 发行版 6.9.0 增加了对在集群缩减期间将日志存档到 Amazon S3 的支持。之前，您只能在集群终止期间将日志文件存档到 Amazon S3。这项新功能可确保即使在节点终止后，集群上生成的日志文件仍保留在 Amazon S3 上。有关更多信息，请参阅配置集群日志记录和调试。
- 为了支持长时间运行的查询，Trino 现在包括容错执行机制。容错执行通过重试失败的查询或其组件任务来减少查询失败。有关更多信息，请参阅Trino 中的容错执行。
- 您可以在 Amazon EMR 上使用 Apache Flink 对 Apache Hive 表或任何 Flink 表源（例如 Iceberg、Kinesis 或 Kafka）的元数据进行统一的 `BATCH` 和 `STREAM` 处理。您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 指定 Amazon Glue 数据目录作为 Flink 的元存储。有关更多信息，请参阅在 Amazon EMR 中配置 Flink。
- 现在，您可以在 EC2 集群上的 Amazon EMR 上使用 Amazon SageMaker Studio，为 Apache Spark、Apache Hive 和 Presto 查询指定 Amazon Identity and Access Management（IAM）运行时角色和基于 Amazon Lake Formation 的访问控制。有关更多信息，请参阅为 Amazon EMR 步骤配置运行时角色。

已知问题

- 对于 Amazon EMR 发行版 6.9.0，Trino 不适用于为 Apache Ranger 启用的集群。如果您需要将 Trino 与 Ranger 结合使用，请联系 Amazon Web Services Support。
- 如果您使用适用于 Apache Spark 的 Amazon Redshift 集成，并且具有 Parquet 格式的时间、timetz、时间戳或 timestamptz（精度为微秒），连接器会将时间值舍入为最接近的毫秒值。解决方法是使用文本卸载格式 `unload_s3_format` 参数。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。

- 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

- 从 Amazon SageMaker Studio 连接到 Amazon EMR 集群可能会间歇性失败，并显示 403 Forbidden（403 禁止访问）响应代码。如果在集群上设置 IAM 角色的时间超过 60 秒，就会发生此错误。解决方法是安装 Amazon EMR 补丁以启用重试，并将超时增加到至少 300 秒。启动集群时，按照以下步骤应用引导操作。

  1. 使用以下 Amazon S3 URI 下载引导脚本和 RPM 文件。

     ```
     s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/gcsc/replace-rpms.sh
     s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/gcsc/emr-secret-agent-1.18.0-SNAPSHOT20221121212949.noarch.rpm
     ```

  2. 将上一步中的文件上传到您自己的 Amazon S3 存储桶中。存储桶必须位于您计划启动集群的同一 Amazon Web Services 区域。
  3. 启动集群时，执行以下引导操作。将 *bootstrap_URI* 和 *RPM_URI* 替换为来自 Amazon S3 的相应 URI。

     ```
     --bootstrap-actions "Path=bootstrap_URI,Args=[RPM_URI]"
     ```

- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

```
--bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
```

- Apache Flink 提供 Native S3 FileSystem 和 Hadoop FileSystem 连接器，允许应用程序创建 FileSink 并将数据写入 Amazon S3。此 FileSink 失败并出现以下两个异常之一。

```
java.lang.UnsupportedOperationException: Recoverable writers on Hadoop are only
 supported for HDFS
```

```
Caused by: java.lang.NoSuchMethodError:
 org.apache.hadoop.io.retry.RetryPolicies.retryOtherThanRemoteAndSaslException(Lorg/
apache/hadoop/io/retry/RetryPolicy;Ljava/util/Map;)Lorg/apache/hadoop/io/retry/
RetryPolicy;
                                        at
 org.apache.hadoop.yarn.client.RMProxy.createRetryPolicy(RMProxy.java:302) ~[hadoop-
yarn-common-3.3.3-amzn-0.jar:?]
```

解决方法是安装 Amazon EMR 补丁，该补丁可以修复 Flink 中的上述问题。要在启动集群时应用引导操作，请完成以下步骤。

1. 将 flink-rpm 下载到 Amazon S3 存储桶中。您的 RPM 路径是 `s3://DOC-EXAMPLE-BUCKET/rpms/flink/`。
2. 使用以下 URI 从 Amazon S3 下载引导脚本和 RPM 文件。将 *regionName* 替换为您计划启动集群的 Amazon Web Services 区域。

   ```
   s3://emr-data-access-control-regionName/customer-bootstrap-actions/gcsc/replace-rpms.sh
   ```

3. Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。在 Amazon EMR 6.8.0 和 6.9.0 中，无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

   在 Amazon EMR 6.10.0 中，有一个解决此问题的方法，可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强和解决的问题

- 对于 Amazon EMR 发行版 6.9.0 及更高版本，Amazon EMR 安装的所有使用 Log4j 库的组件都使用 Log4j 版本 2.17.1 或更高版本。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。Amazon EMR 发行版 6.9.0 修复了此问题。
- 在使用 Spark SQL 读取数据时，Amazon EMR 6.9.0 添加对基于 Lake Formation 的访问控制及 Apache Hudi 的有限支持。支持针对使用 Spark SQL 的 SELECT 查询，并且仅限于列级访问控制。有关更多信息，请参阅 Hudi 和 Lake Formation。
- 当您使用 Amazon EMR 6.9.0 创建启用了节点标签的 Hadoop 集群时，YARN 指标 API 会返回所有分区的聚合信息，而不是默认分区。有关更多信息，请参阅 YARN-11414。
- 在 Amazon EMR 6.9.0 版本中，我们已将 Trino 更新到使用 Java 17 的 398 版本。之前支持的 Amazon EMR 6.8.0 Trino 版本是在 Java 11 上运行的 Trino 388。有关此变更的更多信息，请参阅 Trino 博客上的 Trino updates to Java 17。
- 此版本修复了 Apache BigTop 和 EC2 集群启动序列上的 Amazon EMR 之间的时间序列不匹配的问题。当系统尝试同时执行两个或多个操作而不是按正确的顺序执行它们时，就会发生这种计时序列不匹配。因此，某些集群配置会遇到实例启动超时和较慢的集群启动时间。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使用 `ListReleaseLabels` API 或 `list-release-labels` CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）、以色列（特拉维夫）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 1 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

发行版 6.8.0

以下发布说明包括有关 Amazon EMR 发行版 6.8.0 的信息。更改与 6.7.0 有关。

新功能

- Amazon EMR 步骤功能现支持 Livy 端点和 JDBC/ODBC 客户端。有关更多信息，请参阅为 Amazon EMR 步骤配置运行时角色。

- Amazon EMR 发行版 6.8.0 随附 Apache HBase 发行版 2.4.12。借助此 HBase 发行版，您可以对 HBase 表进行存档和删除。Amazon S3 存档过程会将所有表文件重命名为存档目录。这一过程成本高昂且时间较长。现在，您可以跳过存档过程，快速删除大型表。有关更多信息，请参阅使用 HBase shell。

已知问题

- Hadoop 3.3.3 在 YARN（YARN-9608）中引入了一项更改，即在应用程序完成之前，容器运行所在的节点一直处于停用状态。此更改可确保如随机数据等本地数据不会丢失，并且您无需重新运行作业。在 Amazon EMR 6.8.0 和 6.9.0 中，无论是否启用托管扩展，这种方法还可能导致集群的资源利用不足。

  在 Amazon EMR 6.10.0 中，有一个解决此问题的方法，可以在 `yarn-site.xml` 中将 `yarn.resourcemanager.decommissioning-nodes-watcher.wait-for-applications` 的值设置为 `false`。在 Amazon EMR 版本 6.11.0 及更高版本以及 6.8.1、6.9.1 和 6.10.1 中，默认将配置设置为 `false` 以解决此问题。

更改、增强和解决的问题

- 当 Amazon EMR 发行版 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark Shell 读取 Apache Phoenix 表时，Amazon EMR 会生成 `NoSuchMethodError`。Amazon EMR 发行版 6.8.0 修复了此问题。
- Amazon EMR 发行版 6.8.0 随附 Apache Hudi 0.11.1；但是，Amazon EMR 6.8.0 集群也与 Hudi 0.12.0 中的开源 `hudi-spark3.3-bundle_2.12` 兼容。
- Amazon EMR 发行版 6.8.0 随附 Apache Spar 3.3.0。此 Spark 发行版使用 Apache Log4j 2 和 `log4j2.properties` 文件，在 Spark 进程中配置 Log4j。如果您在集群中使用 Spark 或使用自定义配置参数创建 EMR 集群，并且希望升级到 Amazon EMR 发行版 6.8.0，则必须迁移到新的 `spark-log4j2` 配置分类和 Apache Log4j 2 的密钥格式。有关更多信息，请参阅从 Apache Log4j 1.x 迁移到 Log4j 2.x。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

> **Note**
>
> 此版本不再获得 AMI 自动更新，因为它已被另外 1 个补丁版本取代。补丁版本以第二位小数点后的数字 (`6.8.1`) 表示。要查看您是否使用的是最新补丁版本，请查看 Release Guide 中的可用版本，或者在控制台中创建集群时查看 Amazon EMR 版本下拉列表，或使

用 ListReleaseLabels API 或 list-release-labels CLI 操作。要获取有关新版本的更新，请订阅新增功能页面上的 RSS 源。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）、

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、亚太地区（墨尔本）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 9 月 6 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

已知问题

- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将 `spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

有关发布时间表的更多信息，请参阅更改日志。

发行版 6.7.0

以下发布说明包括有关 Amazon EMR 发行版 6.7.0 的信息。更改与 6.6.0 有关。

首次发布日期：2022 年 7 月 15 日

新功能

- Amazon EMR 现在支持 Apache Spark 3.2.1、Apache Hive 3.1.3、HUDI 0.11、PrestoDB 0.272 和 Trino 0.378。
- 通过 EMR 步骤（Spark、Hive）支持 EC2 集群上的 Amazon EMR 基于 IAM 角色和 Lake Formation 的访问控制。
- 在启用 Apache Ranger 的集群上支持 Apache Spark 数据定义语句。现在，这包括支持 Trino 应用程序在启用 Apache Ranger 的集群上读取和写入 Apache Hive 元数据。有关更多信息，请参阅在 Amazon EMR 上使用 Trino 和 Apache Ranger 启用联合治理。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
			（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 10 月 7 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2022 719.0	4.14.287	2022 年 8 月 10 日	`us-west-1` , `eu-west-3` , `eu-north-1` , `ap-south-1` , `me-south-1`

OsRelea Label (A zon Linux 版本)	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 606.1	4.14.281	2022 年 7 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将

`spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。
- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，`SecretAgent` 和 `RecordServer` 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

```
--bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

```
--bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
```

发行版 6.6.0

以下发布说明包括有关 Amazon EMR 发行版 6.6.0 的信息。更改与 6.5.0 有关。

首次发布日期：2022 年 5 月 9 日

文档更新日期：2022 年 6 月 15 日

新功能

- Amazon EMR 6.6 现在支持 Apache Spark 3.2、Apache Spark RAPIDS 22.02、CUDA 11、Apache Hudi 0.10.1、Apache Iceberg 0.13、Trino 0.367 和 PrestoDB 0.267。
- 当您使用 Amazon EMR 5.36 或更高版本或 6.6 或更高版本的最新补丁版本启动集群时，Amazon EMR 会使用最新的 Amazon Linux 2 版本作为默认 Amazon EMR AMI。如需更多信息，请参阅 Using the default Amazon Linux AMI for Amazon EMR。

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 727.0	4.14.320	2023 年 8 月 14 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲地区（法兰克福）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 719.0	4.14.320	2023 年 8 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、欧洲地区（斯德哥尔摩）、欧洲地区（米兰）、欧洲（西班牙）、欧洲地区（法兰克福）、欧洲（苏黎世）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、亚太地区（香港）、亚太地区（孟买）、亚太地区（海得拉巴）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、亚太地区（雅加达）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）、中东（阿联酋）、加拿大（中部）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 628.0	4.14.318	2023 年 7 月 12 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 612.0	4.14.314	2023 年 6 月 23 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 504.1	4.14.313	2023 年 5 月 16 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 418.0	4.14.311	2023 年 5 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 404.1	4.14.311	2023 年 4 月 18 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2023 404.0	4.14.311	2023 年 4 月 10 日	美国东部（弗吉尼亚州北部）、欧洲地区（巴黎）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 320.0	4.14.309	2023 年 3 月 30 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 307.0	4.14.305	2023 年 3 月 15 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 207.0	4.14.304	2023 年 2 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2023 119.1	4.14.301	2023 年 2 月 3 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 210.1	4.14.301	2023 年 12 月 22 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 103.3	4.14.296	2022 年 12 月 5 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 004.0	4.14.294	2022 年 11 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 912.1	4.14.291	2022 年 10 月 7 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）
2.0.2022 805.0	4.14.287	2022 年 8 月 30 日	us-west-1

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 719.0	4.14.287	2022 年 8 月 10 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 426.0	4.14.281	2022 年 6 月 10 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

OsRelea Label（A zon Linux 版本）	Amazon Linux 内核版本	可用日期	支持的区域
2.0.2022 406.1	4.14.275	2022 年 5 月 2 日	美国东部（弗吉尼亚州北部）、美国东部（俄亥俄州）、美国西部（北加利福尼亚）、美国西部（俄勒冈州）、加拿大（中部）、欧洲地区（斯德哥尔摩）、欧洲地区（爱尔兰）、欧洲地区（伦敦）、欧洲地区（巴黎）、欧洲地区（法兰克福）、欧洲地区（米兰）、亚太地区（香港）、亚太地区（孟买）、亚太地区（雅加达）、亚太地区（东京)、亚太地区（首尔）、亚太地区（大阪）、亚太地区（新加坡）、亚太地区（悉尼）、非洲（开普敦）、南美洲（圣保罗）、中东（巴林）

- 在 Amazon EMR 6.6 及更高版本中，使用 log4J 1.x 和 log4J 2.x 的应用程序将分别升级为使用 log4J 1.2.17（或更高版本）和 log4J 2.17.1（或更高版本），并且不需要使用提供的引导操作来缓解 CVE 问题。
- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管横向缩减和 Spark 编程指南。

- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 `spark.yarn.heterogeneousExecutors.enabled` 配置参数。

更改、增强和解决的问题

- 对于使用 EMR 默认 AMI 选项且仅安装常用应用程序（如 Apache Hadoop、Apache Spark 和 Apache Hive）的集群，Amazon EMR 平均可将启动时间缩短 80 秒。

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 在 Amazon EMR 版本 6.6.0、6.7.0 和 6.8.0 上将 DynamoDB 连接器与 Spark 结合使用时，即使输入拆分引用了非空数据，表中的所有读取都会返回空结果。这是因为 Spark 3.2.0 将 `spark.hadoopRDD.ignoreEmptySplits` 默认设置为 `true`。解决方法是将 `spark.hadoopRDD.ignoreEmptySplits` 显式设置为 `false`。Amazon EMR 发行版 6.9.0 修复了此问题。
- 在 Trino 长时间运行的集群上，Amazon EMR 6.6.0 在 Trino jvm.config 中启用了垃圾回收日志记录参数，以便从垃圾回收日志中获取更好的见解。此更改会将许多垃圾回收日志附加到 launcher.log (/var/log/trino/launcher.log) 文件。如果您在 Amazon EMR 6.6.0 中运行 Trino 集群，由于附加的日志，可能会在集群运行几天后出现节点磁盘空间不足的情况。

  这一问题的解决办法是在为 Amazon EMR 6.6.0 创建或克隆集群时，将以下脚本作为引导操作运行以禁用 jvm.config 中的垃圾回收日志记录参数。

```
#!/bin/bash
  set -ex
  PRESTO_PUPPET_DIR='/var/aws/emr/bigtop-deploy/puppet/modules/trino'
  sudo bash -c "sed -i '/-Xlog/d' ${PRESTO_PUPPET_DIR}/templates/jvm.config"
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。

- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 / 字符 (U+002F)。例如，在 s3://bucket/table/p=a b 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：!"#$%&'()*+,-。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

- 在 Amazon EMR 发行版 5.36.0 和 6.6.0 到 6.9.0 中，SecretAgent 和 RecordServer 服务组件可能会因为 Log4j2 属性中的文件名模式配置不正确而出现日志数据丢失的情况。错误的配置导致组件每天只生成一个日志文件。当应用轮换策略时，它会重写现有文件，而不是按预期生成新的日志文件。应变方法是使用引导操作每小时生成一次日志文件，并在文件名中附加一个自动增量的整数来处理轮换。

  对于 Amazon EMR 发行版 6.6.0 到 6.9.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-6x/replace-puppet.sh,Args=[]"
  ```

  对于 Amazon EMR 发行版 5.36.0，启动集群时，请执行以下引导操作。

  ```
  --bootstrap-actions "Path=s3://emr-data-access-control-us-east-1/customer-bootstrap-actions/log-rotation-emr-5x/replace-puppet.sh,Args=[]"
  ```

发行版 5.35.0

这是 Amazon EMR 发行版 5.35.0 的发布说明。

以下发布说明包括有关 Amazon EMR 发行版 5.35.0 的信息。更改与 5.34.0 有关。

首次发布日期: 2022 年 3 月 30 日

新功能

- 使用 log4J 1.x 和 log4J 2.x 的 Amazon EMR 发行版 5.35 应用程序将分别升级为使用 log4J 1.2.17（或更高版本）和 log4J 2.17.1（或更高版本），并且不需要使用引导操作来缓解之前发行版中的 CVE 问题。请参阅 缓解 CVE-2021-44228 的方法。

更改、增强和解决的问题

Flink 更改

更改类型	描述
升级	• 将 Flink 版本更新到 1.14.2。 • log4j 升级到 2.17.1。

Hadoop 更改

更改类型	描述
自 EMR 5.34.0 以来的 Hadoop 开源逆向移植	• YARN-10438: 处理 ClientRMService#getContainerReport() 中的空 containerId • YARN-7266: 时间轴服务器事件处理程序线程已锁定 • YARN-10438：如果 RollingLevelDb 文件损坏或丢失，ATS 1.5 将无法开启 • HADOOP-13500: 同步配置属性对象的迭代 • YARN-10651: CapacityScheduler 由于 AbstractYarnScheduler.updateNodeResource() 中的 NPE 崩溃 • HDFS-12221: 替换 XmlEditsVisitor 中的 xerces • HDFS-16410: OfflineEditsXmlLoader 中不安全的 Xml 解析
Hadoop 更改和修复	• KMS 和 HttpFS 中使用的 Tomcat 升级到 8.5.75 • 在 FileSystemOptimizedCommitterV2 中，成功标记被写入创建提交程序时定义的 commitJob 输出路径。由于 commitJob 和任务级别输出路径可能不同，因此更正路径以使用清单文件中定义的路径。对于 Hive 任务，这会导致在执行动态分区或 UNION ALL 等操作时正确写入成功标记。

Hive 更改

更改类型	描述
Hive 升级到开源发行版 2.3.9，包括这些 JIRA 修复	• HIVE-17155: HiveConf.java 中的 findConfFile() 存在一些配置路径问题 • HIVE-24797: 在解析 Avro 架构时禁用验证原定设置值 • HIVE-21563: 通过禁用 registerAllFunctionsOnce 提升 Table#getEmptyTable 性能 • HIVE-18147: 测试可能失败，显示 java.net.BindException: 地址已在使用中 • HIVE-24608: 切换回 Hive 2.3.x HMS 客户端中的 get_table • HIVE-21200: 向量化 - 日期列显示 java.lang.UnsupportedOperationException for parquet • HIVE-19228: 删除 commons-httpclient 3.x 使用
自 EMR 5.34.0 以来的 Hive 开源逆向移植	• HIVE-19990: 在联接条件下使用时间间隔文本查询失败 • HIVE-25824: 将 branch-2.3 升级到 log4j 2.17.0 • TEZ-4062: 推测性尝试计划应在任务完成时中止 • TEZ-4108: 在推测性执行竞争条件期间出现 NullPointerException • TEZ-3918: 设置项 tez.task.log.level 无效
Hive 升级和修复	• 将 Log4j 版本升级到 2.17.1 • 将 ORC 版本升级到 1.4.3 • 修复了由于 ShuffleScheduler 中的惩罚线程导致的死锁
新特征	• 添加了在 AM 日志中打印 Hive 查询的功能 默认情况下，将禁用该功能。标记/配置

更改类型	描述
	: `tez.am.emr.print.hive.query.in.log` 。状态（原定设置）: FALSE。

Oozie 更改

更改类型	描述
自 EMR 5.34.0 以来的 Oozie 开源逆向移植	• OOZIE-3652: 当发生 NoSuchFileException 时，Oozie 启动器应重试目录列表

Pig 更改

更改类型	描述
升级	• log4j 升级到 1.2.17。

已知问题

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 5.34.0

以下发布说明包括有关 Amazon EMR 发行版 5.34.0 的信息。更改与 5.33.1 有关。

首次发布日期：2022 年 1 月 20 日

发布更新日期：2022 年 3 月 21 日

新功能

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管扩展 和 Spark 编程指南。
- [Hudi] 简化了 Hudi 配置的改进。预设情况下禁用乐观并发控制。

更改、增强和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 以前，在多主节点集群上手动重启资源管理器会导致 Zookeeper znode 文件中的 Amazon EMR 集群进程守护程序（如 Zookeeper）重新加载以前停用或丢失的所有节点。在某些情况下，这会导致超出默认限制。Amazon EMR 现在会从 Zookeeper 文件中删除已停用或丢失超过一小时的节点记录，并且内部限制也有所提高。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。

- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- Zeppelin 已升级到版本 0.10.0。
- Livy 修复 - 已升级到 0.7.1
- Spark 性能提升 - 当 EMR 5.34.0 中的某些 Spark 配置值被覆盖时禁用异构执行器。
- 默认情况下禁用 WebHDFS 和 HTTFS 服务器。您可以使用 Hadoop 配置重新启用 WebHDFS，`dfs.webhdfs.enabled`。HTTPFS 服务器可以通过使用 `sudo systemctl start hadoop-httpfs` 启动。

已知问题

- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。
- Hue 查询在 Amazon EMR 6.4.0 中不起作用，因为默认情况下 Apache Hadoop HTTPFS 服务器处于禁用状态。要在 Amazon EMR 6.4.0 上使用 Hue，请使用 `sudo systemctl start hadoop-httpfs` 在 Amazon EMR 主节点上手动启动 HTTPFS 服务器，或者使用 Amazon EMR 步骤。
- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

发行版 6.5.0

以下发布说明包括有关 Amazon EMR 发行版 6.5.0 的信息。更改与 6.4.0 有关。

首次发布日期：2022 年 1 月 20 日

发布更新日期：2022 年 3 月 21 日

新功能

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管横向缩减和 Spark 编程指南。
- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 spark.yarn.heterogeneousExecutors.enabled 配置参数。
- 支持 Apache Iceberg 开放表格式，用于大型分析数据集。
- 支持 ranger-trino-plugin 2.0.1-amzn-1
- 支持 toree 0.5.0

更改、增强和解决的问题

- Amazon EMR 6.5 发行版现在支持 Apache Iceberg 0.12.0，并通过适用于 Apache Spark 的 Amazon EMR 运行时、适用于 Presto 的 Amazon EMR 运行时和适用于 Apache Hive 的 Amazon EMR 运行时提供了运行时改进。
- Apache Iceberg 是 Amazon S3 中适用于大型数据集的开放表格式，可提供快速的大型表查询性能、原子提交、并发写入和 SQL 兼容表演进等功能。借助 EMR 6.5，您可以将 Apache Spark 3.1.2 与 Iceberg 表格式结合使用。
- Apache Hudi 0.9 增加了对 Spark SQL DDL 和 DML 的支持。从而让您可以仅使用 SQL 语句创建 upsert Hudi 表。Apache Hudi 0.9 还包括查询端和写入器端的性能改进。
- 适用于 Apache Hive 的 Amazon EMR 运行时取消了暂存操作期间的重命名操作，从而提高了 Apache Hive 在 Amazon S3 上的性能，此外还提高了用于修复表的元数据仓检查（MSCK）命令的性能。

已知问题

- 当 Amazon EMR 版本 6.5.0、6.6.0 或 6.7.0 通过 Apache Spark shell 读取 Apache Phoenix 表时，会出现 `NoSuchMethodError`，因为 Amazon EMR 使用了不正确的 `Hbase.compat.version`。Amazon EMR 发行版 6.8.0 修复了此问题。
- 高可用性（HA）的 Hbase 捆绑集群无法使用默认卷大小和实例类型进行预置。此问题的变通解决方法是增加根卷大小。
- 要将 Spark 操作与 Apache Oozie 一起使用，必须将以下配置添加到 Oozie `workflow.xml` 文件中。否则，Oozie 启动的 Spark 执行器的类路径中将丢失几个诸如 Hadoop 和 EMRFS 之类的关键库。

```
<spark-opts>--conf spark.yarn.populateHadoopClasspath=true</spark-opts>
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 6.4.0

以下发布说明包括有关 Amazon EMR 发行版 6.4.0 的信息。更改与 6.3.0 有关。

首次发布日期：2021 年 9 月 20 日

发布更新日期：2022 年 3 月 21 日

支持的应用程序

- Amazon SDK for Java 1.12.31

- CloudWatch Sink 2.2.0
- DynamoDB 连接器 4.16.0
- EMRFS 2.47.0
- Amazon EMR Goodies 3.2.0
- Amazon EMR Kinesis 连接器 3.5.0
- Amazon EMR 记录服务器 2.1.0
- Amazon EMR Scripts 2.5.0
- Flink 1.13.1
- Ganglia 3.7.2
- Amazon Glue Hive Metastore Client 3.3.0
- Hadoop 3.2.1-amzn-4
- HBase 2.4.4-amzn-0
- HBase-operator-tools 1.1.0
- HCatalog 3.1.2-amzn-5
- Hive 3.1.2-amzn-5
- Hudi 0.8.0-amzn-0
- Hue 4.9.0
- Java JDK Corretto-8.302.08.1（内部 1.8.0_302-b08）
- JupyterHub 1.4.1
- Livy 0.7.1-incubating
- MXNet 1.8.0
- Oozie 5.2.1
- Phoenix 5.1.2
- Pig 0.17.0
- Presto 0.254.1-amzn-0
- Trino 359
- Apache Ranger KMS（多主节点透明加密）版本 2.0.0
- ranger-plugins 2.0.1-amzn-0
- ranger-s3-plugin 1.2.0

- SageMaker Spark SDK 1.4.1
- Scala 2.12.10（OpenJDK 64 位服务器 VM，Java 1.8.0_282）
- Spark 3.1.2-amzn-0
- spark-rapids 0.4.1
- Sqoop 1.4.7
- TensorFlow 2.4.1
- tez 0.9.2
- Zeppelin 0.9.0
- Zookeeper 3.5.7
- 连接器和驱动程序：DynamoDB 连接器 4.16.0

新特征

- [托管式扩展] Spark 随机排序数据托管式扩展优化 – Amazon EMR 5.34.0 及更高版本和 Amazon EMR 6.4.0 及更高版本支持可感知 Spark 随机排序数据（Spark 在分区之间重新分配以执行特定操作的数据）的托管式扩展。有关随机排序操作的更多信息，请参阅《Amazon EMR 管理指南》中的在 Amazon EMR 中使用 EMR 托管扩展 和 Spark 编程指南。
- 在 Apache Ranger 启用的 Amazon EMR 集群上，您可以使用 Apache Spark SQL 将数据插入到 Apache Hive 元数据存储表中或使用`INSERT INTO`、`INSERT OVERWRITE`和`ALTER TABLE`更新 Apache Hive 元数据存储表。将 ALTER TABLE 与 Spark SQL 结合使用时，分区位置必须是表位置的子目录。如果某个分区的分区位置与表位置不同，Amazon EMR 目前不支持将数据插入该分区。
- PrestoSQL 已重命名为 Trino。
- Hive：在获取 LIMIT 子句中提到的记录数目后，通过立即停止查询执行可加快使用 LIMIT 子句执行简单 SELECT 查询的速度。简单 SELECT 查询是没有 GROUP BY/ORDER BY 子句的查询或没有减速阶段的查询。例如，`SELECT * from <TABLE> WHERE <Condition> LIMIT <Number>`。

Hudi 并发控制

- Hudi 目前支持乐观并发控制 (OCC)，它可以与 UPSERT 和 INSERT 等写入操作一起利用，以允许从多个写入器更改为同一 Hudi 表。这是文件级 OCC，因此任何两个提交（或写入器）可以写入同一表内，前提是它们的更改不冲突。有关更多信息，请参阅 Hadi 并发性控制。
- Amazon EMR 集群安装了 Zookeeper，可以利用它作为 OCC 的锁提供商。为了更便捷地使用此功能，Amazon EMR 集群预先配置了以下属性：

```
hoodie.write.lock.provider=org.apache.hudi.client.transaction.lock.ZookeeperBasedLockProvider
hoodie.write.lock.zookeeper.url=<EMR Zookeeper URL>
hoodie.write.lock.zookeeper.port=<EMR Zookeeper Port>
hoodie.write.lock.zookeeper.base_path=/hudi
```

要启用 OCC，您需要使用 Hudi 任务选项或使用 Amazon EMR 配置 API 在集群级别配置以下属性：

```
hoodie.write.concurrency.mode=optimistic_concurrency_control
hoodie.cleaner.policy.failed.writes=LAZY (Performs cleaning of failed writes lazily
 instead of inline with every write)
hoodie.write.lock.zookeeper.lock_key=<Key to uniquely identify the Hudi table> (Table
 Name is a good option)
```

Hudi 监控：Amazon CloudWatch 集成，用于报告 Hudi 指标

- Amazon EMR 支持向 Amazon CloudWatch 发布 Hudi 指标。通过设置以下所需配置来启用：

```
hoodie.metrics.on=true
hoodie.metrics.reporter.type=CLOUDWATCH
```

- 以下是您可以更改的可选 Hudi 配置：

设置	描述	Value
hoodie.metrics.cloudwatch.report.period.seconds	向 Amazon CloudWatch 报告指标的频率（以秒为单位）	默认值为 60 秒，对于 Amazon CloudWatch 提供的默认一分钟分辨率而言是可行的
hoodie.metrics.cloudwatch.metric.prefix	要添加到每个指标名称的前缀	默认值为空（无前缀）
hoodie.metrics.cloudwatch.namespace	以此为发布指标的 Amazon CloudWatch 命名空间	默认值为 Hudi
hoodie.metrics.cloudwatch.maxDatumsPerRequest	向 Amazon CloudWatch 发出的请求中要包含的最大基准数	默认值为 20（与 Amazon CloudWatch 默认值相同）

Amazon EMR Hudi 配置的支持和改进

- 客户目前可以利用 EMR 配置 API 和重新配置功能在集群级别配置 Hudi 配置。与 Spark 和 Hive 等其他应用程序一样，通过 /etc/hudi/CONF/hudi-defaults.conf 引入了基于文件的新配置支持。EMR 配置了几个默认值以改善用户体验：

  —hoodie.datasource.hive_sync.jdbcurl 已配置为集群 Hive 服务器 URL，无需指定。这在 Spark 集群模式下运行任务时十分有效，而您之前必须指定 Amazon EMR 主 IP。

  — HBase 特定的配置，这对于将 HBase 索引与 Hudi 一起使用非常有用。

  — Zookeeper 锁提供商的特定配置，如并发控制下所讨论的内容，这令乐观并发控制 (OCC) 的使用更加方便。
- 还引入了其他更改，以减少需要通过的配置数量，并在可能的情况下自动推断：

  — 该partitionBy 关键字可用于指定分区列。

  — 启用 Hive Sync 时，不再强制通过HIVE_TABLE_OPT_KEY, HIVE_PARTITION_FIELDS_OPT_KEY, HIVE_PARTITION_EXTRACTOR_CLASS_OPT_KEY。这些值可以根据 Hudi 表名称和分区字段推断出来。

  —KEYGENERATOR_CLASS_OPT_KEY不强制通过，可以从更简单的SimpleKeyGenerator和ComplexKeyGenerator情况下推断。

Hudi 注意事项

- Hudi 不支持在 Hive 中用于读取时合并 (MoR) 和 Bootstrap 表格中的矢量化执行。例如，当hive.vectorized.execution.enabled设置为 true 时，Hudi 实时表的count(*)失败。作为解决方法，您可以通过将hive.vectorized.execution.enabled设置为false禁用矢量化读入。
- 多写作器支持与 Hudi 引导启动功能不兼容。
- Flink Streamer 和 Flink SQL 是此发行版中的实验性功能。建议不要在生产部署中使用这些功能。

更改、增强功能和解决的问题

此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。

- 以前，在多主节点集群上手动重启资源管理器会导致 Zookeeper znode 文件中的 Amazon EMR 集群进程守护程序（如 Zookeeper）重新加载以前停用或丢失的所有节点。在某些情况下，这会导致超出默认限制。Amazon EMR 现在会从 Zookeeper 文件中删除已停用或丢失超过一小时的节点记录，并且内部限制也有所提高。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 配置集群以修复 Apache YARN 时间轴服务器 1 和 1.5 版的性能问题

  Apache YARN 时间轴服务器版本 1 和 1.5 可能会对非常活跃的大型 EMR 集群造成性能问题，尤其是 `yarn.resourcemanager.system-metrics-publisher.enabled=true`，这是 Amazon EMR 中的默认设置。开源 YARN 时间轴服务器 v2 解决了与 YARN 时间轴服务器可扩展性相关的性能问题。

  此问题的其他解决方法包括：

  - 配置 yarn.资源管理器.系统指标-发布者.启用=false 在 yarn-site.xml 中。
  - 如下所述，在创建群集时启用此问题的修复程序。

  以下 Amazon EMR 发行版包含针对此 YARN 时间线服务器性能问题的修复。

  EMR 5.30.2、5.31.1、5.32.1、5.33.1、5.34.x、6.0.1、6.1.1、6.2.1、6.3.1、6.4.x

  要对上述任何指定的 Amazon EMR 版本启用修复程序，请使用 `aws emr create-cluster` 命令参数：`--configurations file://./configurations.json` 在传入的配置 JSON 文件中将这些属性设置为 `true`。或者使用重新配置控制台 UI 启用修复程序。

配置 .json 文件内容的示例：

```
[
{
"Classification": "yarn-site",
"Properties": {
"yarn.resourcemanager.system-metrics-publisher.timeline-server-v1.enable-batch":
 "true",
"yarn.resourcemanager.system-metrics-publisher.enabled": "true"
},
"Configurations": []
}
]
```

- 默认情况下禁用 WebHDFS 和 HTTFS 服务器。您可以使用 Hadoop 配置重新启用 WebHDFS，`dfs.webhdfs.enabled`。HTTPFS 服务器可以通过使用 `sudo systemctl start hadoop-httpfs` 启动。
- 现在，默认情况下，已启用 Amazon Linux 存储库的 HTTPS。如果您使用 Amazon S3 VPCE 策略限制对特定存储桶的访问，则必须添加新的 Amazon Linux 存储桶 ARN`arn:aws:s3:::amazonlinux-2-repos-$region/*`到策略（将`$region`替换为终端节点所在的区域）。有关更多信息，请参阅 Amazon 讨论论坛的主题。公告：Amazon Linux 2 目前支持在连接到软件包存储库时使用 HTTPS 的功能。
- Hive：为最后任务，通过启用 HDFS 上的 scratch 目录，从而提高写入查询性能。最终任务的临时数据可写入 HDFS 而不是 Amazon S3，性能可以得到提高，因为数据从 HDFS 移动到最终表位置 (Amazon S3) 而不是在 Amazon S3 设备之间移动。
- Hive：使用 Glue 元存储分区修剪，查询编译时间最多可缩短 2.5 倍。
- 默认情况下，当 Hive 将内置 UDF 传递到 Hive 元存储服务器时，由于 Glue 只支持有限的表达式运算，所以只会将这些内置 UDF 的子集传递到 Glue 元存储。如果您设置`hive.glue.partition.pruning.client=true`，则所有分区修剪发生在客户端。如果您设置`hive.glue.partition.pruning.server=true`，则所有分区修剪发生在服务器端。

已知问题

- Hue 查询在 Amazon EMR 6.4.0 中不起作用，因为默认情况下 Apache Hadoop HTTPFS 服务器处于禁用状态。要在 Amazon EMR 6.4.0 上使用 Hue，请使用 `sudo systemctl start hadoop-httpfs` 在 Amazon EMR 主节点上手动启动 HTTPFS 服务器，或者使用 Amazon EMR 步骤。

- 与 Livy 用户模拟一起使用的 Amazon EMR Notebooks 功能不起作用，因为默认情况下，HTTPFS 处于禁用状态。在这种情况下，EMR 笔记本无法连接到启用了 Livy 模拟的集群。解决方法是在将 EMR 笔记本连接到集群之前使用 `sudo systemctl start hadoop-httpfs` 启动 HTTPFS 服务器。
- 在 Amazon EMR 6.4.0 版本中，Phoenix 不支持 Phoenix 连接器组件。
- 要将 Spark 操作与 Apache Oozie 一起使用，必须将以下配置添加到 Oozie `workflow.xml` 文件中。否则，Oozie 启动的 Spark 执行器的类路径中将丢失几个诸如 Hadoop 和 EMRFS 之类的关键库。

```
<spark-opts>--conf spark.yarn.populateHadoopClasspath=true</spark-opts>
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 5.32.0

以下发布说明包括有关 Amazon EMR 发行版 5.32.0 的信息。更改与 5.31.0 有关。

首次发布日期：2021 年 1 月 8 日

升级

- 已将 Amazon Glue 连接器升级到1.14.0
- 已将 Amazon SageMaker Spark SDK 升级到版本 1.4.1
- 已将 Amazon SDK for Java 升级到版本 1.11.890
- 已将 EMR DynamoDB 连接器升级到版本 4.16.0

- 已将 EMRFS 升级到版本 2.45.0
- 已将 EMR Log Analytics Metrics 升级到版本 1.18.0
- 已将 EMR MetricsAndEventsApiGateway 客户端升级到版本 1.5.0
- 已将 EMR 记录服务器升级到版本 1.8.0
- 已将 EMR S3 Dist CP 升级到版本 2.17.0
- 已将 EMR Secret Agent 升级到版本 1.7.0
- 已将 Flink 升级到版本 1.11.2
- 已将 Hadoop 升级到版本 2.10.1-amzn-0
- 已将 Hive 升级到版本 2.3.7-amzn-3
- 已将 Hue 升级到版本 4.8.0
- 已将 Mxnet 升级到版本 1.7.0
- 已将 OpenCV 升级到版本 4.4.0
- 已将 Presto 升级到版本 0.240.1-amzn-0
- 已将 Spark 升级到版本 2.4.7-amzn-0
- 已将 TensorFlow 升级到版本 2.3.1

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。

- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 升级了组件版本。
- 有关组件版本的列表，请参阅本指南中的关于 Amazon EMR 发行版。

新特征

- 从 Amazon EMR 5.32.0 和 6.5.0 开始，Apache Spark 动态执行程序定型功能会默认启用。要打开或关闭此功能，您可以使用 `spark.yarn.heterogeneousExecutors.enabled` 配置参数。
- 实例元数据服务（IMDS）V2 支持状态：Amazon EMR 5.23.1、5.27.1 和 5.32 或更高版本的组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。对于其它 5.x EMR 版本，禁用 IMDSv1 会导致集群启动失败。
- 从 Amazon EMR 5.32.0 开始，您可以启动与 Apache Ranger 在本地集成的集群。Apache Ranger 是一个开源框架，可跨 Hadoop 平台启用、监控和管理全面的数据安全。有关更多信息，请参阅 Apache Ranger。通过本机集成，您可以自带 Apache Ranger，在 Amazon EMR 上强制实施精细数据访问控制。请参阅《Amazon EMR 版本指南》中的 将 Amazon EMR 与 Apache Ranger 集成。
- Amazon EMR 发行版 5.32.0 支持 Amazon EMR on EKS。有关 EMR on EKS 入门的更多详细信息，请参阅什么是 Amazon EMR on EKS。
- Amazon EMR 发行版 5.32.0 版支持 Amazon EMR Studio（预览版）。有关 EMR Studio 入门的更多详细信息，请参阅 Amazon EMR Studio（预览版）。
- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。

已知问题

- 对于 Amazon EMR 6.3.0 和 6.2.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 sudo

`systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

  使用引导操作 (BA) 设置 ulimit

  您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

  ```
  #!/bin/bash
  for user in hadoop spark hive; do
  sudo tee /etc/security/limits.d/$user.conf << EOF
  $user - nofile 65536
  ```

```
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- **Important**

  运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- GovCloud 区域中目前不支持使用控制台创建指定 Amazon Ranger 集成选项的安全配置。可以使用 CLI 完成安全配置。请参阅《Amazon EMR 管理指南》中的创建 EMR 安全配置。
- 在使用 Amazon EMR 5.31.0 或 5.32.0 的集群上启用了 AtRestEncryption 或 HDFS 加密时，Hive 查询会导致以下运行时系统异常。

```
TaskAttempt 3 failed, info=[Error: Error while running task ( failure ) :
 attempt_1604112648850_0001_1_01_000000_3:java.lang.RuntimeException:
 java.lang.RuntimeException: Hive Runtime Error while closing
 operators: java.io.IOException: java.util.ServiceConfigurationError:
 org.apache.hadoop.security.token.TokenIdentifier: Provider
 org.apache.hadoop.hbase.security.token.AuthenticationTokenIdentifier not found
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：

- 从同一个表扫描两个或多个分区。
- 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
- 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 6.2.0

以下发布说明包括有关 Amazon EMR 发行版 6.2.0 的信息。更改与 6.1.0 有关。

首次发布日期：2020 年 12 月 9 日

上次更新日期：2021 年 10 月 4 日

支持的应用程序

- Amazon SDK for Java 1.11.828
- emr-record-server 1.7.0
- Flink 1.11.2
- Ganglia 3.7.2
- Hadoop 3.2.1-amzn-1
- HBase 2.2.6-amzn-0
- HBase-operator-tools 1.0.0
- HCatalog 3.1.2-amzn-0
- Hive 3.1.2-amzn-3
- Hudi 0.6.0-amzn-1
- Hue 4.8.0
- JupyterHub 1.1.0
- Livy 0.7.0

- MXNet 1.7.0
- Oozie 5.2.0
- Phoenix 5.0.0
- Pig 0.17.0
- Presto 0.238.3-amzn-1
- PrestoSQL 343
- Spark 3.0.1-amzn-0
- spark-rapids 0.2.0
- TensorFlow 2.3.1
- Zeppelin 0.9.0-preview1
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.16.0

新特征

- HBase：删除了提交阶段的重命名，添加了持久性 HFile 跟踪。请参阅《Amazon EMR 版本指南》中的持久性 HFile 跟踪。
- HBase：已逆向移植创建在压缩时强制缓存数据块的配置。
- PrestoDB：改进了动态分区修剪。基于规则的连接重新排序对未分区数据运行。
- 限定范围的托管式策略：为了符合Amazon最佳实践，Amazon EMR 引入了 v2 EMR 范围的默认托管式策略，来替代即将弃用的策略。请参阅 Amazon EMR 托管式策略。
- 实例元数据服务（IMDS）V2 支持状态：对于 Amazon EMR 6.2 或更高版本，Amazon EMR 组件对所有 IMDS 调用都使用 IMDSv2。对于应用程序代码中的 IMDS 调用，您可以同时使用 IMDSv1 和 IMDSv2，或者将 IMDS 配置为仅使用 IMDSv2，以提高安全性。如果您在早于 Amazon EMR 6.x 的发行版中禁用 IMDSv1，则会导致集群启动失败。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。

- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Spark：改进了 Spark 运行时的性能。

已知问题

- Amazon EMR 6.2 对 EMR 6.2.0 中的 /etc/cron.d/libinstance-controller-java 文件设置了错误权限。当文件的权限应为 644 (-rw-r--r--) 时，它们为 645 (-rw-r--r-x)。因此，Amazon EMR 6.2 版本不记录实例状态日志，并且 /emr/instance-log 目录为空。此问题已在 Amazon EMR 6.3.0 及更高版本中得到修复。

  要解决此问题，请在集群启动时将以下脚本作为引导操作运行。

  ```
  #!/bin/bash
  sudo chmod 644 /etc/cron.d/libinstance-controller-java
  ```

- 对于 Amazon EMR 6.2.0 和 6.3.0 私有子网集群，您不能访问 Ganglia Web UI。您将收到“access denied (403)”错误。其它 Web UI（如 Spark、Hue、JupyterHub、Zeppelin、Livy 和 Tez）可正常运行。公有子网集群上的 Ganglia Web UI 访问也正常工作。要解决该问题，请在具有 `sudo systemctl restart httpd` 的主节点上重新启动 httpd 服务。此问题已在 Amazon EMR 6.4.0 中得到修复。

- Amazon EMR 6.2.0 中存在一个问题：httpd 持续失败，导致 Ganglia 不可用。您会收到“cannot connect to the server（无法连接到服务器）”错误。要修复已在运行期间出现此问题的集群，请使用 SSH 连接到集群主节点并将行 `Listen 80` 添加到位于 `/etc/httpd/conf/httpd.conf` 的文件 `httpd.conf` 中。此问题已在 Amazon EMR 6.3.0 中得到修复。
- 使用安全配置时，HTTPD 在 EMR 6.2.0 集群会上失败。因此，Ganglia Web 应用程序用户界面不可用。要访问 Ganglia Web 应用程序用户界面，请将 `Listen 80` 添加到集群主节点上的 `/etc/httpd/conf/httpd.conf` 文件中。有关连接集群的更多信息，请参阅使用 SSH 连接到主节点。

  使用安全配置时，EMR Notebooks 也无法建立与 EMR 6.2.0 集群的连接。笔记本将无法列出内核和提交 Spark 任务。我们建议您改为将 EMR Notebooks 与其它版本的 Amazon EMR 结合使用。
- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```
  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- **Important**

  Amazon EMR 6.1.0 和 6.2.0 包含可能严重影响所有 Hudi 插入、更新插入和删除操作的性能问题。如果您计划将 Hudi 与 Amazon EMR 6.1.0 或 6.2.0 结合使用，请联系 Amazon Support，获取 Hudi RPM 补丁。

- **Important**

  运行 Amazon Linux 或 Amazon Linux 2 AMI（Amazon Linux Machine Image）的 Amazon EMR 集群使用默认的 Amazon Linux 行为，且不会自动下载和安装需要重新启动的重要关键内核更新。这与运行默认 Amazon Linux AMI 的其它 Amazon EC2 实例的行为相同。如果需要重新启动的新 Amazon Linux 软件更新（例如内核、NVIDIA 和 CUDA 更新）在 Amazon EMR 版本发布后可用，则运行默认 AMI 的 Amazon EMR 集群实例不会自动下载和安装这些更新。要获取内核更新，您可以自定义 Amazon EMR AMI，以使用最新的 Amazon Linux AMI。

- Amazon EMR 6.2.0 Maven 构件尚未发布。它们将随 Amazon EMR 未来版本一起发布。
- 使用 HBase 存储文件系统表的持久性 HFile 跟踪不支持 HBase 区域复制功能。有关 HBase 区域复制的更多信息，请参阅时间表一致的高可用读取。
- Amazon EMR 6.x 和 EMR 5.x Hive 分桶版本差异

  EMR 5.x 使用 OOS Apache Hive 2，而 EMR 6.x 使用 OOS Apache Hive 3。开源 Hive2 使用分桶版本 1，而开源 Hive3 使用分桶版本 2。Hive 2（EMR 5.x）和 Hive 3（EMR 6.x）之间的这一分桶版本差异将导致 Hive 分桶哈希函数不同。请参见以下示例。

  下表分别是在 EMR 6.x 和 EMR 5.x 中创建的示例。

```
-- Using following LOCATION in EMR 6.x
CREATE TABLE test_bucketing (id INT, desc STRING)
PARTITIONED BY (day STRING)
CLUSTERED BY(id) INTO 128 BUCKETS
LOCATION 's3://your-own-s3-bucket/emr-6-bucketing/';

-- Using following LOCATION in EMR 5.x
LOCATION 's3://your-own-s3-bucket/emr-5-bucketing/';
```

  在 EMR 6.x 和 EMR 5.x 中插入相同的数据。

```
INSERT INTO test_bucketing PARTITION (day='01') VALUES(66, 'some_data');
INSERT INTO test_bucketing PARTITION (day='01') VALUES(200, 'some_data');
```

  检查 S3 位置，显示分桶文件名不同，这是因为 EMR 6.x（Hive 3）和 EMR 5.x（Hive 2）之间的哈希函数不同。

```
[hadoop@ip-10-0-0-122 ~]$ aws s3 ls s3://your-own-s3-bucket/emr-6-bucketing/day=01/
2020-10-21 20:35:16         13 000025_0
2020-10-21 20:35:22         14 000121_0
[hadoop@ip-10-0-0-122 ~]$ aws s3 ls s3://your-own-s3-bucket/emr-5-bucketing/day=01/
2020-10-21 20:32:07         13 000066_0
2020-10-21 20:32:51         14 000072_0
```

  您还可以通过以下方式查看版本之间的差异：在 EMR 6.x 的 Hive CLI 中运行以下命令。请注意，它将返回分桶版本 2。

```
hive> DESCRIBE FORMATTED test_bucketing;
```

```
...
Table Parameters:
    bucketing_version       2
...
```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

发行版 5.31.0

以下发布说明包括有关 Amazon EMR 发行版 5.31.0 的信息。更改与 5.30.1 有关。

首次发布日期：2020 年 10 月 9 日

上次更新日期：2020 年 10 月 15 日

升级

- 已将 Amazon Glue 连接器升级到版本 1.13.0
- 已将 Amazon SageMaker Spark SDK 升级到版本 1.4.0
- 已将 Amazon Kinesis 连接器升级到版本 3.5.9
- 已将 Amazon SDK for Java 升级到版本 1.11.852
- 已将 Bigtop-tomcat 升级到版本 8.5.56
- 已将 EMR FS 升级到版本 2.43.0
- 已将 EMR MetricsAndEventsApiGateway 客户端升级到版本 1.4.0
- 已将 EMR S3 Dist CP 升级到版本 2.15.0
- 已将 EMR S3 Select 升级到版本 1.6.0
- 已将 Flink 升级到版本 1.11.0
- 已将 Hadoop 升级到版本 2.10.0
- 已将 Hive 升级到版本 2.3.7
- 已将 Hudi 升级到版本 0.6.0
- 已将 Hue 升级到版本 4.7.1
- 已将 JupyterHub 升级到版本 1.1.0
- 已将 Mxnet 升级到版本 1.6.0
- 已将 OpenCV 升级到版本 4.3.0
- 已将 Presto 升级到版本 0.238.3
- 已将 TensorFlow 升级到版本 2.1.0

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Amazon EMR 5.31.0 及更高版本支持 Hive 列统计信息。
- 升级了组件版本。
- Amazon EMR 5.31.0 支持 EMRFS S3EC V2。在 S3 Java SDK 1.11.837 及更高版本中，引入了加密客户端版本 2（S3EC V2），并新增了各种安全增强功能。有关更多信息，请参阅下列内容：
  - S3 博客文章：更新至 Amazon S3 加密客户端。
  - Amazon SDK for Java 开发人员指南：将加密和解密客户端迁移到 V2。
  - EMR 管理指南：Amazon S3 客户端加密。

  为保持向后兼容性，加密客户端 V1 在 SDK 中仍可用。

新特征

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

  从命令行显式设置 ulimit

  1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

     ```
     LimitNOFILE=65536

     LimitNPROC=65536
     ```

  2. 重新启动 InstanceController

     ```
     $ sudo systemctl daemon-reload

     $ sudo systemctl restart instance-controller
     ```

  使用引导操作 (BA) 设置 ulimit

  您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

  ```
  #!/bin/bash
  for user in hadoop spark hive; do
  sudo tee /etc/security/limits.d/$user.conf << EOF
  $user - nofile 65536
  $user - nproc 65536
  ```

```
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- 借助 Amazon EMR 5.31.0，您可以启动与 Lake Formation 集成的集群。该集成提供精细的列级数据筛选功能，用于筛选 Amazon Glue 数据目录中的数据库和表。它还支持从企业身份系统通过联合单点登录的方式登录 EMR Notebooks 或 Apache Zeppelin。有关更多信息，请参阅《Amazon EMR 管理指南》中的 将 Amazon EMR 与 Amazon Lake Formation 集成。

  Amazon EMR（集成 Lake Formation）目前已在 16 个Amazon区域推出：美国东部（俄亥俄和弗吉尼亚北部）、美国西部（加利福尼亚北部和俄勒冈）、亚太地区（孟买、首尔、新加坡、悉尼和东京）、加拿大（中部）、欧洲（法兰克福、爱尔兰、伦敦、巴黎和斯德哥尔摩）、南美洲（圣保罗）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 在使用 Amazon EMR 5.31.0 或 5.32.0 的集群上启用了 AtRestEncryption 或 HDFS 加密时，Hive 查询会导致以下运行时系统异常。

```
TaskAttempt 3 failed, info=[Error: Error while running task ( failure ) :
 attempt_1604112648850_0001_1_01_000000_3:java.lang.RuntimeException:
 java.lang.RuntimeException: Hive Runtime Error while closing
 operators: java.io.IOException: java.util.ServiceConfigurationError:
 org.apache.hadoop.security.token.TokenIdentifier: Provider
 org.apache.hadoop.hbase.security.token.AuthenticationTokenIdentifier not found
```

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 6.1.0

以下发布说明包括有关 Amazon EMR 发行版 6.1.0 的信息。更改与 6.0.0 有关。

首次发布日期：2020 年 9 月 4 日

上次更新日期：2020 年 10 月 15 日

支持的应用程序

- Amazon SDK for Java 1.11.828
- Flink 1.11.0
- Ganglia 3.7.2
- Hadoop 3.2.1-amzn-1
- HBase 2.2.5
- HBase-operator-tools 1.0.0
- HCatalog 3.1.2-amzn-0
- Hive 3.1.2-amzn-1
- Hudi 0.5.2-incubating
- Hue 4.7.1
- JupyterHub 1.1.0
- Livy 0.7.0
- MXNet 1.6.0
- Oozie 5.2.0
- Phoenix 5.0.0
- Presto 0.232
- PrestoSQL 338
- Spark 3.0.0-amzn-0
- TensorFlow 2.1.0
- Zeppelin 0.9.0-preview1
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.14.0

新特征

- 从 Amazon EMR 5.30.0 和 Amazon EMR 6.1.0 开始，支持 ARM 实例类型。
- 从 Amazon EMR 6.1.0 和 5.30.0.开始，支持 M6g 通用型实例类型。有关更多信息，请参阅《Amazon EMR 管理指南》中的支持的实例类型。

- 从 Amazon EMR 5.23.0 开始支持 EC2 置放群组功能，该功能可作为多主节点集群选项。目前，置放群组功能仅支持主节点类型，并会将 SPREAD 策略应用于这些主节点。SPREAD 策略将一小组实例放置在单独的基础硬件上，以防止发生硬件故障时出现多个主节点丢失的问题。有关更多信息，请参阅《Amazon EMR 管理指南》中的 EMR 与 EC2 置放群组的集成。
- 托管扩展 – 使用 Amazon EMR 版本 6.1.0 时，您可以启用 Amazon EMR 托管式自动扩缩功能，以根据工作负载自动增加或减少集群中实例或单位的数量。Amazon EMR 会持续评估集群指标，以便做出扩展决策，从而优化集群的成本和速度。Amazon EMR 5.30.0 及更高版本（但 6.0.0 除外）也提供了托管扩展。有关更多信息，请参阅《Amazon EMR 管理指南》中的扩缩集群资源。
- EMR 6.1.0 支持 PrestoSQL 338。有关更多信息，请参阅 Presto。
  - 仅在 EMR 6.1.0 及更高版本上支持 PrestoSQL，而 EMR 6.0.0 或 EMR 5.x 则不支持。
  - 可以继续使用应用程序名称 `Presto` 在集群上安装 PrestoDB。要在集群上安装 PrestoSQL，请使用应用程序名称 `PrestoSQL`。
  - 您可以安装 PrestoDB 或 PrestoSQL，但不能在同一个集群上同时安装两者。如果在尝试创建集群时同时指定了 PrestoDB 和 PrestoSQL，则会发生验证错误，而且集群创建请求失败。
  - 单主节点集群和多主节点集群均支持 PrestoSQL。在多主节点集群上，需要外部 Hive 元存储才能运行 PrestoSQL 或 PrestoDB。请参阅 Supported applications in an EMR cluster with multiple primary nodes。
- 支持在 Apache Hadoop 和 Apache Spark 上使用 Docker 对 ECR 进行自动身份验证：Spark 用户可以使用 Docker Hub 中的 Docker 镜像和 Amazon Elastic Container Registry（Amazon ECR）来定义环境和库依赖项。

  配置 Docker 和使用 Amazon EMR 6.x 通过 Docker 运行 Spark 应用程序。
- EMR 支持 Apache Hive ACID 事务：Amazon EMR 6.1.0 增加了对 Hive ACID 事务的支持，使其符合数据库的 ACID 属性。借助此功能，您可以使用 Amazon Simple Storage Service（Amazon S3）中的数据在 Hive 托管表中运行`INSERT`, `UPDATE`, `DELETE`,和`MERGE`操作。这是流式提取、数据重述、使用 MERGE 批量更新等使用案例的一项关键功能，并会缓慢更改维度。有关包括配置示例和使用案例在内的更多信息，请参阅 Amazon EMR 支持 Apache Hive ACID 事务。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。

- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。
- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- EMR 6.0.0 上不支持 Apache Fink，但集成了 Flink 1.11.0 的 EMR 6.1.0 可以支持 Apache Fink。这是首个正式支持 Hadoop 3 的 Fink 版本。请参阅 Apache Flink 1.11.0 发布公告。
- 默认 EMR 6.1.0 捆绑包中已经删除了 Ganglia。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 /etc/systemd/system/instance-controller.service，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- > **Important**
  >
  > Amazon EMR 6.1.0 和 6.2.0 包含可能严重影响所有 Hudi 插入、更新插入和删除操作的性能问题。如果您计划将 Hudi 与 Amazon EMR 6.1.0 或 6.2.0 结合使用，请联系 Amazon Support，获取 Hudi RPM 补丁。

- 如果使用 `spark.driver.extraJavaOptions` 和 `spark.executor.extraJavaOptions` 设置自定义垃圾回收配置，将会因为垃圾回收配置冲突导致 EMR 6.1 驱动程序/执行程序启动失败。使用 EMR 发行版 6.1.0 时，您应该使用属性 `spark.driver.defaultJavaOptions` 和 `spark.executor.defaultJavaOptions` 为驱动程序和执行程序指定自定义 Spark 垃圾回收配置。如要了解更多信息，请参阅 Apache Spark 运行时环境和在 Amazon EMR 6.1.0 上配置 Spark 垃圾回收。
- 在 Oozie 中使用 Pig（以及在 Hue 中，因为 Hue 使用 Oozie 操作来运行 Pig 脚本）会生成一个错误，即无法加载 native-lzo 库。此错误消息是信息性的，不会阻止 Pig 运行。
- Hudi 并发支持：目前 Hudi 不支持并发写入单个 Hudi 表。此外，Hudi 会回滚处于运行状态的写入器所做的所有更改后再允许新写入器启动。并发写入可能会干扰此机制并引入竞争条件，这会导致数据损坏。您应确保作为数据处理工作流程的一部分，任何时候都只有一个 Hudi 写入器对 Hudi 表进行操作。Hudi 支持多个并发读取器对同一 Hudi 表进行操作。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- Amazon EMR 6.1.0 中存在一个问题，会影响运行 Presto 的集群。在较长时间（天）后，集群可能会引发错误，例如“su: failed to execute /bin/bash: Resource temporarily unavailable”或“shell request failed on channel 0”。此问题是由内部 Amazon EMR 进程（InstanceController）产生过多的轻量级进程（LWP）导致的，这最终会导致 Hadoop 用户超出其 nproc 限制。这可以阻止用户打开其它进程。此问题的解决方案是：升级到 EMR 6.2.0。

发行版 6.0.0

以下发布说明包括有关 Amazon EMR 发行版 6.0.0 的信息。

首次发布日期：2020 年 3 月 10 日

支持的应用程序

- Amazon SDK for Java 1.11.711
- Ganglia 3.7.2
- Hadoop 3.2.1
- HBase 2.2.3
- HCatalog 3.1.2
- Hive 3.1.2
- Hudi 0.5.0-incubating
- Hue 4.4.0
- JupyterHub 1.0.0
- Livy 0.6.0
- MXNet 1.5.1
- Oozie 5.1.0
- Phoenix 5.0.0
- Presto 0.230

- Spark 2.4.4
- TensorFlow 1.14.0
- Zeppelin 0.9.0-SNAPSHOT
- Zookeeper 3.4.14
- 连接器和驱动程序：DynamoDB 连接器 4.14.0

> **Note**
>
> Flink、Sqoop、Pig 和 Mahout 在 Amazon EMR 6.0.0 中不可用。

新特征

- YARN Docker 运行时支持 - YARN 应用程序（例如 Spark 作业）现在可以在 Docker 容器的上下文中运行。这可让您轻松定义 Docker 镜像中的依赖项，而无需在 Amazon EMR 集群上安装自定义库。有关更多信息，请参阅配置 Docker 集成和使用 Amazon EMR 6.0.0 通过 Docker 运行 Spark 应用程序。
- Hive LLAP 支持 - Hive 现在支持 LLAP 执行模式以提高查询性能。有关更多信息，请参阅使用 Hive LLAP。

更改、增强功能和解决的问题

- 此版本旨在修复 Amazon EMR Scaling 无法成功纵向扩展/缩减集群或导致应用程序故障时出现的问题。
- 修复了当 Amazon EMR 集群上的进程守护程序正在进行运行状况检查活动（例如收集 YARN 节点状态和 HDFS 节点状态）时，针对高利用率的大型集群的扩展请求失败的问题。之所以发生这种情况，是因为集群上的进程守护程序无法将节点的运行状况数据传递给内部 Amazon EMR 组件。
- 改进了 EMR 集群上的进程守护程序，以便在重用 IP 地址时正确跟踪节点状态，从而提高扩缩操作期间的可靠性。
- SPARK-29683。修复了集群缩减期间出现任务失败的问题，因为 Spark 假定所有可用节点都被拒绝列出。
- YARN-9011。修复了集群尝试纵向扩展或缩减时，由于 YARN 停用中的争用条件导致任务失败的问题。
- 通过确保 Amazon EMR 集群上的进程守护程序和 YARN/HDFS 之间的节点状态始终一致，解决了集群扩展期间步骤或任务失败的问题。

- 修复了已启用 Kerberos 身份验证的 Amazon EMR 集群的诸如缩减和步骤提交等集群操作失败的问题。这是因为 Amazon EMR 集群上的进程守护程序没有续订 Kerberos 票证，而该票证是与主节点上运行的 HDFS/YARN 进行安全通信所必需的。
- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- Amazon Linux
  - Amazon Linux 2 是 EMR 6.x 发布版本系列的操作系统。
  - 使用 `systemd` 进行服务管理，而 Amazon Linux 1 中使用的是 `upstart`。
- Java 开发工具包 (JDK)
  - Coretto JDK 8 是 EMR 6.x 版本系列的默认 JDK。
- Scala
  - Scala 2.12 与 Apache Spark 和 Apache Livy 一起使用。
- Python 3
  - Python 3 现在是 EMR 中的默认 Python 版本。
- YARN 节点标注
  - 从 Amazon EMR 6.x 发行版系列开始，默认情况下禁用 YARN 节点标注功能。默认情况下，应用程序主进程可以在核心节点和任务节点上运行。您可以通过配置以下属性来启用 YARN 节点标注功能：`yarn.node-labels.enabled` 和 `yarn.node-labels.am.default-node-label-expression`。有关更多信息，请参阅了解主节点、核心节点和任务节点。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低

ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
```

```
sudo systemctl daemon-reload
```

- Spark 交互式 shell（包括 PySpark、SparkR 和 spark-shell）不支持将 Docker 与其它库一起使用。
- 要在 Amazon EMR 6.0.0 中使用 Python 3，您必须在 `yarn.nodemanager.env-whitelist` 中添加 PATH。
- 使用 Amazon Glue 数据目录作为 Hive 的元存储时，不支持 Live Long and Process（LLAP）功能。
- 将 Amazon EMR 6.0.0 与 Spark 和 Docker 集成使用时，您需要使用同一实例类型和相同数量的 EBS 卷配置集群中的实例，以避免在使用 Docker 运行时提交 Spark 任务时出现故障。
- 在 Amazon EMR 6.0.0 中，HBASE-24286 问题会影响 HBase on Amazon S3 存储模式。使用现有 S3 数据创建集群时，无法初始化 HBase 主服务器。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

发行版 5.30.1

以下发布说明包括有关 Amazon EMR 发行版 5.30.1 的信息。更改与 5.30.0 有关。

首次发布日期：2020 年 6 月 30 日

上次更新时间：2020 年 8 月 24 日

更改、增强功能和解决的问题

- 较新的 Amazon EMR 发行版修复了 Amazon EMR 中较早版本的 AL2 上“最大打开文件数”限制较低的问题。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本现在用更高的“最大打开文件数”设置永久修复了此问题。
- 修复了实例控制器进程生成无限量进程的问题。
- 修复了以下问题：Hue 无法运行 Hive 查询并显示“database is locked（数据库已锁定）”消息、阻止执行查询的问题。
- 修复了一个 Spark 问题，现在可以在 EMR 集群上同时运行更多任务。
- 修复了一个 Jupyter notebook 问题，该问题会导致 Jupyter 服务器中出现“too many files open error（打开过多文件错误）”。
- 修复了集群启动时间的问题。

新特征

- Amazon EMR 版本 6.x 和 EMR 版本 5.30.1 及更高版本提供了 Tez UI 和 YARN 时间线服务器持久性应用程序界面。无需通过 SSH 连接设置 Web 代理，访问永久性应用程序历史记录的一键式链接即可让您快速访问任务历史记录。活动和已终止集群的日志将在应用程序结束后保留 30 天。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看持久性应用程序用户界面。
- 可以使用 EMR Notebooks 执行 API 通过脚本或命令行来执行 EMR Notebooks。无需使用Amazon 控制台以编程方式控制 EMR Notebooks，即可启动、停止、列出和描述 EMR Notebooks 执行。借助参数化笔记本单元，您可以将不同的参数值传递给笔记本，而无需为每组新参数值创建笔记本副本。请参阅 EMR API 操作。有关示例代码，请参阅以编程方式执行 EMR Notebooks 的示例命令。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务

时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 `/etc/systemd/system/instance-controller.service`，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
```

```
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- EMR Notebooks

  在 EMR 版本 5.30.1 上，默认情况下禁用在集群主节点上安装内核和其他 Python 库的功能。有关此功能的更多信息，请参阅在集群主节点上安装内核和 Python 库。

  要启动此功能，请执行以下操作：

  1. 确保附加到 EMR Notebooks 服务角色的权限策略允许执行以下操作：

     ```
     elasticmapreduce:ListSteps
     ```

     有关更多信息，请参阅 EMR Notebooks 的服务角色。

  2. 使用 Amazon CLI 在集群上运行一个设置 EMR Notebooks 的步骤，如以下示例所示。将 *us-east-1* 替换为您的集群所在的区域。有关更多信息，请参阅使用 Amazon CLI 向集群中添加步骤。

     ```
     aws emr add-steps  --cluster-id MyClusterID --steps
      Type=CUSTOM_JAR,Name=EMRNotebooksSetup,ActionOnFailure=CONTINUE,Jar=s3://us-
     east-1.elasticmapreduce/libs/script-runner/script-runner.jar,Args=["s3://
     awssupportdatasvcs.com/bootstrap-actions/EMRNotebooksSetup/emr-notebooks-setup.sh"]
     ```

- 托管扩展

  在未安装 Presto 的 5.30.0 和 5.30.1 的集群上进行托管扩展操作可能会导致应用程序故障或导致统一的实例组或实例集处于 ARRESTED 状态，尤其是在缩减操作之后快速执行扩展操作时。

  解决方法是即使您的任务不需要 Presto，也可以在使用 Amazon EMR 发行版 5.30.0 和 5.30.1 创建集群时，将 Presto 选为要安装的应用程序。

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有其他 14 个非控制字符：`!"#$%&'()*+,-`。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

  解决方法是在 `spark-defaults` 分类中将 `spark.sql.sources.fastS3PartitionDiscovery.enabled` 配置设置为 `false`。

发行版 5.30.0

以下发布说明包括有关 Amazon EMR 发行版 5.30.0 的信息。更改与 5.29.0 有关。

首次发布日期：2020 年 5 月 13 日

上次更新日期：2020 年 6 月 25 日

升级

- 已将 Amazon SDK for Java 升级到版本 1.11.759

- 已将 Amazon SageMaker Spark SDK 升级到版本 1.3.0
- 已将 EMR 记录服务器升级到版本 1.6.0
- 已将 Flink 升级到版本 1.10.0
- 已将 Ganglia 升级到版本 3.7.2
- 已将 HBase 升级到版本 1.4.13
- 已将 Hudi 升级到版本 0.5.2-incubating
- 已将 Hue 升级到版本 4.6.0
- 已将 JupyterHub 升级到版本 1.1.0
- 已将升级 Livy 到版本 0.7.0-incubating
- 已将 Oozie 升级到版本 5.2.0
- 已将 Presto 升级到版本 0.232
- 已将 Spark 升级到版本 2.4.5
- 升级的连接器和驱动程序：Amazon Glue Connector 1.12.0；Amazon Kinesis Connector 3.5.0；EMR DynamoDB Connector 4.14.0

新特征

- EMR Notebooks – 与使用 5.30.0 创建的 EMR 集群结合使用时，EMR Notebooks 内核在集群上运行。这可以提高笔记本的性能，并允许您安装和自定义内核。您还可以在集群主节点上安装 Python 库。有关更多信息，请参阅《EMR 管理指南》中的安装并使用内核和库。
- 托管扩展 – 使用 Amazon EMR 版本 5.30.0 及更高版本时，您可以启用 EMR 托管扩展，以根据工作负载自动增加或减少集群中实例或单位的数量。Amazon EMR 会持续评估集群指标，以便做出扩展决策，从而优化集群的成本和速度。有关更多信息，请参阅《Amazon EMR 管理指南》中的扩缩集群资源。
- 加密 Amazon S3 中存储的日志文件 – 使用 Amazon EMR 版本 5.30.0 及更高版本时，您可以使用 Amazon KMS 客户管理的密钥对 Amazon S3 中存储的日志文件进行加密。有关更多信息，请参阅《Amazon EMR 管理指南》中的加密存储在 Amazon S3 中的日志文件。
- Amazon Linux 2 支持 – 在 EMR 版本 5.30.0 及更高版本中，EMR 使用 Amazon Linux 2 操作系统。新的自定义 AMI（Amazon Machine Image）必须基于 Amazon Linux 2 AMI。有关更多信息，请参阅使用自定义 AMI。
- Presto 正常自动扩展 – 使用 5.30.0 的 EMR 集群可以设置自动扩展超时时段，以便 Presto 任务在其节点停用之前有时间完成运行。有关更多信息，请参阅使用采用 Graceful Decommission 的 Presto 自动扩展配置。

- 使用新的分配策略选项创建队列实例 – EMR 版本 5.12.1 及更高版本中提供了一个新的分配策略选项。它加快了集群预置、提高了 Spot 分配的准确性并减少了竞价型实例中断。需要更新非默认 EMR 服务角色。请查看配置实例集。
- sudo systemctl stop 和 sudo systemctl start 命令 – 在 EMR 版本 5.30.0 及更高版本（使用 Amazon Linux 2 操作系统）中，EMR 使用 `sudo systemctl stop` 和 `sudo systemctl start` 命令重新启动服务。有关更多信息，请参阅如何在 Amazon EMR 中重新启动服务？

更改、增强功能和解决的问题

- 默认情况下，EMR 版本 5.30.0 不安装 Ganglia。您可以在创建集群时明确选择 Ganglia 进行安装。
- Spark 性能优化。
- Presto 性能优化。
- Amazon EMR 版本 5.30.0 及更高版本默认使用 Python 3。
- 用于私有子网中服务访问的默认托管安全组已使用新规则进行更新。如果使用自定义安全组进行服务访问，则必须包含与默认托管安全组相同的规则。有关详细信息，请参阅适用于服务访问（私有子网）的 Amazon EMR 托管安全组。如果您对 Amazon EMR 使用自定义服务角色，则必须向 `ec2:describeSecurityGroups` 授予权限，以便 EMR 可以验证安全组是否已正确创建。如果您使用 `EMR_DefaultRole`，则此权限已包含在默认托管式策略中。

已知问题

- 较早版本的 AL2 上“最大打开文件数”限制较低[此问题已在较新的发行版中修复]。Amazon EMR 发行版 emr-5.30.x、emr-5.31.0、emr-5.32.0、emr-6.0.0、emr-6.1.0 和 emr-6.2.0 基于较早版本的 Amazon Linux 2（AL2）。使用原定设置 AMI 创建 Amazon EMR 集群时，这些版本的“最大打开文件数”ulimit 设置较低。Amazon EMR 发行版 5.30.1、5.30.2、5.31.1、5.32.1、6.0.1、6.1.1、6.2.1、5.33.0、6.3.0 及更高版本使用更高的“最大打开文件数”设置永久修复了此问题。如果使用打开文件数限制较低的发行版，会在提交 Spark 任务时导致“Too many open files”（打开的文件过多）错误。在受影响的发行版中，Amazon EMR 原定设置 AMI 的原定设置“最大打开文件数”ulimit 为 4096，而最新版 Amazon Linux 2 AMI 中的文件限制数为 65536。Spark 驱动程序和执行程序尝试打开超过 4096 个文件时，“打开的最大文件数”的较低 ulimit 设置会导致 Spark 任务失败。要修复此问题，Amazon EMR 使用一个引导操作（BA）脚本，用于在创建集群时调整 ulimit 设置。

  如果您使用没有永久修复此问题的较早版本的 Amazon EMR，则可以通过下面的解决方法，显式将实例控制器 ulimit 设置为最多 65536 个文件。

从命令行显式设置 ulimit

1. 编辑 /etc/systemd/system/instance-controller.service，将以下参数添加到 Service (服务) 部分。

   ```
   LimitNOFILE=65536

   LimitNPROC=65536
   ```

2. 重新启动 InstanceController

   ```
   $ sudo systemctl daemon-reload

   $ sudo systemctl restart instance-controller
   ```

使用引导操作 (BA) 设置 ulimit

您还可以在创建集群时使用引导操作（BA）脚本将实例控制器 ulimit 配置为 65536 个文件。

```
#!/bin/bash
for user in hadoop spark hive; do
sudo tee /etc/security/limits.d/$user.conf << EOF
$user - nofile 65536
$user - nproc 65536
EOF
done
for proc in instancecontroller logpusher; do
sudo mkdir -p /etc/systemd/system/$proc.service.d/
sudo tee /etc/systemd/system/$proc.service.d/override.conf << EOF
[Service]
LimitNOFILE=65536
LimitNPROC=65536
EOF
pid=$(pgrep -f aws157.$proc.Main)
sudo prlimit --pid $pid --nofile=65535:65535 --nproc=65535:65535
done
sudo systemctl daemon-reload
```

- 托管扩展

  在未安装 Presto 的 5.30.0 和 5.30.1 的集群上进行托管扩展操作可能会导致应用程序故障或导致统一的实例组或实例集处于 ARRESTED 状态，尤其是在缩减操作之后快速执行扩展操作时。

解决方法是即使您的任务不需要 Presto，也可以在使用 Amazon EMR 发行版 5.30.0 和 5.30.1 创建集群时，将 Presto 选为要安装的应用程序。

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

- Hue 4.6.0 的默认数据库引擎是 SQLite，Hue 尝试使用外部数据库时，会引发问题。若要解决此问题，请在您的 `hue-ini` 配置分类中将 `engine` 设置为 `mysql`。Amazon EMR 版本 5.30.1 已修复这一问题。
- 当您将 Spark 与 Hive 分区位置格式化结合使用以读取 Amazon S3 中的数据，并在 Amazon EMR 版本 5.30.0 至 5.36.0 以及 6.2.0 至 6.9.0 上运行 Spark 时，可能会遇到导致集群无法正确读取数据的问题。如果您的分区具有以下所有特征，会发生这种情况：
  - 从同一个表扫描两个或多个分区。
  - 至少有一个分区目录路径是至少一个其他分区目录路径的前缀，例如，`s3://bucket/table/p=a` 是 `s3://bucket/table/p=a b` 的前缀。
  - 另一个分区目录中前缀后面的第一个字符的 UTF-8 值小于 `/` 字符 (U+002F)。例如，在 `s3://bucket/table/p=a b` 中，a 和 b 之间出现的空格字符 (U+0020) 就属于此类。请注意，还有

其他 14 个非控制字符：!"#$%&‘()*+,-。有关更多信息，请参阅 UTF-8 encoding table and Unicode characters（UTF-8 编码表和 Unicode 字符）。

解决方法是在 spark-defaults 分类中将 spark.sql.sources.fastS3PartitionDiscovery.enabled 配置设置为 false。

发行版 5.29.0

以下发布说明包括有关 Amazon EMR 发行版 5.29.0 的信息。更改与 5.28.1 有关。

首次发布日期：2020 年 1 月 17 日

升级

- 已将 Amazon SDK for Java 升级到版本 1.11.682
- 已将 Hive 升级到版本 2.3.6
- 已将 Flink 升级到版本 1.9.1
- 已将 EMRFS 升级到版本 2.38.0
- 已将 EMR DynamoDB 连接器升级到版本 4.13.0

更改、增强功能和解决的问题

- Spark
  - Spark 性能优化。
- EMRFS
  - 将管理指南更新为 emrfs-site.xml 默认设置以实现了一致视图。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.28.1

以下发布说明包括有关 Amazon EMR 发行版 5.28.1 的信息。更改与 5.28.0 有关。

首次发布日期：2020 年 1 月 10 日

更改、增强功能和解决的问题

- Spark
  - 修复了 Spark 兼容性问题。
- CloudWatch 指标
  - 修复了在具有多个主节点的 EMR 集群上发布的 Amazon CloudWatch Metrics。
- 已禁用日志消息
  - 已禁用假日志消息“...using old version (<4.5.8) of Apache http client”（使用低于版本 4.5.8 的 Apache http 客户端）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取

决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.28.0

以下发布说明包括有关 Amazon EMR 发行版 5.28.0 的信息。更改与 5.27.0 有关。

首次发布日期：2019 年 11 月 12 日

升级

- 已将 Flink 升级到版本 1.9.0
- 已将 Hive 升级到版本 2.3.6
- 已将 MXNet 升级到版本 1.5.1
- 已将 Phoenix 升级到版本 4.14.3
- 已将 Presto 升级到版本 0.227
- 已将 Zeppelin 升级到版本 0.8.2

新特征

- 创建集群时，Amazon EMR 现在可以安装 Apache Hudi。有关更多信息，请参阅Hudi。

- （2019 年 11 月 25 日）您现在可以选择并行运行多个步骤以提高集群利用率并节省成本。您还可以取消待处理和正在运行的步骤。有关更多信息，请参阅使用 Amazon CLI 和控制台执行步骤。
- （2019 年 12 月 3 日）您现在可以在 Amazon Outposts 上创建和运行 EMR 集群。Amazon Outposts 启用本地设施中的本地Amazon服务、基础设施和操作模型。在 Amazon Outposts 环境中，您可以使用与Amazon云中相同的 Amazon API、工具和基础设施。有关更多信息，请参阅 EMR clusters on Amazon Outposts。
- （2020 年 3 月 11 日）从 Amazon EMR 版本 5.28.0 开始，您可以在 Amazon Local Zones 子网上创建和运行 Amazon EMR 集群，作为支持的Amazon区域的逻辑扩展。本地区域使得 Amazon EMR 功能和Amazon服务的子集（如计算和存储服务）在位置上与用户更近，从而为本地运行的应用程序提供非常低的延迟访问。有关可用的 Local Zones 列表，请参阅 Amazon Local Zones。有关访问可用Amazon Local Zones 的信息，请参阅区域、可用区和 Local Zones。

  Local Zones 目前不支持 Amazon EMR Notebooks，也不支持使用接口 VPC 终端节点（Amazon PrivateLink）直接连接到 Amazon EMR。

更改、增强功能和解决的问题

- 扩展了对高可用性集群的应用程序支持
  - 有关更多信息，请参阅 Amazon EMR Management Guide 中的 Supported applications in an EMR cluster with Multiple Primary Nodes。
- Spark
  - 性能优化
- Hive
  - 性能优化
- Presto
  - 性能优化

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.27.0

以下发布说明包括有关 Amazon EMR 发行版 5.27.0 的信息。更改与 5.26.0 有关。

首次发布日期：2019 年 9 月 23 日

升级

- Amazon SDK for Java 1.11.615
- Flink 1.8.1
- JupyterHub 1.0.0
- Spark 2.4.4
- Tensorflow 1.14.0
- 连接器和驱动程序：
  - DynamoDB 连接器 4.12.0

新特征

- （2019 年 10 月 24 日）所有 Amazon EMR 版本均在 EMR Notebooks 中提供以下新功能。

  - 您可以将 Git 存储库与 EMR Notebooks 关联，以将笔记本存储在版本控制的环境中。您可以通过远程 Git 存储库与同行共享代码，并重复使用现有的 Jupyter notebook。有关更多信息，请参阅《Amazon EMR 管理指南》中的将 Git 存储库与 Amazon EMR Notebooks 关联。
  - nbdime 实用工具现在可在 EMR Notebooks 中使用，简化笔记本比较和合并。
  - EMR Notebooks 现在支持 JupyterLab。JupyterLab 是一个基于 Web 的交互式开发环境，与 Jupyter notebook 完全兼容。现在，您可以选择在 JupyterLab 或 Jupyter notebook 编辑器中打开笔记本。
- （2019 年 10 月 30 日）借助 Amazon EMR 5.25.0 版及更高版本，您可以从控制台中的集群 Summary (摘要) 页面或 Application history (应用程序历史记录) 选项卡连接到 Spark 历史记录服务器 UI。您可以快速访问 Spark 历史记录服务器 UI，来查看应用程序指标并访问活动集群和终止集群的相关日志文件，而无需通过 SSH 连接设置 Web 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群外访问持久性应用程序用户界面。

更改、增强功能和解决的问题

- 具有多个主节点的 Amazon EMR 集群
  - 您可以在具有多个主节点的 Amazon EMR 集群上安装和运行 Flink。有关更多信息，请参阅支持的应用程序和功能。
  - 您可以在具有多个主节点的 Amazon EMR 集群上配置 HDFS 透明加密。有关更多信息，请参阅 HDFS Transparent Encryption on EMR clusters with Multiple Primary Nodes。
  - 现在，您可以修改在具有多个主节点的 Amazon EMR 集群上运行的应用程序的配置。有关更多信息，请参阅为运行的集群中的实例组提供配置。
- Amazon EMR-DynamoDB 连接器
  - Amazon EMR-DynamoDB 连接器现在支持以下 DynamoDB 数据类型：布尔值、列表、映射、项目、空值。有关更多信息，请参阅设置 Hive 表以运行 Hive 命令。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

解决办法：

- 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
- 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

```
kinit -kt <keytab_file> <principal>
```

通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.26.0

以下发布说明包括有关 Amazon EMR 发行版 5.26.0 的信息。更改与 5.25.0 有关。

首次发布日期：2019 年 8 月 8 日

上次更新日期：2019 年 8 月 19 日

升级

- Amazon SDK for Java 1.11.595
- HBase 1.4.10
- Phoenix 4.14.2
- 连接器和驱动程序：
  - DynamoDB 连接器 4.11.0
  - MariaDB 连接器 2.4.2
  - Amazon Redshift JDBC 驱动程序 1.2.32.1056

新特征

- （测试版）借助 Amazon EMR 5.26.0，您可以启动与 Lake Formation 集成的集群。此集成提供了对 Amazon Glue 数据目录中的数据库和表的精细列级别访问。它还支持从企业身份系统通过联合单点登录的方式登录 EMR Notebooks 或 Apache Zeppelin。有关更多信息，请参阅将 Amazon EMR 与 Amazon Lake Formation 集成（测试版）。
- （2019 年 8 月 19 日）所有支持安全组的 Amazon EMR 发行版现在均可提供 Amazon EMR 阻止公有访问功能。可在账户上设置阻止公有访问，适用于所有Amazon区域。如果与集群关联的任何安全组具有一个允许某端口上来自 IPv4 0.0.0.0/0 或 IPv6 ::/0（公有访问）的入站流量的规则，阻止公有访问将阻止集群启动，除非将该端口指定为例外。默认情况下，端口 22 是一个例外。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Amazon EMR 阻止公有访问。

更改、增强功能和解决的问题

- EMR Notebooks
  - 在 EMR 5.26.0 及更高版本中，EMR Notebooks 除了默认的 Python 库外，还支持笔记本范围的 Python 库。无需重新创建集群或重新将笔记本附加到集群，您即可从笔记本编辑器中安装笔记本范围的库。笔记本范围的库是在 Python 虚拟环境中创建的，因此适用于当前笔记本会话。这使得您可以隔离笔记本依赖项。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用笔记本范围的库。
- EMRFS
  - 您可以通过以下方式启用 ETag 验证功能（测试版）：将 `fs.s3.consistent.metadata.etag.verification.enabled` 设置为 `true`。启用后，EMRFS 使用 Amazon S3 ETag 验证所读取的对象是否为最新可用版本。此功能对更新后读取使用案例很有帮助，在这些案例中，将覆盖 Amazon S3 上的文件但保留相同名称。此 ETag 验证功能当前不可用于 S3 Select。有关更多信息，请参阅配置统一视图。
- Spark
  - 现在，默认情况下启用以下优化：动态分区修剪、DISTINCT before INTERSECT、改进了 JPIN（后跟 DISTINCT 查询）的 SQL 计划统计数据推理、展平标量子查询、优化的连接重排序和 Bloom 筛选条件连接。有关更多信息，请参阅优化 Spark 性能。
  - 改进了排序合并连接的整个阶段代码生成。
  - 改进了查询片段和子查询重用。
  - 改进了 Spark 启动时的预分配执行程序。
  - 连接的较小侧包含广播提示时，不再应用 Bloom 筛选条件连接。
- Tez

- 已解决 Tez 中存在的问题。Tez UI 现可用于具有多个主节点的 Amazon EMR 集群。

已知问题

- 改进的“排序合并连接的整个阶段代码生成”功能在启用后会增加内存压力。此优化可提高性能，但如果 `spark.yarn.executor.memoryOverheadFactor` 未调整，不能提供足够的内存，则会导致任务重试或失败。要禁用此功能，请将 `spark.sql.sortMergeJoinExec.extendedCodegen.enabled` 设置为 false。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.25.0

以下发布说明包括有关 Amazon EMR 发行版 5.25.0 的信息。更改与 5.24.1 有关。

首次发布日期：2019 年 7 月 17 日

上次更新日期：2019 年 10 月 30 日

Amazon EMR 5.25.0

升级

- Amazon SDK for Java 1.11.566
- Hive 2.3.5
- Presto 0.220
- Spark 2.4.3
- TensorFlow 1.13.1
- Tez 0.9.2
- Zookeeper 3.4.14

新特征

- （2019 年 10 月 30 日）从 Amazon EMR 版本 5.25.0 开始，您可以从控制台中的集群 Summary (摘要) 页面或 Application history (应用程序历史记录) 选项卡连接到 Spark 历史记录服务器 UI。您可以快速访问 Spark 历史记录服务器 UI，来查看应用程序指标并访问活动集群和终止集群的相关日志文件，而无需通过 SSH 连接设置 Web 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群外访问持久性应用程序用户界面。

更改、增强功能和解决的问题

- Spark
  - 通过使用 Bloom 筛选条件预筛选输入，提高了某些连接的性能。默认情况下，优化处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.bloomFilterJoin.enabled` 设置为 `true`。
  - 改进了按字符串类型列分组的性能。
  - 改进了未安装 HBase 的集群 R4 实例类型的默认 Spark 执行程序内存和内核配置。
  - 解决了动态分区修剪功能之前存在的一个问题，即修剪的表必须位于联接的左侧。
  - 改进了 DISTINCT before INTERSECT 优化，以应用于涉及别名的其它情况。
  - 改进了 JOIN（后跟 DISTING 查询）的 SQL 计划统计数据推理。默认情况下，该改进处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.statsImprovements.enabled` 设置为 `true`。此优化是“Distinct before Intersect”功能所需的，将 `spark.sql.optimizer.distinctBeforeIntersect.enabled` 设置为 `true` 时将自动启用。

  - 根据表格大小和筛选条件优化了联接顺序。默认情况下，该优化处于禁用状态，但可以通过以下方式启用：将 Spark 配置参数 `spark.sql.optimizer.sizeBasedJoinReorder.enabled` 设置为 `true`。

  有关更多信息，请参阅优化 Spark 性能。

- EMRFS
  - 现在，EMRFS 设置 `fs.s3.buckets.create.enabled` 默认处于禁用状态。通过测试，我们发现禁用此设置可提高性能并可防止意外创建 S3 存储桶。如果您的应用程序需使用此功能，则可以通过以下方式启用：将 `emrfs-site` 配置分类中的 `fs.s3.buckets.create.enabled` 设置为 `true`。有关更多信息，请参阅在创建集群时提供配置。
- 安全配置中的本地磁盘加密和 S3 加密改进（2019 年 8 月 5 日）
  - 在安全配置设置中将 Amazon S3 加密设置与本地磁盘加密设置分开。
  - 发行版 5.24.0 及更高版本中添加了一个选项，可启用 EBS 加密。选择此选项后，除了存储卷之外，还会加密根设备卷。之前的版本需要使用自定义 AMI 来加密根设备卷。
  - 有关更多信息，请参阅《Amazon EMR 管理指南》中的加密选项。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.24.1

以下发布说明包括有关 Amazon EMR 发行版 5.24.1 的信息。更改与 5.24.0 有关。

首次发布日期：2019 年 6 月 26 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.24.0

以下发布说明包括有关 Amazon EMR 发行版 5.24.0 的信息。更改与 5.23.0 有关。

首次发布日期：2019 年 6 月 11 日

上次更新时间：2019 年 8 月 5 日

升级

- Flink 1.8.0
- Hue 4.4.0
- JupyterHub 0.9.6
- Livy 0.6.0
- MxNet 1.4.0
- Presto 0.219
- Spark 2.4.2
- Amazon SDK for Java 1.11.546
- 连接器和驱动程序：
  - DynamoDB 连接器 4.9.0
  - MariaDB 连接器 2.4.1
  - Amazon Redshift JDBC 驱动程序 1.2.27.1051

更改、增强功能和解决的问题

- Spark
  - 添加了对动态修剪分区的优化。默认情况下禁用优化。要启用该优化，请将 Spark 参数 `spark.sql.dynamicPartitionPruning.enabled` 设置为 `true`。

- 改进了 INTERSECT 查询的性能。默认情况下禁用此优化。要启用该优化，请将 Spark 参数 `spark.sql.optimizer.distinctBeforeIntersect.enabled` 设置为 `true`。
- 添加了对展平标量子查询的优化，可使用相同关系进行聚合。默认情况下禁用优化。要启用该优化，请将 Spark 参数 `spark.sql.optimizer.flattenScalarSubqueriesWithAggregates.enabled` 设置为 `true`。
- 改进了整个阶段代码生成。

有关更多信息，请参阅优化 Spark 性能。

- 安全配置中的本地磁盘加密和 S3 加密改进（2019 年 8 月 5 日）
  - 在安全配置设置中将 Amazon S3 加密设置与本地磁盘加密设置分开。
  - 添加了一个启用 EBS 加密的选项。选择此选项后，除了存储卷之外，还会加密根设备卷。之前的版本需要使用自定义 AMI 来加密根设备卷。
  - 有关更多信息，请参阅《Amazon EMR 管理指南》中的加密选项。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

> **Note**
>
> 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.23.0

以下发布说明包括有关 Amazon EMR 发行版 5.23.0 的信息。更改与 5.22.0 有关。

首次发布日期：2019 年 4 月 1 日

上次更新时间：2019 年 4 月 30 日

升级

- Amazon SDK for Java 1.11.519

新特征

- （2019 年 4 月 30 日）使用 Amazon EMR 5.23.0 及更高版本，您可以启动具有三个主节点的集群，以支持应用程序（如 YARN Resource Manager、HDFS NameNode、Spark、Hive 和 Ganglia）的高可用性。使用此功能，主节点不再发生潜在的单点故障。如果其中一个主节点出现故障，Amazon EMR 会自动故障转移到备用主节点，并将出现故障的主节点替换为具有相同配置和引导操作的新主节点。有关更多信息，请参阅计划和配置主节点。

已知问题

- Tez UI（已在 Amazon EMR 发行版 5.26.0 中修复）

  Tez UI 不能在具有多个主节点的 EMR 集群上运行。
- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    ```

```
system?user.name=hue&doAs=administrator&wt=json (Caused by
NewConnectionError(': Failed to establish a new connection: [Errno 111]
Connection refused',))
```

要防止显示 Solr 错误消息:

1. 使用 SSH 连接到主节点命令行。
2. 使用文本编辑器打开 hue.ini 文件。例如：

   ```
   sudo vim /etc/hue/conf/hue.ini
   ```

3. 搜索术语 appblacklist，并将该行修改为以下内容：

   ```
   appblacklist = search
   ```

4. 保存更改并重新启动 Hue，如以下示例所示：

   ```
   sudo stop hue; sudo start hue
   ```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 /etc/hadoop.keytab，而 principal 为 hadoop/<hostname>@<REALM> 格式。

    > **Note**
    >
    > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.22.0

以下发布说明包括有关 Amazon EMR 发行版 5.22.0 的信息。更改与 5.21.0 有关。

> **Important**
>
> 从 Amazon EMR 发行版 5.22.0 开始，Amazon EMR 专门使用 Amazon 签名版本 4 来对 Amazon S3 的请求进行身份验证。除非发布说明指出需专门使用签名版本 4，否则早期 Amazon EMR 发行版在某些情况下使用 Amazon 签名版本 2。有关更多信息，请参阅《Amazon Simple Storage Service 开发人员指南》中的对请求进行身份验证（Amazon签名版本 4）和对请求进行身份验证（Amazon签名版本 2）。

首次发布日期：2019 年 3 月 20 日

升级

- Flink 1.7.1
- HBase 1.4.9
- Oozie 5.1.0
- Phoenix 4.14.1
- Zeppelin 0.8.1
- 连接器和驱动程序：
  - DynamoDB 连接器 4.8.0
  - MariaDB 连接器 2.2.6
  - Amazon Redshift JDBC 驱动程序 1.2.20.1043

新特征

- 修改了仅限 EBS 存储的 EC2 实例类型的默认 EBS 配置。在使用 Amazon EMR 发行版 5.22.0 及更高版本创建集群时，默认 EBS 存储量根据实例大小而增加。此外，我们将增加的存储拆分到多个卷，从而提高了 IOPS 性能。如果要使用不同的 EBS 实例存储配置，您可以在创建 EMR 集群或将节点添加到现有集群时指定该配置。有关每个实例类型默认分配的存储容量和卷数的更多信息，请参阅《Amazon EMR 管理指南》中的实例的默认 EBS 存储。

更改、增强功能和解决的问题

- Spark
  - 在 YARN 上引入了一个新的配置属性 `spark.yarn.executor.memoryOverheadFactor`。此属性的值是一个缩放系数，它将内存开销值设置为执行程序内存的百分比，最小为 384 MB。如果内存开销设置为使用 `spark.yarn.executor.memoryOverhead`，则此属性不发挥任何作用。默认值为 `0.1875`，表示 18.75%。与 Spark 内部设置的 10% 的默认值相比，Amazon EMR 的此默认值在 YARN 容器中为执行器内存开销预留了更多空间。根据经验，Amazon EMR 默认值 18.75% 表明 TPC-DS 基准测试中与内存相关的故障较少。
  - 为了改进性能，已逆向移植 SPARK-26316。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

    1. 使用 SSH 连接到主节点命令行。
    2. 使用文本编辑器打开 `hue.ini` 文件。例如：

       ```
       sudo vim /etc/hue/conf/hue.ini
       ```

    3. 搜索术语 `appblacklist`，并将该行修改为以下内容：

       ```
       appblacklist = search
       ```

4. 保存更改并重新启动 Hue，如以下示例所示：

```
sudo stop hue; sudo start hue
```

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

发布版本 5.21.1

以下发布说明包括有关 Amazon EMR 发行版 5.21.1 的信息。更改与 5.21.0 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

已知问题

- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.21.0

以下发布说明包括有关 Amazon EMR 发行版 5.21.0 的信息。更改与 5.20.0 有关。

首次发布日期：2019 年 2 月 18 日

上次更新时间：2019 年 4 月 3 日

升级

- Flink 1.7.0
- Presto 0.215
- Amazon SDK for Java 1.11.479

新特征

- （2019 年 4 月 3 日）对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

更改、增强功能和解决的问题

- Zeppelin
  - 已逆向移植 ZEPPELIN-3878。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

    1. 使用 SSH 连接到主节点命令行。
    2. 使用文本编辑器打开 hue.ini 文件。例如：

       ```
       sudo vim /etc/hue/conf/hue.ini
       ```

    3. 搜索术语 appblacklist，并将该行修改为以下内容：

       ```
       appblacklist = search
       ```

    4. 保存更改并重新启动 Hue，如以下示例所示：

       ```
       sudo stop hue; sudo start hue
       ```

- Tez
  - 此问题已在 Amazon EMR 5.22.0 中得到修复。

    通过 http://*MasterDNS*:8080/tez-ui 连接到 Tez UI 时（通过 SSH 连接到集群主节点），显示错误“Adapter operation failed - Timeline server (ATS) is out of reach。Either it is down, or CORS is not enabled”，或任务不正常地显示为“N/A”。

    这是由于 Tez UI 使用 `localhost`（而没有使用主节点的主机名称）向 YARN 时间线服务器发出请求所致。解决方法：将脚本作为引导操作或步骤运行。脚本更新 Tez `configs.env` 文件中的主机名。有关更多信息以及脚本的位置信息，请参阅引导说明。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 `hadoop` 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 `hadoop` 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.20.0

以下发布说明包括有关 Amazon EMR 发行版 5.20.0 的信息。更改与 5.19.0 有关。

首次发布日期：2018 年 12 月 18 日

上次更新时间：2019 年 1 月 22 日

升级

- Flink 1.6.2
- HBase 1.4.8
- Hive 2.3.4
- Hue 4.3.0
- MXNet 1.3.1
- Presto 0.214
- Spark 2.4.0
- TensorFlow 1.12.0
- Tez 0.9.1
- Amazon SDK for Java 1.11.461

新特征

- （2019 年 1 月 22 日）Amazon EMR 中的 Kerberos 已经得到改进，现在可支持对来自外部 KDC 的委托人进行身份验证。这集中了委托人管理，因为多个集群可以共享单个外部 KDC。此外，外部 KDC 可与 Active Directory 域建立跨领域信任关系。这使得所有集群可以从 Active Directory 对委托人进行身份验证。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Kerberos 身份验证。

更改、增强功能和解决的问题

- Amazon EMR 的默认 Amazon Linux AMI
  - Python 3 软件包已从 Python 3.4 升级到 3.6。
- 经 EMRFS S3 优化的提交程序
  - 现在，已默认启用经 EMRFS S3 优化的提交程序，从而改进写入性能。有关更多信息，请参阅使用经 EMRFS S3 优化的提交程序。

- Hive
  - 已逆向移植 HIVE-16686。
- 集成 Spark 和 Hive 的 Glue
  - 在 EMR 5.20.0 或更高版本中，当使用 Amazon Glue 数据目录作为元存储时，会自动为 Spark 和 Hive 启用并行分区修剪。此更改通过并行执行多个请求来检索分区，显著缩短查询计划时间。可同时执行的分段总数介于 1 到 10 之间。默认值为 5，这是建议的设置。您可以通过以下方式更改该值：指定 `hive-site` 配置分类中的属性 `aws.glue.partition.num.segments`。如果发生节流，则可以通过将值更改为 1 来关闭此功能。有关更多信息，请参阅 Amazon Glue 分段结构。

已知问题

- Hue（已在 Amazon EMR 发行版 5.24.0 中修复）
  - 在 Amazon EMR 上运行的 Hue 不支持 Solr。从 Amazon EMR 发行版 5.20.0 开始，配置错误问题会导致 Solr 启用，并显示类似于以下内容的无害错误消息：

    ```
    Solr server could not be contacted properly:
    HTTPConnectionPool('host=ip-xx-xx-xx-xx.ec2.internal',
    port=1978): Max retries exceeded with url: /solr/admin/info/
    system?user.name=hue&doAs=administrator&wt=json (Caused by
    NewConnectionError(': Failed to establish a new connection: [Errno 111]
    Connection refused',))
    ```

    要防止显示 Solr 错误消息:

    1. 使用 SSH 连接到主节点命令行。
    2. 使用文本编辑器打开 `hue.ini` 文件。例如：

       ```
       sudo vim /etc/hue/conf/hue.ini
       ```

    3. 搜索术语 `appblacklist`，并将该行修改为以下内容：

       ```
       appblacklist = search
       ```

    4. 保存更改并重新启动 Hue，如以下示例所示：

       ```
       sudo stop hue; sudo start hue
       ```

- Tez
  - 此问题已在 Amazon EMR 5.22.0 中得到修复。

通过 http://*MasterDNS*:8080/tez-ui 连接到 Tez UI 时（通过 SSH 连接到集群主节点），显示错误“Adapter operation failed - Timeline server (ATS) is out of reach。Either it is down, or CORS is not enabled”，或任务不正常地显示为“N/A”。

这是由于 Tez UI 使用 `localhost`（而没有使用主节点的主机名称）向 YARN 时间线服务器发出请求所致。解决方法：将脚本作为引导操作或步骤运行。脚本更新 Tez `configs.env` 文件中的主机名。有关更多信息以及脚本的位置信息，请参阅引导说明。

- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标注存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。
- 具有多个主节点的集群和 Kerberos 身份验证中的已知问题

  如果在 Amazon EMR 版本 5.20.0 及更高版本中运行具有多个主节点的集群和 Kerberos 身份验证，则在集群运行一段时间后，您可能在执行集群操作（如缩减或步骤提交）时遇到问题。具体时间段取决于您定义的 Kerberos 票证有效期。缩减问题会影响您提交的自动缩减和显式缩减请求。其它集群操作也可能会受到影响。

  解决办法：

  - 以 hadoop 用户身份通过 SSH 连接到具有多个主节点的 EMR 集群的 lead 主节点。
  - 运行以下命令，为 hadoop 用户续订 Kerberos 票证。

    ```
    kinit -kt <keytab_file> <principal>
    ```

    通常情况下，keytab 文件位于 `/etc/hadoop.keytab`，而 principal 为 `hadoop/<hostname>@<REALM>` 格式。

  > **Note**
  >
  > 此解决方法将在 Kerberos 票证有效期内生效。默认情况下，此持续时间为 10 个小时，但可以通过 Kerberos 设置进行配置。Kerberos 票证过期后，您必须重新运行上述命令。

版本 5.19.0

以下发布说明包括有关 Amazon EMR 发行版 5.19.0 的信息。更改与 5.18.0 有关。

首次发布日期：2018 年 11 月 7 日

上次更新时间：2018 年 11 月 19 日

升级

- Hadoop 2.8.5
- Flink 1.6.1
- JupyterHub 0.9.4
- MXNet 1.3.0
- Presto 0.212
- TensorFlow 1.11.0
- Zookeeper 3.4.13
- Amazon SDK for Java 1.11.433

新特征

- （2018 年 11 月 19 日）EMR Notebooks 是基于 Jupyter notebook 的托管环境。它支持适用于 PySpark、Spark SQL、Spark R 和 Scala 的 Spark magic 内核。EMR Notebooks 可在使用 Amazon EMR 发行版 5.18.0 及更高版本创建的集群上使用。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 EMR Notebooks。
- 使用 Spark 和 EMRFS 编写 Parquet 文件时，可以使用经 EMRFS S3 优化的提交程序。此提交程序改进了写入性能。有关更多信息，请参阅使用经 EMRFS S3 优化的提交程序。

更改、增强功能和解决的问题

- YARN
  - 修改了限制应用程序主进程在核心节点上运行的逻辑。此功能现在可使用 `yarn-site` 和 `capacity-scheduler` 配置分类中的 YARN 节点标注功能和属性。有关信息，请参阅。https://docs.amazonaws.cn/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html#emr-plan-spot-YARN.
- Amazon EMR 的默认 Amazon Linux AMI
  - 默认情况下，不再安装 `ruby18`、`php56` 和 `gcc48`。如果需要，可以使用 `yum` 安装它们。
  - 默认情况下，不再安装 aws-sdk ruby gem。如果需要，可以使用 `gem install aws-sdk` 进行安装。此外，还可以安装特定组件。例如，`gem install aws-sdk-s3`。

已知问题

- EMR Notebooks – 在某些情况下，打开多个笔记本编辑器时，笔记本编辑器可能无法连接到集群。如果发生这种情况，请清除浏览器 Cookie，然后重新打开笔记本编辑器。
- CloudWatch ContainerPending 指标和自动伸缩 – （已在 5.20.0 中修复）Amazon EMR 可能会发出一个 `ContainerPending` 负值。如果在自动伸缩规则中使用 `ContainerPending`，自动伸缩的行为方式可能会不符合预期。请避免在自动伸缩中使用 `ContainerPending`。
- 在 Amazon EMR 版本 5.19.0、5.20.0 和 5.21.0 中，YARN 节点标注存储在 HDFS 目录中。在某些情况下，这会导致核心节点启动延迟，然后导致集群超时和启动失败。从 Amazon EMR 5.22.0 开始，此问题已得到解决。YARN 节点标签存储在每个集群节点的本地磁盘上，避免了对 HDFS 的依赖。

版本 5.18.0

以下发布说明包括有关 Amazon EMR 发行版 5.18.0 的信息。更改与 5.17.0 有关。

首次发布日期：2018 年 10 月 24 日

升级

- Flink 1.6.0
- HBase 1.4.7
- Presto 0.210
- Spark 2.3.2
- Zeppelin 0.8.0

新特征

- 您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发行版 5.18.0 开始）附带的准确版本的库和依赖项的任务代码。有关更多信息，请参阅使用 Amazon EMR 项目存储库检查依赖项。

更改、增强功能和解决的问题

- Hive

  - 添加了对 S3 Select 的支持。有关更多信息，请参阅将 S3 Select 与 Hive 结合使用以提高查询性能。
- Presto
  - 添加了对 S3 Select Pushdown 的支持。有关更多信息，请参阅使用 S3 Select Pushdown 搭配 Presto 提高性能。
- Spark
  - Spark 的默认 log4j 配置已更改为 Spark Streaming 任务每小时的滚动容器日志。这有助于防止删除长时间运行的 Spark Streaming 任务的日志。

发布版本 5.17.1

以下发布说明包括有关 Amazon EMR 发行版 5.17.1 的信息。更改与 5.17.0 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

版本 5.17.0

以下发布说明包括有关 Amazon EMR 发行版 5.17.0 的信息。更改与 5.16.0 有关。

首次发布日期：2018 年 8 月 30 日

升级

- Flink 1.5.2
- HBase 1.4.6
- Presto 0.206

新特征

- 添加了对 Tensorflow 的支持。有关更多信息，请参阅TensorFlow。

更改、增强功能和解决的问题

- JupyterHub
  - Amazon S3 中添加了对笔记本持久性的支持。有关更多信息，请参阅在 Amazon S3 中配置笔记本的持久性。
- Spark
  - 添加了对 S3 Select 的支持。有关更多信息，请参阅将 S3 Select 与 Spark 结合使用以提高查询性能。
- 解决了 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中 Cloudwatch 指标和自动伸缩功能中存在的问题。

已知问题

- 创建使用 Kerberos 的集群时，如果安装了 Livy，Livy 将失败，并显示未启用简单身份验证的错误。重新启动 Livy 服务器可解决此问题。解决方法是在集群创建过程中添加一个在主节点上运行 `sudo restart livy-server` 的步骤。
- 如果您使用基于 Amazon Linux AMI（创建日期为 2018-08-11）的自定义 Amazon Linux AMI，则 Oozie 服务器无法启动。如果您使用 Oozie，请根据具有不同创建日期的 Amazon Linux AMI ID 创建自定义 AMI。您可以使用以下 Amazon CLI 命令返回所有 2018.03 版本的 HVM Amazon Linux AMI 的镜像 ID 列表以及发布日期，以便您可以根据需要选择合适的 Amazon Linux AMI。将 MyRegion 替换为您的区域标识符，如 us-west-2。

```
aws ec2 --region MyRegion describe-images --owner amazon --query 'Images[?
Name!=`null`]|[?starts_with(Name, `amzn-ami-hvm-2018.03`) == `true`].
[CreationDate,ImageId,Name]' --output text | sort -rk1
```

版本 5.16.0

以下发布说明包括有关 Amazon EMR 发行版 5.16.0 的信息。更改与 5.15.0 有关。

首次发布日期：2018 年 7 月 19 日

升级

- Hadoop 2.8.4
- Flink 1.5.0

- Livy 0.5.0
- MXNet 1.2.0
- Phoenix 4.14.0
- Presto 0.203
- Spark 2.3.1
- Amazon SDK for Java 1.11.336
- CUDA 9.2
- Redshift JDBC 驱动程序 1.2.15.1025

更改、增强功能和解决的问题

- HBase
  - 已逆向移植 HBASE-20723。
- Presto
  - 更改了配置，可支持 LDAP 身份验证。有关更多信息，请参阅为 Presto on Amazon EMR 使用 LDAP 身份验证。
- Spark
  - Apache Spark 版本 2.3.1（从 Amazon EMR 发行版 5.16.0 开始提供）解决了 CVE-2018-8024 和 CVE-2018-1334 问题。建议您将 Spark 的早期版本迁移到 Spark 2.3.1 版本或更高版本。

已知问题

- 此发行版不支持 c1.medium 或 m1.small 实例类型。使用这些实例类型的集群将无法启动。解决方法：指定其它实例类型或使用其它发行版。
- 创建使用 Kerberos 的集群时，如果安装了 Livy，Livy 将失败，并显示未启用简单身份验证的错误。重新启动 Livy 服务器可解决此问题。解决方法是在集群创建过程中添加一个在主节点上运行 `sudo restart livy-server` 的步骤。
- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

版本 5.15.0

以下发布说明包括有关 Amazon EMR 发行版 5.15.0 的信息。更改与 5.14.0 有关。

首次发布日期：2018 年 6 月 21 日

升级

- 已将 HBase 升级到 1.4.4
- 已将 Hive 升级到 2.3.3
- 已将 Hue 升级到 4.2.0
- 已将 Oozie 升级到 5.0.0
- 已将 Zookeeper 升级到 3.4.12
- 已将 Amazon SDK 升级到 1.11.333

更改、增强功能和解决的问题

- Hive
  - 已逆向移植 HIVE-18069。
- Hue
  - 更新了 Hue，启用 Kerberos 后可以使用 Livy 正确地进行身份验证。现在，在 Amazon EMR 中使用 Kerberos 时，支持 Livy。
- JupyterHub
  - 更新了 JupyterHub，因此 Amazon EMR 默认安装 LDAP 客户端库。
  - 修复了生成自签名凭证的脚本中的错误。

已知问题

- 此发行版不支持 c1.medium 或 m1.small 实例类型。使用这些实例类型的集群将无法启动。解决方法：指定其它实例类型或使用其它发行版。
- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

版本 5.14.1

以下发布说明包括有关 Amazon EMR 发行版 5.14.1 的信息。更改与 5.14.0 有关。

首次发布日期：2018 年 10 月 17 日

更新了 Amazon EMR 的默认 AMI，解决了潜在的安全漏洞。

版本 5.14.0

以下发布说明包括有关 Amazon EMR 发行版 5.14.0 的信息。更改与 5.13.0 有关。

首次发布日期：2018 年 6 月 4 日

升级

- 已将 Apache Flink 升级到 1.4.2
- 已将 Apache MXnet 升级到 1.1.0
- 已将 Apache Sqoop 升级到 1.4.7

新特征

- 添加了对 JupyterHub 的支持。有关更多信息，请参阅JupyterHub。

更改、增强功能和解决的问题

- EMRFS
  - 更新了对 Amazon S3 的 userAgent 字符串请求，更新为包含调用委托人的用户和组信息。这可以与 Amazon CloudTrail 日志结合使用，来获取更全面的请求跟踪。
- HBase
  - 提供了 HBASE-20447，它解决了可能导致缓存问题的问题，特别是拆分区域。
- MXnet
  - 新增了 OpenCV 库。
- Spark
  - Spark 使用 EMRFS 将 Parquet 文件写入 Amazon S3 位置时，FileOutputCommitter 算法已更新为使用版本 2 而非版本 1。这将减少重命名的数量，从而提高应用程序性能。此更改不会影响：
    - Spark 以外的应用程序。

- 写入其它文件系统的应用程序，例如 HDFS（仍然使用 FileOutputCommitter 版本 1）。
- 使用其它输出格式（如文本或 csv）的应用程序（已使用 EMRFS 直接写入）。

已知问题

- JupyterHub
  - 不支持在创建集群时使用配置分类设置 JupyterHub 和单个 Jupyter notebook。手动编辑每个用户的 jupyterhub_config.py 文件和 jupyter_notebook_config.py 文件。有关更多信息，请参阅配置 JupyterHub。
  - JupyterHub 无法在私有子网内的集群上启动，并显示消息 `Error: ENOENT: no such file or directory, open '/etc/jupyter/conf/server.crt'` 。这由生成自签名凭证的脚本中的错误所致。使用以下解决方法生成自签名凭证。在连接到主节点时执行所有命令。

    1. 将凭证生成脚本从容器复制到主节点：

       ```
       sudo docker cp jupyterhub:/tmp/gen_self_signed_cert.sh ./
       ```

    2. 使用文本编辑器更改第 23 行，将公有主机名更改为本地主机名，如下所示：

       ```
       local hostname=$(curl -s $EC2_METADATA_SERVICE_URI/local-hostname)
       ```

    3. 运行脚本，生成自签名凭证：

       ```
       sudo bash ./gen_self_signed_cert.sh
       ```

    4. 将脚本生成的凭证文件移至 `/etc/jupyter/conf/` 目录：

       ```
       sudo mv /tmp/server.crt /tmp/server.key /etc/jupyter/conf/
       ```

    您可以对 `jupyter.log` 文件执行 `tail`，来验证 JupyterHub 是否重新启动并返回 200 响应代码。例如：

    ```
    tail -f /var/log/jupyter/jupyter.log
    ```

    该命令应返回与以下示例类似的响应：

    ```
    # [I 2018-06-14 18:56:51.356 JupyterHub app:1581] JupyterHub is now running at
     https://:9443/
    ```

```
# 19:01:51.359 - info: [ConfigProxy] 200 GET /api/routes
```

- 在 Amazon EMR 版本 5.14.0、5.15.0 或 5.16.0 中，在主节点重新启动或实例控制器重新启动后，将不会收集 CloudWatch 指标，且不提供自动扩展功能。此问题已在 Amazon EMR 5.17.0 中得到修复。

版本 5.13.0

以下发布说明包括有关 Amazon EMR 发行版 5.13.0 的信息。更改与 5.12.0 有关。

升级

- 已将 Spark 升级到 2.3.0
- 已将 HBase 升级到 1.4.2
- 已将 Presto 升级到 0.194
- 已将 Amazon SDK for Java 升级到 1.11.297

更改、增强功能和解决的问题

- Hive
  - 已逆向移植 HIVE-15436。增强了 Hive API 功能，仅返回视图。

已知问题

- MXNet 目前暂无 OpenCV 库。

版本 5.12.2

以下发布说明包括有关 Amazon EMR 发行版 5.12.2 的信息。更改与 5.12.1 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

版本 5.12.1

以下发布说明包括有关 Amazon EMR 发行版 5.12.1 的信息。更改与 5.12.0 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

版本 5.12.0

以下发布说明包括有关 Amazon EMR 发行版 5.12.0 的信息。更改与 5.11.1 有关。

升级

- Amazon SDK for Java1.11.238 升级到1.11.267。有关更多信息，请参阅 GitHub 上的 Amazon SDK for Java 更改日志。
- Hadoop 2.7.3 升级到 2.8.3。有关更多信息，请参阅 Apache Hadoop 发行版。
- Fink 1.3.2 升级到 1.4.0。有关详细信息，请参阅 Apache Flink 1.4.0 版本公告。
- HBase 1.3.1 升级到 1.4.0。有关详细信息，请参阅 HBase 版本公告。
- Hue 4.0.1 升级到 4.1.0。有关更多信息，请参阅发布说明。
- MxNet 0.12.0 升级到 1.0.0。有关更多信息，请参阅 GitHub 上的 MXNet 更改日志。
- Presto 0.187 升级到 0.188。有关更多信息，请参阅发布说明。

更改、增强功能和解决的问题

- Hadoop
  - `yarn.resourcemanager.decommissioning.timeout` 属性已更改为 `yarn.resourcemanager.nodemanager-graceful-decommission-timeout-secs`。您可以使用此属性自定义集群缩减。有关更多信息，请参阅《Amazon EMR 管理指南》中的集群缩减。
  - Hadoop CLI 向 cp（复制）命令添加了 `-d` 选项，可指定直接复制。可以使用它来避免创建中间 `._COPYING_` 文件，这加快了在 Amazon S3 之间复制数据的速度。有关更多信息，请参阅 HADOOP-12384。
- Pig

  - 添加了 pig-env 配置分类，这简化了 Pig 环境属性的配置。有关更多信息，请参阅配置应用程序。
- Presto
  - 新增 presto-connector-redshift 配置分类，您可以将其用于配置 Presto redshift.properties 配置文件中的值。有关更多信息，请参阅 Presto 文档中 Redshift 连接器以及 配置应用程序。
  - 已添加对 EMRFS 的 Presto 支持，且已设为默认配置。Amazon EMR 早期发行版使用 PrestoS3FileSystem，它是唯一选项。有关更多信息，请参阅EMRFS 和 PrestoS3FileSystem 配置。

    > **Note**
    >
    > 如果您使用 Amazon EMR 版本 5.12.0 查询 Amazon S3 中的底层数据，则可能会出现 Presto 错误。这是因为 Presto 无法从 emrfs-site.xml 提取配置分类值。解决方法是在 usr/lib/presto/plugin/hive-hadoop2/ 下创建一个 emrfs 子目录，并在 usr/lib/presto/plugin/hive-hadoop2/emrfs 中创建一个指向现有 /usr/share/aws/emr/emrfs/conf/emrfs-site.xml 文件的符号链接。然后重新启动 presto-server 进程（首先执行 sudo presto-server stop，然后执行 sudo presto-server start）。
- Spark
  - 已逆向移植 SPARK-22036：BigDecimal 乘法运算有时会返回空值。

已知问题

- MXNet 不包含 OpenCV 库。
- SparkR 不适用于使用自定义 AMI 创建的集群，因为默认情况下不会在集群节点上安装 R。

发布版本 5.11.3

以下发布说明包括有关 Amazon EMR 发行版 5.11.3 的信息。更改与 5.11.2 有关。

首次发布日期：2019 年 7 月 18 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI，提供了重要的 Linux 内核安全更新，其中包括 TCP SACK 拒绝服务问题（AWS-2019-005）。

版本 5.11.2

以下发布说明包括有关 Amazon EMR 发行版 5.11.2 的信息。更改与 5.11.1 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

版本 5.11.1

以下发布说明包括有关 Amazon EMR 发行版 5.11.1 的信息。更改与 Amazon EMR 5.11.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

已知问题

- MXNet 不包含 OpenCV 库。
- 默认情况下，Hive 2.3.2 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

版本 5.11.0

以下发布说明包括有关 Amazon EMR 发行版 5.11.0 的信息。更改与 Amazon EMR 5.10.0 发行版有关。

升级

以下应用程序和组件已在此版本中进行升级以包含以下版本。

- Hive 2.3.2
- Spark 2.2.1
- SDK for Java1.11.238

新特征

- Spark
  - 增加了 `spark.decommissioning.timeout.threshold` 设置，这将改进使用竞价型实例时的 Spark 停用行为。有关更多信息，请参阅配置节点停用行为。
  - 向 Spark 添加了 `aws-sagemaker-spark-sdk` 组件，此组件将安装 Amazon SageMaker Spark 和用于 Spark 与 Amazon SageMaker 集成的关联依赖项。您可以使用 Amazon SageMaker Spark 构造使用 Amazon SageMaker 阶段的 Spark 机器学习 (ML) 管道。有关更多信息，请参阅 GitHub 上的 SageMaker Spark 自述文件和《Amazon SageMaker 开发人员指南》https://docs.amazonaws.cn/sagemaker/latest/dg/apache-spark.html中的将 Apache Spark 与 Amazon SageMaker 结合使用。

已知问题

- MXNet 不包含 OpenCV 库。
- 默认情况下，Hive 2.3.2 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

版本 5.10.0

以下发布说明包括有关 Amazon EMR 发行版 5.10.0 的信息。更改与 Amazon EMR 5.9.0 发行版有关。

升级

以下应用程序和组件已在此版本中进行升级以包含以下版本。

- Amazon SDK for Java 1.11.221
- Hive 2.3.1
- Presto 0.187

新特征

- 添加了对 Kerberos 身份验证的支持。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 Kerberos 身份验证。
- 添加了对用于处理 EMRFS 对 Amazon S3 请求的 IAM 角色的支持。有关更多信息，请参阅《Amazon EMR 管理指南》中的为处理 EMRFS 对 Amazon S3 的请求配置 IAM 角色。
- 添加了对基于 GPU 的 P2 和 P3 实例类型的支持。有关更多信息，请参阅 Amazon EC2 P2 实例和 Amazon EC2 P3 实例。NVIDIA 驱动程序 384.81 和 CUDA 驱动程序 9.0.176 默认安装在这些实例类型上。
- 添加了对 Apache MXNet 的支持。

更改、增强功能和解决的问题

- Presto
  - 添加了对使用 Amazon Glue 数据目录作为默认 Hive 元存储的支持。有关更多信息，请参阅将 Presto 与 Amazon Glue 数据目录结合使用。
  - 添加了对地理空间函数的支持。
  - 为联接添加了溢出到磁盘支持。
  - 添加了对 Redshift 连接器的支持。
- Spark

  - 已逆向移植 SPARK-20640，这使随机注册的 rpc 超时值和重试次数值可使用 `spark.shuffle.registration.timeout` 和 `spark.shuffle.registration.maxAttempts` 属性进行配置。
  - 已逆向移植 SPARK-21549，这更正了在将自定义 OutputFormat 写入非 HDFS 位置时出现的错误。
- 已逆向移植 Hadoop 13270
- 从基本 Amazon EMR AMI 中删除了 Numpy、Scipy 和 Matplotlib 库。如果您的应用程序需要这些库，应用程序存储库中提供了它们，因此您可以通过引导操作使用 `yum install` 在所有节点上安装它们。
- Amazon EMR 基本 AMI 不再包含应用程序 RPM 软件包，因此集群节点上不再存在 RPM 软件包。自定义 AMI 和 Amazon EMR 基本 AMI 现在引用 Amazon S3 中的 RPM 软件包存储库。
- 因为 Amazon EC2 中引入了按秒计费，默认的 Scale down behavior (缩减行为) 现在为 Terminate at task completion (在任务完成时终止) 而非 Terminate at instance hour (在实例小时边界终止)。有关更多信息，请参阅配置集群缩减。

已知问题

- MXNet 不包含 OpenCV 库。
- 默认情况下，Hive 2.3.1 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

版本 5.9.0

以下发布说明包括有关 Amazon EMR 发行版 5.9.0 的信息。更改与 Amazon EMR 5.8.0 发行版有关。

发布日期：2017 年 10 月 5 日

最近功能更新时间：2017 年 10 月 12 日

升级

以下应用程序和组件已在此版本中进行升级以包含以下版本。

- Amazon SDK for Java 1.11.183 版
- Flink 1.3.2
- Hue 4.0.1
- Pig 0.17.0
- Presto 0.184

新特征

- 添加了 Livy 支持 (0.4.0-incubating 版)。有关更多信息，请参阅Apache Livy。
- 添加了对 Hue Notebook for Spark 的支持。
- 添加了对 i3 系列 Amazon EC2 实例的支持 (2017 年 10 月 12 日)。

更改、增强功能和解决的问题

- Spark
  - 添加了一组新功能，有助于确保 Spark 能够更为正常地处理因手动调整大小或自动扩展策略请求导致的节点终止。有关更多信息，请参阅配置节点停用行为。
  - 使用 SSL 取代 3DES 为数据块传输服务提供 in-transit 加密，可在使用带 AES-NI 的 Amazon EC2 实例类型时增强性能。
  - 已逆向移植 SPARK-21494。
- Zeppelin
  - 已逆向移植 ZEPPELIN-2377。
- HBase
  - 添加了补丁 HBASE-18533，因此可以使用 `hbase-site` 配置分类为 HBase BucketCache 配置使用其它值。
- Hue
  - 添加了对 Hue 中 Hive 查询编辑器的 Amazon Glue 数据目录支持。
  - 默认情况下，Hue 中的超级用户可以访问允许 Amazon EMR IAM 角色访问的所有文件。新建用户不会自动拥有对 Amazon S3 filebrowser 的访问权限，并且必须为其组启用 `filebrowser.s3_access` 权限。

- 解决造成使用 Amazon Glue 数据目录创建的底层 JSON 数据不可访问的问题。

已知问题

- 当安装了所有应用程序且未更改默认 Amazon EBS 根卷大小时，集群启动会失败。作为解决方法，请从 Amazon CLI 使用 `aws emr create-cluster` 命令并指定一个更大的 `--ebs-root-volume-size` 参数。
- 默认情况下，Hive 2.3.0 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。

版本 5.8.2

以下发布说明包括有关 Amazon EMR 发行版 5.8.2 的信息。更改与 5.8.1 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

版本 5.8.1

以下发布说明包括有关 Amazon EMR 发行版 5.8.1 的信息。更改与 Amazon EMR 5.8.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

版本 5.8.0

以下发布说明包括有关 Amazon EMR 发行版 5.8.0 的信息。更改与 Amazon EMR 5.7.0 发行版有关。

首次发布日期：2017 年 8 月 10 日

最近功能更新时间：2017 年 9 月 25 日

升级

以下应用程序和组件已在此版本中进行升级以包含以下版本：

- Amazon SDK 1.11.160
- Flink 1.3.1
- Hive 2.3.0。有关更多信息，请参阅 Apache Hive 站点上的发布说明。
- Spark 2.2.0。有关更多信息，请参阅 Apache Spark 站点上的发布说明。

新特征

- 添加了对查看应用程序历史记录的支持 (2017 年 9 月 25 日)。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。

更改、增强功能和解决的问题

- 与 Amazon Glue 数据目录集成
  - 为 Hive 和 Spark SQL 增加了使用 Amazon Glue 数据目录作为 Hive 元数据存储的功能。有关更多信息，请参阅将 Amazon Glue 数据目录用作 Hive 元存储。和使用 Amazon Glue 数据目录作为 Spark SQL 的元存储。
- 已向集群详细信息添加 Application history (应用程序历史记录)，这可让您查看 YARN 应用程序的历史数据以及 Spark 应用程序的其它详细信息。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。
- Oozie
  - 已逆向移植 OOZIE-2748。
- Hue
  - 已逆向移植 HUE-5859

- HBase
  - 添加了补丁，以使用 `getMasterInitializedTime` 通过 Java 管理扩展 (JMX) 公开 HBase 主服务器启动时间。
  - 添加了改进集群启动时间的补丁。

已知问题

- 当安装了所有应用程序且未更改默认 Amazon EBS 根卷大小时，集群启动会失败。作为解决方法，请从 Amazon CLI 使用 `aws emr create-cluster` 命令并指定一个更大的 `--ebs-root-volume-size` 参数。
- 默认情况下，Hive 2.3.0 设置 `hive.compute.query.using.stats=true`。这会导致查询从现有统计数据而不是直接从数据中获取数据，这可能会造成混淆。例如，如果您有一个包含 `hive.compute.query.using.stats=true` 的表并且将新文件上载到表 `LOCATION`，则在该表上运行 `SELECT COUNT(*)` 查询会返回来自统计数据的计数，而不是选择已添加的行。

  作为解决方法，请使用 `ANALYZE TABLE` 命令收集新的统计数据，或者设置 `hive.compute.query.using.stats=false`。有关更多信息，请参阅 Apache Hive 文档中的 Hive 中的统计数据。
- Spark – 在使用 Spark 时，apppusher 进程守护程序存在文件处理程序泄漏问题，长时间运行的 Spark 任务在几个小时或几天后可能会出现此情况。要解决此问题，请连接到主节点并键入 `sudo /etc/init.d/apppusher stop`。这将停止 apppusher 进程守护程序，而 Amazon EMR 将自动重新启动它。
- 应用程序历史记录
  - 死 Spark 执行程序的历史数据不可用。
  - 应用程序历史记录对使用安全配置来启用传输中加密的集群不可用。

版本 5.7.0

以下发布说明包括有关 Amazon EMR 发行版 5.7.0 的信息。更改与 Amazon EMR 5.6.0 发行版有关。

发布日期：2017 年 7 月 13 日

升级

- Flink 1.3.0
- Phoenix 4.11.0

- Zeppelin 0.7.2

新特征

- 添加了创建集群时指定自定义 Amazon Linux AMI 的功能。有关更多信息，请参阅使用自定义 AMI。

更改、增强功能和解决的问题

- HBase
  - 添加了配置 HBase 只读副本集群的功能。请参阅使用只读副本集群。
  - 多个错误修复和增强功能
- Presto – 添加了配置 `node.properties` 的功能。
- YARN – 添加了配置 `container-log4j.properties` 的功能
- Sqoop – 已逆向移植 SQOOP-2880，这将引入一个允许您设置 Sqoop 临时目录的参数。

版本 5.6.0

以下发布说明包括有关 Amazon EMR 发行版 5.6.0 的信息。更改与 Amazon EMR 5.5.0 发行版有关。

发布日期：2017 年 6 月 5 日

升级

- Flink 1.2.1
- HBase 1.3.1
- Mahout 0.13.0。这是 Mahout 在 Amazon EMR 版本 5.0 及更高版本中支持 Spark 2.x 的第一个版本。
- Spark 2.1.1

更改、增强功能和解决的问题

- Presto
  - 添加了通过使用安全配置启用传输中加密，从而在 Presto 节点之间实现 SSL/TLS 安全通信的功能。有关更多信息，请参阅传输中的数据加密。

- 已逆向移植 Presto 7661，它向 EXPLAIN ANALYZE 语句添加了 VERBOSE 选项，以报告有关查询计划的更详细、高低级别的统计数据。

版本 5.5.3

以下发布说明包括有关 Amazon EMR 发行版 5.5.3 的信息。更改与 5.5.2 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- 此版本解决了潜在的安全漏洞。

版本 5.5.2

以下发布说明包括有关 Amazon EMR 发行版 5.5.2 的信息。更改与 5.5.1 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

版本 5.5.1

以下发布说明包括有关 Amazon EMR 发行版 5.5.1 的信息。更改与 Amazon EMR 5.5.0 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

版本 5.5.0

以下发布说明包括有关 Amazon EMR 发行版 5.5.0 的信息。更改与 Amazon EMR 5.4.0 发行版有关。

发布日期：2017 年 4 月 26 日

升级

- Hue 3.12
- Presto 0.170
- Zeppelin 0.7.1
- ZooKeeper 3.4.10

更改、增强功能和解决的问题

- Spark
  - 已将 Spark 补丁 (SPARK-20115) fix DAGScheduler to recompute all the lost shuffle blocks when external shuffle service is unavailable 逆向移植到 2.1.0 版 Spark，此版本包含在本次发布中。
- Flink
  - Flink 现在使用 Scala 2.11 进行构建。如果您使用 Scala API 和库，我们建议您在项目中使用 Scala 2.11。
  - 解决了 HADOOP_CONF_DIR 和 YARN_CONF_DIR 默认值未正确设置，因此 `start-scala-shell.sh` 无法工作的问题。此外，还添加了使用 `env.hadoop.conf.dir` 或 `env.yarn.conf.dir` 配置类别中的 `/etc/flink/conf/flink-conf.yaml` 和 `flink-conf` 设置这些值的功能。
  - 推出了一个新的 EMR 特定的命令 `flink-scala-shell` 作为 `start-scala-shell.sh` 的包装程序。我们建议使用此命令而不是 `start-scala-shell`。新命令可简化执行。例如，`flink-scala-shell -n 2` 将使用任务并行度 2 启动 Flink Scala Shell。
  - 推出了一个新的 EMR 特定的命令 `flink-yarn-session` 作为 `yarn-session.sh` 的包装程序。我们建议使用此命令而不是 `yarn-session`。新命令可简化执行。例如，`flink-yarn-session -d -n 2` 将使用两个任务管理器在分离状态下启动长时间运行的 Flink 会话。
  - 解决了 (FLINK-6125) commons httpclient is not shaded anymore in Flink 1.2 的问题。
- Presto
  - 添加了对 LDAP 身份验证的支持。将 LDAP 与 Presto on Amazon EMR 结合使用，需要您启用对 Presto 协调器的 HTTPS 访问（`config.properties` 中的 `http-server.https.enabled=true`）。有关配置详细信息，请参阅 Presto 文档中的 LDAP 身份验证。
  - 添加了对 `SHOW GRANTS` 的支持。

- Amazon EMR 基本 Linux AMI
  - Amazon EMR 发行版现在基于 Amazon Linux 2017.03。有关更多信息，请参阅 Amazon Linux AMI 2017.03 发布说明。
  - 从 Amazon EMR 基本 Linux 映像中删除了 Python 2.6。默认安装 Python 2.7 和 3.4。如果需要，您可以手动安装 Python 2.6。

版本 5.4.0

以下发布说明包括有关 Amazon EMR 发行版 5.4.0 的信息。更改与 Amazon EMR 5.3.0 发行版有关。

发布日期：2017 年 3 月 8 日

升级

此版本提供以下升级：

- 已升级到 Flink 1.2.0
- 已升级到 Hbase 1.3.0
- 已升级到 Phoenix 4.9.0

> **Note**
>
> 如果您从早期版本的 Amazon EMR 升级到 Amazon EMR 发行版 5.4.0 或更高版本并使用二级索引，请升级本地索引，如 Apache Phoenix 文档中所述。Amazon EMR 将从 `hbase-site` 分类中删除所需配置，但索引需要重新填充。支持在线和离线升级索引。在线升级为默认值，这意味着，在从版本 4.8.0 或更高版本的 Phoenix 客户端初始化时重新填充索引。要指定离线升级，请在 `phoenix.client.localIndexUpgrade` 分类中将 `phoenix-site` 配置设置为 false，然后将 SSH 设置为主节点以运行 `psql [zookeeper] -1`。

- 已升级到 Presto 0.166
- 已升级到 Zeppelin 0.7.0

更改和增强功能

以下是对版本标签 emr-5.4.0 的 Amazon EMR 版本进行的更改：

- 增加了对 r4 实例的支持。请参阅 Amazon EC2 实例类型。

版本 5.3.1

以下发布说明包括有关 Amazon EMR 发行版 5.3.1 的信息。更改与 Amazon EMR 5.3.0 发行版有关。

发布日期：2017 年 2 月 7 日

进行了微小更改：逆向移植 Zeppelin 补丁，并更新了 Amazon EMR 的默认 AMI。

版本 5.3.0

以下发布说明包括有关 Amazon EMR 发行版 5.3.0 的信息。更改与 Amazon EMR 5.2.1 发行版有关。

发布日期：2017 年 1 月 26 日

升级

此版本提供以下升级：

- 已升级到 Hive 2.1.1
- 已升级到 Hue 3.11.0
- 已升级到 Spark 2.1.0
- 已升级到 Oozie 4.3.0
- 已升级到 Flink 1.1.4

更改和增强功能

以下是对版本标签 emr-5.3.0 的 Amazon EMR 版本进行的更改：

- Hue 新增补丁可使您使用 `interpreters_shown_on_wheel` 设置配置解释器在笔记本选择轮盘上最先显示的内容，而不受 `hue.ini` 文件中排序的限制。
- 新增 `hive-parquet-logging` 配置分类，您可以将其用于配置 Hive `parquet-logging.properties` 文件中的值。

版本 5.2.2

以下发布说明包括有关 Amazon EMR 发行版 5.2.2 的信息。更改与 Amazon EMR 5.2.1 发行版有关。

发布日期：2017 年 5 月 2 日

早期版本中已解决的已知问题

- 已逆向移植 SPARK-194459，解决了从包含 char/varchar 列的 ORC 表读取内容时可能失败的问题。

版本 5.2.1

以下发布说明包括有关 Amazon EMR 发行版 5.2.1 的信息。更改与 Amazon EMR 5.2.0 发行版有关。

发布日期：2016 年 12 月 29 日

升级

此版本提供以下升级：

- 已升级到 Presto 0.157.1。有关更多信息，请参阅 Presto 文档中的 Presto 发布说明。
- 已升级到 Zookeeper 3.4.9。有关更多信息，请参阅 Apache ZooKeeper 文档中的 ZooKeeper 发布说明。

更改和增强功能

以下是对版本标签 emr-5.2.1 的 Amazon EMR 版本进行的更改：

- 在 Amazon EMR 版本 4.8.3 及更高版本（但不包括 5.0.0、5.0.3 和 5.2.0）中添加了对 Amazon EC2 m4.16xlarge 实例类型的支持。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。
- 现在，Flink 和 YARN 配置路径的位置默认在 `/etc/default/flink` 中设置，您在运行 `FLINK_CONF_DIR` 或 `HADOOP_CONF_DIR` 驱动程序脚本启动 Flink 作业时，无需设置环境变量 `flink` 和 `yarn-session.sh`。
- 新增对 FlinkKinesisConsumer 类的支持。

早期版本中已解决的已知问题

- 修复了 Hadoop 中的一个问题，即 ReplicationMonitor 线程可能会因为在大型集群中复制和删除同一个文件导致的竞争而卡住很长时间。

- 修复了在作业状态未成功更新时 ControlledJob#toString 出现空指针异常 (NPE) 失败的问题。

版本 5.2.0

以下发布说明包括有关 Amazon EMR 发行版 5.2.0 的信息。更改与 Amazon EMR 5.1.0 发行版有关。

发布日期：2016 年 11 月 21 日

更改和增强功能

此版本中提供了以下更改和增强功能：

- 添加了适用于 HBase 的 Amazon S3 存储模式。
- 使您能够为 HBase rootdir 指定 Amazon S3 位置。有关更多信息，请参阅 Amazon S3 上的 HBase。

升级

此版本提供以下升级：

- 已升级到 Spark 2.0.2

早期版本中已解决的已知问题

- 修复了限制为仅 EBS 实例类型上的 2 TB 的 /mnt 的问题。
- 修复了输出到相应的 .out 文件而不是常规 log4j 配置的 .log 文件 (每小时转动一次) 的 instance-controller 和 logpusher 日志的问题。.out 文件不会轮换，因此这最终将填满 /emr 分区。此问题仅影响硬件虚拟机 (HVM) 实例类型。

版本 5.1.0

以下发布说明包括有关 Amazon EMR 发行版 5.1.0 的信息。更改与 Amazon EMR 5.0.0 发行版有关。

发布日期：2016 年 11 月 3 日

更改和增强功能

此版本中提供了以下更改和增强功能：

- 增加了对 Flink 1.1.3 的支持。
- Presto 已作为 Hue 的记事本部分中的选项添加。

升级

此版本提供以下升级：

- 已升级到 HBase 1.2.3
- 已升级到 Zeppelin 0.6.2

早期版本中已解决的已知问题

- 修复了带 ORC 文件的 Amazon S3 上的 Tez 查询的性能低于早期 Amazon EMR 4.x 版本中的性能的问题。

版本 5.0.3

以下发布说明包括有关 Amazon EMR 发行版 5.0.3 的信息。更改与 Amazon EMR 5.0.0 发行版有关。

发布日期：2016 年 10 月 24 日

升级

此版本提供以下升级：

- 已升级到 Hadoop 2.7.3
- 已升级到 Presto 0.152.3，它包括对 Presto Web 界面的支持。可使用端口 8889 访问 Presto 协调器上的 Presto Web 界面。有关 Presto Web 界面的更多信息，请参阅 Presto 文档中的 Web 界面。
- 已升级到 Spark 2.0.1
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

版本 5.0.0

发布日期：2016 年 7 月 27 日

升级

此版本提供以下升级：

- 已升级到 Hive 2.1
- 已升级到 Presto 0.150
- 已升级到 Spark 2.0
- 已升级到 Hue 3.10.0
- 已升级到 Pig 0.16.0
- 已升级到 Tez 0.8.4
- 已升级到 Zeppelin 0.6.1

更改和增强功能

以下是对版本标签 emr-5.0.0 的 Amazon EMR 版本进行的更改：

- Amazon EMR 支持最新开源版本的 Hive（版本 2.1）和 Pig（版本 0.16.0）。如果您以前使用的是 Amazon EMR 上的 Hive 或 Pig，那么这可能会影响某些使用案例。有关更多信息，请参阅 Hive 和 Pig。
- Hive 和 Pig 的默认执行引擎现在是 Tez。要更改该设置，您可以在 `hive-site` 和 `pig-properties` 配置分类中分别编辑相应的值。
- 添加了增强型步骤调试功能，可让您查看步骤失败的根本原因 (如果服务可以确定原因)。有关更多信息，请参阅《Amazon EMR 管理指南》中的增强型步骤调试。
- 先前以“-Sandbox”结尾的应用程序不再拥有该后缀。这可能会中断您的自动化，例如，如果您使用脚本来启动具有这些应用程序的集群。下表显示了 Amazon EMR 4.7.2 与 Amazon EMR 5.0.0 中的应用程序名称。

  应用程序名称更改

Amazon EMR 4.7.2	Amazon EMR 5.0.0
Oozie-Sandbox	Oozie
Presto-Sandbox	Presto
Sqoop-Sandbox	Sqoop

Amazon EMR 4.7.2	Amazon EMR 5.0.0
Zeppelin-Sandbox	Zeppelin
ZooKeeper-Sandbox	ZooKeeper

- Spark 现在针对 Scala 2.11 进行编译。
- Java 8 现在是默认 JVM。所有应用程序均使用 Java 8 runtime 运行。对任何应用程序的字节代码目标都没有进行更改。大多数应用程序继续运行 Java 7。
- Zeppelin 现在包括身份验证功能。有关更多信息，请参阅 Zeppelin。
- 添加了对安全配置的支持，这使您可以更轻松地创建和应用加密选项。有关更多信息，请参阅数据加密。

版本 4.9.5

以下发布说明包括有关 Amazon EMR 发行版 4.9.5 的信息。更改与 4.9.4 有关。

首次发布日期：2018年 8 月 29 日

更改、增强功能和解决的问题

- HBase
  - 此版本解决了潜在的安全漏洞。

版本 4.9.4

以下发布说明包括有关 Amazon EMR 发行版 4.9.4 的信息。更改与 4.9.3 有关。

首次发布日期：2018 年 3 月 29 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了潜在漏洞。

版本 4.9.3

以下发布说明包括有关 Amazon EMR 发行版 4.9.3 的信息。更改与 Amazon EMR 4.9.2 发行版有关。

首次发布日期：2018 年 1 月 22 日

更改、增强功能和解决的问题

- 更新了 Amazon EMR 的默认 Amazon Linux AMI 的 Amazon Linux 内核，解决了与推测执行相关的漏洞（CVE-2017-5715、CVE-2017-5753 和 CVE-2017-5754）。有关更多信息，请参阅https://www.amazonaws.cn/security/security-bulletins/AWS-2018-013/。

版本 4.9.2

以下发布说明包括有关 Amazon EMR 发行版 4.9.2 的信息。更改与 Amazon EMR 4.9.1 发行版有关。

发布日期：2017 年 7 月 13 日

此版本略微进行了一些改动、错误修复和增强。

版本 4.9.1

以下发布说明包括有关 Amazon EMR 发行版 4.9.1 的信息。更改与 Amazon EMR 4.8.4 发行版有关。

发布日期：2017 年 4 月 10 日

早期版本中已解决的已知问题

- 已逆向移植 HIVE-9976 和 HIVE-10106
- 修复了 YARN 中的一个问题，即，大量节点 (大于 2000 个) 和容器 (大于 5000 个) 会导致内存不足错误，例如："`Exception in thread 'main' java.lang.OutOfMemoryError`"。

更改和增强功能

以下是对版本标签 emr-4.9.1 的 Amazon EMR 版本进行的更改：

- Amazon EMR 发行版现在基于 Amazon Linux 2017.03。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2017.03-release-notes/。
- 从 Amazon EMR 基本 Linux 映像中删除了 Python 2.6。如果需要，您可以手动安装 Python 2.6。

版本 4.8.4

以下发布说明包括有关 Amazon EMR 发行版 4.8.4 的信息。更改与 Amazon EMR 4.8.3 发行版有关。

发布日期：2017 年 2 月 7 日

此版本略微进行了一些改动、错误修复和增强。

版本 4.8.3

以下发布说明包括有关 Amazon EMR 发行版 4.8.3 的信息。更改与 Amazon EMR 4.8.2 发行版有关。

发布日期：2016 年 12 月 29 日

升级

此版本提供以下升级：

- 已升级到 Presto 0.157.1。有关更多信息，请参阅 Presto 文档中的 Presto 发布说明。
- 已升级到 Spark 1.6.3。有关更多信息，请参阅 Apache Spark 文档中的 Spark 发布说明。
- 已升级到 ZooKeeper 3.4.9。有关更多信息，请参阅 Apache ZooKeeper 文档中的 ZooKeeper 发布说明。

更改和增强功能

以下是对版本标签 emr-4.8.3 的 Amazon EMR 版本进行的更改：

- 在 Amazon EMR 版本 4.8.3 及更高版本（但不包括 5.0.0、5.0.3 和 5.2.0）中添加了对 Amazon EC2 m4.16xlarge 实例类型的支持。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

早期版本中已解决的已知问题

- 修复了 Hadoop 中的一个问题，即 ReplicationMonitor 线程可能会因为在大型集群中复制和删除同一个文件导致的竞争而卡住很长时间。
- 修复了在作业状态未成功更新时 ControlledJob#toString 出现空指针异常 (NPE) 失败的问题。

版本 4.8.2

以下发布说明包括有关 Amazon EMR 发行版 4.8.2 的信息。更改与 Amazon EMR 4.8.0 发行版有关。

发布日期：2016 年 10 月 24 日

升级

此版本提供以下升级：

- 已升级到 Hadoop 2.7.3
- 已升级到 Presto 0.152.3，它包括对 Presto Web 界面的支持。可使用端口 8889 访问 Presto 协调器上的 Presto Web 界面。有关 Presto Web 界面的更多信息，请参阅 Presto 文档中的 Web 界面。
- Amazon EMR 发行版现在基于 Amazon Linux 2016.09。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.09-release-notes/。

版本 4.8.0

发布日期：2016 年 9 月 7 日

升级

此版本提供以下升级：

- 已升级到 HBase 1.2.2
- 已升级到 Presto-Sandbox 0.151
- 已升级到 Tez 0.8.4
- 已升级到 Zeppelin-Sandbox 0.6.1

更改和增强功能

以下是对版本标签 emr-4.8.0 的 Amazon EMR 版本进行的更改：

- 修复了 YARN 中的一个问题，ApplicationMaster 将在其中尝试清除不再存在的容器，因为它们的实例已终止。
- 更正了 Oozie 示例中 Hive2 操作的 hive-server2 URL。
- 添加了对其它 Presto 目录的支持。
- 已逆向移植修补程序：HIVE-8948、HIVE-12679、HIVE-13405、PHOENIX-3116、HADOOP-12689
- 添加了对安全配置的支持，这使您可以更轻松地创建和应用加密选项。有关更多信息，请参阅数据加密。

版本 4.7.2

以下发布说明包括有关 Amazon EMR 4.7.2 的信息。

发布日期：2016 年 7 月 15 日

特征

此版本提供以下功能：

- 已升级到 Mahout 0.12.2
- 已升级到 Presto 0.148
- 已升级到 Spark 1.6.2
- 您现在可以使用 URI 作为参数来创建将与 EMRFS 配合使用的 AWSCredentialsProvider。有关更多信息，请参阅为 EMRFS 创建 AWSCredentialsProvider。
- EMRFS 现在允许用户使用 `emrfs-site.xml` 中的 `fs.s3.consistent.dynamodb.endpoint` 属性来为其一致视图元数据配置自定义 DynamoDB 终端节点。
- 在 `/usr/bin` 中添加了一个名为 `spark-example` 的脚本，它将包装 `/usr/lib/spark/spark/bin/run-example`，因此您可以直接运行示例。例如，要运行 Spark 分配的附带的 SparkPi 示例，可从命令行运行 `spark-example SparkPi 100` 或将 `command-runner.jar` 作为 API 中的一个步骤运行。

早期版本中已解决的已知问题

- 修复了 Oozie 在安装 Spark 后拥有的 `spark-assembly.jar` 未位于正确位置 (这导致使用 Oozie 启动 Spark 应用程序失败) 的问题。
- 修复了与 YARN 容器中基于 Spark Log4j 的登录有关的问题。

版本 4.7.1

发布日期：2016 年 6 月 10 日

早期版本中已解决的已知问题

- 修复了延长带有私有子网的 VPC 中启动的集群的启动时间的问题。此错误仅影响使用 Amazon EMR 4.7.0 发行版启动的集群。

- 修复了在 Amazon EMR 中错误处理针对使用 Amazon EMR 4.7.0 发行版启动的集群的文件列表的问题。

版本 4.7.0

> **Important**
>
> Amazon EMR 4.7.0 已弃用。请改用 Amazon EMR 4.7.1 或更高版本。

发布日期：2016 年 6 月 2 日

特征

此版本提供以下功能：

- 已添加 Apache Phoenix 4.7.0
- 已添加 Apache Tez 0.8.3
- 已升级到 HBase 1.2.1
- 已升级到 Mahout 0.12.0
- 已升级到 Presto 0.147
- 已将 Amazon SDK for Java 升级到 1.10.75
- 已从 `mapreduce.cluster.local.dir` 中的 `mapred-site.xml` 属性中删除最终标志以允许用户以本地模式运行 Pig。

集群上可用的 Amazon Redshift JDBC 驱动程序

Amazon Redshift JDBC 驱动程序现在包含在 `/usr/share/aws/redshift/jdbc` 中。`/usr/share/aws/redshift/jdbc/RedshiftJDBC41.jar` 是与 JDBC 4.1 兼容的驱动程序，`/usr/share/aws/redshift/jdbc/RedshiftJDBC4.jar` 是与 JDBC 4.0 兼容的 Amazon Redshift 驱动程序。有关更多信息，请参阅 Amazon Redshift 管理指南中的配置 JDBC 连接。

Java 8

OpenJDK 1.7 是用于所有应用程序 (Presto 除外) 的默认 JDK。但是，将同时安装 OpenJDK 1.7 和 1.8。有关如何为应用程序设置 `JAVA_HOME` 的信息，请参阅配置应用程序以使用 Java 8。

早期版本中已解决的已知问题

- 修复了一个内核问题，该问题已明显影响了 emr-4.6.0 中的 Amazon EMR 的吞吐量优化 HDD（st1）EBS 卷的性能。
- 修复了在不选择 Hadoop 作为应用程序的情况下指定任何 HDFS 加密区域时集群将失败的问题。
- 已将默认 HDFS 编写策略从 `RoundRobin` 更改为 `AvailableSpaceVolumeChoosingPolicy`。未通过 RoundRobin 配置正确利用某些卷，这将导致核心节点失败且 HDFS 不可靠。
- 修复了与 EMRFS CLI 有关的问题，此问题将在创建默认 DynamoDB 元数据表以获得一致视图时导致异常。
- 修复了在分段重命名和复制操作期间可能发生在 EMRFS 中的死锁问题。
- 修复了与 EMRFS 有关的问题，此问题导致 CopyPart 大小默认为 5 MB。默认值现已相应地设置为 128 MB。
- 修复了与 Zeppelin upstart 配置有关的问题，此问题可能会阻止您停止服务。
- 修复了与 Spark 和 Zeppelin 有关的问题，此问题会阻止您使用 `s3a://` URI 方案，因为 `/usr/lib/hadoop/hadoop-aws.jar` 未在其各自的类路径中正确加载。
- 已逆向移植 HUE-2484。
- 已从 Hue 3.9.0 (不存在 JIRA) 逆向移植 commit 来修复与 HBase 浏览器示例有关的问题。
- 已逆向移植 HIVE-9073。

版本 4.6.0

发布日期：2016 年 4 月 21 日

特征

此版本提供以下功能：

- 已添加 HBase 1.2.0
- 已添加 Zookeeper-Sandbox 3.4.8
- 已升级到 Presto-Sandbox 0.143
- Amazon EMR 发行版现在基于 Amazon Linux 2016.03.0。有关更多信息，请参阅https://www.amazonaws.cn/amazon-linux-ami/2016.03-release-notes/。

影响吞吐量优化 HDD（st1）EBS 卷类型的问题

Linux 内核版本 4.2 及更高版本中的问题将显著影响 EMR 的吞吐量优化 HDD (st1) EBS 卷上的性能。此版本 (emr-4.6.0) 使用内核版本 4.4.5，因此会受到影响。因此，如果您打算使用 st1 EBS 卷，我们建议您不要使用 emr-4.6.0。您可将 emr-4.5.0 或早期 Amazon EMR 发行版与 st1 配合使用，而不会产生影响。此外，我们将随将来版本一起提供修复程序。

Python 默认值

现在，默认情况下已安装 Python 3.4，但 Python 2.7 将保留系统默认值。您可以使用引导操作将 Python 3.4 配置为系统默认值；也可以使用配置 API 将 PYSPARK_PYTHON 导出设置为 `/usr/bin/python3.4` 分类中的 `spark-env` 以便影响 PySpark 所使用的 Python 版本。

Java 8

OpenJDK 1.7 是用于所有应用程序 (Presto 除外) 的默认 JDK。但是，将同时安装 OpenJDK 1.7 和 1.8。有关如何为应用程序设置 `JAVA_HOME` 的信息，请参阅配置应用程序以使用 Java 8。

早期版本中已解决的已知问题

- 修复了应用程序预置有时会因生成的密码导致随机失败的问题。
- 之前，`mysqld` 已安装在所有节点上。现在，它仅安装在主实例上，而且仅在所选应用程序将 `mysql-server` 作为组件包含时安装。当前，以下应用程序包含 `mysql-server` 组件：HCatalog、Hive、Hue、Presto-Sandbox 和 Sqoop-Sandbox。
- 已将 `yarn.scheduler.maximum-allocation-vcores` 从默认值 32 更改为 80，这修复了 emr-4.4.0 中引入的一个问题，此问题主要在使用集群（其内核实例类型为具有高于 32 的 YARN 虚拟内核集的几个大型实例类型之一）中的 `maximizeResourceAllocation` 选项时与 Spark 时一起出现；也就是说，此问题影响了 c4.8xlarge、cc2.8xlarge、hs1.8xlarge、i2.8xlarge、m2.4xlarge、r3.8xlarge、d2.8xlarge 或 m4.10xlarge。
- s3-dist-cp 现在对所有 Amazon S3 提名使用 EMRFS，并且不再过渡到临时 HDFS 目录。
- 修复了与针对客户端加密分段上载的异常处理有关的问题。
- 添加了允许用户更改 Amazon S3 存储类的选项。默认情况下，此设置为 `STANDARD`。`emrfs-site` 配置分类设置为 `fs.s3.storageClass`，可能的值为 `STANDARD`、`STANDARD_IA` 和 `REDUCED_REDUNDANCY`。有关存储类的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的存储类。

版本 4.5.0

发布日期：2016 年 4 月 4 日

特征

此版本提供以下功能：

- 已升级到 Spark 1.6.1
- 已升级到 Hadoop 2.7.2
- 已升级到 Presto 0.140
- 添加了对 Amazon S3 服务器端加密的 Amazon KMS 支持。

早期版本中已解决的已知问题

- 修复了重启节点后无法启动 MySQL 和 Apache 服务器的问题。
- 修复了 IMPORT 未正确使用存储在 Amazon S3 中的非分区表的问题
- 修复了与 Presto 有关的问题，此问题导致在写入 Hive 表时要求暂存目录为 `/mnt/tmp` 而不是 `/tmp`。

版本 4.4.0

发布日期：2016 年 3 月 14 日

特征

此版本提供以下功能：

- 已添加 HCatalog 1.0.0
- 已添加 Sqoop-Sandbox 1.4.6
- 已升级到 Presto 0.136
- 已升级到 Zeppelin 0.5.6
- 已升级到 Mahout 0.11.1
- 默认情况下已启用 `dynamicResourceAllocation`。

- 已添加针对此版本的所有配置分类的表。有关更多信息，请参阅配置应用程序中的“配置分类”表。

早期版本中已解决的已知问题

- 修复了 `maximizeResourceAllocation` 设置不为 YARN ApplicationMaster 进程守护程序预留足够内存的问题。
- 修复了遇到的与自定义 DNS 相关的问题。如果 `resolve.conf` 中的任何条目位于提供的自定义条目之前，则自定义条目不可解析。此行为受 VPC 中集群的影响，其中，默认 VPC 名称服务器已作为顶部条目插入 `resolve.conf` 中。
- 修复了默认 Python 已移至版本 2.7 且未为该版本安装 boto 的问题。
- 修复了 YARN 容器和 Spark 应用程序将生成唯一 Ganglia 轮询数据库 (rrd) 文件的问题，此问题会导致第一个附加到实例的磁盘填满。修复后，YARN 容器级别指标和 Spark 应用程序级别指标均已禁用。
- 修复了导致日志推送程序中删除所有空日志文件夹的问题。这会造成 Hive CLI 无法记录日志，因为日志推送程序已删除 `user` 下的 `/var/log/hive` 空文件夹。
- 修复了影响 Hive 导入的问题，此问题影响分区并导致导入期间出现错误。
- 修复了 EMRFS 和 s3-dist-cp 未正确处理包含句点的存储桶名称的问题。
- 更改了 EMRFS 中的行为，以便在启用版本控制的存储桶中，不会持续创建 `_$folder$` 标记文件，从而有助于提高启用版本控制的存储桶的性能。
- 更改了 EMRFS 中的行为，使它不会使用说明文件，已启用客户端加密的情况除外。如果您要在使用客户端加密时删除说明文件，可将 emrfs-site.xml 属性 `fs.s3.cse.cryptoStorageMode.deleteInstructionFiles.enabled` 设置为 true。
- 更改了 YARN 日志聚合以在聚合目标中将日志保留两天。默认目标为您的集群的 HDFS 存储。如果您要更改此持续时间，请在创建集群时使用 `yarn.log-aggregation.retain-seconds` 配置分类来更改 `yarn-site` 的值。与往常一样，您可以在创建集群时使用 `log-uri` 参数将应用程序日志保存到 Amazon S3。

已应用的修补程序

此版本中包含了来自开源项目的以下修补程序：

- HIVE-9655
- HIVE-9183
- HADOOP-12810

版本 4.3.0

发布日期：2016 年 1 月 19 日

特征

此版本提供以下功能：

- 已升级到 Hadoop 2.7.1
- 已升级到 Spark 1.6.0
- 已将 Ganglia 升级到 3.7.2
- 已将 Presto 升级到 0.130

将 spark.dynamicAllocation.enabled 设置为 true 时，Amazon EMR 已对其做出一些更改；默认情况下，此项为 false。如果设置为 true，则会影响由 maximizeResourceAllocation 设置设定的默认值：

- 若spark.dynamicAllocation.enabled设为 true，则 spark.executor.instances 将不被 maximizeResourceAllocation 设置。
- 目前，spark.driver.memory 设置根据集群中的实例类型进行配置，与 spark.executors.memory 设置的方式类似。但是，由于 Spark 驱动应用程序可在主实例或核心实例之一上运行 (例如在 YARN 客户端和集群模式下分别进行)，spark.driver.memory 设置根据更小实例类型的实例类型，在两个实例组之间进行。
- 目前，spark.default.parallelism 设置为 YARN 容器可用的 CPU 内核数的两倍。在上一版本中，这是该值的一半。
- 为 Spark YARN 过程预留的内存开销计算精确性被优化，从而使得 Spark 可用内存总量略有增加 (即spark.executor.memory)。

早期版本中已解决的已知问题

- 默认情况下，现已启用 YARN 日志聚合。
- 修复了在启用 YARN 日志聚合后日志未推送至集群的 Amazon S3 日志存储桶的问题。
- YARN 容器大小现跨所有节点类型具有新的最小值 32。
- 修复了与 Ganglia 有关的问题，此问题已导致大型集群中主节点上的磁盘 I/O 过多。
- 修复了在关闭集群时阻止应用程序日志推送至 Amazon S3 的问题。

- 修复了 EMRFS CLI 中导致某些命令失败的问题。
- 修复了与 Zeppelin 有关的问题，此问题已阻止依赖项在基础 SparkContext 中加载。
- 修复了因发出尝试添加实例的调整大小命令导致的问题。
- 修复了 Hive 中的问题，此问题导致 CREATE TABLE AS SELECT 对 Amazon S3 进行过多的列表调用。
- 修复了在安装 Hue、Oozie 和 Ganglia 时无法正常预置大型集群的问题。
- 修复了 s3-dist-cp 中的问题，此问题导致即使在失败并出现错误的情况下仍将返回零退出代码。

已应用的修补程序

此版本中包含了来自开源项目的以下修补程序：

- OOZIE-2402
- HIVE-12502
- HIVE-10631
- HIVE-12213
- HIVE-10559
- HIVE-12715
- HIVE-10685

版本 4.2.0

发布日期：2015 年 11 月 18 日

特征

此版本提供以下功能：

- 已添加 Ganglia 支持
- 已升级到 Spark 1.5.2
- 已升级到 Presto 0.125
- 已将 Oozie 升级到 4.2.0
- 已将 Zeppelin 升级到 0.5.5
- 已将 Amazon SDK for Java 升级到 1.10.27

早期版本中已解决的已知问题

- 修复了与 EMRFS CLI 有关的问题，此问题发导致不使用默认元数据表名称。
- 修复了在 Amazon S3 中使用 ORC 支持的表时遇到的问题。
- 修复了遇到的 Python 版本在 Spark 配置中不匹配的问题。
- 修复了 YARN 节点状态因 VPC 中集群的 DNS 问题导致无法报告的问题。
- 修复了 YARN 已停用节点从而导致应用程序挂起且无法计划新应用程序时遇到的问题。
- 修复了集群终止且状态为 TIMED_OUT_STARTING 时遇到的问题。
- 修复了在其它内部版本中包含 EMRFS Scala 依赖项时遇到的问题。Scala 依赖项已被删除。

配置应用程序

您可以提供配置对象来覆盖应用程序的默认配置。您也可以使用简写语法提供配置，或者引用 JSON 文件中的配置对象。配置对象包含分类、属性和可选的嵌套配置。属性对应于您想要更改的应用程序设置。您可以在一个 JSON 对象中为多个应用程序指定多个分类。

> **Warning**
>
> Amazon EMR Describe 和 List API 操作以明文形式发出自定义和可配置的设置，用作 Amazon EMR 作业流程的一部分。要在这些设置中提供密码等敏感信息，请参阅将在 Amazon Secrets Manager 中存储敏感配置数据。

可用的配置分类因 Amazon EMR 发行版而异。有关特定发行版中所支持配置分类的列表，请参阅 关于 Amazon EMR 发行版下有关该发行版的页面。

以下是一组配置的示例 JSON 文件。

```
[
  {
    "Classification": "core-site",
    "Properties": {
      "hadoop.security.groups.cache.secs": "250"
    }
  },
  {
    "Classification": "mapred-site",
    "Properties": {
      "mapred.tasktracker.map.tasks.maximum": "2",
      "mapreduce.map.sort.spill.percent": "0.90",
      "mapreduce.tasktracker.reduce.tasks.maximum": "5"
    }
  }
]
```

配置分类通常可以映射到应用程序特定的配置文件。例如，`hive-site` 分类映射到 Hive 的 `hive-site.xml` 配置文件中的设置。此情况的一个例外是不再支持的引导操作 `configure-daemons`，它用于设置 `--namenode-heap-size` 等环境参数。与此类似的选项已归入 `hadoop-env` 和 `yarn-env` 分类，并具有自己的嵌套导出分类。如果任何分类以 `env` 结尾，请使用导出子分类。

另一个例外是 s3get，它用于将客户 EncryptionMaterialsProvider 对象放置在集群中的每个节点上来进行客户端加密。为了实现此目的，已向 emrfs-site 分类添加一个选项。

以下是一个 hadoop-env 分类示例。

```
[
  {
    "Classification": "hadoop-env",
    "Properties": {

    },
    "Configurations": [
      {
        "Classification": "export",
        "Properties": {
          "HADOOP_DATANODE_HEAPSIZE": "2048",
          "HADOOP_NAMENODE_OPTS": "-XX:GCTimeRatio=19"
        },
        "Configurations": [

        ]
      }
    ]
  }
]
```

以下是一个 yarn-env 分类示例。

```
[
  {
    "Classification": "yarn-env",
    "Properties": {

    },
    "Configurations": [
      {
        "Classification": "export",
        "Properties": {
          "YARN_RESOURCEMANAGER_OPTS": "-Xdebug -Xrunjdwp:transport=dt_socket"
        },
        "Configurations": [

        ]
```

```
      }
    ]
  }
]
```

以下设置不属于配置文件，但由 Amazon EMR 用来代表您配置多个设置。

由 Amazon EMR 辅助设置

应用程序	发行版标注分类	有效属性	何时使用
Spark	`spark`	`maximizeResourceAllocation`	配置执行程序以利用每个节点的最大资源。

主题

- 在创建集群时配置应用程序
- 在正在运行的集群中重新配置实例组
- 在 Amazon Secrets Manager 中存储敏感配置数据
- 配置应用程序来使用特定 Java 虚拟机

在创建集群时配置应用程序

在创建集群时，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK 覆盖应用程序的原定设置配置。

要覆盖应用程序的原定设置配置，您需要指定某个配置分类中的自定义值。配置分类对应于应用程序的配置 XML 文件，例如 `hive-site.xml`。

配置分类因 Amazon EMR 发行版而异。有关特定发行版中可用配置分类的列表，请参阅该发行版的详细信息页面。例如，Amazon EMR 发行版 6.4.0。

在创建集群时，在控制台中提供配置

要提供配置，请导航到 Create cluster (创建集群) 页面，然后选择 Edit software settings (编辑软件设置)。然后，您可以在控制台中使用 JSON 或以阴影文本表示的简写语法直接输入配置。否则，您可以为具有 JSON `Configurations` 对象的文件提供一个 Amazon S3 URI。

要为实例组提供配置，请导航到 Hardware Configuration (硬件配置) 页面。在 Node type (节点类型) 表中的 Instance type (实例类型) 列，为每个实例组选择编辑应用程序 Configurations (配置)。

在创建集群时使用 Amazon CLI 提供配置

您可通过提供本地存储或在 Amazon S3 中存储的 JSON 文件的路径来为 create-cluster 提供配置。以下示例假定您使用 Amazon EMR 的默认角色，并且已创建这些角色。如果您需要创建角色，请先运行 `aws emr create-default-roles`。

如果您的配置位于本地目录中，您可以使用示例命令。

```
aws emr create-cluster --use-default-roles --release-label emr-5.36.1 --applications
 Name=Hive \
--instance-type m5.xlarge --instance-count 3 --configurations file://./
configurations.json
```

如果您的配置位于 Amazon S3 路径中，则需要设置以下解决方法，然后才能将 Amazon S3 路径传递给 `create-cluster` 命令。

```
#!/bin/sh
# Assume the ConfigurationS3Path is not public, and its present in the same AWS account
 as the EMR cluster
ConfigurationS3Path="s3://my-bucket/config.json"
# Get a presigned HTTP URL for the s3Path
ConfigurationURL=`aws s3 presign $ConfigurationS3Path --expires-in 300`
# Fetch the presigned URL, and minify the JSON so that it spans only a single line
Configurations=`curl $ConfigurationURL | jq -c .`
aws emr create-cluster --use-default-roles --release-label emr-5.34.0 --instance-type
 m5.xlarge --instance-count 2 --applications Name=Hadoop Name=Spark --configurations
 $Configurations
```

在创建集群时，使用 Java SDK 提供配置

以下程序摘要说明如何使用 Amazon SDK for Java 提供配置。

```
Application hive = new Application().withName("Hive");

Map<String,String> hiveProperties = new HashMap<String,String>();
 hiveProperties.put("hive.join.emit.interval","1000");
 hiveProperties.put("hive.merge.mapfiles","true");
```

```
Configuration myHiveConfig = new Configuration()
 .withClassification("hive-site")
 .withProperties(hiveProperties);

RunJobFlowRequest request = new RunJobFlowRequest()
 .withName("Create cluster with ReleaseLabel")
 .withReleaseLabel("emr-5.20.0")
 .withApplications(hive)
 .withConfigurations(myHiveConfig)
 .withServiceRole("EMR_DefaultRole")
 .withJobFlowRole("EMR_EC2_DefaultRole")
 .withInstances(new JobFlowInstancesConfig()
  .withEc2KeyName("myEc2Key")
  .withInstanceCount(3)
  .withKeepJobFlowAliveWhenNoSteps(true)
  .withMasterInstanceType("m4.large")
  .withSlaveInstanceType("m4.large")
 );
```

在正在运行的集群中重新配置实例组

对于 Amazon EMR 5.21.0 和更高版本，您可以重新配置集群应用程序，并为运行的集群中的每个实例组指定额外的配置分类。为此，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。

当您在新的 Amazon EMR 控制台中更新实例组的应用程序配置时，控制台会尝试将新配置与现有配置合并，从而创建新的活动配置。在 Amazon EMR 无法合并配置的不寻常情况下，控制台会提醒您。

在为实例组提交重新配置请求后，Amazon EMR 为新配置规范分配一个版本号。您可以查看 CloudWatch 事件来跟踪实例组配置版本号或状态。有关更多信息，请参阅监控 CloudWatch Events。

> **Note**
>
> 您只能覆盖（而不能删除）集群创建过程中指定的集群配置。如果现有配置与您提供的文件之间存在差异，Amazon EMR 会将手动修改的配置（例如您在使用 SSH 连接到集群时修改的配置）重置为指定实例组的集群原定设置。

重新配置实例组时的注意事项

重新配置操作

当您使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK 提交重新配置请求时，Amazon EMR 会检查现有集群上的配置文件。如果现有配置与您提供的文件之间存在差异，Amazon EMR 会启动重新配置操作，重新启动某些应用程序，并将任何手动修改的配置（例如您在使用 SSH 连接到集群时修改的配置）重置为指定实例组的集群原定设置。

> **Note**
>
> Amazon EMR 会在每个实例组重新配置期间执行一些默认操作。这些默认操作可能与您所做的集群自定义冲突，并导致重新配置失败。如何对重新配置失败问题进行故障排查的相关信息，请参阅对实例组重新配置问题进行故障排查。

Amazon EMR 还会对您在请求中指定的配置分类启动重新配置操作。有关这些操作的完整列表，请参阅您使用的 Amazon EMR 版本的“配置分类”部分。例如，6.2.0 配置分类。

> **Note**
>
> Amazon EMR 版本指南中仅列出从 Amazon EMR 5.32.0 和 6.2.0 版开始的重新配置操作。

服务中断

Amazon EMR 执行“滚动”过程来重新配置任务和核心实例组中的实例。只能同时修改并重新启动实例组中的 10% 实例。该过程需要更长的时间才能完成，但降低了在运行的集群中发生潜在应用程序故障的可能性。

要在 YARN 重新启动期间运行 YARN 作业，您可以创建一个具有多个主节点的 Amazon EMR 集群，也可以在您的 `yarn-site` 配置分类中将 `yarn.resourcemanager.recovery.enabled` 设置为 `true`。有关使用多个主节点的详细信息，请参阅 YARN ResourceManager 高可用性。

应用程序验证

在重新配置重新启动过程后，Amazon EMR 会检查集群上的每个应用程序是否正在运行。如果任何应用程序不可用，则整个重新配置操作失败。如果重新配置失败，则 Amazon EMR 将配置参数恢复为以前正常工作的版本。

> **Note**
>
> 为避免重新配置失败，我们建议您仅在计划使用的集群上安装应用程序。我们还建议您在提交重新配置请求之前，先确保所有集群应用程序均正在正常运行。

重新配置的类型

您可以以两种方式之一重新配置实例组重新配置实例组：

- 覆盖。默认的重新配置方法和 Amazon EMR 5.35.0 和 6.6.0 之前的版本中提供的唯一方法。此重新配置方法不加区分地使用新提交的配置集覆盖任何集群上的文件。该方法会擦除在重新配置 API 之外对配置文件进行的任何更改。
- 合并。重新配置方法支持 Amazon EMR 版本 5.35.0 以及 6.6.0 和更高版本，Amazon EMR 控制台除外，在该控制台中，没有版本支持它。此重新配置方法合并新提交的配置与已经存在于集群中的配置。此选项仅添加或修改您提交的新配置。它会保留现有的配置。

> **Note**
>
> Amazon EMR 继续覆盖它需要的一些基本 Hadoop 配置，以确保该服务正确运行。

限制

在重新配置正在运行的集群中的实例组时，请考虑以下限制：

- 尤其是在应用程序未正确配置时，非 YARN 应用程序可能会在重新启动期间失败或导致集群问题。接近最大内存和 CPU 使用率的集群可能会在重新启动过程后遇到问题。对于主实例组而言，情况尤其如此。
- 当实例组调整大小时，您无法提交重新配置请求。如果在调整实例组大小时启动重新配置，在实例组完成大小调整后，才能启动重新配置，反之亦然。
- 在重新配置实例组后，Amazon EMR 将重新启动应用程序，从而使新配置生效。如果在重新配置期间正在使用应用程序，则可能会出现作业失败或其它意外应用程序行为。
- 如果实例组重新配置失败，则 Amazon EMR 将配置参数恢复为以前正常工作的版本。如果恢复过程也失败，您必须提交新的 `ModifyInstanceGroup` 请求以从 SUSPENDED 状态中恢复实例组。
- 仅在 Amazon EMR 5.23.0 和更高版本中支持 Phoenix 配置分类的重新配置请求，在 Amazon EMR 5.21.0 或 5.22.0 版本中不支持该请求。

- 仅在 Amazon EMR 5.30.0 和更高版本中支持 HBase 配置分类的重新配置请求，在 Amazon EMR 5.23.0 到 5.29.0 版本中不支持该请求。
- 仅在 Amazon EMR 版本 5.27.0 及更高版本中，Amazon EMR 支持在具有多个主节点的 Amazon EMR 集群上的应用程序重新配置请求。
- 具有多个主节点的 Amazon EMR 集群上不支持重新配置 `hdfs-encryption-zones` 分类或任何 Hadoop KMS 配置分类。
- Amazon EMR 当前不支持需要重新启动 YARN ResourceManager 的容量计划程序的某些重新配置请求。例如，您无法完全删除队列。

在控制台中重新配置实例组

> **Note**
>
> Amazon EMR 控制台不支持合并类型重新配置。

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/elasticmapreduce/。
2. 在集群列表中，在 Name (名称) 下面选择要重新配置的活动集群。
3. 打开集群的集群详细信息页面，然后转到 Configurations (配置) 选项卡。
4. 在 Filter (筛选条件) 下拉列表中，选择要重新配置的实例组。
5. 在 Reconfigure (重新配置) 下拉菜单中，选择 Edit in table (在表中编辑) 或 Edit in JSON file (在 JSON 文件中编辑)。

   - Edit in table (在表中编辑) – 在配置分类表中，编辑现有配置的属性和值，或者选择 Add configuration (添加配置) 来提供额外的配置分类。
   - Edit in JSON file (在 JSON 文件中编辑) – 直接在 JSON 中输入配置，也可以使用简写语法（以阴影文本表示）。否则，请为具有 JSON `Configurations` 对象的文件提供一个 Amazon S3 URI。

   > **Note**
   >
   > 配置分类表中的 Source (源) 列表示是在您创建集群时提供配置，还是在您为该实例组指定额外的配置时提供配置。您可以编辑来自两个来源的实例组配置。您无法删除初始集群配置，但可以覆盖实例组的这些配置。

您还可以直接在表中添加或编辑嵌套的配置分类。例如，要提供 hadoop-env 的额外 export 子分类，请在表中添加一个 hadoop.export 配置分类。然后，为该分类提供特定的属性和值。

6. （可选）选择 Apply this configuration to all active instance groups (将该配置应用于所有活动实例组)。
7. 保存更改。

使用 CLI 重新配置实例组

使用 modify-instance-groups 命令为运行的集群中的一个实例组指定新配置。

> **Note**
>
> 在以下示例中，将 *<j-2AL4XXXXXX5T9>* 替换为您的集群 ID，并将 *<ig-1xxxxxxx9>* 替换为您的实例组 ID。

Example – 替换实例组的配置

以下示例引用了名为 instanceGroups.json 的配置 JSON 文件，来编辑实例组的 YARN NodeManager 磁盘运行状况检查程序的属性。

1. 准备配置分类，并在运行命令的相同目录中将其保存为 instanceGroups.json。

```
[
   {
      "InstanceGroupId":"<ig-1xxxxxxx9>",
      "Configurations":[
         {
            "Classification":"yarn-site",
            "Properties":{
               "yarn.nodemanager.disk-health-checker.enable":"true",
               "yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-
percentage":"100.0"
            },
            "Configurations":[]
         }
```

```
      ]
   }
]
```

2. 运行以下命令。

```
aws emr modify-instance-groups --cluster-id <j-2AL4XXXXXX5T9> \
--instance-groups file://instanceGroups.json
```

Example – 为实例组添加配置

如果要为实例组添加配置，还必须在新的 ModifyInstanceGroup 请求中包含以前为实例组指定的配置。否则，将删除以前指定的配置。

以下示例中，为 YARN NodeManager 虚拟内存检查程序添加了一个属性。此配置还包括 YARN NodeManager 磁盘运行状况检查程序以前指定的值，不会覆盖这些值。

1. 准备 instanceGroups.json 文件中具有的以下内容，并将其保存到您将在其中运行该命令的同一目录中。

```
[
   {
      "InstanceGroupId":"<ig-1xxxxxxx9>",
      "Configurations":[
         {
            "Classification":"yarn-site",
            "Properties":{
               "yarn.nodemanager.disk-health-checker.enable":"true",
               "yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-
percentage":"100.0",
               "yarn.nodemanager.vmem-check-enabled":"true",
               "yarn.nodemanager.vmem-pmem-ratio":"3.0"
            },
            "Configurations":[]
         }
      ]
   }
]
```

2. 运行以下命令。

```
aws emr modify-instance-groups --cluster-id <j-2AL4XXXXXX5T9> \
--instance-groups file://instanceGroups.json
```

Example – 使用 Merge（合并）类型重新配置将配置添加到实例组中

当您想要使用默认的覆盖重新配置方法添加配置时，您必须在新 ModifyInstanceGroup 请求中包括该实例组的所有以前指定的配置。否则，覆盖会删除您以前指定的配置。您不需要使用合并重新配置执行此操作。相反，您必须确保您的请求仅包括新配置。

以下示例中，为 YARN NodeManager 虚拟内存检查程序添加了一个属性。因为这是合并类型重新配置，它不会覆盖以前为 YARN NodeManager 磁盘运行状况检查程序指定的值。

1. 准备 instanceGroups.json 文件中具有的以下内容，并将其保存到您将在其中运行该命令的同一目录中。

```
[
   {"InstanceGroupId":"<ig-1xxxxxxx9>",
    "ReconfigurationType" :"MERGE",
      "Configurations":[
         {"Classification":"yarn-site",
            "Properties":{
               "yarn.nodemanager.vmem-check-enabled":"true",
               "yarn.nodemanager.vmem-pmem-ratio":"3.0"
            },
            "Configurations":[]
         }
      ]
   }
]
```

2. 运行以下命令。

```
aws emr modify-instance-groups --cluster-id <j-2AL4XXXXXX5T9> \
--instance-groups file://instanceGroups.json
```

Example – 删除实例组的配置

要删除实例组的配置，请提交新的重新配置请求以排除以前的配置。

> **Note**
>
> 您只能覆盖初始集群配置。您无法删除该配置。

例如，您可以删除在上一示例中指定的 YARN NodeManager 磁盘运行状况检查程序配置，提交包含以下内容的新 `instanceGroups.json`。

```
[
   {
      "InstanceGroupId":"<ig-1xxxxxxx9>",
      "Configurations":[
         {
            "Classification":"yarn-site",
            "Properties":{
               "yarn.nodemanager.vmem-check-enabled":"true",
               "yarn.nodemanager.vmem-pmem-ratio":"3.0"
            },
            "Configurations":[]
         }
      ]
   }
]
```

> **Note**
>
> 要删除上一重新配置请求中的所有配置，请提交包含一组空配置的重新配置请求。例如：
>
> ```
> [
>    {
>       "InstanceGroupId":"<ig-1xxxxxxx9>",
>       "Configurations":[]
>    }
> ]
> ```

Example – 在一个请求中重新配置实例组并调整其大小

以下示例 JSON 说明了如何在同一请求中重新配置实例组并调整其大小。

```
[
   {
      "InstanceGroupId":"<ig-1xxxxxxx9>",
      "InstanceCount":5,
      "EC2InstanceIdsToTerminate":["i-123"],
      "ForceShutdown":true,
      "ShrinkPolicy":{
         "DecommissionTimeout":10,
         "InstanceResizePolicy":{
            "InstancesToTerminate":["i-123"],
            "InstancesToProtect":["i-345"],
            "InstanceTerminationTimeout":20
         }
      },
      "Configurations":[
         {
            "Classification":"yarn-site",
            "Configurations":[],
            "Properties":{
               "yarn.nodemanager.disk-health-checker.enable":"true",
               "yarn.nodemanager.disk-health-checker.max-disk-utilization-per-disk-
percentage":"100.0"
            }
         }
      ]
   }
]
```

使用 Java SDK 重新配置实例组

> **Note**
>
> 在以下示例中，将 *<j-2AL4XXXXXX5T9>* 替换为您的集群 ID，并将 *<ig-1xxxxxxx9>* 替换为您的实例组 ID。

以下代码段使用 Amazon SDK for Java 为实例组提供新配置。

```
AWSCredentials credentials = new BasicAWSCredentials("access-key", "secret-key");
AmazonElasticMapReduce emr = new AmazonElasticMapReduceClient(credentials);

Map<String,String> hiveProperties = new HashMap<String,String>();
hiveProperties.put("hive.join.emit.interval","1000");
hiveProperties.put("hive.merge.mapfiles","true");

Configuration configuration = new Configuration()
    .withClassification("hive-site")
    .withProperties(hiveProperties);

InstanceGroupModifyConfig igConfig = new InstanceGroupModifyConfig()
    .withInstanceGroupId("<ig-1xxxxxxx9>")
    .withReconfigurationType("MERGE");
    .withConfigurations(configuration);

ModifyInstanceGroupsRequest migRequest = new ModifyInstanceGroupsRequest()
    .withClusterId("<j-2AL4XXXXXX5T9>")
    .withInstanceGroups(igConfig);

emr.modifyInstanceGroups(migRequest);
```

以下代码段通过提供一组空配置来删除以前为实例组指定的配置。

```
List<Configuration> configurations = new ArrayList<Configuration>();

InstanceGroupModifyConfig igConfig = new InstanceGroupModifyConfig()
    .withInstanceGroupId("<ig-1xxxxxxx9>")
    .withConfigurations(configurations);

ModifyInstanceGroupsRequest migRequest = new ModifyInstanceGroupsRequest()
    .withClusterId("<j-2AL4XXXXXX5T9>")
    .withInstanceGroups(igConfig);

emr.modifyInstanceGroups(migRequest);
```

对实例组重新配置问题进行故障排查

如果实例组的重新配置过程失败，Amazon EMR 将恢复重新配置并使用 Amazon CloudWatch 事件记录失败消息。此事件能够提供该重新配置失败的简短摘要。其会列出重新配置失败的实例以及相应的失败消息。下面是一个失败消息示例。

```
The reconfiguration operation for instance group ig-1xxxxxxx9 in Amazon EMR
 cluster j-2AL4XXXXXX5T9 (ExampleClusterName)
failed at 2021-01-01 00:00 UTC and took 2 minutes to fail. Failed configuration version
 is example12345.
Failure message: Instance i-xxxxxxx1, i-xxxxxxx2, i-xxxxxxx3 failed with message "This
 is an example failure message".
```

要收集有关重新配置失败的更多数据，您可以检查节点预配置日志。对于收到类似以下消息时，这样做尤其有用。

```
i-xxxxxxx1 failed with message “Unable to complete transaction and some changes were
 applied.”
```

On the node

通过连接到节点来访问节点预配置日志

1. 使用 SSH 连接到重新配置失败的节点。有关说明，请参阅《适用于 Linux 实例的 Amazon EC2 用户指南》中的连接到您的 Linux 实例。
2. 导航到以下包含节点预置日志文件的目录。

   ```
   /mnt/var/log/provision-node/
   ```

3. 打开 `reports` 子目录并搜索节点预置报告，从而进行重新配置。`reports` 目录根据重新配置版本号、通用唯一标识符（UUID）、Amazon EC2 实例 IP 地址和时间戳来组织日志。每个报告都是一个压缩的 YAML 文件，其中包含有关重新配置过程的详细信息。

   以下为报告文件名和路径的示例。

   ```
   /reports/2/ca598xxx-cxxx-4xxx-bxxx-6dbxxxxxxxxx/ip-10-73-xxx-
   xxx.ec2.internal/202104061715.yaml.gz
   ```

4. 您可以使用文件查看器（如以下示例中的 `zless`）来查看报告。

```
zless 202104061715.yaml.gz
```

Amazon S3

使用 Amazon S3 来访问节点预配置日志

1. 登录到 Amazon Web Services Management Console，然后通过以下网址打开 Amazon S3 控制台：https://console.aws.amazon.com/s3/。
2. 打开您在配置集群时指定的 Amazon S3 存储桶，以便将日志文件存档。
3. 导航到以下包含节点预配置日志文件的文件夹：

   ```
   DOC-EXAMPLE-BUCKET/elasticmapreduce/<cluster id>/node/<instance id>/provision-node/
   ```

4. 打开 `reports` 文件夹并搜索节点预配置报告，从而进行重新配置。`reports` 文件夹根据重新配置版本号、通用唯一标识符（UUID）、Amazon EC2 实例 IP 地址和时间戳来组织日志。每个报告都是一个压缩的 YAML 文件，其中包含有关重新配置过程的详细信息。

   以下为报告文件名和路径的示例。

   ```
   /reports/2/ca598xxx-cxxx-4xxx-bxxx-6dbxxxxxxxxx/ip-10-73-xxx-xxx.ec2.internal/202104061715.yaml.gz
   ```

5. 要查看日志文件，您可以将其作为文本文件从 Amazon S3 下载到本地计算机。有关说明，请参阅下载对象。

每个日志文件都包含关联重新配置的详细预配置报告。要查找错误消息信息，您可以搜索报告的 `err` 日志级别。报告格式取决于集群上的 Amazon EMR 版本。

以下示例显示了早于 Amazon EMR 5.32.0 和 6.2.0 发行版的错误信息。

```
- !ruby/object:Puppet::Util::Log
      level: !ruby/sym err
      tags:
        - err
      message: "Example detailed error message."
      source: Puppet
      time: 2021-01-01 00:00:00.000000 +00:00
```

Amazon EMR 5.32.0 和 6.2.0 及更高发行版使用以下格式。

```
- level: err
  message: 'Example detailed error message.'
  source: Puppet
  tags:
  - err
  time: '2021-01-01 00:00:00.000000 +00:00'
  file:
  line:
```

在 Amazon Secrets Manager 中存储敏感配置数据

Amazon EMR 描述并列出了以明文形式发出自定义配置数据（例如 `DescribeCluster` 和 `ListInstanceGroups`）的 API 操作。Amazon EMR 与 Amazon Secrets Manager 集成，因此您可以将数据存储在 Secrets Manager 中，并在配置中使用密钥 ARN。这样，您就不会将敏感的配置数据以明文形式传递给 Amazon EMR，也不会将其公开给外部 API。如果您指明键值对包含存储在 Secrets Manager 中的密钥 ARN，Amazon EMR 则会在向集群发送配置数据时检索此密钥。Amazon EMR 在使用外部 API 显示配置时不会发送注释。

创建密钥

要创建密钥，请遵循《Amazon Secrets Manager 用户指南》中创建 Amazon Secrets Manager 密钥的说明。在步骤 3 中，必须选择 Plaintext（明文）字段来输入您的敏感值。

请注意，虽然 Secrets Manager 允许密钥最多包含 65536 个字节，但 Amazon EMR 将属性键（不包括注释）和检索到的密钥值的组合长度限制为 1024 个字符。

授予 Amazon EMR 检索密钥的访问权限

Amazon EMR 使用 IAM 服务角色为您预置和管理集群。Amazon EMR 服务角色定义在预置资源，以及执行在集群中运行的特定 Amazon EC2 实例的上下文中不执行的服务级任务时，允许 Amazon EMR 执行的操作。有关服务角色的更多信息，请参阅 Amazon EMR 的服务角色（EMR 角色）和自定义 IAM 角色。

要允许 Amazon EMR 从 Secrets Manager 检索密钥值，请在启动集群时将下面的策略声明添加到您的 Amazon EMR 角色中。

```
{
   "Sid":"AllowSecretsRetrieval",
```

```
   "Effect":"Allow",
   "Action":"secretsmanager:GetSecretValue",
   "Resource":[
      "arn:aws:secretsmanager:<region>:<aws-account-id>:secret:<secret-name>"
   ]
}
```

如果您使用客户托管的 Amazon KMS key 创建密钥，则还必须为 Amazon EMR 角色添加对所用键的 `kms:Decrypt` 权限。有关更多信息，请参阅 Amazon Secrets Manager 用户指南中的 Amazon Secrets Manager 的身份验证和访问控制。

在配置分类中使用密钥

您可以向任何配置属性添加 `EMR.secret@` 注释，以表明其键值对包含存储在 Secrets Manager 中的密钥 ARN。

以下示例演示如何在配置分类中提供密钥 ARN：

```
{
   "Classification":"core-site",
   "Properties":{
      "presto.s3.access-key":"<sensitive-access-key>",
      "EMR.secret@presto.s3.secret-key":"arn:aws:secretsmanager:<region>:<aws-account-id>:secret:<secret-name>"
   }
}
```

在创建集群并提交注释的配置后，Amazon EMR 会验证配置属性。如果您的配置有效，Amazon EMR 将从配置中删除注释并从 Secrets Manager 中检索该密钥以创建实际配置，然后再将其应用于集群：

```
{
   "Classification":"core-site",
   "Properties":{
      "presto.s3.access-key":"<sensitive-access-key>",
      "presto.s3.secret-key":"<my-secret-key-retrieved-from-Secrets-Manager>"
   }
}
```

当您调用类似 `DescribeCluster` 的操作时，Amazon EMR 将返回集群上的当前应用程序配置。如果应用程序配置属性被标记为包含密钥 ARN，则 `DescribeCluster` 调用返回的应用程序配置包含 ARN 而不是密钥值。这样可以确保密钥值仅在集群上可见：

```
{
   "Classification":"core-site",
   "Properties":{
      "presto.s3.access-key":"<sensitive-access-key>",
      "presto.s3.secret-key":"arn:aws:secretsmanager:<region>:<aws-account-
id>:secret:<secret-name>"
   }
}
```

更新密钥值

每当连接的实例组启动、重新配置或调整大小时，Amazon EMR 都会从注释的配置中检索密钥值。您可以使用 Secrets Manager 修改正在运行的集群的配置中使用的密钥的值。完成后，您可以向想要接收更新值的每个实例组提交重新配置请求。如需详细了解如何重新配置实例组以及重新配置时需要考虑的事项，请参阅 在正在运行的集群中重新配置实例组。

配置应用程序来使用特定 Java 虚拟机

Amazon EMR 版本具有不同的默认 Java 虚拟机（JVM）版本。本页介绍了 JVM 对不同版本和应用程序的支持。

注意事项

- 对于 Amazon EMR 5.0.0 及更高版本，默认 Java 虚拟机（JVM）为 Java 8。
- 对于 Amazon EMR 6.9.0 及更高版本，Trino 默认支持 Java 17。有关 Trino 上 Java 17 的更多信息，请参阅 Trino 博客上的 Trino updates to Java 17。
- 对于 Amazon EMR 6.12.0 及更高版本，某些应用程序还支持 Java 11 和 17。有关支持的应用程序 Java 版本的信息，请参阅《Amazon EMR Release Guide》https://docs.amazonaws.cn/emr/latest/ReleaseGuide/emr-release-components.html。请注意，Amazon EMR 仅支持在一个集群中运行一个运行时系统版本，不支持在同一集群上的不同运行时系统版本上运行不同的节点或应用程序。

在选择运行时系统版本时，请记住以下特定于应用程序的注意事项：

特定于应用程序的 Java 配置说明

应用程序	Java 配置说明
Spark	要使用非默认 Java 版本运行 Spark，必须同时配置 Spark 和 Hadoop。有关示例，请参阅 覆盖 JVM。 • 在 `spark-env` 中配置 `JAVA_HOME` 以更新主实例进程的 Java 运行时系统。例如，spark-submit、spark-shell 和 Spark 历史记录服务器。 • 修改 Hadoop 配置以更新 Spark 执行程序和 YARN `ApplicationMaster` 的 Java 运行时系统
Spark RAPIDS	您可以使用为 Spark 配置的 Java 版本运行 RAPIDS。
Iceberg	您可以使用正在使用 Iceberg 的应用程序的已配置 Java 版本来运行 Iceberg。
Delta	您可以使用正在使用 Delta 的应用程序的已配置 Java 版本来运行 Delta。
Hudi	您可以使用正在使用 Hudi 的应用程序的已配置 Java 版本来运行 Hudi。
Hive	要将 Hive 的 Java 版本设置为 11 或 17，请将 Hadoop JVM 设置配置为要使用的 Java 版本。
HBase	要更新适用于 HBase 的 JVM，请修改 `hbase-env` 。默认情况下，除非您覆盖 `hbase-env` 中的设置，否则 Amazon EMR 会根据 Hadoop 的 JVM 配置来设置 HBase JVM。有关示例，请参阅 覆盖 JVM。

应用程序	Java 配置说明
Flink	要更新适用于 Flink 的 JVM，请修改 `flink-conf` 。默认情况下，除非您覆盖 `flink-conf` 中的设置，否则 Amazon EMR 会根据 Hadoop 的 JVM 配置来设置 Flink JVM。有关更多信息，请参阅将 Flink 配置为使用 Java 11 运行。
Oozie	要将 Oozie 配置为在 Java 11 或 17 上运行，请配置 Oozie Server、Oozie LauncherAM Launcher AM，然后更改客户端可执行文件和作业配置。您也可以将 `EmbeddedOozieServer` 配置为在 Java 17 上运行。有关更多信息，请参阅为 Oozie 配置 Java 版本。
Pig	Pig 仅支持 Java 8。您无法在 Hadoop 中使用 Java 11 或 17，也不能在同一个集群上运行 Pig。

覆盖 JVM

要覆盖 Amazon EMR 版本的 JVM 设置 – 例如，在使用 Amazon EMR 版本 6.12.0 的集群中使用 Java 17，请为其环境分类提供 `JAVA_HOME` 设置，该设置对于除 Flink 之外的所有应用程序都为 *`application`*`-env`。对于 Flink 来说，环境分类是 `flink-conf`。有关使用 Flink 配置 Java 运行时系统的步骤，请参阅 将 Flink 配置为使用 Java 11 运行。

主题

- 使用 Apache Spark 覆盖 JVM 设置
- 使用 Apache HBase 覆盖 JVM 设置
- 使用 Apache Hadoop 和 Hive 覆盖 JVM 设置

使用 Apache Spark 覆盖 JVM 设置

在 Amazon EMR 6.12 及更高版本中使用 Spark 时，如果您编写驱动程序以在集群模式下提交，则驱动程序使用 Java 8，但您可以将环境设置为执行程序使用 Java 11 或 17。而且，当您在低于 5.x 的

Amazon EMR 版本中使用 Spark 并编写驱动程序以在集群模式下提交时，驱动程序会使用 Java 7。不过，您可以设置环境以确保执行程序使用 Java 8。

要覆盖 Spark 的 JVM，我们建议您同时设置 Hadoop 和 Spark 分类。

```
{
"Classification": "hadoop-env",
        "Configurations": [
            {
"Classification": "export",
                "Configurations": [],
                "Properties": {
"JAVA_HOME": "/usr/lib/jvm/java-1.8.0"
                }
            }
        ],
        "Properties": {}
    },
    {
"Classification": "spark-env",
        "Configurations": [
            {
"Classification": "export",
                "Configurations": [],
                "Properties": {
"JAVA_HOME": "/usr/lib/jvm/java-1.8.0"
                }
            }
        ],
        "Properties": {}
    }
```

使用 Apache HBase 覆盖 JVM 设置

要将 HBase 配置为使用 Java 11，可以在启动集群时设置以下配置。

```
[
  {
    "Classification": "hbase-env",
    "Configurations": [
      {
        "Classification": "export",
```

```
        "Configurations": [],
        "Properties": {
          "JAVA_HOME": "/usr/lib/jvm/jre-11"
        }
      }
    ],
    "Properties": {}
  }
]
```

使用 Apache Hadoop 和 Hive 覆盖 JVM 设置

以下示例说明如何将 Hadoop 和 Hive 的 JVM 设置为版本 17。

```
[
    {
        "Classification": "hadoop-env",
            "Configurations": [
                {
                    "Classification": "export",
                    "Configurations": [],
                    "Properties": {
                        "JAVA_HOME": "/usr/lib/jvm/jre-17"
                    }
                }
        ],
        "Properties": {}
    }
]
```

服务端口

以下是 YARN 和 HDFS 服务端口。这些设置反映 Hadoop 默认值。其它应用程序服务托管在默认端口上，除非另有指定。有关更多信息，请参阅应用程序的项目文档。

YARN 和 HDFS 的端口设置

设置	主机名/端口
fs.default.name	默认值 (hdfs://*emrDeterminedIP* :8020)
dfs.datanode.address	默认值 (0.0.0.0:50010)

设置	主机名/端口
dfs.datanode.http.address	默认值 (0.0.0.0:50075)
dfs.datanode.https.address	默认值 (0.0.0.0:50475)
dfs.datanode.ipc.address	默认值 (0.0.0.0:50020)
dfs.http.address	默认值 (0.0.0.0:50070)
dfs.https.address	默认值 (0.0.0.0:50470)
dfs.secondary.http.address	默认值 (0.0.0.0:50090)
yarn.nodemanager.address	默认值 (${yarn.nodemanager.hostname}:0)
yarn.nodemanager.localizer.address	默认值 (${yarn.nodemanager.hostname}:8040)
yarn.nodemanager.webapp.address	默认值 (${yarn.nodemanager.hostname}:8042)
yarn.resourcemanager.address	默认值 (${yarn.resourcemanager.hostname}:8032)
yarn.resourcemanager.admin.address	默认值 (${yarn.resourcemanager.hostname}:8033)
yarn.resourcemanager.resource-tracker.address	默认值 (${yarn.resourcemanager.hostname}:8031)
yarn.resourcemanager.scheduler.address	默认值 (${yarn.resourcemanager.hostname}:8030)
yarn.resourcemanager.webapp.address	默认值 (${yarn.resourcemanager.hostname}:8088)
yarn.web-proxy.address	默认值（无值）
yarn.resourcemanager.hostname	*emrDeterminedIP*

> **Note**
>
> 术语 *emrDeterminedIP* 意指由 Amazon EMR 控制面板生成的 IP 地址。在较新的版本中，已删除该约定，但 `yarn.resourcemanager.hostname` 和 `fs.default.name` 设置除外。

应用程序用户

应用程序以自己的用户身份运行进程。例如，Hive JVM 以 `hive` 用户身份运行，MapReduce JVM 以 `mapred` 身份运行，等等。在以下进程状态示例中说明了这一点。

```
USER       PID %CPU %MEM    VSZ   RSS TTY      STAT START   TIME COMMAND
hive      6452  0.2  0.7 853684 218520 ?       Sl   16:32   0:13 /usr/lib/jvm/
java-openjdk/bin/java -Xmx256m -Dhive.log.dir=/var/log/hive -Dhive.log.file=hive-
metastore.log -Dhive.log.threshold=INFO -Dhadoop.log.dir=/usr/lib/hadoop
hive      6557  0.2  0.6 849508 202396 ?       Sl   16:32   0:09 /usr/lib/jvm/java-
openjdk/bin/java -Xmx256m -Dhive.log.dir=/var/log/hive -Dhive.log.file=hive-server2.log
 -Dhive.log.threshold=INFO -Dhadoop.log.dir=/usr/lib/hadoop/l
hbase     6716  0.1  1.0 1755516 336600 ?      Sl   Jun21   2:20 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_master -XX:OnOutOfMemoryError=kill -9 %p -Xmx1024m -ea -XX:
+UseConcMarkSweepGC -XX:+CMSIncrementalMode -Dhbase.log.dir=/var/
hbase     6871  0.0  0.7 1672196 237648 ?      Sl   Jun21   0:46 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_thrift -XX:OnOutOfMemoryError=kill -9 %p -Xmx1024m -ea -XX:
+UseConcMarkSweepGC -XX:+CMSIncrementalMode -Dhbase.log.dir=/var/
hdfs      7491  0.4  1.0 1719476 309820 ?      Sl   16:32   0:22 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_namenode -Xmx1000m -Dhadoop.log.dir=/var/log/hadoop-hdfs -
Dhadoop.log.file=hadoop-hdfs-namenode-ip-10-71-203-213.log -Dhadoo
yarn      8524  0.1  0.6 1626164 211300 ?      Sl   16:33   0:05 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_proxyserver -Xmx1000m -Dhadoop.log.dir=/var/log/hadoop-yarn -
Dyarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-yarn-
yarn      8646  1.0  1.2 1876916 385308 ?      Sl   16:33   0:46 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_resourcemanager -Xmx1000m -Dhadoop.log.dir=/var/log/hadoop-yarn
 -Dyarn.log.dir=/var/log/hadoop-yarn -Dhadoop.log.file=yarn-y
mapred    9265  0.2  0.8 1666628 260484 ?      Sl   16:33   0:12 /usr/lib/jvm/java-
openjdk/bin/java -Dproc_historyserver -Xmx1000m -Dhadoop.log.dir=/usr/lib/hadoop/logs -
Dhadoop.log.file=hadoop.log -Dhadoop.home.dir=/usr/lib/hadoop
```

使用 Amazon EMR 项目存储库检查依赖项

您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发布版 5.18.0 开始）附带的准确版本的库和依赖项的 Apache Hive 和 Apache Hadoop 任务代码。针对存储库中的 Amazon EMR 项目进行构建可确保针对其创建任务的库的版本是在集群上运行时提供的相同版本，从而帮助避免运行时类路径问题。目前，Amazon EMR 项目仅适用于 Maven 构建。

要访问项目存储库，请将存储库 URL 添加到 Maven 设置文件或特定项目的 `pom.xml` 配置文件。之后，您可以在项目配置中指定依赖项。对于依赖项版本，请为 Amazon EMR 5.x 发行版 上所需的发行版使用 Component Versions (组件版本) 下列出的版本。例如，the section called “组件版本” 上提供了最新 Amazon EMR 发行版的组件版本。如果项目的构件未在 Component Versions (组件版本) 下列出，请指定为该发行版中的 Hive 和 Hadoop 列出的版本。例如，对于 Amazon EMR 发行版 5.18.0 中的 Hadoop 组件，版本为 `2.8.4-amzn-1`。

项目存储库 URL 具有以下语法：

```
https://s3-endpoint/region-ID-emr-artifacts/emr-release-label/repos/maven/
```

- *`s3-endpoint`* 是存储库区域的 Amazon Simple Storage Service（Amazon S3）终端节点，而 *`region-ID`* 是对应的区域。例如，`s3.us-west-1.amazonaws.com` 和 `us-west-1`。有关更多信息，请参阅《Amazon Web Services 一般参考》中的 Amazon S3 endpoints。由于区域之间的项目不存在差异，因此，您可以为开发环境指定最方便的区域。
- *`emr-release-label`* 是将运行代码的 Amazon EMR 集群的发行版标注。发行版标注的格式是 `emr-`*`x.x.x`*，例如 `emr-5.36.1`。一个 EMR 版本系列可能包含多个版本。例如，如果您使用的是 EMR 发行版 5.24.1，请在构件存储库 URL 中使用 5.24 系列中的第一个 EMR 发行版标注，即 `emr-5.24.0`：

  ```
  https://s3-endpoint/region-ID-emr-artifacts/emr-5.24.0/repos/maven/
  ```

Example Maven pom.xml 的配置

以下 pom.xml 示例配置 Maven 项目以使用 us-west-1 中的项目存储库针对 emr-5.18.0 Apache Hadoop 和 Apache Hive 项目进行构建。由于快照版本在项目存储库中不可用，因此已在 pom.xml 中禁用快照。以下示例中的省略号 (`...`) 指示忽略其它配置参数。请勿将它们复制到 Maven 项目。

```
<project>
```

```
 ...
 <repositories>
  ...
  <repository>
   <id>emr-5.18.0-artifacts</id>
   <name>EMR 5.18.0 Releases Repository</name>
   <releases>
    <enabled>true</enabled>
   </releases>
   <snapshots>
    <enabled>false</enabled>
   </snapshots>
   <url>https://s3.us-west-1.amazonaws.com/us-west-1-emr-artifacts/emr-5.18.0/repos/
maven/</url>
  </repository>
  ...
 </repositories>
 ...
 <dependencies>
  ...
  <dependency>
   <groupId>org.apache.hive</groupId>
   <artifactId>hive-exec</artifactId>
   <version>2.3.3-amzn-2</version>
  </dependency>
  <dependency>
   <groupId>org.apache.hadoop</groupId>
   <artifactId>hadoop-common</artifactId>
   <version>2.8.4-amzn-1</version>
  </dependency>
  ...
 </dependencies>

</project>
```

EMR 文件系统 (EMRFS)

EMR 文件系统（EMRFS）是 HDFS 的实现，所有 Amazon EMR 集群将其用于直接从 Amazon EMR 读取常规文件并将其写入 Amazon S3。EMRFS 使您能够方便地将持久性数据存储在 Amazon S3 中以便用于 Hadoop，同时它还提供了数据加密等功能。

数据加密可让您对由 EMRFS 写入 Amazon S3 的对象进行加密，并且还允许 EMRFS 处理 Amazon S3 中的加密对象。如果您使用的是 Amazon EMR 发行版 4.8.0 或更高版本，则可使用安全配置设置 Amazon S3 中 EMRFS 对象的加密以及其他加密设置。有关更多信息，请参阅加密选项。如果您使用的是 Amazon EMR 的早期发行版，则可以手动配置加密设置。有关更多信息，请参阅使用 EMRFS 属性指定 Amazon S3 加密。

Amazon S3 在Amazon Web Services 区域为所有 GET、PUT 和 LIST 操作提供了强大的读写后编写一致性。这意味着您使用 EMRFS 编写的内容就是从 Amazon S3 中读取的内容，对性能没有影响。有关更多信息，请参阅 Amazon S3 数据一致性模型。

在使用 Amazon EMR 发行版 5.10.0 或更高版本时，可以根据集群用户、组或 EMRFS 数据在 Amazon S3 中的位置，使用不同的 IAM 角色来处理 EMRFS 对 Amazon S3 的请求。有关更多信息，请参阅为处理 EMRFS 对 Amazon S3 的请求配置 IAM 角色。

> **Warning**
>
> 在为运行 Apache Spark 任务的 Amazon EMR 集群启用推测执行之前，请查看以下信息。EMRFS 包含经 EMRFS S3 优化的提交程序（OutputCommitter 的一种实现替代品），该程序已针对使用 EMRFS 时向 Amazon S3 写入文件进行了优化。如果您对将数据写入 Amazon S3 的应用程序启用 Apache Spark 推测执行功能，并且不使用经 EMRFS S3 优化的提交程序，则可能会遇到 SPARK-10063 中描述的数据正确性问题。如果您使用的是低于 Amazon EMR 5.19 版本的 Amazon EMR 版本，或者正在使用 ORC 和 CSV 等格式将文件写入 Amazon S3，则会发生该情况。EMRFS S3 优化的提交者不支持这些格式。有关使用经 EMRFS S3 优化的提交程序的完整要求列表，请参阅经 EMRFS S3 优化的提交程序的要求。
> 当经 EMRFS S3 优化的提交程序不受支持时（例如在写入以下内容时），通常使用 EMRFS 直接写入：
>
> - Parquet 以外的输出格式（例如 ORC 或文本）。
> - 使用 Spark RDD API 的 Hadoop 文件。
> - 使用 Hive SerDe 的 Parquet。请参阅 Hive 元存储 Parquet 表转换。

以下情形不使用 EMRFS 直接写入

- 启用经 EMRFS S3 优化的提交程序时 请查看经 EMRFS S3 优化的提交程序的要求。
- 写入动态分区时，请将 partitionOverwriteMode 设置为动态。
- 写入自定义分区位置（例如不符合 Hive 默认分区位置约定的位置）时。
- 使用 EMRFS 以外的文件系统（例如写入 HDFS 或使用 S3A 文件系统）时。

要确定您的应用程序是否在 Amazon EMR 5.14.0 或更高版本中使用了直接写入，请启用 Spark INFO 日志记录。如果 Spark 驱动程序日志或 Spark 执行程序容器日志中，存在包含文本“Direct Write: ENABLED”的日志行，则 Spark 应用程序会使用直接写入的方式进行写入。默认情况下，Amazon EMR 集群上的推测执行处于 OFF 状态。如果以下两个条件都为真，我们强烈建议您不要启用推测执行：

- 您正将数据写入 Amazon S3。
- 数据以 Apache Parque 以外的格式写入，或者以 Apache Parquet 格式写入但不使用经 EMRFS S3 优化的提交程序。

如果您启用 Spark 推测执行并使用 EMRFS 直接写入，将数据写入 Amazon S3，您可能会遇到间歇性数据丢失问题。将数据写入 HDFS，或使用经 EMRFS S3 优化的提交程序以 Parquet 格式写入数据时，Amazon EMR 不使用直接写入，也不会发生此问题。
如果您需要以使用 EMRFS 直接写入的格式从 Spark 将数据写入 Amazon S3，并使用推测执行，我们建议您写入 HDFS，然后使用 S3DistCP 将输出文件传输到 Amazon S3。

主题

一致视图

> **Warning**
>
> 2023 年 6 月 1 日，EMRFS 一致视图将终止对未来 Amazon EMR 发行版的标准支持。EMRFS 一致视图将继续支持现有发行版。

2020 年 12 月 1 日，Amazon S3 发布了强大的先写后读一致性，此后您不再需要对 Amazon EMR 集群使用 EMRFS 一致视图（EMRFS CV）。EMRFS CV 是一项可选功能，允许 Amazon EMR 集群检查 Amazon S3 对象的列表和先写后读一致性。在创建集群并开启了 EMRFS CV 时，Amazon EMR 将创建一个 Amazon DynamoDB 数据库来存储用于跟踪 S3 对象的列表和先写后读一致性的对象元数据。现在，您可以关闭 EMRFS CV 并删除它使用的 DynamoDB 数据库，这样就不会产生额外费用。以下过程说明了如何检查、关闭 CV 功能以及删除该功能使用的 DynamoDB 数据库。

检查您是否在使用 EMRFS CV 功能

1. 导航到 Configuration（配置）选项卡。如果您的集群具有以下配置，它将使用 EMRFS CV。

   ```
   Classification=emrfs-site,Property=fs.s3.consistent,Value=true
   ```

2. 或者，请使用 Amazon CLI 来通过 `describe-cluster` API 描述您的集群。如果输出包含 `fs.s3.consistent: true`，则您的集群使用 EMRFS CV。

在您的 Amazon EMR 集群上关闭 EMRFS CV

要关闭 EMRFS CV 功能，请使用以下三个选项之一。在将这些选项应用到生产环境之前，应先在测试环境中测试它们。

1. 停止现有集群并启动没有 EMRFS CV 选项的新集群。

   a. 在停止集群之前，请务必备份数据并通知用户。

   b. 要停止集群，请按照终止集群中的说明操作。

   c. 如果您使用 Amazon EMR 控制台创建新集群，请导航到 Advanced Options（高级选项）。在 Edit software settings（编辑软件设置）部分中，取消选中该选项以打开 EMRFS CV。如果 EMRFS consistent view（EMRFS 一致视图）复选框可用，请保持其未选中。

   d. 如果您使用 Amazon CLI 通过 `create-cluster` API 创建新集群，请勿使用 `--emrfs` 选项，该选项将打开 EMRFS CV。

e. 如果您使用开发工具包或 Amazon CloudFormation 创建新集群，请不要使用 Configure consistent view（配置一致视图）中列出的任何配置。

2. 克隆集群并删除 EMRFS CV

a. 在 Amazon EMR 控制台中，选择使用 EMRFS CV 的集群。

b. 在 Cluster Details（集群详细信息）页面顶部，选择 Clone（克隆）。

c. 选择 Previous（上一步）并导航至 Step 1: Software and Steps（步骤 1：软件和步骤）。

d. 在 Edit software settings（编辑软件设置）中，删除 EMRFS CV。在 Edit configuration（编辑配置）椎间盘每个。删除 `emrfs-site` 分类中的以下配置。如果您要从 S3 存储桶加载 JSON，则必须修改您的 S3 对象。

```
[
 {"classification":
  "emrfs-site",
  "properties": {
   "fs.s3.consistent.retryPeriodSeconds":"10",
   "fs.s3.consistent":"true",
   "fs.s3.consistent.retryCount":"5",
   "fs.s3.consistent.metadata.tableName":"EmrFSMetadata"
  }
 }
]
```

3. 从使用实例组的集群中删除 EMRFS CV

a. 使用以下命令检查是否有一个 EMR 集群使用与 EMRFS CV 关联的 DynamoDB 表，或者是否有多个集群共享该表。表名称在 `fs.s3.consistent.metadata.tableName` 中指定，如 Configure consistent view（配置一致视图）中所述。EMRFS CV 使用的默认表名称为 `EmrFSMetadata`。

```
aws emr describe-cluster --cluster-id j-XXXXX | grep
 fs.s3.consistent.metadata.tableName
```

b. 如果您的集群未与其他集群共享您的 DynamoDB 数据库，请使用以下命令重新配置集群并停用 EMRFS CV。有关更多信息，请参阅重新配置正在运行的集群中的实例组。

```
aws emr modify-instance-groups --cli-input-json file://disable-emrfs-1.json
```

此命令打开要修改的文件。请使用以下配置修改此文件。

```
{
 "ClusterId": "j-xxxx",
 "InstanceGroups": [
  {
   "InstanceGroupId": "ig-xxxx",
   "Configurations": [
    {
     "Classification": "emrfs-site",
     "Properties": {
      "fs.s3.consistent": "false"
     },
     "Configurations": []
    }
   ]
  }
 ]
}
```

c. 如果您的集群与其他集群共享 DynamoDB 表，请在没有集群修改共享 S3 位置中的任何对象时关闭所有集群上的 EMRFS CV。

删除与 EMRFS CV 关联的 Amazon DynamoDB 资源

从 Amazon EMR 集群中删除 EMRFS CV 后，请删除与 EMRFS CV 关联的 DynamoDB 资源。在您执行此操作之前，您将继续承担与 EMRFS CV 相关的 DynamoDB 费用。

1. 检查您的 DynamoDB 表的 CloudWatch 指标，并确认该表未被任何集群使用。
2. 删除 DynamoDB 表。

```
aws dynamodb delete-table --table-name <your-table-name>
```

删除与 EMRFS CV 关联的 Amazon SQS 资源

1. 如果您将集群配置为向 Amazon SQS 推送不一致通知，则可以删除所有 SQS 队列。
2. 查找 `fs.s3.consistent.notification.SQS.queueName` 中指定的 Amazon SQS 队列名称，如 Configure consistent view（配置一致视图）中所述。默认队列名称格式为 `EMRFS-Inconsistency-<j-cluster ID>`。

```
aws sqs list-queues | grep 'EMRFS-Inconsistency'
```

```
aws sqs delete-queue -queue-url <your-queue-url>
```

停止使用 EMRFS CLI

- EMRFS CLI 管理 EMRFS CV 生成的元数据。随着对 EMRFS CV 的标准支持在 Amazon EMR 的未来版本中即将结束，对 EMRFS CLI 的支持也将结束。

主题

启用一致视图

您可以使用 Amazon Web Services Management Console、Amazon CLI，或 `emrfs-site` 配置分类，为 EMRFS 启用 Amazon S3 服务器端加密或一致视图。

使用控制台配置一致视图

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 选择 Step 1: Software and Steps (步骤 1: 软件和步骤) 和 Step 2: Hardware (步骤 2: 硬件) 的设置。
4. 对于 Step 3: General Cluster Settings (步骤 3: 常规集群设置)，在 Additional Options (附加选项) 下选择 EMRFS consistent view (EMRFS 一致视图)。
5. 对于 EMRFS Metadata store (EMRFS 元数据存储)，键入您的元数据存储的名称。默认值为 **`EmrFSMetadata`**。如果 EmrFSMetadata 表不存在，则在 DynamoDB 中为您创建它。

> **Note**
>
> 集群终止时，Amazon EMR 不会自动从 DynamoDB 中删除 EMRFS 元数据。

6. 对于 Number of retries (重试次数)，键入一个整数值。如果检测到不一致，EMRFS 会尝试进行此次数的 Amazon S3 调用。默认值为 **5**。
7. 对于 Retry period (in seconds) (重试期间 (秒))，键入一个整数值。这是 EMRFS 在重试尝试之间等待的时间量。默认值为 **10**。

> **Note**
>
> 后续重试会使用指数退避。

使用 Amazon CLI 启动一个启用一致视图的集群

建议您安装最新版本的 Amazon CLI。要下载最新版本，请访问 https://aws.amazon.com/cli/。

-

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --instance-type m5.xlarge --instance-count 3 --emrfs
 Consistent=true \
--release-label emr-5.36.1 --ec2-attributes KeyName=myKey
```

使用 Amazon Web Services Management Console检查是否启用了一致视图

- 要在控制台中检查是否启用了一致视图，请导航到 Cluster List (集群列表)，然后选择您的集群名称以查看 Cluster Details (集群详细信息)。“EMRFS consistent view (EMRFS 一致视图)”字段的值为 `Enabled` (已启用) 或 `Disabled` (已禁用)。

通过检查 **emrfs-site.xml** 文件查看是否启用了一致视图

- 您可以通过检查集群主节点上的 emrfs-site.xml 配置文件，来查看是否启用了一致性。如果 fs.s3.consistent 的布尔值设置为 true，则表示已为涉及 Amazon S3 的文件系统操作启用了一致视图。

了解 EMRFS 一致视图如何跟踪 Amazon S3 中的对象

EMRFS 通过向 EMRFS 元数据添加有关 Amazon S3 中对象的信息，创建这些对象的一致视图。EMRFS 会在以下情况下向其元数据添加这些列表：

- EMRFS 在 Amazon EMR 任务执行期间写入对象。
- 使用 EMRFS CLI 将对象与 EMRFS 元数据同步或导入到元数据。

EMRFS 读取的对象不会自动添加到元数据。当 EMRFS 删除对象时，一个具有已删除状态的列表仍保留在元数据中，直到使用 EMRFS CLI 清除该列表。要了解有关 CLI 的更多信息，请参阅EMRFS CLI 命令参考。有关在 EMRFS 元数据中清除列表的更多信息，请参阅EMRFS 一致视图元数据。

对于每个 Amazon S3 操作，EMRFS 都会在元数据中检查有关一致视图中的对象集的信息。如果 EMRFS 在其中一个操作执行过程中发现 Amazon S3 是不一致的，则会根据 emrfs-site 配置属性中定义的参数重试该操作。在 EMRFS 用尽重试次数后，它会引发 ConsistencyException 或记录异常并继续执行工作流程。有关重试逻辑的更多信息，请参阅重试逻辑。您可以在日志中找到 ConsistencyExceptions，例如：

- listStatus: No Amazon S3 object for metadata item /S3_bucket/dir/object
- getFileStatus: Key dir/file is present in metadata but not Amazon S3

如果直接从 Amazon S3 中删除 EMRFS 一致视图跟踪的对象，则 EMRFS 会将该对象视为不一致，这是因为 Amazon S3 中显示它仍存在于元数据中。如果您的元数据与 EMRFS 在 Amazon S3 中跟踪的对象不同步，则可以使用 EMRFS CLI 的 sync 子命令重置元数据以使其反映 Amazon S3 的情况。要了解元数据与 Amazon S3 之间的差异，请使用 diff。最后，EMRFS 只有在元数据中引用的对象的一致视图；相同 Amazon S3 路径中可能存在未进行跟踪的其它对象。EMRFS 在列出 Amazon S3 路径中的对象时，将返回在元数据中进行跟踪的对象与该 Amazon S3 路径中的对象的超集。

重试逻辑

EMRFS 将尝试针对其元数据中跟踪的对象验证列表一致性，并重试特定次数。默认 为 5。如果超过重试次数，则发起任务会返回错误，除非 `fs.s3.consistent.throwExceptionOnInconsistency` 设置为 `false` (此时仅将跟踪的对象记录为不一致)。EMRFS 默认使用指数退避重试策略，但您也可以将它设置为固定策略。用户还可能希望在重试特定一段时间之后继续任务的其余操作，而不是引发异常。可通过将 `fs.s3.consistent.throwExceptionOnInconsistency` 设置为 `false`，将 `fs.s3.consistent.retryPolicyType` 设置为 `fixed`，将 `fs.s3.consistent.retryPeriodSeconds` 设置为所需的值，来实现此目的。以下示例创建一个启用了一致性的集群，将记录不一致并设置 10 秒的固定重试间隔：

Example 将重试期间设置为固定量

```
aws emr create-cluster --release-label emr-5.36.1 \
--instance-type m5.xlarge --instance-count 1 \
--emrfs Consistent=true,Args=[fs.s3.consistent.throwExceptionOnInconsistency=false,
 fs.s3.consistent.retryPolicyType=fixed,fs.s3.consistent.retryPeriodSeconds=10] --ec2-
attributes KeyName=myKey
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

有关更多信息，请参阅一致视图。

为 IMDS 的 EMRFS 配置获取区域调用

EMRFS 通过 IMDS (实例元数据服务) 获取实例区域和 Amazon S3、DynamoDB 或Amazon KMS 终端节点。但是，IMDS 对可以处理的请求数量有限制，以及超过限制失败的请求。此 IMDS 限制可能会导致 EMRFS 初始化失败，查询或命令失败。您可以使用以下随机的指数回退重试机制和 emrfs-site.xml 中的回退区域配置属性来解决所有重试都失败的情况。

```
<property>
    <name>fs.s3.region.retryCount</name>
    <value>3</value>
    <description>
    Maximum retries that would be attempted to get Amazon region.
```

```
    </description>
</property>
<property>
    <name>fs.s3.region.retryPeriodSeconds</name>
    <value>3</value>
    <description>
    Base sleep time in second for each get-region retry.
    </description>
</property>
<property>
    <name>fs.s3.region.fallback</name>
    <value>us-east-1</value>
    <description>
    Fallback to this region after maximum retries for getting Amazon region have been
 reached.
    </description>
</property>
```

EMRFS 一致视图元数据

EMRFS 一致视图使用 DynamoDB 表跟踪 Amazon S3 中已与 EMRFS 同步或已由 EMRFS 创建的对象，从而跟踪一致性。元数据用于跟踪所有操作 (读取、写入、更新和复制)。其中不存储任何实际内容。此元数据用于验证从 Amazon S3 接收的对象或元数据是否与预期内容匹配。通过这种确认，EMRFS 能够针对 EMRFS 向 Amazon S3 写入的新对象或与 EMRFS 同步的对象检查列表一致性和先写后读一致性。多个集群可共享相同的元数据。

如何向元数据添加条目

您可以使用 `sync` 或 `import` 子命令向元数据添加条目。`sync` 反映路径中 Amazon S3 对象的状态，而 `import` 用于向元数据添加新条目。有关更多信息，请参阅EMRFS CLI 命令参考。

如何检查元数据与 Amazon S3 中的对象之间的差异

要检查元数据与 Amazon S3 之间的差异，请使用 EMRFS CLI 的 `diff` 子命令。有关更多信息，请参阅EMRFS CLI 命令参考。

如何了解元数据操作是否受限制

EMRFS 针对元数据的读取和写入操作，分别设置了默认 500 和 100 个单位的吞吐量容量限制。大量对象或存储桶可能会导致操作超过此容量，此时 DynamoDB 会对它们进行限制。例如，如果您执行的操作超过这些容量限制，则应用程序可能会导致 EMRFS 引发 `ProvisionedThroughputExceededException`。施加节流时，EMRFS CLI 工具将尝试使用指

数回退重试对 DynamoDB 表进行写入操作，直到操作完成，或是达到将对象从 Amazon EMR 写入到 Amazon S3 的最大重试次数值。

您可以配置自己的吞吐容量限制。但是，DynamoDB 对读取和写入操作具有严格的分区限制：每秒 3000 个读取容量单位（RCU）和 1000 个写入容量单位（WCU）。为避免因节流而导致的 sync 故障，我们建议您将读取操作的吞吐量限制为低于 3000 RCU，并将写入操作限制在 1000 WCU 以下。有关设置自定义吞吐容量限制的说明，请参阅 配置一致视图。

您可以在 DynamoDB 控制台中查看 EMRFS 元数据的 Amazon CloudWatch 指标，还可以查看受限制的读取和写入请求。如果受限制的请求数不为零值，则增加为读取或写入操作分配的吞吐量容量可能会使应用程序受益。如果您发现操作长时间接近分配的最大读取或写入吞吐量容量，则这样做也可能会获得性能好处。

重要 EMRFS 操作的吞吐量特征

读取和写入操作的默认值分别为 400 和 100 个吞吐量容量单位。您可以通过以下性能特征了解特定操作所需的吞吐量。这些测试是使用单节点 m3.large 集群执行的。所有操作都是单线程执行。特定应用程序特征会对性能造成很大影响，可能需要通过实验来优化文件系统操作。

操作	每秒平均读取量	每秒平均写入量
create（对象）	26.79	6.70
delete（对象）	10.79	10.79
delete（包含 1000 个对象的目录）	21.79	338.40
getFileStatus（对象）	34.70	0
getFileStatus（目录）	19.96	0
listStatus（包含 1 个对象的目录）	43.31	0
listStatus（包含 10 个对象的目录）	44.34	0
listStatus（包含 100 个对象的目录）	84.44	0

操作	每秒平均读取量	每秒平均写入量
listStatus（包含 1000 个对象的目录）	308.81	0
listStatus（包含 10000 个对象的目录）	416.05	0
listStatus（包含 100000 个对象的目录）	823.56	0
listStatus（包含 1000000 个对象的目录）	882.36	0
mkdir（持续 120 秒）	24.18	4.03
mkdir	12.59	0
rename（对象）	19.53	4.88
rename（包含 1000 个对象的目录）	23.22	339.34

提交从元数据存储中清除旧数据的步骤

用户可能希望在基于 DynamoDB 的元数据中删除特定条目。这样可以帮助降低与表关联的存储成本。用户可以使用 EMRFS CLI `delete` 子命令，以手动或编程方式清除特定条目。但是，如果从元数据中删除条目，则 EMRFS 不再进行任何一致性检查。

可以通过向集群提交在 EMRFS CLI 中执行命令的最终步骤，以编程方式在任务完成之后进行清除。例如，键入以下命令可向集群提交删除两天之前的所有条目的步骤。

```
aws emr add-steps --cluster-id j-2AL4XXXXXX5T9 --steps Name="emrfsCLI",Jar="command-
runner.jar",Args=["emrfs","delete","--time","2","--time-unit","days"]
{
    "StepIds": [
        "s-B12345678902"
    ]
}
```

可使用返回的 StepId 值检查日志以了解操作结果。

为 CloudWatch 和 Amazon SQS 配置一致性通知

您可以针对 Amazon S3 最终一致性问题在 EMRFS 中启用 CloudWatch 指标和 Amazon SQS 消息。

CloudWatch

启用 CloudWatch 指标后，当 `FileSystem` API 调用因 Amazon S3 最终一致性问题而失败时，系统会将推送名为 Inconsistency (不一致） 的指标。

查看针对 Amazon S3 最终一致性问题的 CloudWatch 指标

查看 CloudWatch 中的 Inconsistency (不一致） 指标时，选择 EMRFS 指标，然后选择 JobFlowId/Metric Name 对。例如：`j-162XXXXXXM2CU ListStatus`、`j-162XXXXXXM2CU GetFileStatus` 等。

1. 通过以下网址打开 CloudWatch 控制台：https://console.aws.amazon.com/cloudwatch/。
2. 在 Dashboard (控制面板) 的 Metrics (指标) 部分中，选择 EMRFS。
3. 在 Job Flow Metrics (任务流程指标) 窗格中，选择一个或多个 JobFlowId (任务流程 ID)/Metric Name (指标名称) 对。将在下面的窗口中显示指标的图形表示。

Amazon SQS

启用 Amazon SQS 通知后，系统会在初始化 EMRFS 时创建一个名为 `EMRFS-Inconsistency-<jobFlowId>` 的 Amazon SQS 队列。当 `FileSystem` API 调用因 Amazon S3 最终一致性问题而失败时，系统会将 Amazon SQS 消息推送到该队列中。消息包含诸如 JobFlowId、API、不一致路径的列表、堆栈跟踪等的信息。可以使用 Amazon SQS 控制台或 EMRFS `read-sqs` 命令读取消息。

管理针对 Amazon S3 最终一致性问题的 Amazon SQS 消息

可使用 EMRFS CLI 读取针对 Amazon S3 最终一致性问题的 Amazon SQS 消息。要从 EMRFS Amazon SQS 队列读取消息，请键入 `read-sqs` 命令并为生成的输出文件在主节点的本地文件系统上指定输出位置。

还可以使用 `delete-sqs` 命令删除 EMRFS Amazon SQS 队列。

1. 要从 Amazon SQS 队列读取消息，请键入以下命令。将 *`queuename`* 替换为您配置的 Amazon SQS 队列的名称，并将 *`/path/filename`* 替换为输出文件的路径：

```
emrfs read-sqs --queue-name queuename --output-file /path/filename
```

例如，要从默认队列读取和输出 Amazon SQS 消息，请键入：

```
emrfs read-sqs --queue-name EMRFS-Inconsistency-j-162XXXXXXM2CU --output-file /
path/filename
```

> **Note**
>
> 还可以分别使用 -q 和 -o 快捷方式代替 --queue-name 和 --output-file。

2. 要删除 Amazon SQS 队列，请键入以下命令：

```
emrfs delete-sqs --queue-name queuename
```

例如，要删除默认队列，请键入：

```
emrfs delete-sqs --queue-name EMRFS-Inconsistency-j-162XXXXXXM2CU
```

> **Note**
>
> 还可以使用 -q 快捷方式代替 --queue-name。

配置一致视图

您可以为一致视图配置其它设置，方法是使用 emrfs-site 属性的配置属性来提供这些设置。例如，您可以选择其它的默认 DynamoDB 吞吐量（方式是将以下参数提供给 CLI --emrfs 选项，使用 emrfs-site 配置分类（仅限 Amazon EMR 发行版 4.x 及更高版本））或引导操作来配置主节点上的 emrfs-site.xml 文件：

Example 在集群启动时更改默认元数据读取和写入值

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge \
--emrfs Consistent=true,Args=[fs.s3.consistent.metadata.read.capacity=600,\
fs.s3.consistent.metadata.write.capacity=300] --ec2-attributes KeyName=myKey
```

或者，使用下面的配置文件并将其保存到本地或 Amazon S3 中：

```
[
    {
      "Classification": "emrfs-site",
      "Properties": {
        "fs.s3.consistent.metadata.read.capacity": "600",
        "fs.s3.consistent.metadata.write.capacity": "300"
      }
    }
 ]
```

按照下面的语法使用您创建的配置：

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Hive \
--instance-type m5.xlarge --instance-count 2 --configurations file://./myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

可以使用配置或 Amazon CLI `--emrfs` 参数设置以下选项。有关这些实参的信息，请参阅 Amazon CLI 命令参考。

一致视图的 **emrfs-site.xml** 属性

属性	默认值	描述
fs.s3.consistent	**false**	设置为 **true** 时，此属性会将 EMRFS 配置为使用 DynamoDB 提供一致性。
fs.s3.consistent.retryPolicyType	**exponential**	此属性标识针对一致性问题进行重试时要使用的策略。选项包括：exponential、fixed 和 none。

属性	默认值	描述
fs.s3.consistent.retryPeriodSeconds	**1**	此属性设置两次一致性重试尝试之间等待的时间长度。
fs.s3.consistent.retryCount	**10**	此属性设置检测到不一致时的最大重试次数。
fs.s3.consistent.throwExceptionOnInconsistency	**true**	此属性确定是引发还是记录一致性异常。设置为 **true** 时，会引发 ConsistencyException 。
fs.s3.consistent.metadata.autoCreate	**true**	设置为 **true** 时，此属性会启用元数据表的自动创建。
fs.s3.consistent.metadata.etag.verification.enabled	**true**	对于 Amazon EMR 5.29.0，此属性已默认启用。启用后，EMRFS 使用 S3 ETag 验证所读取的对象为最新可用版本。此功能对更新后读取使用案例很有帮助，此时将覆盖 S3 上的文件但保留相同名称。此 ETag 验证功能当前不可用于 S3 Select。
fs.s3.consistent.metadata.tableName	**EmrFSMetadata**	此属性指定 DynamoDB 中元数据表的名称。
fs.s3.consistent.metadata.read.capacity	**500**	此属性指定创建元数据表时要配置的 DynamoDB 读取容量。
fs.s3.consistent.metadata.write.capacity	**100**	此属性指定创建元数据表时要配置的 DynamoDB 写入容量。
fs.s3.consistent.fastList	**true**	设置为 **true** 时，此属性会使用多个线程列出目录（需要时）。必须启用一致性才能使用此属性。

属性	默认值	描述
fs.s3.consistent.fastList.prefetchMetadata	**false**	设置为 **true** 时，此属性会为包含 20000 个以上的项目的目录启用元数据预取。
fs.s3.consistent.notification.CloudWatch	**false**	设置为 **true** 时，为因 Amazon S3 最终一致性问题而失败的 FileSystem API 调用启用 CloudWatch 指标。
fs.s3.consistent.notification.SQS	**false**	设置为 **true** 时，向 Amazon SQS 队列推送最终一致性通知。
fs.s3.consistent.notification.SQS.queueName	**EMRFS-Inconsistency-<jobFlowId>**	通过更改此属性可以为有关 Amazon S3 最终一致性问题的消息指定您自己的 SQS 队列名称。
fs.s3.consistent.notification.SQS.customMsg	**none**	通过此属性可以指定有关 Amazon S3 最终一致性问题的 SQS 消息中包含的自定义信息。如果没有为此属性指定值，则消息中的对应字段为空。
fs.s3.consistent.dynamodb.endpoint	**none**	此属性允许您为一致性视图元数据指定自定义 DynamoDB 终端节点。
fs.s3.useRequesterPaysHeader	**false**	当设置为**true**，此属性允许在启用付款人选项请求的情况下向存储桶发出Amazon S3 请求。

EMRFS CLI 命令参考

默认情况下，EMRFS CLI 安装在使用 Amazon EMR 发行版 3.2.1 或更高版本创建的所有集群主节点上。您可以使用 EMRFS CLI 管理一致视图的元数据。

> **Note**
>
> emrfs 命令仅支持 VT100 终端仿真。但是，它可能适用于其它终端仿真器模式。

emrfs 顶级命令

emrfs 顶级命令支持以下结构。

```
emrfs [describe-metadata | set-metadata-capacity | delete-metadata | create-metadata |
 \
list-metadata-stores | diff | delete | sync | import ] [options] [arguments]
```

指定 [选项]，根据需要使用或不使用下表中描述的 [参数]。有关特定于子命令 (describe-metadata、set-metadata-capacity 等) 的 [选项]，请参阅下面的每个子命令。

emrfs 的 [选项]

选项	描述	必填
-a *AWS_ACCESS_KEY_ID* | --access-key *AWS_ACCESS_KEY_ID*	用于将对象写入 Amazon S3 以及在 DynamoDB 中创建或访问元数据存储的Amazon访问密钥。默认情况下，*AWS_ACCESS_KEY_ID* 设置为用于创建集群的访问密钥。	否
-s *AWS_SECRET_ACCESS_KEY* | --secret-key *AWS_SECRET_ACCESS_KEY*	与用于将对象写入 Amazon S3 以及在 DynamoDB 中创建或访问元数据存储的访问密钥关联的Amazon私有密钥。默认情况下，*AWS_SECRET_ACCESS_KEY* 设置为与用于创建集群的访问密钥关联的私有密钥。	否
-v | --verbose	使输出为详细模式。	否
-h | --help	通过用法语句显示 emrfs 命令的帮助消息。	否

emrfs describe-metadata 子命令

emrfs describe-metadata 的 [选项]

选项	描述	必填
-m *METADATA_NAME* | --metadata-name *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 EmrFSMetadata 。	否

Example emrfs describe-metadata 示例

以下示例描述默认元数据表。

```
$ emrfs describe-metadata
EmrFSMetadata
  read-capacity: 400
  write-capacity: 100
  status: ACTIVE
  approximate-item-count (6 hour delay): 12
```

emrfs set-metadata-capacity 子命令

emrfs set-metadata-capacity 的 [选项]

选项	描述	必填
-m *METADATA_NAME* | --metadata-name *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 EmrFSMetadata 。	否
-r *READ_CAPACITY* | --read-capacity *READ_CAPACITY*	对元数据表请求的读取吞吐容量。如果未提供 *READ_CAPACITY* 参数，则默认值为 400。	否
	对元数据表请求的写入吞吐容量。如果未提供 *WRITE_CAPACITY* 参数，则默认值为 100。	否

选项	描述	必填
-w *WRITE_CAPACITY* | --write-capacity *WRITE_CAPACITY*		

Example emrfs set-metadata-capacity 示例

以下示例将名为 600 的元数据表的读取吞吐容量设置为 150，将写入容量设置为 EmrMetadataAlt。

```
$ emrfs set-metadata-capacity --metadata-name EmrMetadataAlt  --read-capacity 600 --
write-capacity 150
  read-capacity: 400
  write-capacity: 100
  status: UPDATING
  approximate-item-count (6 hour delay): 0
```

emrfs delete-metadata 子命令

emrfs delete-metadata 的 [选项]

选项	描述	必填
-m *METADATA_NAME* | --metadata-name *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 EmrFSMetadata 。	否

Example emrfs delete-metadata 示例

以下示例删除默认元数据表。

```
$ emrfs delete-metadata
```

emrfs create-metadata 子命令

emrfs create-metadata 的 [选项]

选项	描述	必填
-m *METADATA_NAME* | --metadata-name *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 EmrFSMetadata 。	否
-r *READ_CAPACITY* | --read-capacity *READ_CAPACITY*	对元数据表请求的读取吞吐容量。如果未提供 *READ_CAPACITY* 参数，则默认值为 400。	否
-w *WRITE_CAPACITY* | --write-capacity *WRITE_CAPACITY*	对元数据表请求的写入吞吐容量。如果未提供 *WRITE_CAPACITY* 参数，则默认值为 100。	否

Example emrfs create-metadata 示例

以下示例创建一个名为 EmrFSMetadataAlt 的元数据表。

```
$ emrfs create-metadata -m EmrFSMetadataAlt
Creating metadata: EmrFSMetadataAlt
EmrFSMetadataAlt
  read-capacity: 400
  write-capacity: 100
  status: ACTIVE
  approximate-item-count (6 hour delay): 0
```

emrfs list-metadata-stores 子命令

emrfs list-metadata-stores 子命令没有任何 [选项]。

Example List-metadata-stores 示例

以下示例列出您的元数据表。

```
$ emrfs list-metadata-stores
  EmrFSMetadata
```

emrfs diff 子命令

emrfs diff 的 [选项]

选项	描述	必填
`-m` *METADATA_NAME* | `--metadata-name` *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 `EmrFSMetadata` 。	否
s3://s3Path	与元数据表进行比较的 Amazon S3 存储桶的路径。存储桶以递归方式同步。	是

Example emrfs diff 示例

以下示例将默认元数据表与 Amazon S3 存储桶进行比较。

```
$ emrfs diff s3://elasticmapreduce/samples/cloudfront
BOTH | MANIFEST ONLY | S3 ONLY
DIR elasticmapreduce/samples/cloudfront
DIR elasticmapreduce/samples/cloudfront/code/
DIR elasticmapreduce/samples/cloudfront/input/
DIR elasticmapreduce/samples/cloudfront/logprocessor.jar
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-14.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-15.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-16.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-17.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-18.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-19.WxYz1234
DIR elasticmapreduce/samples/cloudfront/input/XABCD12345678.2009-05-05-20.WxYz1234
DIR elasticmapreduce/samples/cloudfront/code/cloudfront-loganalyzer.tgz
```

emrfs delete 子命令

emrfs delete 的 [选项]

选项	描述	必填
`-m METADATA_NAME | --metadata-name METADATA_NAME`	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 `EmrFSMetadata` 。	否
s3://s3Path	为实现一致视图而跟踪的 Amazon S3 存储桶的路径。存储桶以递归方式同步。	是
`-t TIME | --time TIME`	过期时间 (使用时间单位参数进行解释)。对于指定存储桶，早于 *TIME* 参数的所有元数据条目都会被删除。	
`-u UNIT | --time-unit UNIT`	用于解释时间参数的度量值 (纳秒、微秒、毫秒、秒、分钟、小时或天)。如果未指定参数，则默认值为 `days`。	
`--read-consumption READ_CONSUMPTION`	请求用于 delete 操作的可用读取吞吐量。如果未指定 *READ_CONSUMPTION* 参数，则默认值为 `400`。	否
`--write-consumption WRITE_CONSUMPTION`	请求用于 delete 操作的可用写入吞吐量。如果未指定 *WRITE_CONSUMPTION* 参数，则默认值为 `100`。	否

Example emrfs delete 示例

以下示例从一致视图的跟踪元数据中删除一个 Amazon S3 存储桶中的所有对象。

```
$ emrfs delete s3://elasticmapreduce/samples/cloudfront
```

```
entries deleted: 11
```

emrfs import 子命令

for emrfs import 的 [选项]

选项	描述	必填
`-m` *METADATA_NAME* | `--metadata-name` *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 `EmrFSMetadata` 。	否
s3://s3Path	为实现一致视图而跟踪的 Amazon S3 存储桶的路径。存储桶以递归方式同步。	是
`--read-consumption` *READ_CONSUMPTION*	请求用于 delete 操作的可用读取吞吐量。如果未指定 *READ_CONSUMPTION* 参数，则默认值为 `400`。	否
`--write-consumption` *WRITE_CONSUMPTION*	请求用于 delete 操作的可用写入吞吐量。如果未指定 *WRITE_CONSUMPTION* 参数，则默认值为 `100`。	否

Example emrfs import 示例

以下示例随一致视图的跟踪元数据导入一个 Amazon S3 存储桶中的所有对象。忽略所有未知键。

```
$ emrfs import s3://elasticmapreduce/samples/cloudfront
```

emrfs sync 子命令

emrfs sync 的 [选项]

选项	描述	必填

选项	描述	必填
-m *METADATA_NAME* | --metadata-name *METADATA_NAME*	*METADATA_NAME* 是 DynamoDB 元数据表的名称。如果未提供 *METADATA_NAME* 参数，则默认值为 EmrFSMetadata 。	否
s3://s3Path	为实现一致视图而跟踪的 Amazon S3 存储桶的路径。存储桶以递归方式同步。	是
--read-consumption *READ_CONSUMPTION*	请求用于 delete 操作的可用读取吞吐量。如果未指定 *READ_CONSUMPTION* 参数，则默认值为 400。	否
--write-consumption *WRITE_CONSUMPTION*	请求用于 delete 操作的可用写入吞吐量。如果未指定 *WRITE_CONSUMPTION* 参数，则默认值为 100。	否

Example emrfs sync 命令示例

以下示例随一致视图的跟踪元数据导入一个 Amazon S3 存储桶中的所有对象。删除所有未知键。

```
$ emrfs sync s3://elasticmapreduce/samples/cloudfront
Synching samples/cloudfront                                    0 added | 0 updated |
 0 removed | 0 unchanged
Synching samples/cloudfront/code/                              1 added | 0 updated |
 0 removed | 0 unchanged
Synching samples/cloudfront/                                   2 added | 0 updated |
 0 removed | 0 unchanged
Synching samples/cloudfront/input/                             9 added | 0 updated |
 0 removed | 0 unchanged
Done synching s3://elasticmapreduce/samples/cloudfront         9 added | 0 updated |
 1 removed | 0 unchanged
creating 3 folder key(s)
folders written: 3
```

emrfs read-sqs 子命令

emrfs read-sqs 的 [选项]

选项	描述	必填
`-q QUEUE_NAME | --queue-name QUEUE_NAME`	*QUEUE_NAME* 是在 emrfs-site.xml 中配置的 Amazon SQS 队列的名称。默认值为 **EMRFS-Inconsistency-<jobFlowId>** 。	是
`-o OUTPUT_FILE | --output-file OUTPUT_FILE`	*OUTPUT_FILE* 是主节点本地文件系统上的输出文件的路径。从队列读取的消息会写入此文件。	是

emrfs delete-sqs 子命令

emrfs delete-sqs 的 [选项]

选项	描述	必填
`-q QUEUE_NAME | --queue-name QUEUE_NAME`	*QUEUE_NAME* 是在 emrfs-site.xml 中配置的 Amazon SQS 队列的名称。默认值为 **EMRFS-Inconsistency-<jobFlowId>** 。	是

以步骤形式提交 EMRFS CLI 命令

以下示例说明如何通过利用 Amazon CLI 或 API 和 `emrfs` 以步骤形式运行 `command-runner.jar` 命令，在主节点上使用 `emrfs` 实用工具。该示例使用 Amazon SDK for Python (Boto3) 向集群添加一个步骤，该步骤向默认 EMRFS 元数据表添加 Amazon S3 存储桶中的对象。

```
import boto3
from botocore.exceptions import ClientError


def add_emrfs_step(command, bucket_url, cluster_id, emr_client):
    """
    Add an EMRFS command as a job flow step to an existing cluster.
```

```
    :param command: The EMRFS command to run.
    :param bucket_url: The URL of a bucket that contains tracking metadata.
    :param cluster_id: The ID of the cluster to update.
    :param emr_client: The Boto3 Amazon EMR client object.
    :return: The ID of the added job flow step. Status can be tracked by calling
             the emr_client.describe_step() function.
    """
    job_flow_step = {
        "Name": "Example EMRFS Command Step",
        "ActionOnFailure": "CONTINUE",
        "HadoopJarStep": {
            "Jar": "command-runner.jar",
            "Args": ["/usr/bin/emrfs", command, bucket_url],
        },
    }

    try:
        response = emr_client.add_job_flow_steps(
            JobFlowId=cluster_id, Steps=[job_flow_step]
        )
        step_id = response["StepIds"][0]
        print(f"Added step {step_id} to cluster {cluster_id}.")
    except ClientError:
        print(f"Couldn't add a step to cluster {cluster_id}.")
        raise
    else:
        return step_id


def usage_demo():
    emr_client = boto3.client("emr")
    # Assumes the first waiting cluster has EMRFS enabled and has created metadata
    # with the default name of 'EmrFSMetadata'.
    cluster = emr_client.list_clusters(ClusterStates=["WAITING"])["Clusters"][0]
    add_emrfs_step(
        "sync", "s3://elasticmapreduce/samples/cloudfront", cluster["Id"], emr_client
    )


if __name__ == "__main__":
    usage_demo()
```

可以使用返回的 `step_id` 值检查日志以了解操作结果。

授予对 Amazon S3 中的 EMRFS 数据的访问权

默认情况下，EC2 的 EMR 角色确定访问 Amazon S3 中 EMRFS 数据的权限。无论是用户还是组通过 EMRFS 提出请求，附加到此角色的 IAM policy 都适用。默认为 `EMR_EC2_DefaultRole`。有关更多信息，请参阅集群 EC2 实例的服务角色（EC2 实例配置文件）。

从 Amazon EMR 发行版 5.10.0 开始，可以使用安全配置来指定 EMRFS 的 IAM 角色。这样可以为多用户集群自定义 EMRFS 对 Amazon S3 的请求的权限。您可以为不同用户和组指定不同的 IAM 角色，也可根据在 Amazon S3 中的前缀为不同的 Amazon S3 存储桶位置进行指定。当 EMRFS 向 Amazon S3 发出的请求与您指定的用户、组或位置匹配时，集群将使用您指定的相应角色，而不是 EC2 的 EMR 角色。有关更多信息，请参阅为处理 EMRFS 对 Amazon S3 的请求配置 IAM 角色。

或者，如果您的 Amazon EMR 解决方案的需求超出了 EMRFS 的 IAM 角色所能提供的权限，您也可以定义自定义凭证提供程序类，从而让您能够自定义对 Amazon S3 中的 EMRFS 数据的访问。

为 Amazon S3 中的 EMRFS 数据创建自定义凭证提供程序

要创建自定义凭证提供程序，您可以实现 AWSCredentialsProvider 和 Hadoop 可配置类。

有关此方法的详细说明，请参阅Amazon大数据博客中的 Securely analyze data from another Amazon account with EMRFS。博文中包含了全流程分步教程，涵盖从创建 IAM 角色到启动集群。其中还提供了实施自定义凭证提供程序类的 Java 代码示例。

基本步骤如下所示：

指定自定义凭证提供程序

1. 创建编译为 JAR 文件的自定义凭证提供程序类。
2. 将脚本作为引导操作运行，从而将自定义凭证提供程序 JAR 文件复制到集群主节点的 `/usr/share/aws/emr/emrfs/auxlib` 位置。有关引导操作的更多信息，请参阅（可选）创建引导操作以安装其它软件。
3. 自定义 `emrfs-site` 分类，以指定在 JAR 文件中实施的类。有关指定要自定义应用程序的配置对象的更多信息，请参阅《Amazon EMR 版本指南》中的配置应用程序。

   以下示例演示了启动包含常见配置参数的 Hive 集群的 `create-cluster` 命令，并包括：

   - 运行脚本 *`copy_jar_file.sh`* 的引导操作，该脚本已保存到 Amazon S3 中的 *`mybucket`*。

- 将 JAR 文件中定义的自定义凭证提供程序指定为 `emrfs-site` 的 *`MyCustomCredentialsProvider`* 分类

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --applications Name=Hive \
--bootstrap-actions '[{"Path":"s3://mybucket/copy_jar_file.sh","Name":"Custom
 action"}]' \
--ec2-attributes '{"KeyName":"MyKeyPair","InstanceProfile":"EMR_EC2_DefaultRole",\
"SubnetId":"subnet-xxxxxxxx","EmrManagedSlaveSecurityGroup":"sg-xxxxxxxx",\
"EmrManagedMasterSecurityGroup":"sg-xxxxxxxx"}' \
--service-role EMR_DefaultRole_V2 --enable-debugging --release-label emr-5.36.1 \
--log-uri 's3n://my-emr-log-bucket/' --name 'test-awscredentialsprovider-emrfs' \
--instance-type=m5.xlarge --instance-count 3  \
--configurations '[{"Classification":"emrfs-site",\
"Properties":
{"fs.s3.customAWSCredentialsProvider":"MyAWSCredentialsProviderWithUri"},\
"Configurations":[]}]'
```

管理默认的 Amazon Security Token Service 终端节点

EMRFS 使用 Amazon Security Token Service（STS）检索临时安全凭证，以便访问您的Amazon资源。早期的 Amazon EMR 发行版将所有 Amazon STS 请求发送到 `https://sts.amazonaws.com` 的一个全局终端节点。而 Amazon EMR 发行版 5.31.0 和 6.1.0 及更高版本向区域 Amazon STS 终端节点发出请求。这可以降低延迟并提高会话令牌的有效性。有关 Amazon STS 终端节点的更多信息，请参阅《Amazon Identity and Access Management IAM 用户指南》中的在Amazon区域中管理 Amazon STS。

使用 Amazon EMR 发行版 5.31.0 和 6.1.0 及更高版本时，您可以覆盖默认的 Amazon STS 终端节点。为此，您必须更改 `emrfs-site` 配置中的 `fs.s3.sts.endpoint` 属性。

以下 Amazon CLI 示例，将 EMRF 使用的默认的 Amazon STS 终端节点设置为全局终端节点。

```
aws emr create-cluster --release-label <emr-5.33.0> --instance-type m5.xlarge \
--emrfs Args=[fs.s3.sts.endpoint=https://sts.amazonaws.com]
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

或者，您也可以使用以下示例创建 JSON 配置文件，并使用 `emr create-cluster` 的 `--configurations` 实参指定它。有关使用 `--configurations`, 的更多信息，请参阅 Amazon CLI 命令参考。

```
[
  {
    "classification": "emrfs-site",
    "properties": {
      "fs.s3.sts.endpoint": "https://sts.amazonaws.com"
    }
  }
]
```

使用 EMRFS 属性指定 Amazon S3 加密

> **Important**
>
> 从 Amazon EMR 发行版 4.8.0 开始，您可以使用安全配置以更轻松的方式应用安全设置，并获得更多选项。建议您使用安全配置。有关更多信息，请参阅配置数据加密。此部分中所述的控制台说明适用于 4.8.0 之前的版本。如果您使用 Amazon CLI 来配置后续版本中的集群配置和安全配置中的 Amazon S3 加密，则安全配置会覆盖集群配置。

在创建集群时，您可以使用控制台或通过 Amazon CLI 或 EMR SDK 使用 `emrfs-site` 分类属性为 Amazon S3 中的 EMRFS 数据指定服务器端加密（SSE）或客户端加密（CSE）。Amazon S3 SSE 和 CSE 是互斥的；您可以任选其一，但不能同时选择两者。

有关 Amazon CLI 描述，请参阅加密类型的相应部分。

使用 Amazon Web Services Management Console指定 EMRFS 加密选项

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 选择 Release (版本) 4.7.2 或更早版本。
4. 为 Software and Steps (软件和步骤) 选择适用于您的应用程序的其它选项，然后选择 Next (下一步)。
5. 在 Hardware (硬件) 和 General Cluster Settings (常规集群设置) 窗格中选择适用于您的应用程序的设置。
6. 在 Security (安全) 窗格上的 Authentication and encryption (身份验证和加密) 下，选择要使用的 S3 Encryption (with EMRFS) 选项。

   > **Note**
   >
   > 在使用 Amazon EMR 发行版 4.4 或更早版本时，S3 server-side encryption with KMS Key Management (利用 KMS Key Management 进行 S3 服务器端加密) (SSE-KMS) 不可用。

   - 如果您选择一个使用 Amazon Key Management 的选项，请选择一个 Amazon KMS Key ID。有关更多信息，请参阅使用 Amazon KMS keys 进行 EMRFS 加密。
   - 如果您选择 S3 client-side encryption with custom materials provider (利用自定义材料提供程序进行 S3 客户端加密)，请提供 Class name (类名称) 和 JAR location (AR 位置)。有关更多信息，请参阅Amazon S3 客户端加密。
7. 选择适用于您的应用程序的其它选项，然后选择 Create Cluster (创建集群)。

使用 Amazon KMS keys 进行 EMRFS 加密

Amazon KMS 加密密钥，必须在与您的 Amazon EMR 集群实例和与 EMRFS 一起使用的 Amazon S3 存储桶相同的区域创建。如果指定的密钥没有位于用于配置集群的账户中，则必须使用它的 ARN 指定密钥。

Amazon EC2 实例配置文件的角色必须具有使用您指定的 KMS 密钥的权限。Amazon EMR 中实例配置文件的默认角色是 `EMR_EC2_DefaultRole`。如果您对实例配置文件使用不同的角色，或者对 Amazon S3 的 EMRFS 请求使用 IAM 角色，请确保根据需要将每个角色添加为密钥用户。这会为该角

色授予使用该 KMS 密钥的权限。有关更多信息，请参阅《Amazon Key Management Service开发人员指南》和为向 Amazon S3 发出的 EMRFS 请求配置 IAM 角色中的使用密钥策略。

您可以使用 Amazon Web Services Management Console 将实例配置文件或 EC2 实例配置文件添加到指定 KMS 密钥的密钥用户列表中，也可以使用 Amazon CLI 或 Amazon SDK 来附加适当的密钥策略。

请注意 Amazon EMR 仅支持对称 KMS 密钥。不能使用非对称 KMS 密钥加密 Amazon EMR 集群中的静态数据。要获取确定 KMS 密钥是对称还是非对称的帮助，请参阅识别对称密钥和非对称密钥。

以下步骤介绍了如何使用 Amazon Web Services Management Console 将 Amazon EMR 实例配置文件 `EMR_EC2_DefaultRole` 作为密钥用户添加。它假定您已创建一个 KMS 密钥。要创建新的 KMS 密钥，请参阅《Amazon Key Management Service 开发人员指南》中的创建密钥。

将 Amazon EMR 的 EC2 实例配置文件添加到加密密钥用户列表中

1. 登录到 Amazon Web Services Management Console，然后通过以下网址打开 Amazon Key Management Service (Amazon KMS) 控制台：https://console.aws.amazon.com/kms。
2. 要更改 Amazon Web Services 区域，请使用页面右上角的区域选择器。
3. 选择要修改的 KMS 密钥的别名。
4. 在密钥详细信息页面的 Key Users (密钥用户) 下，选择 Add (添加)。
5. 在 Add key users (添加密钥用户) 对话框中，选择适当的角色。默认角色的名称为 `EMR_EC2_DefaultRole`。
6. 选择 Add（添加）。

Amazon S3 服务器端加密

设置 Amazon S3 服务器端加密时，Amazon S3 在向磁盘写入数据时会在对象级别加密数据，并在访问数据时对数据进行解密。有关 SSE 的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的使用服务器端加密保护数据。

在 Amazon EMR 中指定 SSE 时，可以在两个不同的密钥管理系统之间进行选择：

- SSE-S3 – Amazon S3 为您管理密钥。
- SSE-KMS - 您使用适用于 Amazon EMR 的策略设置一个 Amazon KMS key。有关 Amazon EMR 的密钥要求的更多信息，请参阅使用 Amazon KMS keys 进行加密。

客户提供密钥的 SSE（SSE-C）不能用于 Amazon EMR。

使用 Amazon CLI 创建启用了 SSE-S3 的集群

- 键入以下命令：

```
aws emr create-cluster --release-label emr-4.7.2 or earlier \
--instance-count 3 --instance-type m5.xlarge --emrfs Encryption=ServerSide
```

您也可以通过在 `emrfs-site` 属性中将 fs.s3.enableServerSideEncryption 属性设置为 true 来启用 SSE-S3。请参阅下面的 SSE-KMS 示例并忽略密钥 ID 的属性。

使用 Amazon CLI 创建启用了 SSE-KMS 的集群

> **Note**
>
> SSE-KMS 仅在 Amazon EMR 发行版 4.5.0 及更高版本中可用。

- 键入以下 Amazon CLI 命令以创建使用 SSE-KMS 的集群，其中 *keyID* 是一个 Amazon KMS key，例如 *a4567b8-9900-12ab-1234-123a45678901*：

```
aws emr create-cluster --release-label emr-4.7.2 or earlier --instance-count 3 \
--instance-type m5.xlarge --use-default-roles \
--emrfs Encryption=ServerSide,Args=[fs.s3.serverSideEncryption.kms.keyId=keyId]
```

--OR--

使用 `emrfs-site` 分类键入以下 Amazon CLI 命令，并为配置 JSON 文件提供内容，如以下示例中的 `myConfig.json` 所示：

```
aws emr create-cluster --release-label emr-4.7.2 or earlier --instance-count 3
 --instance-type m5.xlarge --applications Name=Hadoop --configurations file://
myConfig.json --use-default-roles
```

myConfig.json 的示例内容：

```
[
  {
    "Classification":"emrfs-site",
    "Properties": {
```

```
        "fs.s3.enableServerSideEncryption": "true",
        "fs.s3.serverSideEncryption.kms.keyId":"a4567b8-9900-12ab-1234-123a45678901"
    }
  }
]
```

SSE-S3 和 SSE-KMS 的配置属性

可使用 `emrfs-site` 配置分类来配置这些属性。SSE-KMS 仅在 Amazon EMR 发行版 4.5.0 及更高版本中可用。

属性	默认值	描述
`fs.s3.enableServerSideEncryption`	**`false`**	设置为 **`true`** 时，使用服务器端加密对 Amazon S3 中存储的对象进行加密。如果未指定密钥，则使用 SSE-S3。
`fs.s3.serverSideEncryption.kms.keyId`	**`n/a`**	指定 Amazon KMS 密钥 ID 或 ARN。如果未指定密钥，则使用 SSE-KMS。

Amazon S3 客户端加密

对于 Amazon S3 客户端加密，Amazon S3 加密和解密过程在您的 EMR 集群上的 EMRFS 客户端中进行。在对象上载到 Amazon S3 之前对其进行加密，并在下载后对其进行解密。您指定的提供程序会提供客户端使用的加密密钥。客户端可以使用 Amazon KMS 提供的密钥（CSE-KMS）或提供客户端根密钥（CSE-C）的自定义 Java 类。CSE-KMS 和 CSE-C 之间的加密细节略有不同，具体取决于指定的提供程序以及正在解密或加密对象的元数据。有关这些区别的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的使用客户端加密保护数据。

> **Note**
>
> Amazon S3 CSE 仅确保与 Amazon S3 交换的 EMRFS 数据已加密；不确保集群实例卷上的所有数据都已加密。此外，由于 Hue 不使用 EMRFS，因此 Hue S3 文件浏览器写入 Amazon S3 的对象没有加密。

使用 Amazon CLI 为 Amazon S3 中的 EMRFS 数据指定 CSE-KMS

- 键入以下命令并将 *MyKMSKeyID* 替换为要使用的 KMS 密钥的密钥 ID 或 ARN：

```
aws emr create-cluster --release-label emr-4.7.2 or earlier
--emrfs Encryption=ClientSide,ProviderType=KMS,KMSKeyId=MyKMSKeyId
```

创建自定义密钥提供程序

创建自定义密钥提供程序时，应用程序应实施 EncryptionMaterialsProvider 接口，它在 Amazon SDK for Java 1.11.0 版及更高版本中可用。为提供加密材料，该实施可以使用任何策略。例如，您可以选择提供静态加密材料，也可以选择与更复杂的密钥管理系统集成。

用于自定义加密材料的加密算法必须是 AES/GCM/NoPadding。

EncryptionMaterialsProvider 类通过加密上下文获取加密材料。Amazon EMR 在运行时填充加密上下文信息，以帮助调用者确定要返回的正确加密材料。

Example 示例：通过 EMRFS 使用自定义密钥提供程序对 Amazon S3 进行加密

当 Amazon EMR 从 EncryptionMaterialsProvider 类中获取加密材料以执行加密时，EMRFS 可以选择使用两个字段填充 materialsDescription 参数：对象的 Amazon S3 URI 和集群的 JobFlowId，这两个字段可供 EncryptionMaterialsProvider 类使用来选择性地返回加密材料。

例如，提供程序可能会为不同的 Amazon S3 URI 前缀返回不同的密钥。它是最终与 Amazon S3 对象一起存储的返回加密材料的描述，而不是 EMRFS 生成并传递给提供程序的 materialsDescription 值。解密 Amazon S3 对象时，加密材料描述将传递给 EncryptionMaterialsProvider 类，以便它可以再次选择性地返回匹配的密钥以解密对象。

以下是 EncryptionMaterialsProvider 参考实施。另一个自定义提供程序 EMRFSRSAEncryptionMaterialsProvider 可从 GitHub 获取。

```
import com.amazonaws.services.s3.model.EncryptionMaterials;
```

```
import com.amazonaws.services.s3.model.EncryptionMaterialsProvider;
import com.amazonaws.services.s3.model.KMSEncryptionMaterials;
import org.apache.hadoop.conf.Configurable;
import org.apache.hadoop.conf.Configuration;

import java.util.Map;

/**
 * Provides KMSEncryptionMaterials according to Configuration
 */
public class MyEncryptionMaterialsProviders implements EncryptionMaterialsProvider,
 Configurable{
  private Configuration conf;
  private String kmsKeyId;
  private EncryptionMaterials encryptionMaterials;

  private void init() {
    this.kmsKeyId = conf.get("my.kms.key.id");
    this.encryptionMaterials = new KMSEncryptionMaterials(kmsKeyId);
  }

  @Override
  public void setConf(Configuration conf) {
    this.conf = conf;
    init();
  }

  @Override
  public Configuration getConf() {
    return this.conf;
  }

  @Override
  public void refresh() {

  }

  @Override
  public EncryptionMaterials getEncryptionMaterials(Map<String, String>
 materialsDescription) {
    return this.encryptionMaterials;
  }

  @Override
```

```
  public EncryptionMaterials getEncryptionMaterials() {
    return this.encryptionMaterials;
  }
}
```

使用 Amazon CLI 指定自定义材料提供程序

要使用 Amazon CLI，请将 Encryption、ProviderType、CustomProviderClass 和 CustomProviderLocation 参数传递给 emrfs 选项。

```
aws emr create-cluster --instance-type m5.xlarge --release-label emr-4.7.2 or earlier
 --emrfs Encryption=ClientSide,ProviderType=Custom,CustomProviderLocation=s3://
mybucket/myfolder/provider.jar,CustomProviderClass=classname
```

将 Encryption 设置为 ClientSide 会启用客户端加密，CustomProviderClass 是您的 EncryptionMaterialsProvider 对象的名称，而 CustomProviderLocation 是本地或 Amazon S3 位置，Amazon EMR 从此位置将 CustomProviderClass 复制到集群中的每个节点并放置在类路径中。

使用 SDK 指定自定义材料提供程序

要使用 SDK，您可以将属性 fs.s3.cse.encryptionMaterialsProvider.uri 设置为将存储在 Amazon S3 中的自定义 EncryptionMaterialsProvider 类下载到集群中的每个节点。您在 emrfs-site.xml 文件中配置此设置，同时启用 CSE 并提供自定义提供程序的正确位置。

例如，在 Amazon SDK for Java 中使用 RunJobFlowRequest 时，代码如下所示：

```
<snip>
  Map<String,String> emrfsProperties = new HashMap<String,String>();
      emrfsProperties.put("fs.s3.cse.encryptionMaterialsProvider.uri","s3://mybucket/
MyCustomEncryptionMaterialsProvider.jar");
      emrfsProperties.put("fs.s3.cse.enabled","true");
      emrfsProperties.put("fs.s3.consistent","true");

 emrfsProperties.put("fs.s3.cse.encryptionMaterialsProvider","full.class.name.of.EncryptionMate

  Configuration myEmrfsConfig = new Configuration()
      .withClassification("emrfs-site")
      .withProperties(emrfsProperties);
```

```
  RunJobFlowRequest request = new RunJobFlowRequest()
   .withName("Custom EncryptionMaterialsProvider")
   .withReleaseLabel("emr-5.36.1")
   .withApplications(myApp)
   .withConfigurations(myEmrfsConfig)
   .withServiceRole("EMR_DefaultRole_V2")
   .withJobFlowRole("EMR_EC2_DefaultRole")
   .withLogUri("s3://myLogUri/")
   .withInstances(new JobFlowInstancesConfig()
    .withEc2KeyName("myEc2Key")
    .withInstanceCount(2)
    .withKeepJobFlowAliveWhenNoSteps(true)
    .withMasterInstanceType("m5.xlarge")
    .withSlaveInstanceType("m5.xlarge")
   );

  RunJobFlowResult result = emr.runJobFlow(request);
</snip>
```

带参数的自定义 EncryptionMaterialsProvider

您可能需要将参数直接传递给提供程序。要执行此操作，您可以将 emrfs-site 配置分类与定义为属性的自定义参数结合使用。下面显示了一个示例配置，该示例配置将另存为 myConfig.json 文件：

```
[
    {
      "Classification": "emrfs-site",
      "Properties": {
        "myProvider.arg1":"value1",
     "myProvider.arg2":"value2"
      }
    }
 ]
```

在 Amazon CLI 中，您可以使用 create-cluster 命令中的 --configurations 选项指定文件，如下所示：

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge
 --instance-count 2 --configurations file://myConfig.json --emrfs
 Encryption=ClientSide,CustomProviderLocation=s3://mybucket/myfolder/
myprovider.jar,CustomProviderClass=classname
```

配置 EMRFS S3EC V2 支持

S3 Java SDK 版本（1.11.837 及更高版本）支持带有各种安全增强功能的加密客户端版本 2（S3EC V2）。有关更多信息，请参阅 S3 博客文章 Updates to the Amazon S3 encryption client。此外，还可以参阅《Amazon SDK for Java 开发人员指南》中的 Amazon S3 加密客户端迁移。

为保持向后兼容性，加密客户端 V1 在 SDK 中仍可用。默认情况下，启用 CSE 后，EMRFS 将使用 S3EC V1 加密和解密 S3 对象。

在发行版早于 emr-5.31.0（emr-5.30.1 及更早版本、emr-6.1.0 及更早版本）的 EMR 集群上，使用 S3EC V2 加密的 S3 对象无法通过 EMRFS 来解密。

Example 将 EMRFS 配置为使用 S3EC V2

要将 EMRFS 配置为使用 S3EC V2，请添加以下配置：

```
{
  "Classification": "emrfs-site",
  "Properties": {
    "fs.s3.cse.encryptionV2.enabled": "true"
  }
}
```

Amazon S3 客户端加密的 **emrfs-site.xml** 属性

属性	默认值	描述
`fs.s3.cse.enabled`	**false**	设置为 **true** 时，使用客户端加密对 Amazon S3 中存储的 EMRFS 对象进行加密。
`fs.s3.cse.encryptionV2.enabled`	**false**	设置为 `true` 时，EMRFS 使用 S3 加密客户端版本 2 来加密和解密 S3 上的对象。在 EMR 版本 5.31.0 及更高版本中提供。
`fs.s3.cse.encryptionMaterialsProvider.uri`	**N/A**	在使用自定义加密材料时适用。带 `EncryptionMaterial`

属性	默认值	描述
		sProvider 的 JAR 所在的 Amazon S3 URI。如果您提供此 URI，Amazon EMR 将此 JAR 自动下载到集群中的所有节点。
fs.s3.cse.encryptionMaterialsProvider	**N/A**	用于客户端加密的 EncryptionMaterialsProvider 类路径。在使用 CSE-KMS 时，请指定 com.amazon.ws.emr.hadoop.fs.cse.KMSEncryptionMaterialsProvider 。
fs.s3.cse.materialsDescription.enabled	**false**	在设置为 true 时，使用对象的 Amazon S3 URI 和 JobFlowId 填充加密对象的 materialsDescription。在使用自定义加密材料时设置为 true。
fs.s3.cse.kms.keyId	**N/A**	在使用 CSE-KMS 时适用。KeyId 的值、ARN 或用于加密的 KMS 密钥的别名。
fs.s3.cse.cryptoStorageMode	**ObjectMetadata**	Amazon S3 存储模式。默认情况下，加密信息的描述存储在对象元数据中。也可以将描述存储在指令文件中。有效值为 ObjectMetadata 和 InstructionFile。有关更多信息，请参阅使用 Amazon SDK for Java 和 Amazon S3 进行客户端数据加密。

Delta Lake

Delta Lake 是一种存储层框架，用于构建通常在 Amazon S3 上构建的数据湖仓一体架构。从 Amazon EMR 版本 6.9.0 开始，您可以在具有 Delta Lake 表的 Amazon EMR 集群上使用 Apache Spark 3.x。有关使用 Delta Lake 构建湖仓一体架构的更多信息，请参阅 https://delta.io/。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Delta 版本，以及 Amazon EMR 随 Delta 一起安装的组件。

有关此发行版中随 Delta 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Delta 版本信息

Amazon EMR 发行版标签	Delta 版本	随 Delta 安装的组件
emr-6.14.0	Delta 2.4.0	Not available.

Delta Lake 简介

Delta Lake 是一个开源项目，可帮助实施通常构建在 Amazon S3 上的现代数据湖架构。Delta Lake 提供以下功能：

- Spark 上的原子、一致、隔离、持久（ACID）事务。在 Spark 作业期间，读者可以看到一致的表格视图。
- 可扩展的元数据处理，由 Spark 进行分布式处理。
- 使用相同的 Delta 表结合流处理和批处理使用案例。
- 强制执行自动架构以避免数据摄取期间出现错误记录。
- 使用数据版本控制进行时空旅行。
- 支持合并、更新和删除操作，以支持复杂的使用案例，例如更改数据捕获（CDC）、流插入等等。

使用安装有 Delta Lake 的集群

主题

- 将 Delta Lake 集群与 Flink 结合使用
- 将 Delta Lake 集群与 Trino 结合使用

- 将 Delta Lake 集群与 Spark 结合使用
- 将 Delta Lake 集群与 Spark 和 Amazon Glue 结合使用

将 Delta Lake 集群与 Flink 结合使用

从 Amazon EMR 6.11 版本开始，您可以将 Delta Lake 与您的 Flink 集群结合使用。以下示例使用 Amazon CLI 在 Amazon EMR Flink 集群上使用 Delta Lake。

> **Note**
>
> 当您将 Delta Lake 与 Flink 集群结合使用时，Amazon EMR 支持 Flink DataStream API。

创建 Delta Lake 集群

1. 创建文件 `delta_configurations.json` 并输入以下内容：

```
[{"Classification":"delta-defaults",
    "Properties":{"delta.enabled":"true"}}]
```

2. 使用以下配置创建集群。在该 URL 中，将 `example Amazon S3 bucket path` 和 `subnet ID` 替换为您自己的值。

```
aws emr create-cluster
--release-label emr-6.11.0
--applications Name=Flink
--configurations file://delta_configurations.json
--region us-east-1  --name My_Spark_Delta_Cluster
--log-uri  s3://DOC-EXAMPLE-BUCKET/
--instance-type m5.xlarge
--instance-count 3
--service-role EMR_DefaultRole_V2
--ec2-attributes
 InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef0
```

初始化 Flink yarn 会话

要初始化 Flink yarn 会话，请运行以下命令：

```
flink-yarn-session -d
```

使用 Delta Lake 创建 Flink 作业

以下示例展示如何使用 sbt 或 Maven 在 Delta Lake 中构建 Flink 作业。

sbt

sbt 是 Scala 的构建工具，当您处理小型项目时，只需很少甚至不需要配置即可使用。

```
libraryDependencies ++= Seq(
  "io.delta" %% "delta-flink" % deltaConnectorsVersion % "provided",
  "io.delta" %% "delta-standalone" % deltaConnectorsVersion % "provided",
  "org.apache.flink" %% "flink-clients" % flinkVersion % "provided",
  "org.apache.flink" %% "flink-parquet" % flinkVersion % "provided",
  "org.apache.hadoop" % "hadoop-client" % hadoopVersion % "provided",
  "org.apache.flink" % "flink-table-common" % flinkVersion % "provided",
  "org.apache.flink" %% "flink-table-runtime" % flinkVersion % "provided")
```

Maven

Maven 是 Apache Software Foundation 推出的开源构建自动化工具。使用 Maven，您可以在 Amazon EMR 上使用 Delta Lake 构建、发布和部署 Flink 作业。

```
<project>
<properties>
    <scala.main.version>2.12</scala.main.version>
    <delta-connectors-version>0.6.0</delta-connectors-version>
    <flink-version>1.16.1</flink-version>
    <hadoop-version>3.1.0</hadoop-version>
</properties>

<dependencies>
    <dependency>
        <groupId>io.delta</groupId>
        <artifactId>delta-flink</artifactId>
        <version>$delta-connectors-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>io.delta</groupId>
        <artifactId>delta-standalone_$scala-main-version</artifactId>
```

```
        <version>$delta-connectors-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-clients</artifactId>
        <version>$flink-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-parquet</artifactId>
        <version>$flink-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-client</artifactId>
        <version>$hadoop-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-table-common</artifactId>
        <version>$flink-version</version>
        <scope>provided</scope>
    </dependency>
    <dependency>
        <groupId>org.apache.flink</groupId>
        <artifactId>flink-table-runtime</artifactId>
        <version>$flink-version</version>
        <scope>provided</scope>
    </dependency>
</dependencies>
```

通过 Flink Datastream API 写入 Delta 表

使用以下示例创建一个 DeltaSink，以通过 `deltaTablePath:` 写入表

```
public static DataStream<RowData> createDeltaSink(
        DataStream<RowData> stream,
        String deltaTablePath,
```

```
        RowType rowType) {
    Configuration configuration = new Configuration();
    DeltaSink<RowData> deltaSink = DeltaSink
            .forRowData(
                    new org.apache.flink.core.fs.Path(deltaTablePath),
                    configuration,
                    rowType)
            .build();
    stream.sinkTo(deltaSink);
    return stream;
}
```

通过 Flink Datastream API 从 Delta 表中读取

使用以下示例创建一个有界的 DeltaSource，以通过 `deltaTablePath`：从表中读取

```
public static DataStream<RowData> createBoundedDeltaSourceAllColumns(
        StreamExecutionEnvironment env,
        String deltaTablePath) {
    Configuration configuration = new Configuration();
    DeltaSource<RowData> deltaSource = DeltaSource
            .forBoundedRowData(
                    new org.apache.flink.core.fs.Path(deltaTablePath),
                    configuration)
            .build();

    return env.fromSource(deltaSource, WatermarkStrategy.noWatermarks(), "delta-
source");
}
```

使用对 Delta Lake 独立版的多集群支持创建接收器

使用以下示例创建一个 DeltaSink，以通过 `deltaTablePath` 和多集群支持写入表：

```
public DataStream<RowData> createDeltaSink(
        DataStream<RowData> stream,
        String deltaTablePath) {
    Configuration configuration = new Configuration();
    configuration.set("spark.delta.logStore.s3.impl",
 "io.delta.storage.S3DynamoDBLogStore");
    configuration.set("spark.io.delta.storage.S3DynamoDBLogStore.ddb.tableName",
 "delta_log");
```

```
    configuration.set("spark.io.delta.storage.S3DynamoDBLogStore.ddb.region", "us-
east-1");

    DeltaSink<RowData> deltaSink = DeltaSink
        .forRowData(
            new Path(deltaTablePath),
            configuration,
            rowType)
        .build();
    stream.sinkTo(deltaSink);
    return stream;
}
```

运行 Flink 作业

使用下列命令以运行您的作业：

```
flink run FlinkJob.jar
```

将 Delta Lake 集群与 Trino 结合使用

从 Amazon EMR 6.9.0 及更高版本开始，您可以将 Delta Lake 与您的 Trino 集群结合使用。

在本教程中，我们将通过 Amazon CLI 在 Amazon EMR Trino 集群上使用 Delta Lake。

创建 Delta Lake 集群

1. 创建文件 `delta_configurations.json`，然后为您选择的目录设置值。例如，假设您想将 Hive 元存储作为目录使用，则您的文件应包含以下内容：

   ```
   [{"Classification":"delta-defaults",
       "Properties":{"delta.enabled":"true"}},
       {"Classification":"trino-connector-delta",
       "Properties":{"hive.metastore.uri":"thrift://localhost:9083"}}]
   ```

   如果您想将 Amazon Glue 数据目录作为存储使用，则您的文件应包含以下内容：

   ```
   [{"Classification":"delta-defaults",
       "Properties":{"delta.enabled":"true"}},
       {"Classification":"trino-connector-delta",
   ```

```
    "Properties":{"hive.metastore":"glue"}}]
```

2. 使用以下配置创建集群，将 **example Amazon S3 bucket path** 和 **subnet ID** 替换为您自己的值。

```
aws emr create-cluster
    --release-label emr-6.9.0
    --applications Name=Trino
    --configurations file://delta_configurations.json
    --region us-east-1  --name My_Spark_Delta_Cluster
    --log-uri  s3://DOC-EXAMPLE-BUCKET/
    --instance-type m5.xlarge
    --instance-count 2
    --service-role EMR_DefaultRole_V2
    --ec2-attributes
 InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef0
```

初始化 Delta Lake 的 Trino 会话

要初始化 Trino 会话，请运行以下命令。

```
trino-cli --catalog delta
```

写入 Delta Lake 表

使用以下 SQL 命令创建并写入您的表：

```
SHOW SCHEMAS;

CREATE TABLE default.delta_table (id  int, data varchar, category varchar) WITH
( location =  's3://DOC-EXAMPLE-BUCKET/<prefix>');

INSERT INTO default.delta_table VALUES  (1,'a','c1'), (2,'b','c2'), (3,'c','c3');
```

从 Delta Lake 表中读取

使用以下 SQL 命令从您的表中读取：

```
SELECT * from default.delta_table;
```

将 Delta Lake 集群与 Spark 结合使用

从 Amazon EMR 版本 6.9.0 开始，您可以将 Delta Lake 与 Spark 集群结合使用，无需引导操作。对于 Amazon EMR 6.8.0 及更早版本，您可以使用引导操作来预安装需要的依赖项。

以下示例使用 Amazon CLI 在 Amazon EMR Spark 集群上使用 Delta Lake。

要在 Amazon EMR 上将 Delta Lake 与 Amazon Command Line Interface 结合使用，请首先创建集群。有关使用 Amazon Command Line Interface 指定 Delta Lake 分类的信息，请参阅 Supply a configuration using the Amazon Command Line Interface when you create a cluster 或 Supply a configuration with the Java SDK when you create a cluster。

1. 创建文件 `configurations.json` 并输入以下内容：

   ```
   [{"Classification":"delta-defaults",  "Properties":{"delta.enabled":"true"} }]
   ```

2. 使用以下配置创建集群，将示例 Amazon S3 **`bucket path`** 和 **`subnet ID`** 替换为您自己的值。

   ```
   aws emr create-cluster
        --release-label  emr-6.9.0
        --applications Name=Spark
        --configurations file://delta_configurations.json
        --region us-east-1
        --name My_Spark_Delta_Cluster
        --log-uri  s3://DOC-EXAMPLE-BUCKET/
        --instance-type m5.xlarge
        --instance-count 2
        --service-role EMR_DefaultRole_V2
        --ec2-attributes
    InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef0
   ```

   或者，您可以创建一个 Amazon EMR 集群和 Spark 应用程序，并在 Spark 作业中使用以下文件作为 JAR 依赖项：

   ```
   /usr/share/aws/delta/lib/delta-core.jar,
       /usr/share/aws/delta/lib/delta-storage.jar,
       /usr/share/aws/delta/lib/delta-storage-s3-dynamodb.jar
   ```

   有关更多信息，请参阅提交应用程序。

要将 jar 依赖项包含在 Spark 任务中，您可以将以下配置属性添加到 Spark 应用程序中：

```
--conf “spark.jars=/usr/share/aws/delta/lib/delta-core.jar,
     /usr/share/aws/delta/lib/delta-storage.jar,
     /usr/share/aws/delta/lib/delta-storage-s3-dynamodb.jar"
```

有关 Spark 任务依赖项的更多信息，请参阅 Dependency Management（依赖项管理）。

初始化 Delta Lake 的 Spark 会话

以下示例演示如何启动交互式 Spark Shell、使用 Spark 提交，或如何使用 Amazon EMR Notebooks 在 Amazon EMR 上使用 Delta Lake。

spark-shell

1. 使用 SSH 连接到主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以启动 Spark shell。要使用 PySpark Shell，请使用 `pyspark` 替换 `spark-shell`。

```
spark-shell \
   --conf "spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension" \
   --conf
 "spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog"
```

spark-submit

1. 使用 SSH 连接到主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以启动 Delta Lake 的 Spark 会话。

```
spark-submit
    –conf "spark.sql.extensions=io.delta.sql.DeltaSparkSessionExtension"
    –conf
 "spark.sql.catalog.spark_catalog=org.apache.spark.sql.delta.catalog.DeltaCatalog"
```

EMR Studio notebooks

要使用 Amazon EMR Studio Notebooks 初始化 Spark 会话，请使用 Amazon EMR Notebook 中的 %%configure 魔术命令配置 Spark 会话，如下例所示。有关更多信息，请参阅 Amazon EMR 管理指南中的使用 EMR Notebooks 魔法命令。

```
%%configure -f
{
  "conf": {
    "spark.sql.extensions":  "io.delta.sql.DeltaSparkSessionExtension",
     "spark.sql.catalog.spark_catalog":
 "org.apache.spark.sql.delta.catalog.DeltaCatalog"
  }
}
```

写入 Delta Lake 表

以下示例演示如何创建 DataFrame 并将其作为 Delta Lake 数据集写入。此示例演示如何使用 Spark Shell 处理数据集，同时使用 SSH 作为默认 hadoop 用户连接到主节点。

> **Note**
>
> 要将代码示例粘贴到 Spark Shell 中，请在提示符处键入 :paste，粘贴示例，然后按 CTRL + D。

PySpark

Spark 包含基于 Python 的 Shell `pyspark`，您可以用它来设计以 Python 编写的 Spark 程序的原型。就像使用 `spark-shell` 一样，在主节点上调用 `pyspark`。

```
## Create a DataFrame
data =  spark.createDataFrame([("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
("101",  "2015-01-01", "2015-01-01T12:14:58.597216Z"),
("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
("103",  "2015-01-01",  "2015-01-01T13:51:40.519832Z")],
["id", "creation_date",  "last_update_time"])

## Write a DataFrame as a Delta Lake dataset to the S3  location
spark.sql("""CREATE  TABLE IF NOT EXISTS delta_table (id string, creation_date
 string,
```

```
last_update_time string)
USING delta location
's3://DOC-EXAMPLE-BUCKET/example-prefix/db/delta_table'""");

data.writeTo("delta_table").append()
```

Scala

```
import org.apache.spark.sql.SaveMode
import org.apache.spark.sql.functions._

// Create a DataFrame
val data = Seq(("100",  "2015-01-01",  "2015-01-01T13:51:39.340396Z"),
("101",  "2015-01-01",  "2015-01-01T12:14:58.597216Z"),
("102",  "2015-01-01", "2015-01-01T13:51:40.417052Z"),
("103",  "2015-01-01",  "2015-01-01T13:51:40.519832Z")).toDF("id", "creation_date",
 "last_update_time")

// Write a DataFrame as a Delta Lake dataset to the S3  location
spark.sql("""CREATE  TABLE IF NOT EXISTS delta_table (id string,
creation_date string,
last_update_time string)
USING delta location
's3://DOC-EXAMPLE-BUCKET/example-prefix/db/delta_table'""");

data.write.format("delta").mode("append").saveAsTable("delta_table")
```

SQL

```
-- Create a Delta  Lake table with the S3 location
CREATE TABLE delta_table(id string,
creation_date string,
last_update_time string)
USING delta LOCATION
's3://DOC-EXAMPLE-BUCKET/example-prefix/db/delta_table';

-- insert data into the table
INSERT INTO delta_table VALUES  ("100", "2015-01-01",
 "2015-01-01T13:51:39.340396Z"),
("101",  "2015-01-01",  "2015-01-01T12:14:58.597216Z"),
("102",  "2015-01-01", "2015-01-01T13:51:40.417052Z"),
("103",  "2015-01-01", "2015-01-01T13:51:40.519832Z");
```

从 Delta Lake 表中读取

PySpark

```
ddf = spark.table("delta_table")
ddf.show()
```

Scala

```
val ddf =  spark.table("delta_table")
ddf.show()
```

SQL

```
SELECT * FROM delta_table;
```

将 Delta Lake 集群与 Spark 和 Amazon Glue 结合使用

要将 Amazon Glue 数据目录作为 Delta Lake 表的元存储，请按如下步骤创建集群。有关使用 Amazon Command Line Interface 指定 Delta Lake 分类的信息，请参阅在创建集群时使用 Amazon Command Line Interface 提供配置或 Supply a configuration using the Java SDK when you create a cluster（在创建集群时使用 Java SDK 提供配置）。

创建 Delta Lake 集群

1. 创建文件 `configurations.json` 并输入以下内容：

```
[{"Classification":"delta-defaults",
"Properties":{"delta.enabled":"true"}},
{"Classification":"spark-hive-site",
"Properties":
{"hive.metastore.client.factory.class":"com.amazonaws.glue.catalog.metastore.AWSGlueDataCat
```

2. 使用以下配置创建集群，将 **`example Amazon S3 bucket path`** 和 **`subnet ID`** 替换为您自己的值。

```
aws emr create-cluster
```

```
    --release-label  emr-6.9.0
    --applications Name=Spark
    --configurations file://delta_configurations.json
    --region us-east-1
    --name My_Spark_Delta_Cluster
    --log-uri  s3://DOC-EXAMPLE-BUCKET/
    --instance-type m5.xlarge
    --instance-count 2
    --service-role EMR_DefaultRole_V2
    --ec2-attributes
InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef0
```

注意事项和限制

- 在 Amazon EMR 6.9.0 及更高版本上支持使用 Delta Lake。您可以在包含 Delta 表的 Amazon EMR 集群上使用 Apache Spark 3.x。
- 建议您对 S3 位置路径使用 s3 URI 方案而不是 s3a，以获得最佳性能、安全性和可靠性。有关更多信息，请参阅使用存储和文件系统。
- 在 Amazon EMR 6.9 和 6.10 中，当 Delta Lake 表数据存储在 Amazon S3 中时，列数据在列重命名操作后会变为 NULL。从 Amazon EMR 6.11 开始，此问题已得到解决。有关此实验性列重命名操作的更多信息，请参阅《Delta Lake User Guide》中的 Column rename operation。
- 如果您在 Apache Spark 之外的 Amazon Glue 数据目录中创建数据库，则该数据库可能有一个空的 LOCATION 字段。由于 Spark 不允许使用空位置属性创建数据库，因此如果您在 Amazon EMR 中使用 Spark 在 Glue 数据库中创建 Delta 表，并且该数据库具有空的 LOCATION 属性，则会出现以下错误：

  ```
  IllegalArgumentException: Can not create a Path from an empty string
  ```

  要解决此问题，请在数据目录中创建数据库，并且 LOCATION 字段使用有效的非空路径。有关实现此解决方案的步骤，请参阅《Amazon Athena 用户指南》中的 创建表时出现非法参数异常。

Delta 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Delta 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Delta 版本信息

Amazon EMR 发行版标签	Delta 版本	随 Delta 安装的组件
emr-6.14.0	2.4.0	Not available.
emr-6.13.0	2.4.0	Not available.
emr-6.12.0	2.4.0	Not available.
emr-6.11.1	2.2.0	Not available.
emr-6.11.0	2.2.0	Not available.
emr-6.10.1	2.2.0	Not available.
emr-6.10.0	2.2.0	Not available.
emr-6.9.1	2.1.0	Not available.
emr-6.9.0	2.1.0	Not available.

Apache Flink

Apache Flink 是一个流式处理数据流引擎，您可以使用此引擎在高吞吐量数据源上轻松运行实时流处理。Flink 支持无序事件的事件时间语义、确切一次语义、反向压力控制以及已为写入流和批处理应用程序优化的 API。

此外，Flink 具有适用于第三方数据源的连接器，例如以下内容：

- Amazon Kinesis Data Streams
- Apache Kafka
- Flink Elasticsearch Connector
- Twitter Streaming API
- Cassandra

Amazon EMR 支持 Flink 作为 YARN 应用程序，以便您能管理资源以及集群中的其它应用程序。利用 Flink-on-YARN，您可以提交临时 Flink 作业，也可以创建一个长时间运行的集群，该集群接受多个作业并根据整体 YARN 预留分配资源。

Flink 包含在 Amazon EMR发行版 5.1.0 及更高版本中。

> **Note**
>
> 在 Amazon EMR 5.2.1 发行版本中增加了对 `FlinkKinesisConsumer` 类的支持。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Flink 的版本，以及 Amazon EMR 随 Flink 一起安装的组件。

有关此发行版中随 Flink 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Flink 版本信息

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.14.0	Flink 1.17.1-amzn-0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode,

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
		hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Flink 的版本，以及 Amazon EMR 随 Flink 一起安装的组件。

有关此发行版中随 Flink 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Flink 版本信息

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.36.1	Flink 1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

主题

- 使用 Flink 创建集群
- 在 Amazon EMR 中配置 Flink
- 在 Amazon EMR 中使用 Flink 作业
- 使用 Scala Shell
- 查找 Flink Web 界面

- 在 Amazon EMR 中通过 Zeppelin 使用 Flink 作业
- Flink 发布历史记录

使用 Flink 创建集群

您可以使用Amazon Web Services Management Console、Amazon CLI 或 Amazon SDK 启动集群。

使用控制台启动安装了 Flink 的集群

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 对于 Software Configuration (软件配置)，选择 EMR Release emr-5.1.0 (EMR 版本 emr-5.1.0) 或更高版本。
4. 选择 Flink 作为应用程序（与要安装的任何其它应用程序一起）。
5. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

使用 Amazon CLI 启动带有 Flink 的集群

- 使用下面的命令创建集群：

```
aws emr create-cluster --release-label emr-5.36.1 \
--applications Name=Flink \
--configurations file://./configurations.json \
--region us-east-1 \
--log-uri s3://myLogUri \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole_V2 \
--ec2-attributes KeyName=MyKeyName,InstanceProfile=EMR_EC2_DefaultRole \
--steps Type=CUSTOM_JAR,Jar=command-runner.jar,Name=Flink_Long_Running_Session,\
Args=flink-yarn-session,-d
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

在 Amazon EMR 中配置 Flink

使用 Hive 元存储和 Glue 目录配置 Flink

Amazon EMR 版本 6.9.0 及更高版本支持 Hive 元存储和 Amazon Glue 目录使用 Apache Flink 连接器连接到 Hive。本部分概括介绍了使用 Flink 配置 Amazon Glue 目录和 Hive 元存储所需的步骤。

主题

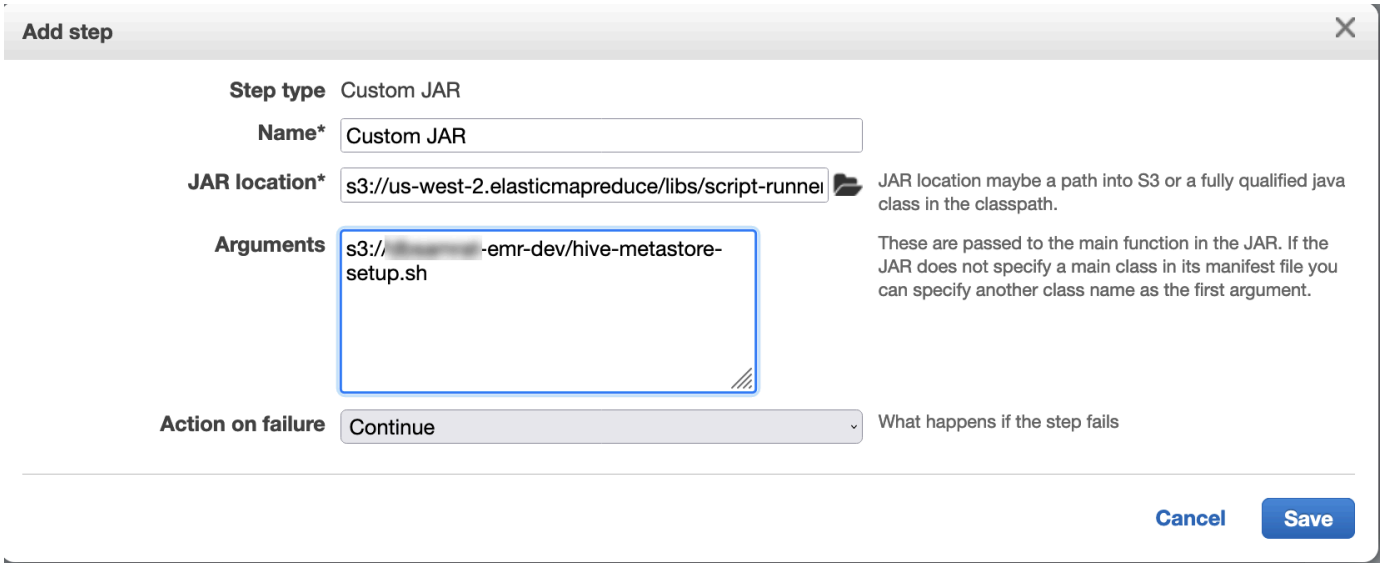
- 使用 Hive 元存储
- 使用 Amazon Glue 数据目录

使用 Hive 元存储

1. 创建 EMR 集群，其中包含版本 6.9.0 或更高版本，并至少包含两个应用程序：Hive 和 Flink。
2. 使用脚本运行程序将以下脚本作为步骤函数执行：

   `hive-metastore-setup.sh`

```
sudo cp /usr/lib/hive/lib/antlr-runtime-3.5.2.jar /usr/lib/flink/lib
sudo cp /usr/lib/hive/lib/hive-exec-3.1.3*.jar /lib/flink/lib
sudo cp /usr/lib/hive/lib/libfb303-0.9.3.jar /lib/flink/lib
sudo cp /usr/lib/flink/opt/flink-connector-hive_2.12-1.15.2.jar /lib/flink/lib
sudo chmod 755 /usr/lib/flink/lib/antlr-runtime-3.5.2.jar
sudo chmod 755 /usr/lib/flink/lib/hive-exec-3.1.3*.jar
sudo chmod 755 /usr/lib/flink/lib/libfb303-0.9.3.jar
sudo chmod 755 /usr/lib/flink/lib/flink-connector-hive_2.12-1.15.2.jar
```

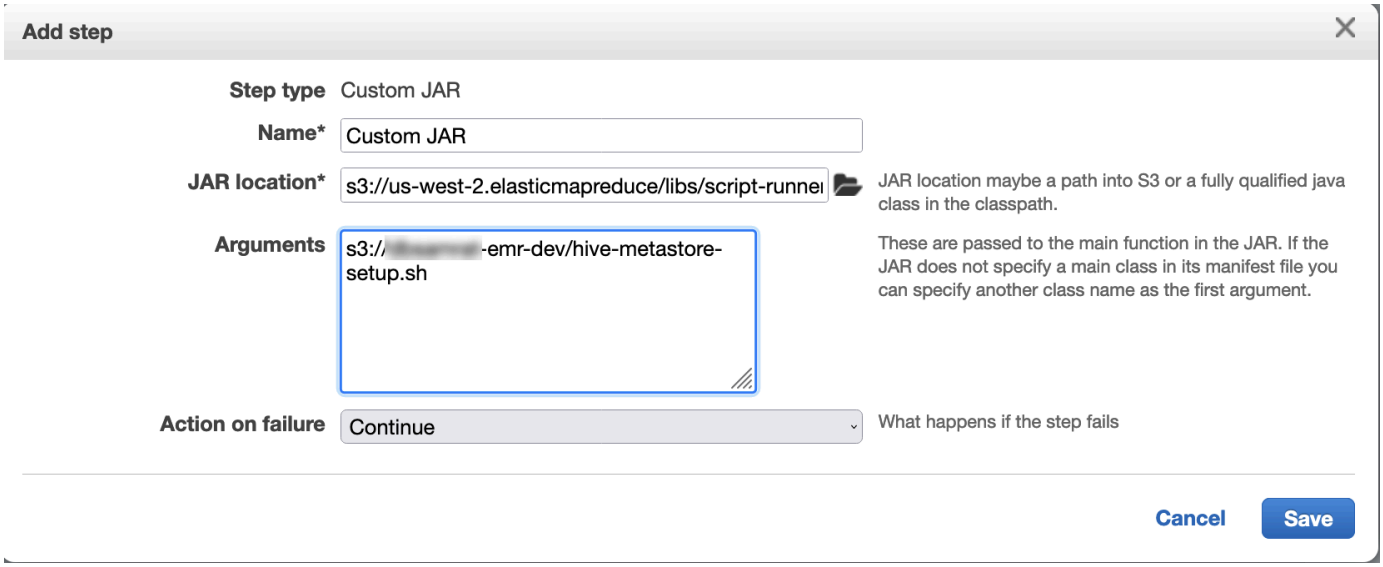


使用 Amazon Glue 数据目录

1. 创建 EMR 集群，其中包含版本 6.9.0 或更高版本，并至少包含两个应用程序：Hive 和 Flink。
2. 在 Amazon Glue 数据目录设置中选择用于 Hive 表元数据，以在集群中启用数据目录。
3. 使用脚本运行程序并将以下脚本作为阶跃函数执行：在 Amazon EMR 集群上运行命令和脚本：

   glue-catalog-setup.sh

```
sudo cp /usr/lib/hive/auxlib/aws-glue-datacatalog-hive3-client.jar /usr/lib/flink/
lib
sudo cp /usr/lib/hive/lib/antlr-runtime-3.5.2.jar /usr/lib/flink/lib
sudo cp /usr/lib/hive/lib/hive-exec-3.1.3*.jar /lib/flink/lib
sudo cp /usr/lib/hive/lib/libfb303-0.9.3.jar /lib/flink/lib
sudo cp /usr/lib/flink/opt/flink-connector-hive_2.12-1.15.2.jar /lib/flink/lib
sudo chmod 755 /usr/lib/flink/lib/aws-glue-datacatalog-hive3-client.jar
sudo chmod 755 /usr/lib/flink/lib/antlr-runtime-3.5.2.jar
sudo chmod 755 /usr/lib/flink/lib/hive-exec-3.1.3*.jar
sudo chmod 755 /usr/lib/flink/lib/libfb303-0.9.3.jar
sudo chmod 755 /usr/lib/flink/lib/flink-connector-hive_2.12-1.15.2.jar
```



使用配置文件配置 Flink

您可以使用 Amazon EMR 配置 API 通过配置文件配置 Flink。目前，可在 API 中配置的文件包括：

- `flink-conf.yaml`
- `log4j.properties`
- `flink-log4j-session`
- `log4j-cli.properties`

Flink 的主配置文件的名称为 `flink-conf.yaml`。

从 Amazon CLI 配置用于 Flink 的任务槽的数目

1. 创建文件 `configurations.json` 并输入以下内容：

```
[
    {
      "Classification": "flink-conf",
      "Properties": {
        "taskmanager.numberOfTaskSlots":"2"
      }
    }
]
```

2. 接下来，使用以下配置创建集群：

```
aws emr create-cluster --release-label emr-5.36.1 \
--applications Name=Flink \
--configurations file://./configurations.json \
--region us-east-1 \
--log-uri s3://myLogUri \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole_V2 \
--ec2-attributes KeyName=YourKeyName,InstanceProfile=EMR_EC2_DefaultRole
```

> **Note**
>
> 您也可以使用 Flink API 更改某些配置。有关更多信息，请参阅 Flink 文档中的概念。对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

并行选项

作为应用程序所有者，您最了解应将哪些资源分配给 Flink 中的任务。对于本文档中的示例，请使用与您用于应用程序的任务实例相同的任务数量。通常，我们建议对初始并行级别执行此操作，但您也可以使用任务槽来增加并行粒度，它一般不应超过每实例虚拟内核数量。有关 Flink 架构的更多信息，请参阅 Flink 文档中的 Concepts。

在包括多个主节点的 EMR 集群中配置 Flink

在包含多个主节点的 Amazon EMR 集群中进行主节点失效转移的过程中，Flink 的 JobManager 仍然可用。从 Amazon EMR 版本 5.28.0 开始，JobManager 的高可用性也会自动启用。无需手动配置。

对于 Amazon EMR 5.27.0 或更早版本，JobManager 是单点故障。当 JobManager 失败时，它会失去所有作业状态，并且不会恢复正在运行的作业。通过配置应用程序尝试计数、开展检查点检验并启用 ZooKeeper 作为 Flink 的状态存储，您可以启用 JobManager 高可用性，如以下示例所示：

```
[
```

```
  {
    "Classification": "yarn-site",
    "Properties": {
      "yarn.resourcemanager.am.max-attempts": "10"
    }
  },
  {
    "Classification": "flink-conf",
    "Properties": {
        "yarn.application-attempts": "10",
        "high-availability": "zookeeper",
        "high-availability.zookeeper.quorum": "%{hiera('hadoop::zk')}",
        "high-availability.storageDir": "hdfs:///user/flink/recovery",
        "high-availability.zookeeper.path.root": "/flink"
    }
  }
]
```

您必须同时为 YARN 和 Flink 配置最大的应用程序主尝试次数。有关更多信息，请参阅 YARN 集群高可用性的配置。您可能还需要配置 Flink 检查点，以使重新启动的 JobManager 从先前完成的检查点恢复正在运行的作业。有关更多信息，请参阅开展 Flink 检查点检验。

配置内存进程大小

对于使用 Flink 1.11.x 的 Amazon EMR 版本，您必须在 `flink-conf.yaml` 中为 JobManager (`jobmanager.memory.process.size`) 和 TaskManager (`taskmanager.memory.process.size`) 配置总内存进程大小。您可以通过使用配置 API 来配置集群或通过 SSH 手动取消这些字段来设置这些值。Flink 提供以下默认值。

- `jobmanager.memory.process.size`：1600m
- `taskmanager.memory.process.size`：1728m

要排除 JVM 元空间和开销，请使用 Flink 总内存大小 (`taskmanager.memory.flink.size`) 而非 `taskmanager.memory.process.size`。`taskmanager.memory.process.size` 的默认值为 1280m。不建议同时设置 `taskmanager.memory.process.size` 和 `taskmanager.memory.process.size`。

所有使用 Flink 1.12.0 及更高版本的 Amazon EMR 版本，都将 Flink 的开源设置中列出的默认值作为 Amazon EMR 上的默认值，因此您无需自行配置。

配置日志输出文件大小

Flink 应用程序容器创建并写入三种类型的日志文件：.out 文件、.log 文件和 .err 文件。仅限将 .err 文件压缩并从文件系统中删除，而将 .log 和 .out 日志文件保留在文件系统中。为确保这些输出文件保持可管理以及集群保持稳定，您可以在 log4j.properties 设置文件的上限数量并限制其大小。

Amazon EMR 版本 5.30.0 及更高版本

从 Amazon EMR 5.30.0 开始，Flink 使用带有配置分类名称 flink-log4j. 的 log4j2 日志记录框架。以下示例配置演示 log4j2 格式。

```
[
  {
    "Classification": "flink-log4j",
    "Properties": {
      "appender.main.name": "MainAppender",
      "appender.main.type": "RollingFile",
      "appender.main.append" : "false",
      "appender.main.fileName" : "${sys:log.file}",
      "appender.main.filePattern" : "${sys:log.file}.%i",
      "appender.main.layout.type" : "PatternLayout",
      "appender.main.layout.pattern" : "%d{yyyy-MM-dd HH:mm:ss,SSS} %-5p %-60c %x - %m
%n",
      "appender.main.policies.type" : "Policies",
      "appender.main.policies.size.type" : "SizeBasedTriggeringPolicy",
      "appender.main.policies.size.size" : "100MB",
      "appender.main.strategy.type" : "DefaultRolloverStrategy",
      "appender.main.strategy.max" : "10"
    },
  }
]
```

Amazon EMR 版本 5.29.0 及较早版本

对于 Amazon EMR 5.29.0 及更早版本，Flink 使用 log4j 日志记录框架。下面的示例配置演示了 log4j 格式。

```
[
  {
    "Classification": "flink-log4j",
```

```
    "Properties": {
      "log4j.appender.file": "org.apache.log4j.RollingFileAppender",
      "log4j.appender.file.append":"true",
      # keep up to 4 files and each file size is limited to 100MB
      "log4j.appender.file.MaxFileSize":"100MB",
      "log4j.appender.file.MaxBackupIndex":4,
      "log4j.appender.file.layout":"org.apache.log4j.PatternLayout",
      "log4j.appender.file.layout.ConversionPattern":"%d{yyyy-MM-dd HH:mm:ss,SSS} %-5p
 %-60c %x - %m%n"
    },
  }
]
```

将 Flink 配置为使用 Java 11 运行

Amazon EMR 6.12.0 及更高版本为 Flink 提供 Java 11 运行时系统支持。以下各节介绍如何配置集群以为 Flink 提供 Java 11 运行时系统支持。

主题

- 在创建集群时配置 Flink for Java 11
- 在正在运行的集群上配置 Flink for Java 11
- 在正在运行的集群上确认 Flink 的 Java 运行时系统

在创建集群时配置 Flink for Java 11

使用以下步骤创建包含 Flink 和 Java 11 运行时系统的 EMR 集群。添加 Java 11 运行时系统支持所在的配置文件是 `flink-conf.yaml`。

New console

在新控制台中创建包含 Flink 和 Java 11 运行时系统的集群

1. 登录 Amazon Web Services Management Console 并打开 Amazon EMR 控制台，网址为 https://console.aws.amazon.com/emr。
2. 在导航窗格中的 EC2 上的 EMR 下，选择集群，然后选择创建集群。
3. 选择 Amazon EMR 6.12.0 或更高版本，然后选择安装 Flink 应用程序。选择要在集群上安装的任何其他应用程序。
4. 继续设置您的集群。在可选的软件设置部分，使用默认的输入配置选项，并输入以下配置：

```
[
    {
      "Classification": "flink-conf",
      "Properties": {
        "containerized.taskmanager.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
        "containerized.master.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
        "env.java.home":"/usr/lib/jvm/jre-11"
      }
    }
]
```

5. 继续设置并启动您的集群。

Amazon CLI

从 CLI 创建包含 Flink 和 Java 11 运行时系统的集群

1. 创建一个将 Flink 配置为使用 Java 11 的配置文件 `configurations.json`。

```
[
    {
      "Classification": "flink-conf",
      "Properties": {
        "containerized.taskmanager.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
        "containerized.master.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
        "env.java.home":"/usr/lib/jvm/jre-11"
      }
    }
]
```

2. 从 Amazon CLI 中，使用 Amazon EMR 6.12.0 或更高版本创建新 EMR 集群，然后安装 Flink 应用程序，如以下示例所示：

```
aws emr create-cluster --release-label emr-6.12.0 \
--applications Name=Flink \
--configurations file://./configurations.json \
--region us-east-1 \
--log-uri s3://myLogUri \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole_V2 \
```

```
--ec2-attributes KeyName=YourKeyName,InstanceProfile=EMR_EC2_DefaultRole
```

在正在运行的集群上配置 Flink for Java 11

使用以下步骤更新包含 Flink 和 Java 11 运行时系统的 EMR 集群。添加 Java 11 运行时系统支持所在的配置文件是 `flink-conf.yaml`。

New console

在新控制台中更新包含 Flink 和 Java 11 运行时系统的正在运行的集群

1. 登录 Amazon Web Services Management Console 并打开 Amazon EMR 控制台，网址为 https://console.aws.amazon.com/emr。
2. 在导航窗格中的 EC2 上的 EMR 下，选择集群，然后选择要更新的集群。

> **Note**
>
> 集群必须使用 Amazon EMR 6.12.0 或更高版本才能支持 Java 11。

3. 选择配置选项卡。
4. 在实例组配置部分，选择要更新的正在运行的实例组，然后从列表操作菜单中选择重新配置。
5. 使用编辑属性选项重新配置实例组，如下所示。在每个配置之后选择添加新配置。

分类	属性	Value
flink-conf	containerized.taskmanager.env.JAVA_HOME	/usr/lib/jvm/jre-11
flink-conf	containerized.master.env.JAVA_HOME	/usr/lib/jvm/jre-11
flink-conf	env.java.home	/usr/lib/jvm/jre-11

6. 选择保存更改以添加配置。

Amazon CLI

从 CLI 中更新正在运行的集群，以使用 Flink 和 Java 11 运行时系统

使用 `modify-instance-groups` 命令为运行的集群中的一个实例组指定新配置。

1. 首先，创建一个将 Flink 配置为使用 Java 11 的配置文件 `configurations.json`。在以下示例中，将 *ig-1xxxxxxx9* 替换为您要重新配置的实例组的 ID。将文件保存在您将要运行 `modify-instance-groups` 命令的同一目录中。

```
[
   {
      "InstanceGroupId":"ig-1xxxxxxx9",
      "Configurations":[
         {
            "Classification":"flink-conf",
            "Properties":{
              "containerized.taskmanager.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
              "containerized.master.env.JAVA_HOME":"/usr/lib/jvm/jre-11",
              "env.java.home":"/usr/lib/jvm/jre-11"
            },
            "Configurations":[]
         }
      ]
   }
]
```

2. 从 Amazon CLI 运行以下命令。替换您要重新配置的实例组的 ID：

```
aws emr modify-instance-groups --cluster-id j-2AL4XXXXXX5T9 \
--instance-groups file://configurations.json
```

在正在运行的集群上确认 Flink 的 Java 运行时系统

要确定正在运行的集群的 Java 运行时系统，请使用 SSH 登录主节点，如 Connect to the primary node with SSH 中所述。然后运行以下命令：

```
ps -ef | grep flink
```

包含 -ef 选项的 ps 命令列出了系统上所有正在运行的进程。您可以使用 grep 过滤该输出，以查找提及 flink 字符串的内容。查看 Java 运行时环境（JRE）值的输出 jre-XX。在以下输出中，jre-11 表示在运行时系统为 Flink 选择了 Java 11。

```
flink    19130     1  0 09:17 ?        00:00:15 /usr/lib/jvm/jre-11/bin/
java -Djava.io.tmpdir=/mnt/tmp -Dlog.file=/usr/lib/flink/log/flink-flink-
historyserver-0-ip-172-31-32-127.log -Dlog4j.configuration=file:/usr/lib/flink/conf/
log4j.properties -Dlog4j.configurationFile=file:/usr/lib/flink/conf/log4j.properties
 -Dlogback.configurationFile=file:/usr/lib/flink/conf/logback.xml -classpath /usr/lib/
flink/lib/flink-cep-1.17.0.jar:/usr/lib/flink/lib/flink-connector-files-1.17.0.jar:/
usr/lib/flink/lib/flink-csv-1.17.0.jar:/usr/lib/flink/lib/flink-json-1.17.0.jar:/usr/
lib/flink/lib/flink-scala_2.12-1.17.0.jar:/usr/lib/flink/lib/flink-table-api-java-
uber-1.17.0.jar:/usr/lib/flink/lib/flink-table-api-scala-bridge_2.12-1.17.0.
```

或者，使用 SSH 登录主节点，然后使用命令 flink-yarn-session -d 启动 Flink YARN 会话。输出显示了 Flink 的 Java 虚拟机（JVM），如以下 java-11-amazon-corretto 示例所示：

```
2023-05-29 10:38:14,129 INFO  org.apache.flink.configuration.GlobalConfiguration
     [] - Loading configuration property: containerized.master.env.JAVA_HOME, /usr/lib/
jvm/java-11-amazon-corretto.x86_64
```

在 Amazon EMR 中使用 Flink 作业

可通过多种方式与 Amazon EMR 上的 Flink 交互：通过控制台、在 ResourceManager 跟踪 UI 中找到的 Flink 接口，以及在命令行。您可通过以上任一方式将 JAR 文件提交到 Flink 应用程序。提交 JAR 文件后，它就会变成由 Flink JobManager 管理的作业。JobManager 位于托管 Flink 会话 Application Master 进程守护程序的 YARN 节点上。

可以将 Flink 应用程序作为长时间运行集群或临时集群上的 YARN 作业。在长时间运行集群上，您可以将多个 Flink 作业提交给 Amazon EMR 上运行的一个 Flink 集群。如果您在临时集群上运行 Flink 作业，则 Amazon EMR 集群仅在其运行 Flink 应用程序的时间内存在，因此您只需为使用的资源和时间付费。您可以使用 Amazon EMR AddSteps API 操作，并通过 Amazon CLI add-steps 或 create-cluster 命令来提交 Flink 作业，作为 RunJobFlow 操作的步骤参数。

启动 Flink YARN 应用程序，作为长时间运行集群上的步骤

要启动 Flink 应用程序，使多个客户端能够通过 YARN API 操作向其提交工作，需要您创建集群或将 Flink 应用程序添加到现有集群中。有关如何创建新集群的说明，请参阅使用 Flink 创建集群。要在现有集群上启动 YARN 会话，可从控制台、Amazon CLI 或 Java SDK 使用以下步骤。

> **Note**
>
> 向 Amazon EMR 5.5.0 版本中添加了 `flink-yarn-session` 命令作为 `yarn-session.sh` 脚本的包装程序以简化执行。如果您使用 Amazon EMR 的更早版本，请将 `bash -c "/usr/lib/flink/bin/yarn-session.sh -d"` 在控制台中替换为 Arguments (参数) 或在 Amazon CLI 命令中替换为 `Args`。

使用控制台在现有集群上提交 Flink 作业

使用 `flink-yarn-session` 命令在现有集群中提交 Flink 会话。

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 在集群列表中，选择先前已启动的集群。
3. 在集群详细信息页面上，选择 Steps (步骤)，再选择 Add Step (添加步骤)。
4. 使用随后提供的指南输入参数，然后选择添加。

参数	描述
Step type (步骤类型)	自定义 JAR
名称	可帮助您标识步骤的名称。例如，*<example-flink-step-name>* 。
Jar location (Jar 位置)	command-runner.jar
Arguments (参数)	带适合您的应用的参数的 `flink-yarn-session` 命令。例如，*`flink-yarn-session -d`* 在 YARN 集群中以分离状态 (`-d`) 启动 Flink 会话。有关参数详细信息，请参阅新版 Flink 文档中的 YARN 设置。

使用 Amazon CLI 在现有集群上提交 Flink 作业

- 使用 `add-steps` 命令将 Flink 任务添加到长时间运行的集群。以下示例命令指定 `Args="flink-yarn-session", "-d"` 在分离状态下 (-d) 在 YARN 集群中启动 Flink 会话。有关参数详细信息，请参阅新版 Flink 文档中的 YARN 设置。

```
aws emr add-steps --cluster-id <j-XXXXXXXX> --steps Type=CUSTOM_JAR,Name=<example-
flink-step-name>,Jar=command-runner.jar,Args="flink-yarn-session","-d"
```

将工作提交到长时间运行集群上的现有 Flink 应用程序

如果您在长时间运行的集群上已有 Flink 应用程序，则可以指定集群的 Flink 应用程序 ID，以便向其提交工作。要获取应用程序 ID，请在 Amazon CLI 上运行 `yarn application -list` 或通过 YarnClient API 操作：

```
$ yarn application -list
16/09/07 19:32:13 INFO client.RMProxy: Connecting to ResourceManager at
 ip-10-181-83-19.ec2.internal/10.181.83.19:8032
Total number of applications (application-types: [] and states: [SUBMITTED, ACCEPTED,
 RUNNING]):1
Application-Id    Application-Name    Application-Type    User    Queue    State
 Final-State    Progress    Tracking-URL
application_1473169569237_0002    Flink session with 14 TaskManagers (detached)
   Apache Flink     hadoop    default            RUNNING          UNDEFINED
 100% http://ip-10-136-154-194.ec2.internal:33089
```

此 Flink 会话的应用程序 ID 为 `application_1473169569237_0002`，支持您使用 Amazon CLI 或 SDK 来将作业提交到应用程序。

Example SDK for Java

```
List<StepConfig> stepConfigs = new ArrayList<StepConfig>();

HadoopJarStepConfig flinkWordCountConf = new HadoopJarStepConfig()
    .withJar("command-runner.jar")
    .withArgs("flink", "run", "-m", "yarn-cluster", "-yid",
 "application_1473169569237_0002", "-yn", "2", "/usr/lib/flink/examples/streaming/
WordCount.jar",
      "--input", "s3://myBucket/pg11.txt", "--output", "s3://myBucket/alice2/");

StepConfig flinkRunWordCount = new StepConfig()
  .withName("Flink add a wordcount step")
  .withActionOnFailure("CONTINUE")
  .withHadoopJarStep(flinkWordCountConf);

stepConfigs.add(flinkRunWordCount);
```

```
AddJobFlowStepsResult res = emr.addJobFlowSteps(new AddJobFlowStepsRequest()
   .withJobFlowId("myClusterId")
   .withSteps(stepConfigs));
```

Example Amazon CLI

```
aws emr add-steps --cluster-id <j-XXXXXXXX> \
--steps Type=CUSTOM_JAR,Name=Flink_Submit_To_Long_Running,Jar=command-runner.jar,\
Args="flink","run","-m","yarn-cluster","-yid","application_1473169569237_0002",\
"/usr/lib/flink/examples/streaming/WordCount.jar",\
"--input","s3://myBucket/pg11.txt","--output","s3://myBucket/alice2/" \
--region <region-code>
```

提交临时 Flink 作业

以下示例启动一个临时集群，它运行 Flink 作业并在完成时将其终止。

Example SDK for Java

```
import java.util.ArrayList;
import java.util.List;
import com.amazonaws.AmazonClientException;
import com.amazonaws.auth.AWSCredentials;
import com.amazonaws.auth.AWSStaticCredentialsProvider;
import com.amazonaws.auth.profile.ProfileCredentialsProvider;
import com.amazonaws.services.elasticmapreduce.AmazonElasticMapReduce;
import com.amazonaws.services.elasticmapreduce.AmazonElasticMapReduceClientBuilder;
import com.amazonaws.services.elasticmapreduce.model.*;

public class Main_test {

 public static void main(String[] args) {
  AWSCredentials credentials_profile = null;
  try {
   credentials_profile = new ProfileCredentialsProvider("default").getCredentials();
        } catch (Exception e) {
            throw new AmazonClientException(
                    "Cannot load credentials from .aws/credentials file. " +
                    "Make sure that the credentials file exists and the profile name is
 specified within it.",
                    e);
        }
```

```
  AmazonElasticMapReduce emr = AmazonElasticMapReduceClientBuilder.standard()
    .withCredentials(new AWSStaticCredentialsProvider(credentials_profile))
    .withRegion(Regions.US_WEST_1)
    .build();

  List<StepConfig> stepConfigs = new ArrayList<StepConfig>();
     HadoopJarStepConfig flinkWordCountConf = new HadoopJarStepConfig()
       .withJar("command-runner.jar")
       .withArgs("bash","-c", "flink", "run", "-m", "yarn-cluster", "-yn", "2", "/usr/
lib/flink/examples/streaming/WordCount.jar", "--input", "s3://path/to/input-file.txt",
 "--output", "s3://path/to/output/");

     StepConfig flinkRunWordCountStep = new StepConfig()
       .withName("Flink add a wordcount step and terminate")
       .withActionOnFailure("CONTINUE")
       .withHadoopJarStep(flinkWordCountConf);

     stepConfigs.add(flinkRunWordCountStep);

     Application flink = new Application().withName("Flink");

     RunJobFlowRequest request = new RunJobFlowRequest()
       .withName("flink-transient")
       .withReleaseLabel("emr-5.20.0")
       .withApplications(flink)
       .withServiceRole("EMR_DefaultRole")
       .withJobFlowRole("EMR_EC2_DefaultRole")
       .withLogUri("s3://path/to/my/logfiles")
       .withInstances(new JobFlowInstancesConfig()
           .withEc2KeyName("myEc2Key")
           .withEc2SubnetId("subnet-12ab3c45")
           .withInstanceCount(3)
           .withKeepJobFlowAliveWhenNoSteps(false)
           .withMasterInstanceType("m4.large")
           .withSlaveInstanceType("m4.large"))
       .withSteps(stepConfigs);

     RunJobFlowResult result = emr.runJobFlow(request);
  System.out.println("The cluster ID is " + result.toString());

 }
```

```
}
```

Example Amazon CLI

使用 `create-cluster` 子命令创建一个临时集群，该集群在 Flink 作业完成时终止：

```
aws emr create-cluster --release-label emr-5.2.1 \
--name "Flink_Transient" \
--applications Name=Flink \
--configurations file://./configurations.json \
--region us-east-1 \
--log-uri s3://myLogUri \
--auto-terminate
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole_V2 \
--ec2-attributes KeyName=<YourKeyName>,InstanceProfile=EMR_EC2_DefaultRole \
--steps Type=CUSTOM_JAR,Jar=command-runner.jar,Name=Flink_Long_Running_Session,\
Args="bash","-c","\"flink run -m yarn-cluster /usr/lib/flink/examples/streaming/
WordCount.jar
--input s3://myBucket/pg11.txt --output s3://myBucket/alice/""
```

使用 Scala Shell

适用于 EMR 集群的 Flink Scala Shell 仅配置为启动新的 YARN 会话。您可以通过以下过程使用 Scala Shell。

在主节点上使用 Flink Scala Shell

1. 使用 SSH 登录主节点，如 Connect to the primary node with SSH 中所述。
2. 键入以下命令启动 Shell：

   在 Amazon EMR 5.5.0 及更高版本中，您可以借助一个 TaskManager 来使用以下命令启动 Scala Shell Yarn 集群。

   ```
   % flink-scala-shell yarn 1
   ```

   在 Amazon EMR 的更早版本中，使用：

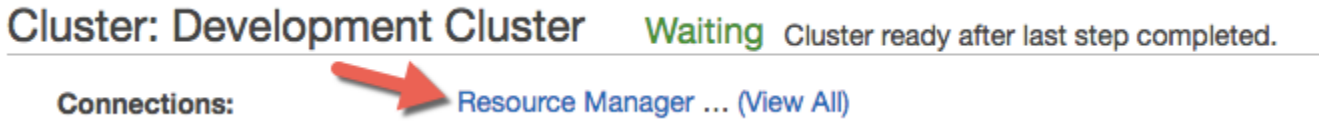
   ```
   % /usr/lib/flink/bin/start-scala-shell.sh yarn 1
   ```

这将启动 Flink Scala Shell，以便您能以交互方式使用 Flink。与使用其它接口和选项一样，您可以基于要从 Shell 运行的任务数来缩放示例中使用的 -n 选项值。

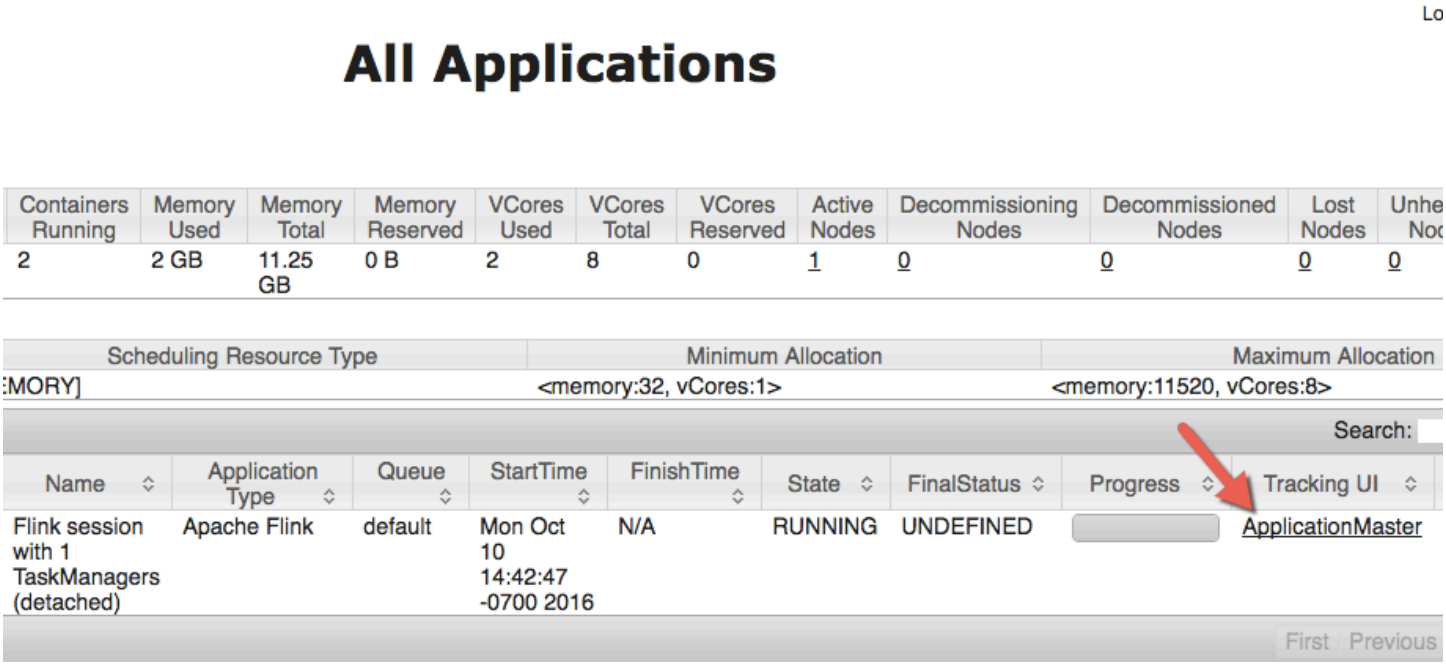
有关更多信息，请参阅官方 Apache Flink 文档中的 Scala REPL。

查找 Flink Web 界面

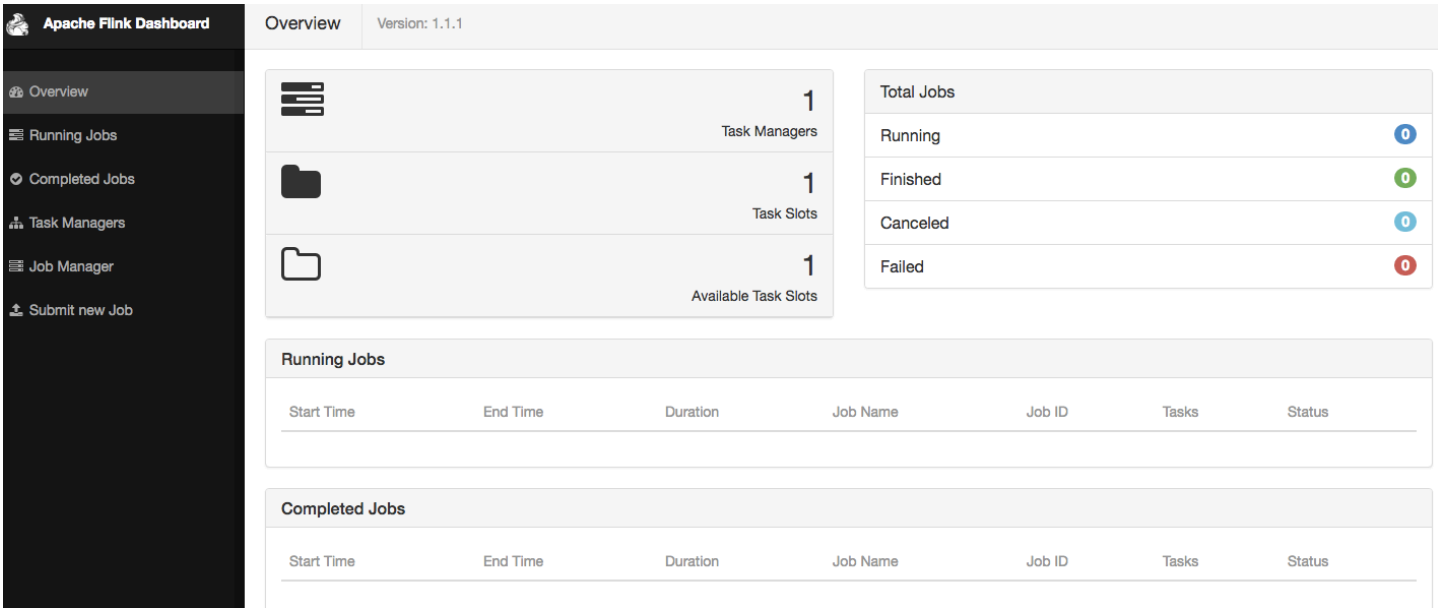
属于 Flink 应用程序的 Application Master 托管 Flink Web 界面。这是将 JAR 作为作业提交或查看其他作业当前状态的另一种方式。只要有 Flink 会话在运行，Flink Web 界面就处于活动状态。如果您有长时间运行的 YARN 作业并且已处于活动状态，则可以按照 Amazon EMR Management Guide 中的 Connect to the primary node with SSH 主题中的说明，来连接到 YARN ResourceManager。例如，如果您已设置 SSH 隧道并且已在浏览器中激活代理，则可在 EMR 集群详细信息页面中的 Connections (连接) 下选择 ResourceManager 连接。



在找到 ResourceManager 后，选择正在托管 Flink 会话的 YARN 应用程序。选择 Tracking UI (跟踪 UI) 列下的链接。



在 Flink Web 界面中，您可以查看配置，将自己的自定义 JAR 作为作业提交或监视正在进行的作业。



在 Amazon EMR 中通过 Zeppelin 使用 Flink 作业

简介

Amazon EMR 发布了 6.10.0 及更高版本，支持与 Apache Flink 的 Apache Zeppelin 集成。您可以通过 Zeppelin 笔记本以交互方式提交 Flink 作业。使用 Flink 解释器，您可以执行 Flink 查询、定义 Flink 流媒体和批处理作业，以及在 Zeppelin 笔记本中可视化输出。Flink 解释器基于 Flink REST API 构建。这使您可以从 Zeppelin 环境中访问和操作 Flink 作业，以执行实时数据处理和分析。

Flink 解释器中有四个子解释器。它们的用途不同，但都在 JVM 中，与 Flink 共享相同的预配置入口点（ExecutionEnviroment、StreamExecutionEnvironment、BatchTableEnvironment、StreamTa
解释器如下：

- %flink – 创建 ExecutionEnvironment、StreamExecutionEnvironment、BatchTableEnvironment、StreamT 并提供 Scala 环境
- %flink.pyflink – 提供一个 Python 环境
- %flink.ssql – 提供流式 SQL 环境
- %flink.bsql – 提供批处理 SQL 环境

先决条件

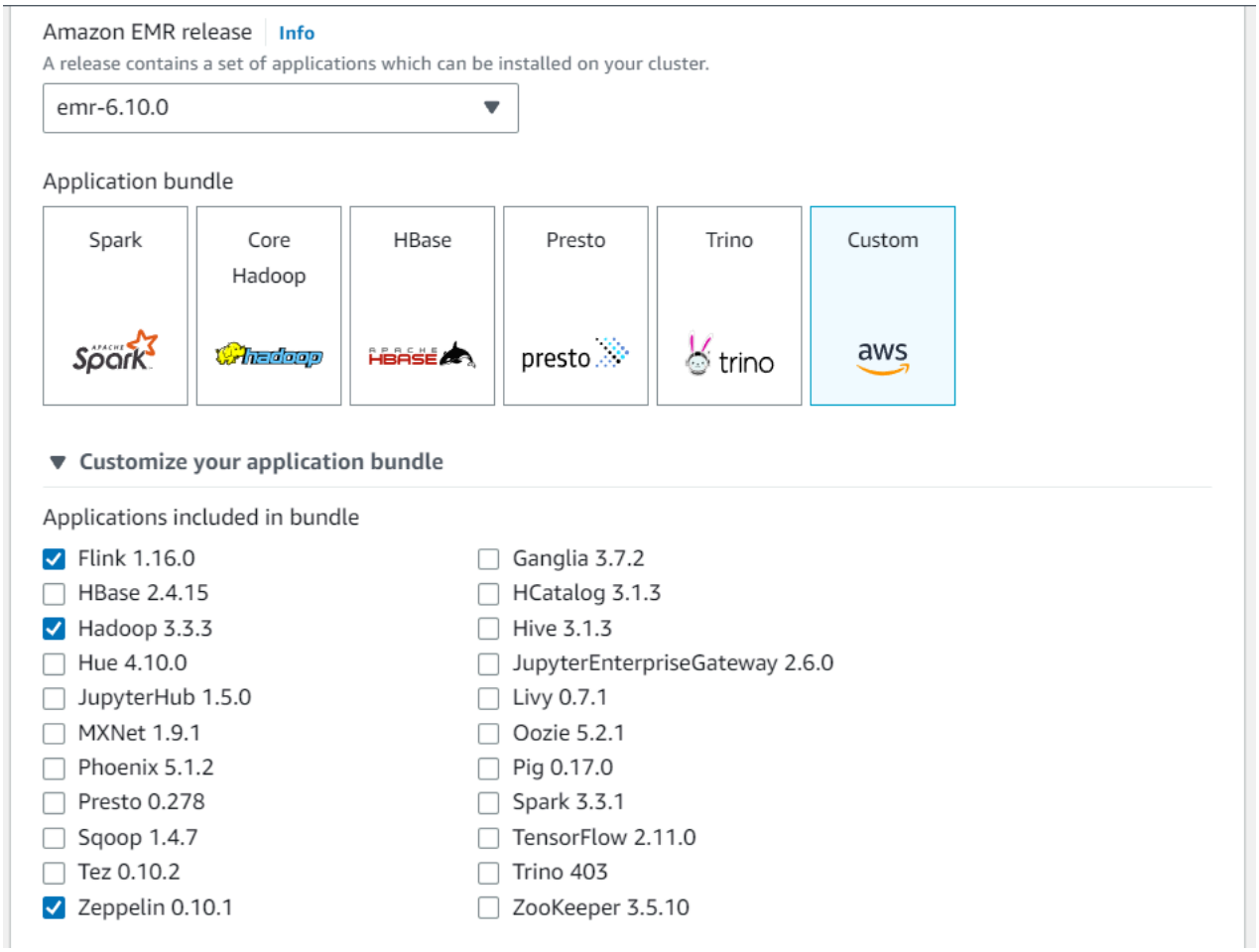
- 使用 Amazon EMR 6.10.0 及更高版本创建的集群支持 Zeppelin 与 Flink 集成。

- 要根据这些步骤的要求查看 EMR 集群上托管的 Web 界面，必须配置 SSH 隧道以允许入站访问。有关更多信息，请参阅 Configure proxy settings to view websites hosted on the primary node。

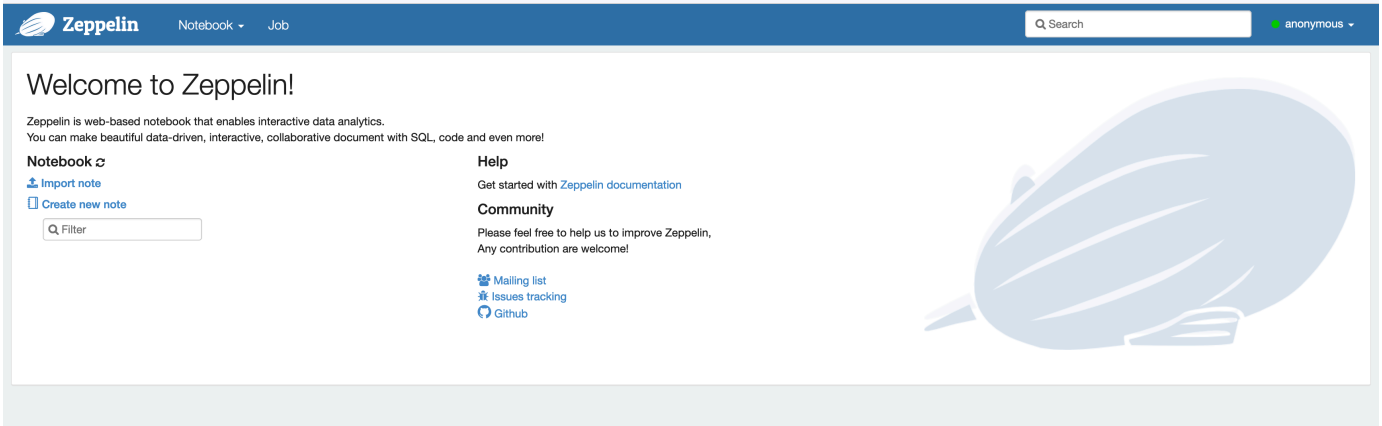
在 EMR 集群上配置 Zeppelin-Flink

使用以下步骤将 Apache Zeppelin 上的 Apache Flink 配置为在 EMR 集群上运行：

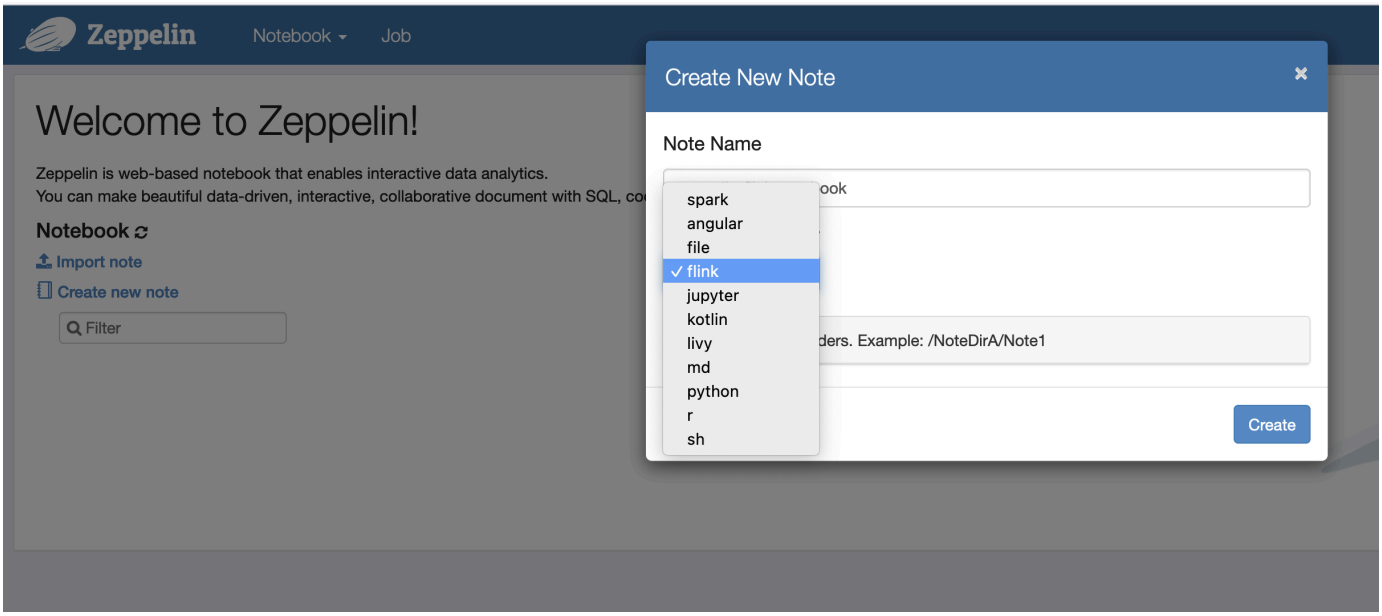
1. 从 Amazon EMR 控制台创建新集群。为 Amazon EMR 版本选择 emr-6.10.0 或更高版本。然后，选择使用“自定义”选项自定义您的应用程序捆绑包。在您的捆绑包中至少包含 Flink、Hadoop 和 Zeppelin。



2. 使用您首选的设置创建集群的其余部分。
3. 一旦集群开始运行，在控制台中选择集群以查看其详细信息并打开“应用程序”选项卡。从“应用程序”用户界面部分选择“Zeppelin”，以打开 Zeppelin 网页界面。请确保您已设置了对 Zeppelin Web 界面的访问，包含连接到主节点的 SSH 隧道和代理连接，如 先决条件 中所述。



4. 现在，您可以使用 Flink 作为默认解释器在 Zeppelin 笔记本中创建新笔记。



5. 请参阅以下代码示例，这些示例演示了如何从 Zeppelin 笔记本运行 Flink 作业。

在 EMR 集群上使用 Zeppelin-Flink 运行 Flink 作业

- 示例 1，Flink Scala

  a) 批处理字数示例（SCALA）

```
%flink

val data = benv.fromElements("hello world", "hello flink", "hello hadoop")
data.flatMap(line => line.split("\\s"))
            .map(w => (w, 1))
```

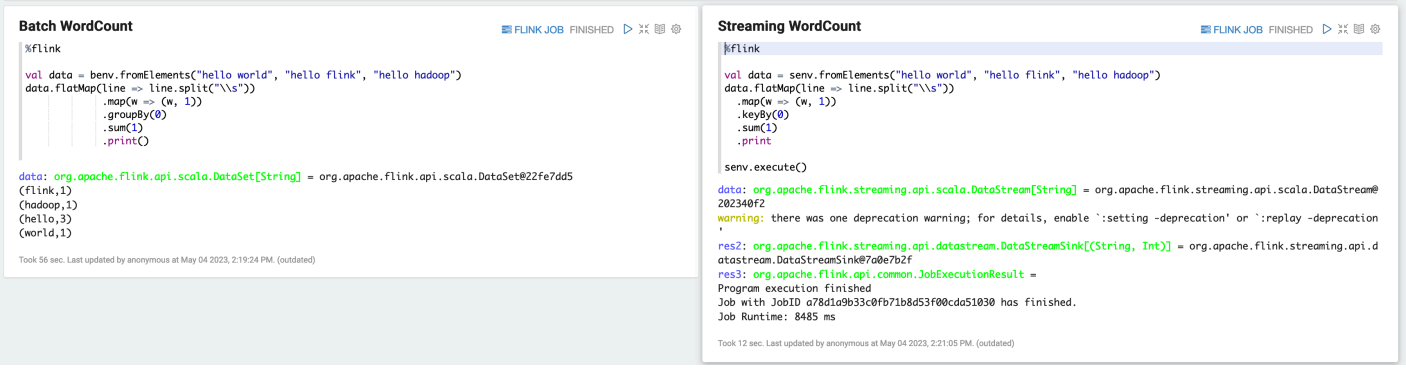
```
            .groupBy(0)
            .sum(1)
            .print()
```

b) 流式传输字数示例（SCALA）

```
%flink

val data = senv.fromElements("hello world", "hello flink", "hello hadoop")
data.flatMap(line => line.split("\\s"))
  .map(w => (w, 1))
  .keyBy(0)
  .sum(1)
  .print

senv.execute()
```



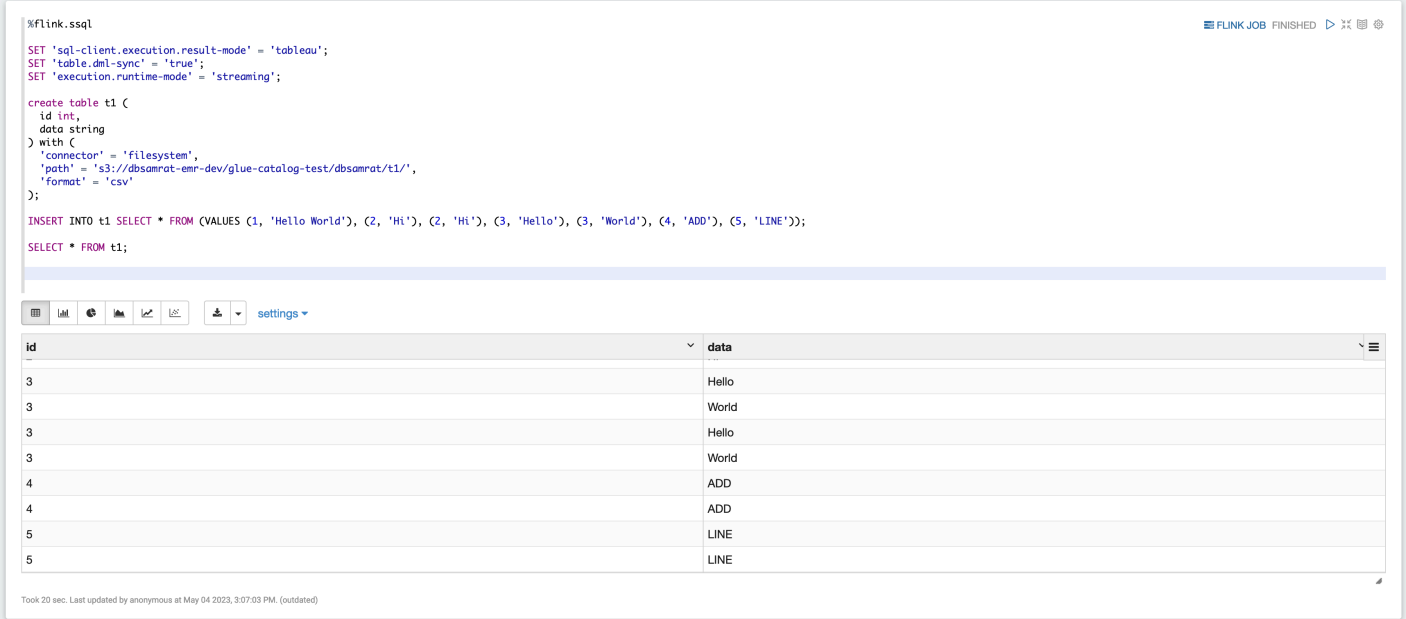
- 示例 2，Flink 流式传输 SQL

```
%flink.ssql
SET 'sql-client.execution.result-mode' = 'tableau';
SET 'table.dml-sync' = 'true';
SET 'execution.runtime-mode' = 'streaming';

create table dummy_table (
  id int,
  data string
) with (
  'connector' = 'filesystem',
  'path' = 's3://<s3-bucket>/glue-catalog-test/dbsamrat/t1/',
  'format' = 'csv'
);
```

```
INSERT INTO dummy_table SELECT * FROM (VALUES (1, 'Hello World'), (2, 'Hi'), (2,
 'Hi'), (3, 'Hello'), (3, 'World'), (4, 'ADD'), (5, 'LINE'));

SELECT * FROM dummy_table;
```



- 示例 3，Pyflink

```
%flink.pyflink

import argparse
import logging
import sys

from pyflink.common import Row
from pyflink.table import (EnvironmentSettings, TableEnvironment, TableDescriptor,
 Schema,
                           DataTypes, FormatDescriptor)
from pyflink.table.expressions import lit, col
from pyflink.table.udf import udtf

def word_count(input_path, output_path):
    t_env = TableEnvironment.create(EnvironmentSettings.in_streaming_mode())
    # write all the data to one file
    t_env.get_config().set("parallelism.default", "1")

    # define the source
    if input_path is not None:
```

```
        t_env.create_temporary_table(
            'source',
            TableDescriptor.for_connector('filesystem')
                           .schema(Schema.new_builder()
                                   .column('word', DataTypes.STRING())
                                   .build())
                           .option('path', input_path)
                           .format('csv')
                           .build())
        tab = t_env.from_path('source')
    else:
        print("Executing word_count example with default input data set.")
        print("Use --input to specify file input.")
        tab = t_env.from_elements(map(lambda i: (i,), word_count_data),
                                  DataTypes.ROW([DataTypes.FIELD('line',
DataTypes.STRING())]))

    # define the sink
    if output_path is not None:
        t_env.create_temporary_table(
            'sink',
            TableDescriptor.for_connector('filesystem')
                           .schema(Schema.new_builder()
                                   .column('word', DataTypes.STRING())
                                   .column('count', DataTypes.BIGINT())
                                   .build())
                           .option('path', output_path)
                           .format(FormatDescriptor.for_format('canal-json')
                                   .build())
                           .build())
    else:
        print("Printing result to stdout. Use --output to specify output path.")
        t_env.create_temporary_table(
            'sink',
            TableDescriptor.for_connector('print')
                           .schema(Schema.new_builder()
                                   .column('word', DataTypes.STRING())
                                   .column('count', DataTypes.BIGINT())
                                   .build())
                           .build())

    @udtf(result_types=[DataTypes.STRING()])
    def split(line: Row):
        for s in line[0].split():
```

```
            yield Row(s)

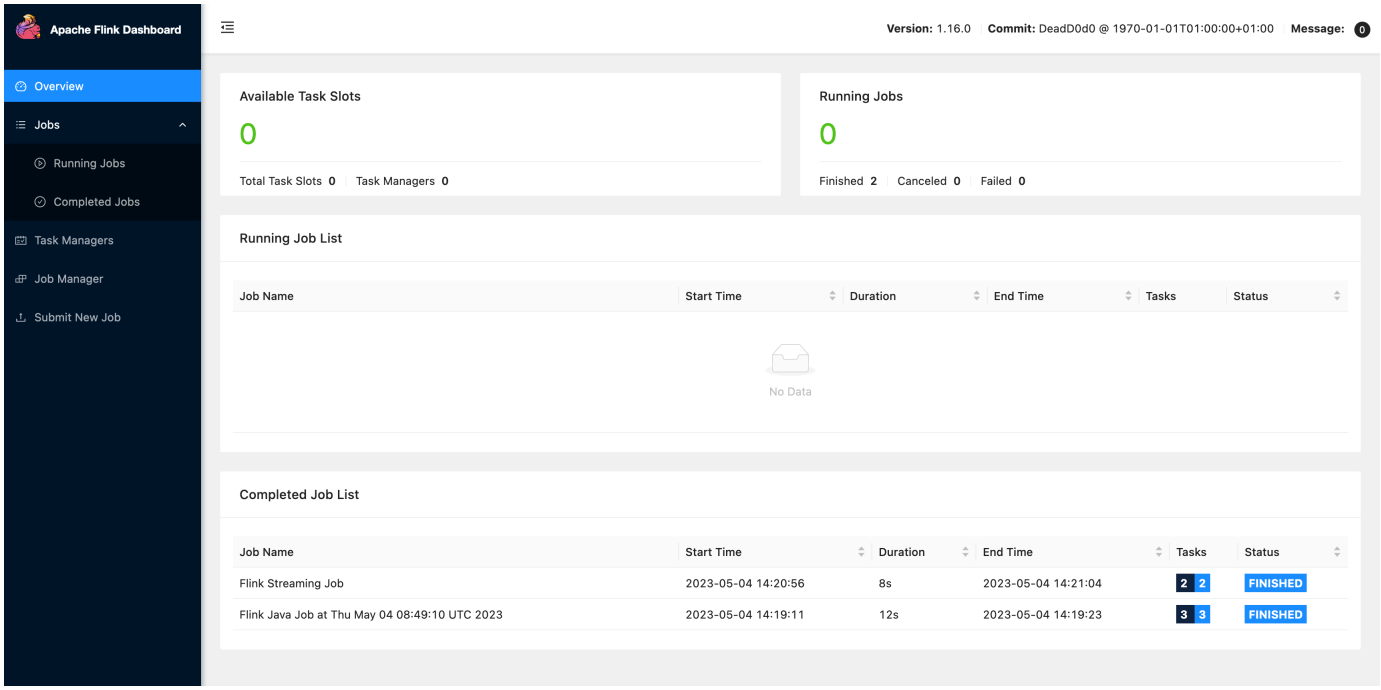
    # compute word count
    tab.flat_map(split).alias('word') \
       .group_by(col('word')) \
       .select(col('word'), lit(1).count) \
       .execute_insert('sink') \
       .wait()


logging.basicConfig(stream=sys.stdout, level=logging.INFO, format="%(message)s")


word_count("s3://<s3_bucket>/word.txt", "s3://<s3_bucket>/demo_output.txt")
```

1. 在 Zeppelin 用户界面中选择 FLINK 作业即可访问和查看 Flink Web 用户界面。

2. 在浏览器的另一个选项卡中选择 FLINK 作业，会路由到 Flink Web 控制台。



Flink 发布历史记录

下表列出了 Amazon EMR 每个发行版本中包含的 Flink 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Flink 版本信息

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.14.0	1.17.1-amzn-0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.13.0	1.17.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors
emr-6.12.0	1.17.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors
emr-6.11.1	1.16.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.11.0	1.16.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi, delta-standalone-connectors
emr-6.10.1	1.16.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.10.0	1.16.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.9.1	1.15.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.9.0	1.15.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.8.1	1.15.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.8.0	1.15.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.7.0	1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-5.36.1	1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.36.0	1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-6.6.0	1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-5.35.0	1.14.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.5.0	1.14.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.4.0	1.13.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config, hudi
emr-6.3.1	1.12.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.3.0	1.12.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-6.2.1	1.11.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-6.2.0	1.11.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-6.1.1	1.11.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-6.1.0	1.11.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.34.0	1.13.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.33.1	1.12.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-5.33.0	1.12.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-5.32.1	1.11.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.32.0	1.11.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-5.31.1	1.11.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config
emr-5.31.0	1.11.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client, flink-jobmanager-config

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.30.2	1.10.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.30.1	1.10.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.30.0	1.10.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.29.0	1.9.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.28.1	1.9.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.28.0	1.9.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.27.1	1.8.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.27.0	1.8.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.26.0	1.8.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.25.0	1.8.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.24.1	1.8.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.24.0	1.8.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.23.1	1.7.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.23.0	1.7.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.22.0	1.7.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.21.2	1.7.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.21.1	1.7.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.21.0	1.7.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.20.1	1.6.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.20.0	1.6.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.19.1	1.6.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.19.0	1.6.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.18.1	1.6.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.18.0	1.6.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.17.2	1.5.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.17.1	1.5.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.17.0	1.5.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.16.1	1.5.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.16.0	1.5.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.15.1	1.4.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.15.0	1.4.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.14.2	1.4.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.14.1	1.4.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.14.0	1.4.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.13.1	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.13.0	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.12.3	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.12.2	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.12.1	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.12.0	1.4.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.11.4	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.11.3	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.11.2	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.11.1	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.11.0	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.10.1	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.10.0	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.9.1	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.9.0	1.3.2	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.8.3	1.3.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.8.2	1.3.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.8.1	1.3.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.8.0	1.3.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.7.1	1.3.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.7.0	1.3.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.6.1	1.2.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client
emr-5.6.0	1.2.1	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.5.4	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.5.3	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.5.2	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.5.1	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.5.0	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.4.1	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.4.0	1.2.0	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.3.2	flink-client	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.3.1	flink-client	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.3.0	flink-client	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.2.3	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.2.2	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client

Amazon EMR 发行版标签	Flink 版本	随 Flink 安装的组件
emr-5.2.1	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.2.0	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.1.1	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client
emr-5.1.0	1.1.3	emrfs, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, flink-client

Ganglia

Ganglia 开源项目是一个可扩展的分布式系统，旨在监控集群和网格，同时尽量减少对其性能的影响。当您在集群上启用 Ganglia 时，您可以生成报告并查看整个集群的性能，还可以检查单个节点实例的性能。还配置 Ganglia 以提取和可视化 Hadoop 和 Spark 指标。有关 Ganglia 开源项目的更多信息，请转到 http://ganglia.info/。

当您在浏览器中查看 Ganglia Web UI 时，可以看到集群的性能概览，通过图形详细介绍了负载、内存使用率、CPU 使用率和集群的网络流量。位于集群统计数据下方的是集群中每个单独服务器的图形。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Ganglia 的版本，以及 Amazon EMR 随 Ganglia 一起安装的组件。

有关此发行版中随 Ganglia 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Ganglia 版本信息

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.14.0	Ganglia 3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Ganglia 的版本，以及 Amazon EMR 随 Ganglia 一起安装的组件。

有关此发行版中随 Ganglia 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Ganglia 版本信息

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.36.1	Ganglia 3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

主题

使用 Ganglia 创建集群

使用控制台通过 Ganglia 创建集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 选择 Create cluster (创建集群)。
3. 在 Software configuration (软件配置) 中，选择 All Applications (所有应用程序)、Core Hadoop (核心 Hadoop) 或者 Spark。
4. 根据需要继续利用配置创建集群。

使用 Amazon CLI 向集群添加 Ganglia

在 Amazon CLI 中，可以使用带有 `--applications` 参数的 `create-cluster` 向集群添加 Ganglia。如果使用 `--applications` 参数仅指定 Ganglia，则 Ganglia 是唯一安装的应用程序。

- 键入以下命令以在创建集群时添加 Ganglia，将 *myKey* 替换为您的 EC2 密钥对的名称。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\)。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --name "Spark cluster with Ganglia" --release-
label emr-5.36.1 \
--applications Name=Spark Name=Ganglia \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge \
--instance-count 3 --use-default-roles
```

如果不使用 `--instance-groups` 参数指定实例计数，则将启动单个主节点，其余实例将作为核心节点启动。所有节点都使用该命令中指定的实例类型。

> **Note**
>
> 如果您之前未创建默认 EMR 服务角色和 EC2 实例配置文件，请先键入 `aws emr create-default-roles` 创建它们，然后再键入 `create-cluster` 子命令。

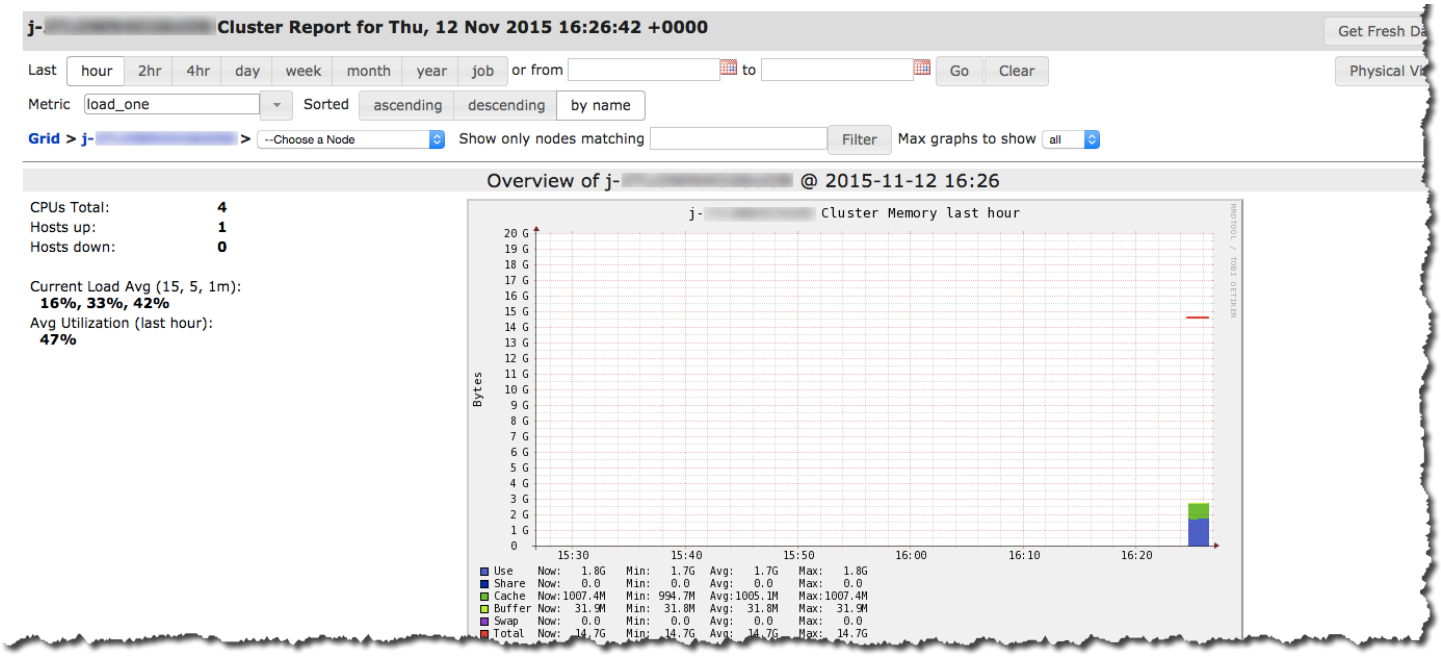
有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅https://docs.amazonaws.cn/cli/latest/reference/emr。

查看 Ganglia 指标

Ganglia 提供基于 Web 的用户界面，您可以使用该界面查看 Ganglia 收集的指标。当您在 Amazon EMR 上运行 Ganglia 时，Web 界面会在主节点上运行，并可以使用端口转发（也称为创建 SSH 隧道）进行查看。有关在 Amazon EMR 上查看 Web 界面的更多信息，请参阅《Amazon EMR 管理指南》中的查看 Amazon EMR 集群上托管的 Web 界面。

查看 Ganglia Web 界面

1. 使用 SSH 隧道进入主节点并创建安全连接。有关如何创建到主节点的 SSH 隧道的信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 1 部分：使用动态端口转发设置到主节点的 SSH 隧道。
2. 使用代理工具（如 Firefox 的 FoxyProxy 插件）安装 Web 浏览器，为 *ec2*.amazonaws.com* 类型的域创建 SOCKS 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 2 部分：配置代理设置以查看主节点上托管的网站。
3. 通过设置代理并打开 SSH 连接，您可以打开浏览器窗口，通过 http://*master-public-dns-name*/ganglia/ 查看 Ganglia UI，其中 *master-public-dns-name* 是 EMR 集群中主服务器的公有 DNS 地址。



Ganglia 中的 Hadoop 和 Spark 指标

Ganglia 报告每个实例的 Hadoop 指标。各种类型的指标按类别作为前缀：分布式文件系统（dfs.*）、Java 虚拟机（jvm.*）、MapReduce（jvm.*）和远程过程调用（rpc.*）。

基于 YARN 的 Ganglia 指标（如 Spark 和 Hadoop）对于 EMR 发行版 4.4.0 和 4.5.0 不可用。利用更高版本来使用这些指标。

Ganglia 中的 Spark 指标通常具有 YARN 应用程序 ID 和 Spark DAGScheduler 的前缀。前缀遵循以下形式：

- DAGScheduler.*
- application_xxxxxxxxxx_xxxx.driver.*
- application_xxxxxxxxxx_xxxx.executor.*

Ganglia 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Ganglia 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Ganglia 版本信息

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.14.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.13.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.12.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.11.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.11.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.10.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.10.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.9.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.9.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.8.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.8.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.7.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.36.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.36.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.6.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.35.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.5.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.4.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.3.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.3.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.2.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.2.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.1.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-6.1.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.0.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-6.0.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.34.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.33.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.33.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.32.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.32.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.31.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.31.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.30.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.30.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.30.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.29.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.28.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.28.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.27.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.27.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.26.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.25.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.24.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.24.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.23.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.23.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.22.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.21.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.21.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.21.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.20.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.20.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.19.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.19.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.18.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.18.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.17.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.17.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.17.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.16.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.16.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.15.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.15.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.14.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.14.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.14.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.13.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.13.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.12.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.12.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.12.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.12.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.11.4	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.11.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.11.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.11.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.11.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.10.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.10.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.9.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.9.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.8.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.8.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.8.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.8.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.7.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.7.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.6.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.6.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, webserver
emr-5.5.4	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.5.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.5.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.5.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.5.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.4.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.4.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.3.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.3.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.3.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.2.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.2.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.2.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.2.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.1.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.1.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-5.0.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-5.0.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.9.6	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.9.5	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.9.4	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.9.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.9.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.9.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.8.5	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.8.4	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.8.3	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.8.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.8.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.7.4	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.7.2	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.7.1	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.7.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.6.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.5.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Amazon EMR 发行版标签	Ganglia 版本	随 Ganglia 安装的组件
emr-4.4.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.3.0	3.7.2	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver
emr-4.2.0	3.6.0	emrfs, emr-goodies, ganglia-monitor, ganglia-metadata-collector, ganglia-web, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, webserver

Apache Hadoop

Apache Hadoop 是一种开源 Java 软件框架，支持跨越实例集群处理大量数据。它可以在单个实例或数千个实例上运行。Hadoop 使用各种处理模型（如 MapReduce 和 Tez）在多个实例之间分发处理，还使用名为 HDFS 的分布式文件系统在多个实例之间存储数据。Hadoop 监控集群中实例的运行状况，并可从一个或多个节点的故障中恢复。通过这种方式，Hadoop 可增加处理和存储容量以及高可用性。有关更多信息，请参阅 Hadoop 文档。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Hadoop 的版本，以及 Amazon EMR 随 Hadoop 一起安装的组件。

有关此发行版中随 Hadoop 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Hadoop 版本信息

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.14.0	Hadoop 3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Hadoop 的版本，以及 Amazon EMR 随 Hadoop 一起安装的组件。

有关此发行版中随 Hadoop 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Hadoop 版本信息

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.36.1	Hadoop 2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发行版 5.18.0 开始）附带的准确版本的库和依赖项的任务代码。有关更多信息，请参阅使用 Amazon EMR 项目存储库检查依赖项。

主题

配置 Hadoop

下列各节提供了 Hadoop 守护程序、任务和 HDFS 的默认配置设置。

主题

任务配置

您可以设置配置变量来优化 MapReduce 任务的性能。本节提供了一些重要设置的默认值。这些默认值因集群中使用的节点的 EC2 实例类型而异。在使用 Amazon EMR 发行版 4.6.0 和更高版本时，HBase 可用。安装 HBase 时使用不同的默认值。这些值将与初始默认值一起提供。

Hadoop 2 使用两个参数（`mapreduce.map.java.opts` 和 `mapreduce.reduce.java.opts`）分别配置用于映射和缩减 JVM 的内存。它们替代之前的 Hadoop 版本中的单个 `mapreduce.map.java.opts` 配置选项。

同样，在 Hadoop 2.7.2 及更高版本中，`mapred.job.jvm.num.tasks` 替换 `mapred.job.reuse.jvm.num.tasks`。Amazon EMR 将此值设置为 20，无论 EC2 实例类型如何。您可以使用 `mapred-site` 配置分类覆盖此设置。设置值 `-1` 表示可对单个作业内无线数量的任务重新使用 JVM，1 的值表示为每个任务生成一个新的 JVM。

例如，要将 `mapred.job.jvm.num.tasks` 的值设置为 -1，您可以创建一个包含以下内容的文件：

```
[
    {
      "Classification": "mapred-site",
      "Properties": {
        "mapred.job.jvm.num.tasks": "-1"
      }
    }
  ]
```

然后，当您通过 Amazon CLI 使用命令 `create-cluster` 或 `modify-instance-groups` 时，可以引用 JSON 配置文件。在以下示例中，该配置文件将另存为 `myConfig.json` 并存储在 Amazon S3 中。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge \
--instance-count 3 --applications Name=Hadoop --configurations https://
s3.amazonaws.com/mybucket/myfolder/myConfig.json \
--use-default-roles
```

您可以通过相同方式使用 `mapred-site` 配置分类更改下面列出的默认值，并使用一个 JSON 文件设置多个值和多个配置分类。有关更多信息，请参阅配置应用程序。

对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

任务配置设置的默认值

实例类型

- c1 实例
- c3 实例
- c4 实例
- c5 实例
- c5a 实例
- c5ad 实例
- c5d 实例
- c5n 实例
- c6a 实例
- c6g 实例
- c6gd 实例
- c6gn 实例
- c6i 实例
- c6id 实例
- c6in 实例
- c7g 实例
- c7gd 实例
- c7gn 实例
- c7i 实例
- d2 实例

- d3 实例
- d3en 实例
- g3 实例
- g3s 实例
- g4dn 实例
- g5 实例
- h1 实例
- i2 实例
- i3 实例
- i3en 实例
- i4g 实例
- i4i 实例
- im4gn 实例
- is4gen 实例
- m1 实例
- m2 实例
- m3 实例
- m4 实例
- m5 实例
- m5a 实例
- m5ad 实例
- m5d 实例
- m5dn 实例
- m5n 实例
- m5zn 实例
- m6a 实例
- m6g 实例
- m6gd 实例
- m6i 实例
- m6id 实例

- m6idn 实例
- m6in 实例
- m7a 实例
- m7g 实例
- m7gd 实例
- m7i 实例
- m7i-flex 实例
- p2 实例
- p3 实例
- p5 实例
- r3 实例
- r4 实例
- r5 实例
- r5a 实例
- r5ad 实例
- r5b 实例
- r5d 实例
- r5dn 实例
- r5n 实例
- r6a 实例
- r6g 实例
- r6gd 实例
- r6i 实例
- r6id 实例
- r6idn 实例
- r6in 实例
- r7a 实例
- r7g 实例
- r7gd 实例
- r7iz 实例

- x1 实例
- x1e 实例
- x2gd 实例
- x2idn 实例
- x2iedn 实例
- z1d 实例

c1 实例

c1.medium

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx288m	-Xmx288m
mapreduce.java.opts	-Xmx288m	-Xmx288m
mapreduce.map.memory.mb	512	512
mapreduce.reduce.memory.mb	512	512
yarn.app.mapreduce.am.resource.mb	512	512
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	512	512
yarn.nodemanager.resource.memory-mb	1024	512

c1.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx864m	-Xmx864m

配置选项	默认值	安装了 HBase
mapreduce.java.opts	-Xmx1536m	-Xmx1536m
mapreduce.map.memory.mb	1024	1024
mapreduce.reduce.memory.mb	2048	2048
yarn.app.mapreduce.am.resource.mb	2048	2048
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	2048	2560
yarn.nodemanager.resource.memory-mb	5120	2560

c3 实例

c3.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32

配置选项	默认值	安装了 HBase
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11520	5760
yarn.nodemanager.resource.memory-mb	11520	5760

c3.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m

配置选项	默认值	安装了 HBase
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23040	11520
yarn.nodemanager.resource.memory-mb	23040	11520

c3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1331m	-Xmx1331m
mapreduce.java.opts	-Xmx2662m	-Xmx2662m
mapreduce.map.memory.mb	1664	1664
mapreduce.reduce.memory.mb	3328	3328
yarn.app.mapreduce.am.resource.mb	3328	3328
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	53248	26624

配置选项	默认值	安装了 HBase
yarn.nodemanager.resource.memory-mb	53248	26624

c4 实例

c4.large

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx717m	-Xmx717m
mapreduce.java.opts	-Xmx1434m	-Xmx1434m
mapreduce.map.memory.mb	896	896
mapreduce.reduce.memory.mb	1792	1792
yarn.app.mapreduce.am.resource.mb	1792	1792
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1792	896
yarn.nodemanager.resource.memory-mb	1792	896

c4.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m

配置选项	默认值	安装了 HBase
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c4.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11520	5760

配置选项	默认值	安装了 HBase
yarn.nodemanager.resource.memory-mb	11520	5760

c4.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23040	11520
yarn.nodemanager.resource.memory-mb	23040	11520

c4.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1183m	-Xmx1183m
mapreduce.java.opts	-Xmx2366m	-Xmx2366m
mapreduce.map.memory.mb	1479	1479

配置选项	默认值	安装了 HBase
mapreduce.reduce.memory.mb	2958	2958
yarn.app.mapreduce.am.resource.mb	2958	2958
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	53248	26624
yarn.nodemanager.resource.memory-mb	53248	26624

c5 实例

c5.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	6144	3072

配置选项	默认值	安装了 HBase
yarn.nodemanager.resource.memory-mb	6144	3072

c5.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

c5.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536

配置选项	默认值	安装了 HBase
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

c5.9xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1456m	-Xmx1456m
mapreduce.java.opts	-Xmx2912m	-Xmx2912m
mapreduce.map.memory.mb	1820	1820
mapreduce.reduce.memory.mb	3640	3640
yarn.app.mapreduce.am.resource.mb	3640	3640
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	65536	32768
yarn.nodemanager.resource.memory-mb	65536	32768

c5.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1502m	-Xmx1502m
mapreduce.java.opts	-Xmx3004m	-Xmx3004m
mapreduce.map.memory.mb	1877	1877
mapreduce.reduce.memory.mb	3754	3754
yarn.app.mapreduce.am.resource.mb	3754	3754
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30048
yarn.nodemanager.resource.memory-mb	90112	30048

c5.18xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1547m	-Xmx1547m
mapreduce.java.opts	-Xmx3094m	-Xmx3094m
mapreduce.map.memory.mb	1934	1934
mapreduce.reduce.memory.mb	3868	3868
yarn.app.mapreduce.am.resource.mb	3868	3868

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	139264	30960
yarn.nodemanager.resource.memory-mb	139264	30960

c5.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1570m	-Xmx1570m
mapreduce.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.map.memory.mb	1963	1963
mapreduce.reduce.memory.mb	3926	3926
yarn.app.mapreduce.am.resource.mb	3926	3926
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31376
yarn.nodemanager.resource.memory-mb	188416	31376

c5a 实例

c5a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c5a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c5a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c5a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c5a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1502m	-Xmx1502m
mapreduce.java.opts	-Xmx3004m	-Xmx3004m
mapreduce.map.memory.mb	1877	1877
mapreduce.reduce.memory.mb	3754	3754
yarn.app.mapreduce.am.resource.mb	3754	3754

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30048
yarn.nodemanager.resource.memory-mb	90112	30048

c5a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c5a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c5ad 实例

c5ad.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c5ad.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c5ad.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c5ad.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c5ad.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c5ad.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c5ad.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c5d 实例

c5d.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	6144	3072
yarn.nodemanager.resource.memory-mb	6144	3072

c5d.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

c5d.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

c5d.9xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1456m	-Xmx1456m
mapreduce.java.opts	-Xmx2912m	-Xmx2912m
mapreduce.map.memory.mb	1820	1820
mapreduce.reduce.memory.mb	3640	3640
yarn.app.mapreduce.am.resource.mb	3640	3640
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	65536	32768
yarn.nodemanager.resource.memory-mb	65536	32768

c5d.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1502m	-Xmx1502m
mapreduce.java.opts	-Xmx3004m	-Xmx3004m
mapreduce.map.memory.mb	1877	1877
mapreduce.reduce.memory.mb	3754	3754
yarn.app.mapreduce.am.resource.mb	3754	3754
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30048
yarn.nodemanager.resource.memory-mb	90112	30048

c5d.18xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1547m	-Xmx1547m
mapreduce.java.opts	-Xmx3094m	-Xmx3094m
mapreduce.map.memory.mb	1934	1934
mapreduce.reduce.memory.mb	3868	3868
yarn.app.mapreduce.am.resource.mb	3868	3868

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	139264	30960
yarn.nodemanager.resource.memory-mb	139264	30960

c5d.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1570m	-Xmx1570m
mapreduce.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.map.memory.mb	1963	1963
mapreduce.reduce.memory.mb	3926	3926
yarn.app.mapreduce.am.resource.mb	3926	3926
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31376
yarn.nodemanager.resource.memory-mb	188416	31376

c5n 实例

c5n.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1613m	-Xmx1613m
mapreduce.java.opts	-Xmx3226m	-Xmx3226m
mapreduce.map.memory.mb	2016	2016
mapreduce.reduce.memory.mb	4032	4032
yarn.app.mapreduce.am.resource.mb	4032	4032
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	8064	4032
yarn.nodemanager.resource.memory-mb	8064	4032

c5n.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1613m	-Xmx1613m
mapreduce.java.opts	-Xmx3226m	-Xmx3226m
mapreduce.map.memory.mb	2016	2016
mapreduce.reduce.memory.mb	4032	4032
yarn.app.mapreduce.am.resource.mb	4032	4032

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	16128	8064
yarn.nodemanager.resource.memory-mb	16128	8064

c5n.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1741m	-Xmx1741m
mapreduce.java.opts	-Xmx3482m	-Xmx3482m
mapreduce.map.memory.mb	2176	2176
mapreduce.reduce.memory.mb	4352	4352
yarn.app.mapreduce.am.resource.mb	4352	4352
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	34816	17408
yarn.nodemanager.resource.memory-mb	34816	17408

c5n.9xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2002m	-Xmx2002m
mapreduce.java.opts	-Xmx4004m	-Xmx4004m
mapreduce.map.memory.mb	2503	2503
mapreduce.reduce.memory.mb	5006	5006
yarn.app.mapreduce.am.resource.mb	5006	5006
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30040
yarn.nodemanager.resource.memory-mb	90112	30040

c5n.18xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2094m	-Xmx2094m
mapreduce.java.opts	-Xmx4188m	-Xmx4188m
mapreduce.map.memory.mb	2617	2617
mapreduce.reduce.memory.mb	5234	5234
yarn.app.mapreduce.am.resource.mb	5234	5234

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31396
yarn.nodemanager.resource.memory-mb	188416	31396

c6a 实例

c6a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c6a.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1510m	-Xmx1510m
mapreduce.java.opts	-Xmx3020m	-Xmx3020m
mapreduce.map.memory.mb	1888	1888
mapreduce.reduce.memory.mb	3776	3776
yarn.app.mapreduce.am.resource.mb	3776	3776
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

c6a.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1527m	-Xmx1527m
mapreduce.java.opts	-Xmx3054m	-Xmx3054m
mapreduce.map.memory.mb	1909	1909
mapreduce.reduce.memory.mb	3818	3818
yarn.app.mapreduce.am.resource.mb	3818	3818

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30608
yarn.nodemanager.resource.memory-mb	366592	30608

c6g 实例

c6g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6gd 实例

c6gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6gn 实例

c6gn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6gn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6gn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6gn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6gn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6gn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6i 实例

c6i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6i.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6i.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c6i.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1510m	-Xmx1510m
mapreduce.java.opts	-Xmx3020m	-Xmx3020m
mapreduce.map.memory.mb	1888	1888
mapreduce.reduce.memory.mb	3776	3776
yarn.app.mapreduce.am.resource.mb	3776	3776
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

c6id 实例

c6id.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6id.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6id.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6id.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6id.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6id.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6id.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c6id.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1510m	-Xmx1510m
mapreduce.java.opts	-Xmx3020m	-Xmx3020m
mapreduce.map.memory.mb	1888	1888
mapreduce.reduce.memory.mb	3776	3776
yarn.app.mapreduce.am.resource.mb	3776	3776
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

c6in 实例

c6in.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c6in.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c6in.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c6in.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c6in.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c6in.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c6in.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c6in.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1510m	-Xmx1510m
mapreduce.java.opts	-Xmx3020m	-Xmx3020m
mapreduce.map.memory.mb	1888	1888
mapreduce.reduce.memory.mb	3776	3776
yarn.app.mapreduce.am.resource.mb	3776	3776

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

c7g 实例

c7g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c7g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c7g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c7g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c7g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c7g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c7gd 实例

c7gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c7gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c7gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c7gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c7gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c7gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c7gn 实例

c7gn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c7gn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c7gn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c7gn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c7gn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c7gn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c7i 实例

c7i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1126m	-Xmx1126m
mapreduce.java.opts	-Xmx2252m	-Xmx2252m
mapreduce.map.memory.mb	1408	1408
mapreduce.reduce.memory.mb	2816	2816
yarn.app.mapreduce.am.resource.mb	2816	2816
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	5632	2816
yarn.nodemanager.resource.memory-mb	5632	2816

c7i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

c7i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1171m	-Xmx1171m
mapreduce.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.map.memory.mb	1464	1464
mapreduce.reduce.memory.mb	2928	2928
yarn.app.mapreduce.am.resource.mb	2928	2928

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

c7i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1357m	-Xmx1357m
mapreduce.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.map.memory.mb	1696	1696
mapreduce.reduce.memory.mb	3392	3392
yarn.app.mapreduce.am.resource.mb	3392	3392
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

c7i.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1425m	-Xmx1425m
mapreduce.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.map.memory.mb	1781	1781
mapreduce.reduce.memory.mb	3562	3562
yarn.app.mapreduce.am.resource.mb	3562	3562
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32074
yarn.nodemanager.resource.memory-mb	85504	32074

c7i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1459m	-Xmx1459m
mapreduce.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.map.memory.mb	1824	1824
mapreduce.reduce.memory.mb	3648	3648
yarn.app.mapreduce.am.resource.mb	3648	3648

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

c7i.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1494m	-Xmx1494m
mapreduce.java.opts	-Xmx2988m	-Xmx2988m
mapreduce.map.memory.mb	1867	1867
mapreduce.reduce.memory.mb	3734	3734
yarn.app.mapreduce.am.resource.mb	3734	3734
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29840
yarn.nodemanager.resource.memory-mb	179200	29840

c7i.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1527m	-Xmx1527m
mapreduce.java.opts	-Xmx3054m	-Xmx3054m
mapreduce.map.memory.mb	1909	1909
mapreduce.reduce.memory.mb	3818	3818
yarn.app.mapreduce.am.resource.mb	3818	3818
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30608
yarn.nodemanager.resource.memory-mb	366592	30608

d2 实例

d2.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

d2.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

d2.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

d2.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2417m	-Xmx2417m
mapreduce.java.opts	-Xmx4834m	-Xmx4834m
mapreduce.map.memory.mb	3021	3021
mapreduce.reduce.memory.mb	6042	6042
yarn.app.mapreduce.am.resource.mb	6042	6042

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30194
yarn.nodemanager.resource.memory-mb	241664	30194

d3 实例

d3.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

d3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

d3.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

d3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

d3en 实例

d3en.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

d3en.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

d3en.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

d3en.6xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.java.opts	-Xmx5700m	-Xmx5700m
mapreduce.map.memory.mb	3563	3563
mapreduce.reduce.memory.mb	7126	7126
yarn.app.mapreduce.am.resource.mb	7126	7126
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	28496
yarn.nodemanager.resource.memory-mb	85504	28496

d3en.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

d3en.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

g3 实例

g3.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

g3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

g3.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

g3s 实例

g3s.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

g4dn 实例

g4dn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144

配置选项	默认值	安装了 HBase
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

g4dn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

g4dn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2867m	-Xmx2867m
mapreduce.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.map.memory.mb	3584	3584
mapreduce.reduce.memory.mb	7168	7168
yarn.app.mapreduce.am.resource.mb	7168	7168
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

g4dn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

g4dn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.java.opts	-Xmx6280m	-Xmx6280m
mapreduce.map.memory.mb	3925	3925
mapreduce.reduce.memory.mb	7850	7850
yarn.app.mapreduce.am.resource.mb	7850	7850
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31416
yarn.nodemanager.resource.memory-mb	188416	31416

g4dn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.java.opts	-Xmx6348m	-Xmx6348m
mapreduce.map.memory.mb	3968	3968
mapreduce.reduce.memory.mb	7936	7936
yarn.app.mapreduce.am.resource.mb	7936	7936
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

g5 实例

g5.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

g5.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

g5.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

g5.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

g5.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

g5.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

g5.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

g5.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3089m	-Xmx3089m
mapreduce.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.map.memory.mb	3861	3861
mapreduce.reduce.memory.mb	7722	7722
yarn.app.mapreduce.am.resource.mb	7722	7722
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30952
yarn.nodemanager.resource.memory-mb	741376	30952

h1 实例

h1.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

h1.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2867m	-Xmx2867m
mapreduce.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.map.memory.mb	3584	3584
mapreduce.reduce.memory.mb	7168	7168
yarn.app.mapreduce.am.resource.mb	7168	7168

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

h1.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

h1.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.java.opts	-Xmx6348m	-Xmx6348m
mapreduce.map.memory.mb	3968	3968
mapreduce.reduce.memory.mb	7936	7936
yarn.app.mapreduce.am.resource.mb	7936	7936
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

i2 实例

i2.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

i2.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

i2.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

i2.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

i3 实例

i3.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

i3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

i3.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

i3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

i3.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

i3en 实例

i3en.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4915m	-Xmx4915m
mapreduce.java.opts	-Xmx9830m	-Xmx9830m
mapreduce.map.memory.mb	6144	6144
mapreduce.reduce.memory.mb	12288	12288
yarn.app.mapreduce.am.resource.mb	12288	12288

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

i3en.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.java.opts	-Xmx11468m	-Xmx11468m
mapreduce.map.memory.mb	7168	7168
mapreduce.reduce.memory.mb	14336	14336
yarn.app.mapreduce.am.resource.mb	14336	14336
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

i3en.3xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6007m	-Xmx6007m
mapreduce.java.opts	-Xmx12014m	-Xmx12014m
mapreduce.map.memory.mb	7509	7509
mapreduce.reduce.memory.mb	15018	15018
yarn.app.mapreduce.am.resource.mb	15018	15018
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30040
yarn.nodemanager.resource.memory-mb	90112	30040

i3en.6xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6281m	-Xmx6281m
mapreduce.java.opts	-Xmx12562m	-Xmx12562m
mapreduce.map.memory.mb	7851	7851
mapreduce.reduce.memory.mb	15702	15702
yarn.app.mapreduce.am.resource.mb	15702	15702

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31396
yarn.nodemanager.resource.memory-mb	188416	31396

i3en.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6417m	-Xmx6417m
mapreduce.java.opts	-Xmx12834m	-Xmx12834m
mapreduce.map.memory.mb	8021	8021
mapreduce.reduce.memory.mb	16042	16042
yarn.app.mapreduce.am.resource.mb	16042	16042
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32100
yarn.nodemanager.resource.memory-mb	385024	32100

i3en.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6486m	-Xmx6486m
mapreduce.java.opts	-Xmx12972m	-Xmx12972m
mapreduce.map.memory.mb	8107	8107
mapreduce.reduce.memory.mb	16214	16214
yarn.app.mapreduce.am.resource.mb	16214	16214
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	778240	32396
yarn.nodemanager.resource.memory-mb	778240	32396

i4g 实例

i4g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

i4g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

i4g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

i4g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

i4g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

i4i 实例

i4i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

i4i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

i4i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

i4i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

i4i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

i4i.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6195m	-Xmx6195m
mapreduce.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.map.memory.mb	7744	7744
mapreduce.reduce.memory.mb	15488	15488
yarn.app.mapreduce.am.resource.mb	15488	15488
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

im4gn 实例

im4gn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

im4gn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

im4gn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

im4gn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

im4gn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

is4gen 实例

is4gen.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3514m	-Xmx3514m
mapreduce.java.opts	-Xmx7028m	-Xmx7028m
mapreduce.map.memory.mb	4393	4393
mapreduce.reduce.memory.mb	8786	8786
yarn.app.mapreduce.am.resource.mb	8786	8786
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	17572.12	8786.06
yarn.nodemanager.resource.memory-mb	17572.12	8786.06

is4gen.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3866m	-Xmx3866m
mapreduce.java.opts	-Xmx7732m	-Xmx7732m
mapreduce.map.memory.mb	4832	4832
mapreduce.reduce.memory.mb	9664	9664
yarn.app.mapreduce.am.resource.mb	9664	9664
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	38656	19328
yarn.nodemanager.resource.memory-mb	38656	19328

is4gen.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4275m	-Xmx4275m
mapreduce.java.opts	-Xmx8550m	-Xmx8550m
mapreduce.map.memory.mb	5344	5344
mapreduce.reduce.memory.mb	10688	10688
yarn.app.mapreduce.am.resource.mb	10688	10688

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	32064
yarn.nodemanager.resource.memory-mb	85504	32064

is4gen.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4480m	-Xmx4480m
mapreduce.java.opts	-Xmx8960m	-Xmx8960m
mapreduce.map.memory.mb	5600	5600
mapreduce.reduce.memory.mb	11200	11200
yarn.app.mapreduce.am.resource.mb	11200	11200
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	22400
yarn.nodemanager.resource.memory-mb	179200	22400

m1 实例

m1.small

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx288m	-Xmx288m
mapreduce.java.opts	-Xmx288m	-Xmx288m
mapreduce.map.memory.mb	512	512
mapreduce.reduce.memory.mb	512	512
yarn.app.mapreduce.am.resource.mb	512	512
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	512	512
yarn.nodemanager.resource.memory-mb	1024	512

m1.medium

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx512m	-Xmx512m
mapreduce.java.opts	-Xmx768m	-Xmx768m
mapreduce.map.memory.mb	768	768
mapreduce.reduce.memory.mb	1024	1024
yarn.app.mapreduce.am.resource.mb	1024	1024

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	2048	1024
yarn.nodemanager.resource.memory-mb	2048	1024

m1.large

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx512m	-Xmx512m
mapreduce.java.opts	-Xmx1024m	-Xmx1024m
mapreduce.map.memory.mb	768	768
mapreduce.reduce.memory.mb	1536	1536
yarn.app.mapreduce.am.resource.mb	1536	1536
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	3072	2560
yarn.nodemanager.resource.memory-mb	5120	2560

m1.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx512m	-Xmx512m
mapreduce.java.opts	-Xmx1536m	-Xmx1536m
mapreduce.map.memory.mb	768	768
mapreduce.reduce.memory.mb	2048	2048
yarn.app.mapreduce.am.resource.mb	2048	2048
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	8192	6144
yarn.nodemanager.resource.memory-mb	12288	6144

m2 实例

m2.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx864m	-Xmx864m
mapreduce.java.opts	-Xmx1536m	-Xmx1536m
mapreduce.map.memory.mb	1024	1024
mapreduce.reduce.memory.mb	2048	2048
yarn.app.mapreduce.am.resource.mb	2048	2048

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	7168	7168
yarn.nodemanager.resource.memory-mb	14336	7168

m2.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1280m	-Xmx1280m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	2560	2560
yarn.app.mapreduce.am.resource.mb	2560	2560
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	8192	15360
yarn.nodemanager.resource.memory-mb	30720	15360

m2.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1280m	-Xmx1280m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	2560	2560
yarn.app.mapreduce.am.resource.mb	2560	2560
yarn.scheduler.minimum-allocation-mb	256	256
yarn.scheduler.maximum-allocation-mb	8192	30720
yarn.nodemanager.resource.memory-mb	61440	30720

m3 实例

m3.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11520	5760
yarn.nodemanager.resource.memory-mb	11520	5760

m3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1152m	-Xmx1152m
mapreduce.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.map.memory.mb	1440	1440
mapreduce.reduce.memory.mb	2880	2 880
yarn.app.mapreduce.am.resource.mb	2 880	2 880
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23040	11520
yarn.nodemanager.resource.memory-mb	23040	11520

m4 实例

m4.large

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	6144	3072
yarn.nodemanager.resource.memory-mb	6144	3072

m4.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

m4.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1229m	-Xmx1229m
mapreduce.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.map.memory.mb	1536	1536
mapreduce.reduce.memory.mb	3072	3072
yarn.app.mapreduce.am.resource.mb	3072	3072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

m4.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1434m	-Xmx1434m
mapreduce.java.opts	-Xmx2868m	-Xmx2868m
mapreduce.map.memory.mb	1792	1792
mapreduce.reduce.memory.mb	3584	3584
yarn.app.mapreduce.am.resource.mb	3584	3584
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

m4.10xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1557m	-Xmx1557m
mapreduce.java.opts	-Xmx3114m	-Xmx3114m
mapreduce.map.memory.mb	1946	1946
mapreduce.reduce.memory.mb	3892	3892
yarn.app.mapreduce.am.resource.mb	3892	3892

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	155648	31104
yarn.nodemanager.resource.memory-mb	155648	31104

m4.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx1587m	-Xmx1587m
mapreduce.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.map.memory.mb	1984	1984
mapreduce.reduce.memory.mb	3968	3968
yarn.app.mapreduce.am.resource.mb	3968	3968
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

m5 实例

m5.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

m5.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

m5.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2867m	-Xmx2867m
mapreduce.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.map.memory.mb	3584	3584
mapreduce.reduce.memory.mb	7168	7168
yarn.app.mapreduce.am.resource.mb	7168	7168
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

m5.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

m5.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.java.opts	-Xmx6280m	-Xmx6280m
mapreduce.map.memory.mb	3925	3925
mapreduce.reduce.memory.mb	7850	7850
yarn.app.mapreduce.am.resource.mb	7850	7850

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31416
yarn.nodemanager.resource.memory-mb	188416	31416

m5.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.java.opts	-Xmx6348m	-Xmx6348m
mapreduce.map.memory.mb	3968	3968
mapreduce.reduce.memory.mb	7936	7936
yarn.app.mapreduce.am.resource.mb	7936	7936
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

m5.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3209m	-Xmx3209m
mapreduce.java.opts	-Xmx6418m	-Xmx6418m
mapreduce.map.memory.mb	4011	4011
mapreduce.reduce.memory.mb	8022	8022
yarn.app.mapreduce.am.resource.mb	8022	8022
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32056
yarn.nodemanager.resource.memory-mb	385024	32056

m5a 实例

m5a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

m5a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

m5a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2867m	-Xmx2867m
mapreduce.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.map.memory.mb	3584	3584
mapreduce.reduce.memory.mb	7168	7168
yarn.app.mapreduce.am.resource.mb	7168	7168
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

m5a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

m5a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.java.opts	-Xmx6280m	-Xmx6280m
mapreduce.map.memory.mb	3925	3925
mapreduce.reduce.memory.mb	7850	7850
yarn.app.mapreduce.am.resource.mb	7850	7850
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31416
yarn.nodemanager.resource.memory-mb	188416	31416

m5a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.java.opts	-Xmx6348m	-Xmx6348m
mapreduce.map.memory.mb	3968	3968
mapreduce.reduce.memory.mb	7936	7936
yarn.app.mapreduce.am.resource.mb	7936	7936
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

m5a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3209m	-Xmx3209m
mapreduce.java.opts	-Xmx6418m	-Xmx6418m
mapreduce.map.memory.mb	4011	4011
mapreduce.reduce.memory.mb	8022	8022
yarn.app.mapreduce.am.resource.mb	8022	8022

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32056
yarn.nodemanager.resource.memory-mb	385024	32056

m5ad 实例

m5ad.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m5ad.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m5ad.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m5ad.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m5ad.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m5ad.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m5ad.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m5d 实例

m5d.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	12288	6144
yarn.nodemanager.resource.memory-mb	12288	6144

m5d.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2458m	-Xmx2458m
mapreduce.java.opts	-Xmx4916m	-Xmx4916m
mapreduce.map.memory.mb	3072	3072
mapreduce.reduce.memory.mb	6144	6144
yarn.app.mapreduce.am.resource.mb	6144	6144

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

m5d.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2867m	-Xmx2867m
mapreduce.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.map.memory.mb	3584	3584
mapreduce.reduce.memory.mb	7168	7168
yarn.app.mapreduce.am.resource.mb	7168	7168
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

m5d.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

m5d.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3140m	-Xmx3140m
mapreduce.java.opts	-Xmx6280m	-Xmx6280m
mapreduce.map.memory.mb	3925	3925
mapreduce.reduce.memory.mb	7850	7850
yarn.app.mapreduce.am.resource.mb	7850	7850

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31416
yarn.nodemanager.resource.memory-mb	188416	31416

m5d.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3174m	-Xmx3174m
mapreduce.java.opts	-Xmx6348m	-Xmx6348m
mapreduce.map.memory.mb	3968	3968
mapreduce.reduce.memory.mb	7936	7936
yarn.app.mapreduce.am.resource.mb	7936	7936
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

m5d.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3209m	-Xmx3209m
mapreduce.java.opts	-Xmx6418m	-Xmx6418m
mapreduce.map.memory.mb	4011	4011
mapreduce.reduce.memory.mb	8022	8022
yarn.app.mapreduce.am.resource.mb	8022	8022
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32056
yarn.nodemanager.resource.memory-mb	385024	32056

m5dn 实例

m5dn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m5dn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m5dn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m5dn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m5dn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m5dn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m5dn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m5n 实例

m5n.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m5n.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m5n.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m5n.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m5n.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m5n.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m5n.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m5zn 实例

m5zn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2304m	-Xmx2304m
mapreduce.java.opts	-Xmx4608m	-Xmx4608m
mapreduce.map.memory.mb	2880	2 880
mapreduce.reduce.memory.mb	5760	5760
yarn.app.mapreduce.am.resource.mb	5760	5760
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11520	5760
yarn.nodemanager.resource.memory-mb	11520	5760

m5zn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m5zn.3xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2577m	-Xmx2577m
mapreduce.java.opts	-Xmx5154m	-Xmx5154m
mapreduce.map.memory.mb	3221	3221
mapreduce.reduce.memory.mb	6442	6442
yarn.app.mapreduce.am.resource.mb	6442	6442
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	38656	19328
yarn.nodemanager.resource.memory-mb	38656	19328

m5zn.6xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2850m	-Xmx2850m
mapreduce.java.opts	-Xmx5700m	-Xmx5700m
mapreduce.map.memory.mb	3563	3563
mapreduce.reduce.memory.mb	7126	7126
yarn.app.mapreduce.am.resource.mb	7126	7126
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	85504	28496
yarn.nodemanager.resource.memory-mb	85504	28496

m5zn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m6a 实例

m6a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m6a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m6a.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

m6a.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3089m	-Xmx3089m
mapreduce.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.map.memory.mb	3861	3861
mapreduce.reduce.memory.mb	7722	7722
yarn.app.mapreduce.am.resource.mb	7722	7722

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30952
yarn.nodemanager.resource.memory-mb	741376	30952

m6g 实例

m6g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	181248	30208
yarn.nodemanager.resource.memory-mb	181248	30208

m6g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6gd 实例

m6gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	181248	30208
yarn.nodemanager.resource.memory-mb	181248	30208

m6gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6i 实例

m6i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6i.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	181248	30208
yarn.nodemanager.resource.memory-mb	181248	30208

m6i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6i.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m6i.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

m6id 实例

m6id.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6id.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6id.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6id.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6id.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m6id.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6id.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m6id.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

m6idn 实例

m6idn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6idn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6idn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6idn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6idn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m6idn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6idn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m6idn.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

m6in 实例

m6in.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m6in.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m6in.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m6in.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m6in.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	181248	30208
yarn.nodemanager.resource.memory-mb	181248	30208

m6in.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m6in.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m6in.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3072m	-Xmx3072m
mapreduce.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.map.memory.mb	3840	3840
mapreduce.reduce.memory.mb	7680	7680
yarn.app.mapreduce.am.resource.mb	7680	7680
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

m7a 实例

m7a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m7a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m7a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m7g 实例

m7g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m7g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m7g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m7g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m7g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m7g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m7gd 实例

m7gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m7gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m7gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m7gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m7gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2986m	-Xmx2986m
mapreduce.java.opts	-Xmx5972m	-Xmx5972m
mapreduce.map.memory.mb	3733	3733
mapreduce.reduce.memory.mb	7466	7466
yarn.app.mapreduce.am.resource.mb	7466	7466
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	179200	29880
yarn.nodemanager.resource.memory-mb	179200	29880

m7gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m7i 实例

m7i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m7i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m7i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m7i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

m7i.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	181248	30208
yarn.nodemanager.resource.memory-mb	181248	30208

m7i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

m7i.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3055m	-Xmx3055m
mapreduce.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.map.memory.mb	3819	3819
mapreduce.reduce.memory.mb	7638	7638
yarn.app.mapreduce.am.resource.mb	7638	7638

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30520
yarn.nodemanager.resource.memory-mb	366592	30520

m7i.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3089m	-Xmx3089m
mapreduce.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.map.memory.mb	3861	3861
mapreduce.reduce.memory.mb	7722	7722
yarn.app.mapreduce.am.resource.mb	7722	7722
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30952
yarn.nodemanager.resource.memory-mb	741376	30952

m7i-flex 实例

m7i-flex.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	11712	5856
yarn.nodemanager.resource.memory-mb	11712	5856

m7i-flex.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

m7i-flex.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

m7i-flex.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

p2 实例

p2.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.java.opts	-Xmx21708m	-Xmx21708m
mapreduce.map.memory.mb	13568	13568
mapreduce.reduce.memory.mb	27136	27136
yarn.app.mapreduce.am.resource.mb	27136	27136

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

p2.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.java.opts	-Xmx24576m	-Xmx24576m
mapreduce.map.memory.mb	15360	15360
mapreduce.reduce.memory.mb	30720	30720
yarn.app.mapreduce.am.resource.mb	30720	30720
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

p2.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx9267m	-Xmx9267m
mapreduce.java.opts	-Xmx18534m	-Xmx18534m
mapreduce.map.memory.mb	11584	11584
mapreduce.reduce.memory.mb	23168	23168
yarn.app.mapreduce.am.resource.mb	23168	23168
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	23168
yarn.nodemanager.resource.memory-mb	741376	23168

p3 实例

p3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

p3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

p3.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

p5 实例

p5.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx8294m	-Xmx8294m
mapreduce.java.opts	-Xmx16588m	-Xmx16588m
mapreduce.map.memory.mb	10368	10368
mapreduce.reduce.memory.mb	20736	20736
yarn.app.mapreduce.am.resource.mb	20736	20736

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1990656	20736
yarn.nodemanager.resource.memory-mb	1990656	20736

r3 实例

r3.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2342m	-Xmx2342m
mapreduce.java.opts	-Xmx4684m	-Xmx4684m
mapreduce.map.memory.mb	2928	2928
mapreduce.reduce.memory.mb	5856	5856
yarn.app.mapreduce.am.resource.mb	5856	5856
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r3.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2714m	-Xmx2714m
mapreduce.java.opts	-Xmx5428m	-Xmx5428m
mapreduce.map.memory.mb	3392	3392
mapreduce.reduce.memory.mb	6784	6784
yarn.app.mapreduce.am.resource.mb	6784	6784
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r3.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx2918m	-Xmx2918m
mapreduce.java.opts	-Xmx5836m	-Xmx5836m
mapreduce.map.memory.mb	3648	3648
mapreduce.reduce.memory.mb	7296	7296
yarn.app.mapreduce.am.resource.mb	7296	7296

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r3.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx3021m	-Xmx3021m
mapreduce.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.map.memory.mb	3776	3776
mapreduce.reduce.memory.mb	7552	7552
yarn.app.mapreduce.am.resource.mb	7552	7552
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r4 实例

r4.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r4.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r4.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r4.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r4.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r5 实例

r5.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4915m	-Xmx4915m
mapreduce.java.opts	-Xmx9830m	-Xmx9830m
mapreduce.map.memory.mb	6144	6144
mapreduce.reduce.memory.mb	12288	12288
yarn.app.mapreduce.am.resource.mb	12288	12288
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

r5.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.java.opts	-Xmx11468m	-Xmx11468m
mapreduce.map.memory.mb	7168	7168
mapreduce.reduce.memory.mb	14336	14336
yarn.app.mapreduce.am.resource.mb	14336	14336
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

r5.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

r5.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6349m	-Xmx6349m
mapreduce.java.opts	-Xmx12698m	-Xmx12698m
mapreduce.map.memory.mb	7936	7936
mapreduce.reduce.memory.mb	15872	15872
yarn.app.mapreduce.am.resource.mb	15872	15872
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

r5.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6417m	-Xmx6417m
mapreduce.java.opts	-Xmx12834m	-Xmx12834m
mapreduce.map.memory.mb	8021	8021
mapreduce.reduce.memory.mb	16042	16042
yarn.app.mapreduce.am.resource.mb	16042	16042
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32100
yarn.nodemanager.resource.memory-mb	385024	32100

r5.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6451m	-Xmx6451m
mapreduce.java.opts	-Xmx12902m	-Xmx12902m
mapreduce.map.memory.mb	8064	8064
mapreduce.reduce.memory.mb	16128	16128
yarn.app.mapreduce.am.resource.mb	16128	16128

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	516096	32256
yarn.nodemanager.resource.memory-mb	516096	32256

r5.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6486m	-Xmx6486m
mapreduce.java.opts	-Xmx12972m	-Xmx12972m
mapreduce.map.memory.mb	8107	8107
mapreduce.reduce.memory.mb	16214	16214
yarn.app.mapreduce.am.resource.mb	16214	16214
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	778240	32396
yarn.nodemanager.resource.memory-mb	778240	32396

r5a 实例

r5a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4915m	-Xmx4915m
mapreduce.java.opts	-Xmx9830m	-Xmx9830m
mapreduce.map.memory.mb	6144	6144
mapreduce.reduce.memory.mb	12288	12288
yarn.app.mapreduce.am.resource.mb	12288	12288
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

r5a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.java.opts	-Xmx11468m	-Xmx11468m
mapreduce.map.memory.mb	7168	7168
mapreduce.reduce.memory.mb	14336	14336
yarn.app.mapreduce.am.resource.mb	14336	14336

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

r5a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

r5a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6349m	-Xmx6349m
mapreduce.java.opts	-Xmx12698m	-Xmx12698m
mapreduce.map.memory.mb	7936	7936
mapreduce.reduce.memory.mb	15872	15872
yarn.app.mapreduce.am.resource.mb	15872	15872
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

r5a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6417m	-Xmx6417m
mapreduce.java.opts	-Xmx12834m	-Xmx12834m
mapreduce.map.memory.mb	8021	8021
mapreduce.reduce.memory.mb	16042	16042
yarn.app.mapreduce.am.resource.mb	16042	16042

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32100
yarn.nodemanager.resource.memory-mb	385024	32100

r5a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6451m	-Xmx6451m
mapreduce.java.opts	-Xmx12902m	-Xmx12902m
mapreduce.map.memory.mb	8064	8064
mapreduce.reduce.memory.mb	16128	16128
yarn.app.mapreduce.am.resource.mb	16128	16128
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	516096	32256
yarn.nodemanager.resource.memory-mb	516096	32256

r5a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6486m	-Xmx6486m
mapreduce.java.opts	-Xmx12972m	-Xmx12972m
mapreduce.map.memory.mb	8107	8107
mapreduce.reduce.memory.mb	16214	16214
yarn.app.mapreduce.am.resource.mb	16214	16214
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	778240	32396
yarn.nodemanager.resource.memory-mb	778240	32396

r5ad 实例

r5ad.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r5ad.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r5ad.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r5ad.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r5ad.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r5ad.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6246m	-Xmx6246m
mapreduce.java.opts	-Xmx12492m	-Xmx12492m
mapreduce.map.memory.mb	7808	7808
mapreduce.reduce.memory.mb	15616	15616
yarn.app.mapreduce.am.resource.mb	15616	15616
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	499712	31232
yarn.nodemanager.resource.memory-mb	499712	31232

r5ad.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r5b 实例

r5b.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r5b.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r5b.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r5b.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r5b.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r5b.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r5b.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r5d 实例

r5d.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4915m	-Xmx4915m
mapreduce.java.opts	-Xmx9830m	-Xmx9830m
mapreduce.map.memory.mb	6144	6144
mapreduce.reduce.memory.mb	12288	12288
yarn.app.mapreduce.am.resource.mb	12288	12288
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

r5d.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.java.opts	-Xmx11468m	-Xmx11468m
mapreduce.map.memory.mb	7168	7168
mapreduce.reduce.memory.mb	14336	14336
yarn.app.mapreduce.am.resource.mb	14336	14336

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

r5d.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	122880	30720
yarn.nodemanager.resource.memory-mb	122880	30720

r5d.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6349m	-Xmx6349m
mapreduce.java.opts	-Xmx12698m	-Xmx12698m
mapreduce.map.memory.mb	7936	7936
mapreduce.reduce.memory.mb	15872	15872
yarn.app.mapreduce.am.resource.mb	15872	15872
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	253952	31744
yarn.nodemanager.resource.memory-mb	253952	31744

r5d.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6417m	-Xmx6417m
mapreduce.java.opts	-Xmx12834m	-Xmx12834m
mapreduce.map.memory.mb	8021	8021
mapreduce.reduce.memory.mb	16042	16042
yarn.app.mapreduce.am.resource.mb	16042	16042

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32100
yarn.nodemanager.resource.memory-mb	385024	32100

r5d.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6451m	-Xmx6451m
mapreduce.java.opts	-Xmx12902m	-Xmx12902m
mapreduce.map.memory.mb	8064	8064
mapreduce.reduce.memory.mb	16128	16128
yarn.app.mapreduce.am.resource.mb	16128	16128
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	516096	32256
yarn.nodemanager.resource.memory-mb	516096	32256

r5d.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6486m	-Xmx6486m
mapreduce.java.opts	-Xmx12972m	-Xmx12972m
mapreduce.map.memory.mb	8107	8107
mapreduce.reduce.memory.mb	16214	16214
yarn.app.mapreduce.am.resource.mb	16214	16214
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	778240	32396
yarn.nodemanager.resource.memory-mb	778240	32396

r5dn 实例

r5dn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r5dn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r5dn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r5dn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r5dn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r5dn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r5dn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r5n 实例

r5n.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r5n.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r5n.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r5n.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r5n.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r5n.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r5n.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6a 实例

r6a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6a.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6a.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6a.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6a.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6a.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6195m	-Xmx6195m
mapreduce.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.map.memory.mb	7744	7744
mapreduce.reduce.memory.mb	15488	15488
yarn.app.mapreduce.am.resource.mb	15488	15488

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

r6a.48xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6212m	-Xmx6212m
mapreduce.java.opts	-Xmx12424m	-Xmx12424m
mapreduce.map.memory.mb	7765	7765
mapreduce.reduce.memory.mb	15530	15530
yarn.app.mapreduce.am.resource.mb	15530	15530
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1490944	31124
yarn.nodemanager.resource.memory-mb	1490944	31124

r6g 实例

r6g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6gd 实例

r6gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6i 实例

r6i.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6i.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6i.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6i.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6i.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6i.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6i.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6i.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6029m	-Xmx6029m
mapreduce.java.opts	-Xmx12058m	-Xmx12058m
mapreduce.map.memory.mb	7536	7536
mapreduce.reduce.memory.mb	15072	15072
yarn.app.mapreduce.am.resource.mb	15072	15072

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	964608	30144
yarn.nodemanager.resource.memory-mb	964608	30144

r6id 实例

r6id.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6id.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6id.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6id.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6id.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6id.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6id.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6195m	-Xmx6195m
mapreduce.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.map.memory.mb	7744	7744
mapreduce.reduce.memory.mb	15488	15488
yarn.app.mapreduce.am.resource.mb	15488	15488
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

r6idn 实例

r6idn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6idn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6idn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6idn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6idn.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6idn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6idn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6idn.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6195m	-Xmx6195m
mapreduce.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.map.memory.mb	7744	7744
mapreduce.reduce.memory.mb	15488	15488
yarn.app.mapreduce.am.resource.mb	15488	15488
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

r6in 实例

r6in.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r6in.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r6in.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r6in.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r6in.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r6in.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r6in.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6178m	-Xmx6178m
mapreduce.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.map.memory.mb	7723	7723
mapreduce.reduce.memory.mb	15446	15446
yarn.app.mapreduce.am.resource.mb	15446	15446
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30860
yarn.nodemanager.resource.memory-mb	741376	30860

r6in.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6195m	-Xmx6195m
mapreduce.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.map.memory.mb	7744	7744
mapreduce.reduce.memory.mb	15488	15488
yarn.app.mapreduce.am.resource.mb	15488	15488

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

r7a 实例

r7a.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r7a.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r7a.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r7g 实例

r7g.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r7g.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r7g.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r7g.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r7g.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r7g.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r7gd 实例

r7gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r7gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r7gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r7gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r7gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r7gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r7iz 实例

r7iz.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4685m	-Xmx4685m
mapreduce.java.opts	-Xmx9370m	-Xmx9370m
mapreduce.map.memory.mb	5856	5856
mapreduce.reduce.memory.mb	11712	11712
yarn.app.mapreduce.am.resource.mb	11712	11712
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	23424	11712
yarn.nodemanager.resource.memory-mb	23424	11712

r7iz.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5427m	-Xmx5427m
mapreduce.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.map.memory.mb	6784	6784
mapreduce.reduce.memory.mb	13568	13568
yarn.app.mapreduce.am.resource.mb	13568	13568
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

r7iz.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5837m	-Xmx5837m
mapreduce.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.map.memory.mb	7296	7296
mapreduce.reduce.memory.mb	14592	14592
yarn.app.mapreduce.am.resource.mb	14592	14592

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

r7iz.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6042m	-Xmx6042m
mapreduce.java.opts	-Xmx12084m	-Xmx12084m
mapreduce.map.memory.mb	7552	7552
mapreduce.reduce.memory.mb	15104	15104
yarn.app.mapreduce.am.resource.mb	15104	15104
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

r7iz.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6110m	-Xmx6110m
mapreduce.java.opts	-Xmx12220m	-Xmx12220m
mapreduce.map.memory.mb	7637	7637
mapreduce.reduce.memory.mb	15274	15274
yarn.app.mapreduce.am.resource.mb	15274	15274
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	366592	30564
yarn.nodemanager.resource.memory-mb	366592	30564

r7iz.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6144m	-Xmx6144m
mapreduce.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.map.memory.mb	7680	7680
mapreduce.reduce.memory.mb	15360	15360
yarn.app.mapreduce.am.resource.mb	15360	15360

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

r7iz.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6029m	-Xmx6029m
mapreduce.java.opts	-Xmx12058m	-Xmx12058m
mapreduce.map.memory.mb	7536	7536
mapreduce.reduce.memory.mb	15072	15072
yarn.app.mapreduce.am.resource.mb	15072	15072
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	964608	30144
yarn.nodemanager.resource.memory-mb	964608	30144

x1 实例

x1.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12058m	-Xmx12058m
mapreduce.java.opts	-Xmx24116m	-Xmx24116m
mapreduce.map.memory.mb	15072	15072
mapreduce.reduce.memory.mb	30144	30144
yarn.app.mapreduce.am.resource.mb	30144	30144
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	964608	30144
yarn.nodemanager.resource.memory-mb	964608	30144

x1.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12109m	-Xmx12109m
mapreduce.java.opts	-Xmx24218m	-Xmx24218m
mapreduce.map.memory.mb	15136	15136
mapreduce.reduce.memory.mb	30272	30272
yarn.app.mapreduce.am.resource.mb	30272	30272

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1937408	30272
yarn.nodemanager.resource.memory-mb	1937408	30272

x1e 实例

x1e.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx22682m	-Xmx22682m
mapreduce.java.opts	-Xmx45364m	-Xmx45364m
mapreduce.map.memory.mb	28352	28352
mapreduce.reduce.memory.mb	56704	56704
yarn.app.mapreduce.am.resource.mb	56704	56704
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	113408	0
yarn.nodemanager.resource.memory-mb	113408	0

x1e.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx23501m	-Xmx23501m
mapreduce.java.opts	-Xmx47002m	-Xmx47002m
mapreduce.map.memory.mb	29376	29376
mapreduce.reduce.memory.mb	58752	58752
yarn.app.mapreduce.am.resource.mb	58752	58752
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	235008	0
yarn.nodemanager.resource.memory-mb	235008	0

x1e.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx23910m	-Xmx23910m
mapreduce.java.opts	-Xmx47820m	-Xmx47820m
mapreduce.map.memory.mb	29888	29888
mapreduce.reduce.memory.mb	59776	59776
yarn.app.mapreduce.am.resource.mb	59776	59776

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	478208	0
yarn.nodemanager.resource.memory-mb	478208	0

x1e.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24115m	-Xmx24115m
mapreduce.java.opts	-Xmx48230m	-Xmx48230m
mapreduce.map.memory.mb	30144	30144
mapreduce.reduce.memory.mb	60288	60288
yarn.app.mapreduce.am.resource.mb	60288	60288
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	964608	0
yarn.nodemanager.resource.memory-mb	964608	0

x1e.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24218m	-Xmx24218m
mapreduce.java.opts	-Xmx48436m	-Xmx48436m
mapreduce.map.memory.mb	30272	30272
mapreduce.reduce.memory.mb	60544	60544
yarn.app.mapreduce.am.resource.mb	60544	60544
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1937408	0
yarn.nodemanager.resource.memory-mb	1937408	0

x1e.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24269m	-Xmx24269m
mapreduce.java.opts	-Xmx48538m	-Xmx48538m
mapreduce.map.memory.mb	30336	30336
mapreduce.reduce.memory.mb	60672	60672
yarn.app.mapreduce.am.resource.mb	60672	60672

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	3883008	0
yarn.nodemanager.resource.memory-mb	3883008	0

x2gd 实例

x2gd.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx10854m	-Xmx10854m
mapreduce.java.opts	-Xmx21708m	-Xmx21708m
mapreduce.map.memory.mb	13568	13568
mapreduce.reduce.memory.mb	27136	27136
yarn.app.mapreduce.am.resource.mb	27136	27136
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	54272	27136
yarn.nodemanager.resource.memory-mb	54272	27136

x2gd.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx11674m	-Xmx11674m
mapreduce.java.opts	-Xmx23348m	-Xmx23348m
mapreduce.map.memory.mb	14592	14592
mapreduce.reduce.memory.mb	29184	29184
yarn.app.mapreduce.am.resource.mb	29184	29184
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	29184
yarn.nodemanager.resource.memory-mb	116736	29184

x2gd.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12083m	-Xmx12083m
mapreduce.java.opts	-Xmx24166m	-Xmx24166m
mapreduce.map.memory.mb	15104	15104
mapreduce.reduce.memory.mb	30208	30208
yarn.app.mapreduce.am.resource.mb	30208	30208

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	30208
yarn.nodemanager.resource.memory-mb	241664	30208

x2gd.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12288m	-Xmx12288m
mapreduce.java.opts	-Xmx24576m	-Xmx24576m
mapreduce.map.memory.mb	15360	15360
mapreduce.reduce.memory.mb	30720	30720
yarn.app.mapreduce.am.resource.mb	30720	30720
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	30720
yarn.nodemanager.resource.memory-mb	491520	30720

x2gd.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12356m	-Xmx12356m
mapreduce.java.opts	-Xmx24712m	-Xmx24712m
mapreduce.map.memory.mb	15445	15445
mapreduce.reduce.memory.mb	30890	30890
yarn.app.mapreduce.am.resource.mb	30890	30890
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	741376	30906
yarn.nodemanager.resource.memory-mb	741376	30906

x2gd.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.java.opts	-Xmx24780m	-Xmx24780m
mapreduce.map.memory.mb	15488	15488
mapreduce.reduce.memory.mb	30976	30976
yarn.app.mapreduce.am.resource.mb	30976	30976

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

x2idn 实例

x2idn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12390m	-Xmx12390m
mapreduce.java.opts	-Xmx24780m	-Xmx24780m
mapreduce.map.memory.mb	15488	15488
mapreduce.reduce.memory.mb	30976	30976
yarn.app.mapreduce.am.resource.mb	30976	30976
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	30976
yarn.nodemanager.resource.memory-mb	991232	30976

x2idn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12425m	-Xmx12425m
mapreduce.java.opts	-Xmx24850m	-Xmx24850m
mapreduce.map.memory.mb	15531	15531
mapreduce.reduce.memory.mb	31062	31062
yarn.app.mapreduce.am.resource.mb	31062	31062
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1490944	31030
yarn.nodemanager.resource.memory-mb	1490944	31030

x2idn.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx12442m	-Xmx12442m
mapreduce.java.opts	-Xmx24884m	-Xmx24884m
mapreduce.map.memory.mb	15552	15552
mapreduce.reduce.memory.mb	31104	31104
yarn.app.mapreduce.am.resource.mb	31104	31104

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1990656	31104
yarn.nodemanager.resource.memory-mb	1990656	31104

x2iedn 实例

x2iedn.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx23347m	-Xmx23347m
mapreduce.java.opts	-Xmx46694m	-Xmx46694m
mapreduce.map.memory.mb	29184	29184
mapreduce.reduce.memory.mb	58368	58368
yarn.app.mapreduce.am.resource.mb	58368	58368
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	116736	0
yarn.nodemanager.resource.memory-mb	116736	0

x2iedn.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24166m	-Xmx24166m
mapreduce.java.opts	-Xmx48332m	-Xmx48332m
mapreduce.map.memory.mb	30208	30208
mapreduce.reduce.memory.mb	60416	60416
yarn.app.mapreduce.am.resource.mb	60416	60416
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	241664	0
yarn.nodemanager.resource.memory-mb	241664	0

x2iedn.4xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24576m	-Xmx24576m
mapreduce.java.opts	-Xmx49152m	-Xmx49152m
mapreduce.map.memory.mb	30720	30720
mapreduce.reduce.memory.mb	61440	61440
yarn.app.mapreduce.am.resource.mb	61440	61440

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	491520	0
yarn.nodemanager.resource.memory-mb	491520	0

x2iedn.8xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24781m	-Xmx24781m
mapreduce.java.opts	-Xmx49562m	-Xmx49562m
mapreduce.map.memory.mb	30976	30976
mapreduce.reduce.memory.mb	61952	61952
yarn.app.mapreduce.am.resource.mb	61952	61952
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	991232	0
yarn.nodemanager.resource.memory-mb	991232	0

x2iedn.16xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24883m	-Xmx24883m
mapreduce.java.opts	-Xmx49766m	-Xmx49766m
mapreduce.map.memory.mb	31104	31104
mapreduce.reduce.memory.mb	62208	62208
yarn.app.mapreduce.am.resource.mb	62208	62208
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	1990656	0
yarn.nodemanager.resource.memory-mb	1990656	0

x2iedn.24xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24918m	-Xmx24918m
mapreduce.java.opts	-Xmx49836m	-Xmx49836m
mapreduce.map.memory.mb	31147	31147
mapreduce.reduce.memory.mb	62294	62294
yarn.app.mapreduce.am.resource.mb	62294	62294

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	2990080	-32
yarn.nodemanager.resource.memory-mb	2990080	-32

x2iedn.32xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx24934m	-Xmx24934m
mapreduce.java.opts	-Xmx49868m	-Xmx49868m
mapreduce.map.memory.mb	31168	31168
mapreduce.reduce.memory.mb	62336	62336
yarn.app.mapreduce.am.resource.mb	62336	62336
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	3989504	0
yarn.nodemanager.resource.memory-mb	3989504	0

z1d 实例

z1d.xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx4915m	-Xmx4915m
mapreduce.java.opts	-Xmx9830m	-Xmx9830m
mapreduce.map.memory.mb	6144	6144
mapreduce.reduce.memory.mb	12288	12288
yarn.app.mapreduce.am.resource.mb	12288	12288
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	24 576	12288
yarn.nodemanager.resource.memory-mb	24 576	12288

z1d.2xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx5734m	-Xmx5734m
mapreduce.java.opts	-Xmx11468m	-Xmx11468m
mapreduce.map.memory.mb	7168	7168
mapreduce.reduce.memory.mb	14336	14336
yarn.app.mapreduce.am.resource.mb	14336	14336

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	57344	28672
yarn.nodemanager.resource.memory-mb	57344	28672

z1d.3xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6007m	-Xmx6007m
mapreduce.java.opts	-Xmx12014m	-Xmx12014m
mapreduce.map.memory.mb	7509	7509
mapreduce.reduce.memory.mb	15018	15018
yarn.app.mapreduce.am.resource.mb	15018	15018
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	90112	30040
yarn.nodemanager.resource.memory-mb	90112	30040

z1d.6xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6281m	-Xmx6281m
mapreduce.java.opts	-Xmx12562m	-Xmx12562m
mapreduce.map.memory.mb	7851	7851
mapreduce.reduce.memory.mb	15702	15702
yarn.app.mapreduce.am.resource.mb	15702	15702
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	188416	31396
yarn.nodemanager.resource.memory-mb	188416	31396

z1d.12xlarge

配置选项	默认值	安装了 HBase
mapreduce.map.java.opts	-Xmx6417m	-Xmx6417m
mapreduce.java.opts	-Xmx12834m	-Xmx12834m
mapreduce.map.memory.mb	8021	8021
mapreduce.reduce.memory.mb	16042	16042
yarn.app.mapreduce.am.resource.mb	16042	16042

配置选项	默认值	安装了 HBase
yarn.scheduler.minimum-allocation-mb	32	32
yarn.scheduler.maximum-allocation-mb	385024	32100
yarn.nodemanager.resource.memory-mb	385024	32100

Hadoop 守护进程配置设置

Hadoop 守护进程设置因集群节点使用的 EC2 实例类型而异。以下表格列出了每种 EC2 实例类型的默认配置设置。

要自定义这些设置，请使用 hadoop-env 配置分类。有关更多信息，请参阅配置应用程序。

实例类型

- c1 实例
- c3 实例
- c4 实例
- c5 实例
- c5a 实例
- c5ad 实例
- c5d 实例
- c5n 实例
- c6a 实例
- c6g 实例
- c6gd 实例
- c6gn 实例
- c6i 实例
- c6id 实例
- c6in 实例

- c7g 实例
- c7gd 实例
- c7gn 实例
- d2 实例
- d3 实例
- d3en 实例
- g2 实例
- g3 实例
- g3s 实例
- g4dn 实例
- g5 实例
- h1 实例
- i2 实例
- i3 实例
- i3en 实例
- i4g 实例
- i4i 实例
- im4gn 实例
- is4gen 实例
- m1 实例
- m2 实例
- m3 实例
- m4 实例
- m5 实例
- m5a 实例
- m5ad 实例
- m5d 实例
- m5dn 实例
- m5n 实例
- m5zn 实例

- m6a 实例
- m6g 实例
- m6gd 实例
- m6i 实例
- m6id 实例
- m6idn 实例
- m6in 实例
- m7g 实例
- m7gd 实例
- m7i 实例
- m7i-flex 实例
- p2 实例
- p3 实例
- p5 实例
- r3 实例
- r4 实例
- r5 实例
- r5a 实例
- r5ad 实例
- r5b 实例
- r5d 实例
- r5dn 实例
- r5n 实例
- r6a 实例
- r6g 实例
- r6gd 实例
- r6i 实例
- r6id 实例
- r6idn 实例
- r6in 实例

- r7g 实例
- r7gd 实例
- x1 实例
- x1e 实例
- x2gd 实例
- x2idn 实例
- x2iedn 实例
- z1d 实例

c1 实例

c1.medium

参数	Value
YARN_RESOURCEMANAGER_HEAPSIZE	192
YARN_PROXYSERVER_HEAPSIZE	96
YARN_NODEMANAGER_HEAPSIZE	128
HADOOP_JOB_HISTORYSERVER_HEAPSIZE	128
HADOOP_NAMENODE_HEAPSIZE	192
HADOOP_DATANODE_HEAPSIZE	96

c1.xlarge

参数	Value
YARN_RESOURCEMANAGER_HEAPSIZE	768
YARN_PROXYSERVER_HEAPSIZE	384

参数	Value
YARN_NODEMANAGER_HEAPSIZE	512
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	512
HADOOP_NAMENODE_HEAPSIZE	768
HADOOP_DATANODE_HEAPSIZE	384

c3 实例

c3.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2396
YARN_PROXYSERVER_HEAPSIZE	2396

参数	Value
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2396
HADOOP_NAMENODE_HEAPSIZE	1740
HADOOP_DATANODE_HEAPSIZE	757

c3.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2703
YARN_PROXYSERVER_HEAPSIZE	2703
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2703
HADOOP_NAMENODE_HEAPSIZE	3276
HADOOP_DATANODE_HEAPSIZE	1064

c3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3317
YARN_PROXYSERVER_HEAPSIZE	3317
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3317
HADOOP_NAMENODE_HEAPSIZE	6348
HADOOP_DATANODE_HEAPSIZE	1679

c4 实例

c4.large

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	1152
YARN_PROXYSERVER_HEAPSIZE	1152
YARN_NODEMANAGER_HEAPSIZE	1152
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1152
HADOOP_NAMENODE_HEAPSIZE	576
HADOOP_DATANODE_HEAPSIZE	384

c4.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c4.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2396
YARN_PROXYSERVER_HEAPSIZE	2396
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2396
HADOOP_NAMENODE_HEAPSIZE	1740
HADOOP_DATANODE_HEAPSIZE	757

c4.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2703
YARN_PROXYSERVER_HEAPSIZE	2703
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2703
HADOOP_NAMENODE_HEAPSIZE	3276
HADOOP_DATANODE_HEAPSIZE	1064

c4.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3317
YARN_PROXYSERVER_HEAPSIZE	3317
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3317
HADOOP_NAMENODE_HEAPSIZE	6348
HADOOP_DATANODE_HEAPSIZE	1679

c5 实例

c5.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2252
YARN_PROXYSERVER_HEAPSIZE	2252
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2252
HADOOP_NAMENODE_HEAPSIZE	1024
HADOOP_DATANODE_HEAPSIZE	614

c5.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

c5.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

c5.9xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3563
YARN_PROXYSERVER_HEAPSIZE	3563
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3563
HADOOP_NAMENODE_HEAPSIZE	7577
HADOOP_DATANODE_HEAPSIZE	1925

c5.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

c5.18xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5038
YARN_PROXYSERVER_HEAPSIZE	5038
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5038
HADOOP_NAMENODE_HEAPSIZE	14950
HADOOP_DATANODE_HEAPSIZE	3399

c5.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

c5a 实例

c5a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c5a.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c5a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c5a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c5a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

c5a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c5a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c5ad 实例

c5ad.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c5ad.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c5ad.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c5ad.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c5ad.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c5ad.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c5ad.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c5d 实例

c5d.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2252
YARN_PROXYSERVER_HEAPSIZE	2252
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2252
HADOOP_NAMENODE_HEAPSIZE	1024
HADOOP_DATANODE_HEAPSIZE	614

c5d.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

c5d.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

c5d.9xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3563
YARN_PROXYSERVER_HEAPSIZE	3563
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3563
HADOOP_NAMENODE_HEAPSIZE	7577
HADOOP_DATANODE_HEAPSIZE	1925

c5d.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

c5d.18xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5038
YARN_PROXYSERVER_HEAPSIZE	5038
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5038
HADOOP_NAMENODE_HEAPSIZE	14950
HADOOP_DATANODE_HEAPSIZE	3399

c5d.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

c5n 实例

c5n.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2304
YARN_PROXYSERVER_HEAPSIZE	2304
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2304
HADOOP_NAMENODE_HEAPSIZE	1280
HADOOP_DATANODE_HEAPSIZE	665

c5n.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2519
YARN_PROXYSERVER_HEAPSIZE	2519
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2519
HADOOP_NAMENODE_HEAPSIZE	2355
HADOOP_DATANODE_HEAPSIZE	880

c5n.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2949
YARN_PROXYSERVER_HEAPSIZE	2949
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2949
HADOOP_NAMENODE_HEAPSIZE	4505
HADOOP_DATANODE_HEAPSIZE	1310

c5n.9xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

c5n.18xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

c6a 实例

c6a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6a.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c6a.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

c6a.48xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

c6g 实例

c6g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6gd 实例

c6gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6gn 实例

c6gn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6gn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6gn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6gn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6gn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6gn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6i 实例

c6i.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6i.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6i.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6i.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6i.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6i.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6i.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c6i.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

c6id 实例

c6id.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6id.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6id.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6id.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6id.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6id.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6id.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c6id.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

c6in 实例

c6in.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c6in.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c6in.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c6in.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c6in.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c6in.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c6in.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

c6in.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

c7g 实例

c7g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c7g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c7g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c7g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c7g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c7g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c7gd 实例

c7gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c7gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c7gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c7gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c7gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c7gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

c7gn 实例

c7gn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2124
YARN_PROXYSERVER_HEAPSIZE	2124
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2124
HADOOP_NAMENODE_HEAPSIZE	972
HADOOP_DATANODE_HEAPSIZE	588

c7gn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

c7gn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

c7gn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

c7gn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

c7gn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

d2 实例

d2.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

d2.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

d2.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

d2.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

d3 实例

d3.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

d3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

d3.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

d3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

d3en 实例

d3en.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

d3en.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

d3en.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

d3en.6xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

d3en.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

d3en.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

g2 实例

g2.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	1536
YARN_PROXYSERVER_HEAPSIZE	1024
YARN_NODEMANAGER_HEAPSIZE	1024

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1024
HADOOP_NAMENODE_HEAPSIZE	2304
HADOOP_DATANODE_HEAPSIZE	384

g3 实例

g3.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

g3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

g3.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

g3s 实例

g3s.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

g4dn 实例

g4dn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

g4dn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

g4dn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

g4dn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

g4dn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

g4dn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

g5 实例

g5.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

g5.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

g5.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

g5.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

g5.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

g5.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

g5.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

g5.48xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

h1 实例

h1.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

h1.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

h1.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

h1.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

i2 实例

i2.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

i2.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

i2.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

i2.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

i3 实例

i3.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

i3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

i3.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

i3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

i3.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

i3en 实例

i3en.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

i3en.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

i3en.3xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

i3en.6xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

i3en.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

i3en.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17817
YARN_PROXYSERVER_HEAPSIZE	17817
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17817
HADOOP_NAMENODE_HEAPSIZE	78848
HADOOP_DATANODE_HEAPSIZE	4096

i4g 实例

i4g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

i4g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

i4g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

i4g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

i4g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

i4i 实例

i4i.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

i4i.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

i4i.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

i4i.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

i4i.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

i4i.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

im4gn 实例

im4gn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

im4gn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

im4gn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

im4gn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

im4gn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

is4gen 实例

is4gen.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2557
YARN_PROXYSERVER_HEAPSIZE	2557
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2557
HADOOP_NAMENODE_HEAPSIZE	2547
HADOOP_DATANODE_HEAPSIZE	919

is4gen.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3025
YARN_PROXYSERVER_HEAPSIZE	3025
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3025
HADOOP_NAMENODE_HEAPSIZE	4889
HADOOP_DATANODE_HEAPSIZE	1387

is4gen.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

is4gen.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m1 实例

m1.small

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	256
YARN_PROXYSERVER_HEAPSIZE	96
YARN_NODEMANAGER_HEAPSIZE	192

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	128
HADOOP_NAMENODE_HEAPSIZE	192
HADOOP_DATANODE_HEAPSIZE	96

m1.medium

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	384
YARN_PROXYSERVER_HEAPSIZE	192
YARN_NODEMANAGER_HEAPSIZE	256
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	256
HADOOP_NAMENODE_HEAPSIZE	384
HADOOP_DATANODE_HEAPSIZE	192

m1.large

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	768
YARN_PROXYSERVER_HEAPSIZE	384
YARN_NODEMANAGER_HEAPSIZE	512

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	512
HADOOP_NAMENODE_HEAPSIZE	768
HADOOP_DATANODE_HEAPSIZE	384

m1.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	1024
YARN_PROXYSERVER_HEAPSIZE	512
YARN_NODEMANAGER_HEAPSIZE	768
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1024
HADOOP_NAMENODE_HEAPSIZE	2304
HADOOP_DATANODE_HEAPSIZE	384

m2 实例

m2.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	1536
YARN_PROXYSERVER_HEAPSIZE	1024
YARN_NODEMANAGER_HEAPSIZE	1024

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1024
HADOOP_NAMENODE_HEAPSIZE	3072
HADOOP_DATANODE_HEAPSIZE	384

m2.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	1536
YARN_PROXYSERVER_HEAPSIZE	1024
YARN_NODEMANAGER_HEAPSIZE	1024
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1536
HADOOP_NAMENODE_HEAPSIZE	6144
HADOOP_DATANODE_HEAPSIZE	384

m2.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2048
YARN_PROXYSERVER_HEAPSIZE	1024
YARN_NODEMANAGER_HEAPSIZE	1536

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	1536
HADOOP_NAMENODE_HEAPSIZE	12288
HADOOP_DATANODE_HEAPSIZE	384

m3 实例

m3.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2396
YARN_PROXYSERVER_HEAPSIZE	2396
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2396
HADOOP_NAMENODE_HEAPSIZE	1740
HADOOP_DATANODE_HEAPSIZE	757

m3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2703
YARN_PROXYSERVER_HEAPSIZE	2703
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2703
HADOOP_NAMENODE_HEAPSIZE	3276
HADOOP_DATANODE_HEAPSIZE	1064

m4 实例

m4.large

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2252
YARN_PROXYSERVER_HEAPSIZE	2252
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2252
HADOOP_NAMENODE_HEAPSIZE	1024
HADOOP_DATANODE_HEAPSIZE	614

m4.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

m4.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

m4.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

m4.10xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5365
YARN_PROXYSERVER_HEAPSIZE	5365
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5365
HADOOP_NAMENODE_HEAPSIZE	16588
HADOOP_DATANODE_HEAPSIZE	3727

m4.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

m5 实例

m5.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

m5.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

m5.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

m5.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

m5.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

m5.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

m5.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

m5a 实例

m5a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

m5a.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

m5a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

m5a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

m5a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

m5a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

m5a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

m5ad 实例

m5ad.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m5ad.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m5ad.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m5ad.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m5ad.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m5ad.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m5ad.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m5d 实例

m5d.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2416
YARN_PROXYSERVER_HEAPSIZE	2416
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2416
HADOOP_NAMENODE_HEAPSIZE	1843
HADOOP_DATANODE_HEAPSIZE	778

m5d.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

m5d.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

m5d.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

m5d.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

m5d.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

m5d.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

m5dn 实例

m5dn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m5dn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m5dn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m5dn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m5dn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m5dn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m5dn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m5n 实例

m5n.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m5n.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m5n.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m5n.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m5n.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m5n.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m5n.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m5zn 实例

m5zn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2396
YARN_PROXYSERVER_HEAPSIZE	2396
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2396
HADOOP_NAMENODE_HEAPSIZE	1740
HADOOP_DATANODE_HEAPSIZE	757

m5zn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m5zn.3xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3025
YARN_PROXYSERVER_HEAPSIZE	3025
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3025
HADOOP_NAMENODE_HEAPSIZE	4889
HADOOP_DATANODE_HEAPSIZE	1387

m5zn.6xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3962
YARN_PROXYSERVER_HEAPSIZE	3962
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3962
HADOOP_NAMENODE_HEAPSIZE	9574
HADOOP_DATANODE_HEAPSIZE	2324

m5zn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m6a 实例

m6a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6a.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m6a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m6a.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

m6a.48xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

m6g 实例

m6g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5877
YARN_PROXYSERVER_HEAPSIZE	5877
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5877
HADOOP_NAMENODE_HEAPSIZE	19148
HADOOP_DATANODE_HEAPSIZE	4096

m6g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6gd 实例

m6gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5877
YARN_PROXYSERVER_HEAPSIZE	5877
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5877
HADOOP_NAMENODE_HEAPSIZE	19148
HADOOP_DATANODE_HEAPSIZE	4096

m6gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6i 实例

m6i.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6i.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6i.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6i.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6i.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5877
YARN_PROXYSERVER_HEAPSIZE	5877
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5877
HADOOP_NAMENODE_HEAPSIZE	19148
HADOOP_DATANODE_HEAPSIZE	4096

m6i.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6i.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m6i.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

m6id 实例

m6id.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6id.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6id.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6id.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6id.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m6id.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6id.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m6id.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

m6idn 实例

m6idn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6idn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6idn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6idn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6idn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m6idn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6idn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m6idn.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

m6in 实例

m6in.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m6in.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m6in.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m6in.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m6in.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5877
YARN_PROXYSERVER_HEAPSIZE	5877
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5877
HADOOP_NAMENODE_HEAPSIZE	19148
HADOOP_DATANODE_HEAPSIZE	4096

m6in.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m6in.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

m6in.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

m7g 实例

m7g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m7g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m7g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m7g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m7g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m7g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m7gd 实例

m7gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m7gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m7gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m7gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m7gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	5836
YARN_PROXYSERVER_HEAPSIZE	5836
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	5836
HADOOP_NAMENODE_HEAPSIZE	18944
HADOOP_DATANODE_HEAPSIZE	4096

m7gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

m7i 实例

m7i.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m7i.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m7i.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

m7i.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

m7i-flex 实例

m7i-flex.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2401
YARN_PROXYSERVER_HEAPSIZE	2401
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2401
HADOOP_NAMENODE_HEAPSIZE	1766
HADOOP_DATANODE_HEAPSIZE	762

m7i-flex.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

m7i-flex.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

p2 实例

p2.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

p2.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

p2.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

p3 实例

p3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

p3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

p3.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

p5 实例

p5.48xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	42065
YARN_PROXYSERVER_HEAPSIZE	42065
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	42065
HADOOP_NAMENODE_HEAPSIZE	200089
HADOOP_DATANODE_HEAPSIZE	4096

r3 实例

r3.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r3.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r3.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r3.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r4 实例

r4.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r4.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r4.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r4.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r4.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r5 实例

r5.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

r5.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

r5.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

r5.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

r5.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

r5.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12574
YARN_PROXYSERVER_HEAPSIZE	12574
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12574
HADOOP_NAMENODE_HEAPSIZE	52633
HADOOP_DATANODE_HEAPSIZE	4096

r5.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17817
YARN_PROXYSERVER_HEAPSIZE	17817
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17817
HADOOP_NAMENODE_HEAPSIZE	78848
HADOOP_DATANODE_HEAPSIZE	4096

r5a 实例

r5a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

r5a.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

r5a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

r5a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

r5a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

r5a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12574
YARN_PROXYSERVER_HEAPSIZE	12574
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12574
HADOOP_NAMENODE_HEAPSIZE	52633
HADOOP_DATANODE_HEAPSIZE	4096

r5a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17817
YARN_PROXYSERVER_HEAPSIZE	17817
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17817
HADOOP_NAMENODE_HEAPSIZE	78848
HADOOP_DATANODE_HEAPSIZE	4096

r5ad 实例

r5ad.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r5ad.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r5ad.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r5ad.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r5ad.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r5ad.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12247
YARN_PROXYSERVER_HEAPSIZE	12247
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12247
HADOOP_NAMENODE_HEAPSIZE	50995
HADOOP_DATANODE_HEAPSIZE	4096

r5ad.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r5b 实例

r5b.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r5b.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r5b.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r5b.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r5b.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r5b.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r5b.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r5d 实例

r5d.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

r5d.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

r5d.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4710
YARN_PROXYSERVER_HEAPSIZE	4710
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4710
HADOOP_NAMENODE_HEAPSIZE	13312
HADOOP_DATANODE_HEAPSIZE	3072

r5d.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7331
YARN_PROXYSERVER_HEAPSIZE	7331
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7331
HADOOP_NAMENODE_HEAPSIZE	26419
HADOOP_DATANODE_HEAPSIZE	4096

r5d.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

r5d.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12574
YARN_PROXYSERVER_HEAPSIZE	12574
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12574
HADOOP_NAMENODE_HEAPSIZE	52633
HADOOP_DATANODE_HEAPSIZE	4096

r5d.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17817
YARN_PROXYSERVER_HEAPSIZE	17817
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17817
HADOOP_NAMENODE_HEAPSIZE	78848
HADOOP_DATANODE_HEAPSIZE	4096

r5dn 实例

r5dn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r5dn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r5dn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r5dn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r5dn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r5dn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r5dn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r5n 实例

r5n.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r5n.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r5n.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r5n.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r5n.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r5n.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r5n.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6a 实例

r6a.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6a.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6a.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6a.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6a.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6a.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6a.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

r6a.48xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	32071
YARN_PROXYSERVER_HEAPSIZE	32071
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	32071
HADOOP_NAMENODE_HEAPSIZE	150118
HADOOP_DATANODE_HEAPSIZE	4096

r6g 实例

r6g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6gd 实例

r6gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6i 实例

r6i.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6i.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6i.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6i.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6i.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6i.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6i.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6i.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	21544
YARN_PROXYSERVER_HEAPSIZE	21544
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	21544
HADOOP_NAMENODE_HEAPSIZE	97484
HADOOP_DATANODE_HEAPSIZE	4096

r6id 实例

r6id.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6id.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6id.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6id.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6id.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6id.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6id.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

r6idn 实例

r6idn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6idn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6idn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6idn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6idn.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6idn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6idn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6idn.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

r6in 实例

r6in.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r6in.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r6in.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r6in.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r6in.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r6in.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r6in.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

r6in.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

r7g 实例

r7g.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r7g.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r7g.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r7g.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r7g.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r7g.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

r7gd 实例

r7gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2713
YARN_PROXYSERVER_HEAPSIZE	2713
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2713
HADOOP_NAMENODE_HEAPSIZE	3328
HADOOP_DATANODE_HEAPSIZE	1075

r7gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

r7gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

r7gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

r7gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9584
YARN_PROXYSERVER_HEAPSIZE	9584
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9584
HADOOP_NAMENODE_HEAPSIZE	37683
HADOOP_DATANODE_HEAPSIZE	4096

r7gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

x1 实例

x1.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	21544
YARN_PROXYSERVER_HEAPSIZE	21544
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	21544
HADOOP_NAMENODE_HEAPSIZE	97484
HADOOP_DATANODE_HEAPSIZE	4096

x1.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	41000
YARN_PROXYSERVER_HEAPSIZE	41000
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	41000
HADOOP_NAMENODE_HEAPSIZE	194764
HADOOP_DATANODE_HEAPSIZE	4096

x1e 实例

x1e.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4520
YARN_PROXYSERVER_HEAPSIZE	4520
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4520
HADOOP_NAMENODE_HEAPSIZE	12364
HADOOP_DATANODE_HEAPSIZE	2882

x1e.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6952
YARN_PROXYSERVER_HEAPSIZE	6952
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6952
HADOOP_NAMENODE_HEAPSIZE	24524
HADOOP_DATANODE_HEAPSIZE	4096

x1e.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	11816
YARN_PROXYSERVER_HEAPSIZE	11816
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	11816
HADOOP_NAMENODE_HEAPSIZE	48844
HADOOP_DATANODE_HEAPSIZE	4096

x1e.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	21544
YARN_PROXYSERVER_HEAPSIZE	21544
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	21544
HADOOP_NAMENODE_HEAPSIZE	97484
HADOOP_DATANODE_HEAPSIZE	4096

x1e.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	41000
YARN_PROXYSERVER_HEAPSIZE	41000
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	41000
HADOOP_NAMENODE_HEAPSIZE	194764
HADOOP_DATANODE_HEAPSIZE	4096

x1e.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	79912
YARN_PROXYSERVER_HEAPSIZE	79912
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	79912
HADOOP_NAMENODE_HEAPSIZE	389324
HADOOP_DATANODE_HEAPSIZE	4096

x2gd 实例

x2gd.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3338
YARN_PROXYSERVER_HEAPSIZE	3338
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3338
HADOOP_NAMENODE_HEAPSIZE	6451
HADOOP_DATANODE_HEAPSIZE	1699

x2gd.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

x2gd.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

x2gd.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

x2gd.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	17080
YARN_PROXYSERVER_HEAPSIZE	17080
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	17080
HADOOP_NAMENODE_HEAPSIZE	75161
HADOOP_DATANODE_HEAPSIZE	4096

x2gd.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

x2idn 实例

x2idn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

x2idn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	32071
YARN_PROXYSERVER_HEAPSIZE	32071
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	32071
HADOOP_NAMENODE_HEAPSIZE	150118
HADOOP_DATANODE_HEAPSIZE	4096

x2idn.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	42065
YARN_PROXYSERVER_HEAPSIZE	42065
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	42065
HADOOP_NAMENODE_HEAPSIZE	200089
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn 实例

x2iedn.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4587
YARN_PROXYSERVER_HEAPSIZE	4587
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4587
HADOOP_NAMENODE_HEAPSIZE	12697
HADOOP_DATANODE_HEAPSIZE	2949

x2iedn.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	7086
YARN_PROXYSERVER_HEAPSIZE	7086
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	7086
HADOOP_NAMENODE_HEAPSIZE	25190
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn.4xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	12083
YARN_PROXYSERVER_HEAPSIZE	12083
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	12083
HADOOP_NAMENODE_HEAPSIZE	50176
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn.8xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	22077
YARN_PROXYSERVER_HEAPSIZE	22077
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	22077
HADOOP_NAMENODE_HEAPSIZE	100147
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn.16xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	42065
YARN_PROXYSERVER_HEAPSIZE	42065
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	42065
HADOOP_NAMENODE_HEAPSIZE	200089
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn.24xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	62054
YARN_PROXYSERVER_HEAPSIZE	62054
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	62054
HADOOP_NAMENODE_HEAPSIZE	300032
HADOOP_DATANODE_HEAPSIZE	4096

x2iedn.32xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	82042
YARN_PROXYSERVER_HEAPSIZE	82042
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	82042
HADOOP_NAMENODE_HEAPSIZE	399974
HADOOP_DATANODE_HEAPSIZE	4096

z1d 实例

z1d.xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	2744
YARN_PROXYSERVER_HEAPSIZE	2744
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	2744
HADOOP_NAMENODE_HEAPSIZE	3481
HADOOP_DATANODE_HEAPSIZE	1105

z1d.2xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	3399
YARN_PROXYSERVER_HEAPSIZE	3399
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	3399
HADOOP_NAMENODE_HEAPSIZE	6758
HADOOP_DATANODE_HEAPSIZE	1761

z1d.3xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	4055
YARN_PROXYSERVER_HEAPSIZE	4055
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	4055
HADOOP_NAMENODE_HEAPSIZE	10035
HADOOP_DATANODE_HEAPSIZE	2416

z1d.6xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	6021
YARN_PROXYSERVER_HEAPSIZE	6021
YARN_NODEMANAGER_HEAPSIZE	2048
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	6021
HADOOP_NAMENODE_HEAPSIZE	19865
HADOOP_DATANODE_HEAPSIZE	4096

z1d.12xlarge

参数	Value
YARN_RESOURCEMANAG ER_HEAPSIZE	9953
YARN_PROXYSERVER_HEAPSIZE	9953
YARN_NODEMANAGER_HEAPSIZE	2048

参数	Value
HADOOP_JOB_HISTORY SERVER_HEAPSIZE	9953
HADOOP_NAMENODE_HEAPSIZE	39526
HADOOP_DATANODE_HEAPSIZE	4096

HDFS 配置

下表描述了默认 Hadoop Distributed File System（HDFS）参数及其设置。您可以使用 `hdfs-site` 配置分类更改这些值。有关更多信息，请参阅配置应用程序。

> **Warning**
>
> 1. 如果单个节点出现故障，则在少于四个节点的集群上将 `dfs.replication` 设置为 1 可能会导致 HDFS 数据丢失。如果您的集群有 HDFS 存储，我们建议您将集群配置为至少四个用于生产工作负载的核心节点，以避免出现数据丢失情况。
> 2. Amazon EMR 不允许集群扩展 `dfs.replication` 下方的核心节点。例如，如果是 `dfs.replication = 2`，则最小核心节点数为 2。
> 3. 当您使用托管式自动扩缩功能、自动扩缩功能或选择手动调整集群大小时，建议您将 `dfs.replication` 设置为 2 或更高。

参数	定义	默认值
`dfs.block.size`	HDFS 数据块的大小。当对 HDFS 中存储的数据进行操作时，拆分大小通常是 HDFS 数据块的大小。数字越大，提供的任务粒度越小，但集群 NameNode 受到的压力也越小。	134217728 (128 MB)
`dfs.replication`	要持久性存储的每个数据块的副本数量。Amazon EMR 根据集群预置的核心节点数量设置该值。调整该值以满足您的需求。要覆盖默认值，请使用 `hdfs-site` 分类。	1 适用于预置少于四个核心节点的集群

参数	定义	默认值
		2 适用于预置少于 10 个核心节点的集群 3 适用于所有其他集群

Amazon EMR 上的 HDFS 中的透明加密

透明加密是通过使用 HDFS 加密区域（您定义的 HDFS 路径）实现的。每个加密区域都有其自己的密钥 (存储在使用 `hdfs-site` 配置分类指定的密钥服务器中)。

从 Amazon EMR 发行版 4.8.0 开始，您可以使用 Amazon EMR 安全配置更轻松地为集群配置数据加密设置。安全配置提供用于为 Amazon S3 中 Amazon Elastic Block Store（Amazon EBS）存储卷和 EMRFS 数据中的传输中的数据和静态数据增强安全性的设置。有关更多信息，请参阅《Amazon EMR 管理指南》中的加密传输中的数据和静态数据。

默认情况下，Amazon EMR 使用 Hadoop KMS；不过您可以使用其它实现 KeyProvider API 操作的 KMS。HDFS 加密区域中的每个文件都有其自己唯一的数据加密密钥（通过加密区域密钥加密）。当 HDFS 数据写入加密区域时，将对数据进行端到端加密 (静态和传输中)。因为加密和解密活动仅在客户端中进行。

您无法在加密区域之间移动文件，也无法将文件从加密区域移至未加密路径。

NameNode 和 HDFS 客户端通过 KeyProvider API 操作与 Hadoop KMS (或您配置的替代 KMS) 交互。KMS 负责将加密密钥存储在后备密钥存储中。此外，Amazon EMR 包含 JCE 无限制强度策略，以便您能够创建具有所需长度的密钥。

有关更多信息，请参阅 Hadoop 文档中的 HDFS 中的透明加密。

> **Note**
>
> 在 Amazon EMR 中，对于 Hadoop KMS，默认不启用通过 HTTPS 的 KMS。有关如何启用通过 HTTPS 的 KMS 的更多信息，请参阅 Hadoop KMS 文档。

配置 HDFS 透明加密

您可通过创建密钥并添加加密区域在 Amazon EMR 中配置透明加密。有几种方式可以实现：

- 在创建集群时使用 Amazon EMR 配置 API 操作
- 使用 Hadoop JAR 步骤与 command-runner.jar
- 登录到 Hadoop 集群的主节点并使用 `hadoop key` 和 `hdfs crypto` 命令行客户端
- 对 Hadoop KMS 和 HDFS 使用 REST API

有关 REST API 的更多信息，请参阅 Hadoop KMS 和 HDFS 各自的文档。

使用 CLI 在创建集群时创建加密区域及其密钥

配置 API 操作中的 `hdfs-encryption-zones` 分类允许您在创建集群时指定密钥名称和加密区。Amazon EMR 在您的集群的 Hadoop KMS 中创建此密钥并配置加密区域。

- 使用以下命令创建集群。

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge --
instance-count 2 \
--applications Name=App1 Name=App2 --configurations https://s3.amazonaws.com/
mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

`myConfig.json`:

```
[
  {
    "Classification": "hdfs-encryption-zones",
    "Properties": {
      "/myHDFSPath1": "path1_key",
      "/myHDFSPath2": "path2_key"
    }
  }
```

```
]
```

在主节点上手动创建加密区域及其密钥

1. 使用高于 4.1.0 的 Amazon EMR 发行版启动您的集群。
2. 使用 SSH 连接到集群的主节点。
3. 在 Hadoop KMS 中创建密钥。

```
$ hadoop key create path2_key
path2_key has been successfully created with options Options{cipher='AES/CTR/
NoPadding', bitLength=256, description='null', attributes=null}.
KMSClientProvider[http://ip-x-x-x-x.ec2.internal:16000/kms/v1/] has been updated.
```

> **Important**
>
> Hadoop KMS 要求您的密钥名称为小写。如果您使用的密钥包含大写字符，则您的集群将在启动过程中失败。

4. 在 HDFS 中创建加密区域路径。

```
$ hadoop fs -mkdir /myHDFSPath2
```

5. 使用您创建的密钥使 HDFS 路径成为加密区域。

```
$ hdfs crypto -createZone -keyName path2_key -path /myHDFSPath2
Added encryption zone /myHDFSPath2
```

使用 Amazon CLI 手动创建加密区域及其密钥

- 使用以下命令添加步骤以手动创建 KMS 密钥和加密区域。

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF --steps Type=CUSTOM_JAR,Name="Create
 First Hadoop KMS Key",Jar="command-runner.jar",ActionOnFailure=CONTINUE,Args=[/
bin/bash,-c,"\"hadoop key create path1_key\""] \
Type=CUSTOM_JAR,Name="Create First Hadoop HDFS Path",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[/bin/bash,-c,"\"hadoop fs -mkdir /
myHDFSPath1\""] \
```

```
Type=CUSTOM_JAR,Name="Create First Encryption Zone",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[/bin/bash,-c,"\"hdfs crypto -createZone
 -keyName path1_key -path /myHDFSPath1\""] \
Type=CUSTOM_JAR,Name="Create Second Hadoop KMS Key",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[/bin/bash,-c,"\"hadoop key create
 path2_key\""] \
Type=CUSTOM_JAR,Name="Create Second Hadoop HDFS Path",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[/bin/bash,-c,"\"hadoop fs -mkdir /
myHDFSPath2\""] \
Type=CUSTOM_JAR,Name="Create Second Encryption Zone",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[/bin/bash,-c,"\"hdfs crypto -createZone
 -keyName path2_key -path /myHDFSPath2\""]
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^) 。

HDFS 透明加密注意事项

最佳实践是为可能写入文件的每个应用程序创建一个加密区域。此外，您可使用配置 API 中的 hdfs-encryption-zones 分类加密所有 HDFS，并指定根路径 (/) 作为加密区域。

Hadoop 密钥管理服务器

Hadoop KMS 是一个密钥管理服务器，可用于为 Hadoop 集群实施加密服务，并且可充当 Amazon EMR 上的 HDFS 中的透明加密 的密钥供应商。如果您在启动 EMR 集群时选择 Hadoop 应用程序，则默认情况下将安装并启用 Amazon EMR 中的 Hadoop KMS。Hadoop KMS 不存储密钥 (临时缓存时除外)。Hadoop KMS 充当密钥提供程序和客户端受信任者之间的代理服务，它不是密钥库。为 Hadoop KMS 创建的默认密钥存储是 Java Cryptography Extension KeyStore (JCEKS)。还包括 JCE 无限制强度策略，以便您能够创建具有所需长度的密钥。Hadoop KMS 还支持一系列 ACL，后者独立于其它客户端应用程序 (如 HDFS) 控制对密钥和密钥操作的访问。Amazon EMR 中的默认密钥长度为 256 位。

要配置 Hadoop KMS，请使用 hadoop-kms-site 分类来更改设置。要配置 ACL，请使用分类 kms-acls。

有关更多信息，请参阅 Hadoop KMS 文档。Hadoop KMS 用于 Hadoop HDFS 透明加密。要了解有关 HDFS 透明加密的更多信息，请参阅 Apache Hadoop 文档中的 HDFS 透明加密主题。

> **Note**
>
> 在 Amazon EMR 中，对于 Hadoop KMS，默认不启用通过 HTTPS 的 KMS。要了解如何通过 HTTPS 启用 KMS，请参阅 Hadoop KMS 文档。

> **Important**
>
> Hadoop KMS 要求您的密钥名称为小写。如果您使用的密钥包含大写字符，则您的集群将在启动过程中失败。

在 Amazon EMR 中配置 Hadoop KMS

如果使用 Amazon EMR 发行版 4.6.0 或更高版本，`kms-http-port` 为 9700，`kms-admin-port` 为 9701。

您可使用 Amazon EMR 发行版的配置 API 在创建集群时配置 Hadoop KMS。下面是对 Hadoop KMS 可用的配置对象分类：

Hadoop KMS 配置分类

分类	文件名
hadoop-kms-site	`kms-site.xml`
hadoop-kms-acls	`kms-acls.xml`
hadoop-kms-env	`kms-env.sh`
hadoop-kms-log4j	`kms-log4j.properties`

使用 CLI 设置 Hadoop KMS ACL

- 通过以下命令使用带 ACL 的 Hadoop KMS 创建集群：

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge --
instance-count 2 \
```

```
--applications Name=App1 Name=App2 --configurations https://s3.amazonaws.com/
mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

myConfig.json:

```
[
    {
      "Classification": "hadoop-kms-acls",
      "Properties": {
        "hadoop.kms.blacklist.CREATE": "hdfs,foo,myBannedUser",
        "hadoop.kms.acl.ROLLOVER": "myAllowedUser"
      }
    }
  ]
```

使用 CLI 禁用 Hadoop KMS 缓存

- 要在 Hadoop KMS hadoop.kms.cache.enable 设置为 false 的情况下创建集群，请使用以下命令：

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge --
instance-count 2 \
--applications Name=App1 Name=App2 --configurations https://s3.amazonaws.com/
mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

myConfig.json:

```
[
    {
      "Classification": "hadoop-kms-site",
      "Properties": {
        "hadoop.kms.cache.enable": "false"
      }
    }
  ]
```

使用 CLI 在 **kms-env.sh** 脚本中设置环境变量

- 通过 kms-env.sh 配置更改 hadoop-kms-env 中的设置。使用以下命令通过 Hadoop KMS 创建集群：

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge --
instance-count 2 \
--applications Name=App1 Name=App2 --configurations https://s3.amazonaws.com/
mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

myConfig.json:

```
[
  {
    "Classification": "hadoop-kms-env",
    "Properties": {
    },
    "Configurations": [
      {
        "Classification": "export",
        "Properties": {
          "JAVA_LIBRARY_PATH": "/path/to/files",
          "KMS_SSL_KEYSTORE_FILE": "/non/Default/Path/.keystore",
          "KMS_SSL_KEYSTORE_PASS": "myPass"
```

```
        },
        "Configurations": [
        ]
      }
    ]
  }
]
```

有关配置 Hadoop KMS 的信息，请参阅 Hadoop KMS 文档。

具有多个主节点 的 EMR 集群上的 HDFS 透明加密

Apache Ranger KMS 可在具有多个主节点的 Amazon EMR 集群中使用，以在 HDFS 中进行透明加密。

Apache Ranger KMS 将其根密钥和加密区域（EZ）密钥存储在 Amazon RDS 中，用于具有多个主节点的 Amazon EMR 集群。要在具有多个主节点的 Amazon EMR 集群上的 HDFS 中启用透明加密，必须提供以下配置。

- Amazon RDS 或您自己的 MySQL 服务器连接 URL，用于存储 Ranger KMS 根密钥和 EZ 密钥
- MySQL 的用户名和密码
- Ranger KMS 根密钥的密码
- 用于到 MySQL 服务器的 SSL 连接的凭证颁发机构 (CA) PEM 文件

您可以使用 `ranger-kms-dbks-site` 分类和 `ranger-kms-db-ca` 分类提供这些配置，如以下示例所示。

```
[
  {
    "Classification": "ranger-kms-dbks-site",
    "Properties": {
      "ranger.ks.jpa.jdbc.url": "jdbc:log4jdbc:mysql://mysql-host-url.xx-
xxx-1.xxx.amazonaws.com:3306/rangerkms",
      "ranger.ks.jpa.jdbc.user": "mysql-user-name",
      "ranger.ks.jpa.jdbc.password": "mysql-password",
      "ranger.db.encrypt.key.password": "password-for-encrypting-a-master-key"
    }
  },
  {
```

```
    "Classification": "ranger-kms-db-ca",
    "Properties": {
      "ranger.kms.trust.ca.file.s3.url": "s3://rds-downloads/rds-ca-2019-root.pem"
    }
     }
]
```

以下是 Apache Ranger KMS 的配置对象分类。

Hadoop KMS 配置分类

分类	描述
ranger-kms-dbks-site	更改 Ranger KMS 的 dbks-site.xml 文件中的值。
ranger-kms-site	更改 Ranger KMS 的 ranger-kms-site.xml 文件中的值。
ranger-kms-env	更改 Ranger KMS 环境中的值。
ranger-kms-log4j	更改 Ranger KMS 的 kms-log4j.properties 文件中的值。
ranger-kms-db-ca	更改 S3 上用于与 Ranger KMS 进行 MySQL SSL 连接的 CA 文件的值。

注意事项

- 强烈建议您加密 Amazon RDS 实例以提高安全性。有关更多信息，请参阅加密 Amazon RDS 资源概览。
- 强烈建议您为每个具有多个主节点的 Amazon EMR 集群使用单独的 MySQL 数据库以提高安全性。
- 要在具有多个主节点的 Amazon EMR 集群上的 HDFS 中配置透明加密，必须在创建集群时指定 `hdfs-encryption-zones` 分类。否则，Ranger KMS 将不会配置或启动。具有多个主节点的 Amazon EMR 集群上不支持在运行的集群上重新配置 `hdfs-encryption-zones` 分类或任何 Hadoop KMS 配置分类。

创建或运行 Hadoop 应用程序

主题

使用 Amazon EMR 构建二进制文件

您可以使用 Amazon EMR 作为构建环境，以编译用于您的集群的程序。在 Amazon EMR 中使用的程序必须在运行 Linux 的系统上进行编译且 Linux 版本与 Amazon EMR 所用的相同。对于 32 位版本，您应在 32 位机器上或在打开 32 位交叉编译选项的情况下进行编译。对于 64 位版本，您需要在 64 位机器上编译或打开 64 位交叉编译选项。有关 EC2 实例版本的更多信息，请参阅《Amazon EMR 管理指南》中的计划和配置 EC2 实例。支持的编程语言包括 C++、Python 和 C#。

下表概览了使用 Amazon EMR 构建和测试您的应用程序所涉及的步骤。

构建模块的过程

1	连接到集群的主节点。
2	将源文件复制到主节点。
3	使用任何必要的优化方法构建二进制文件。
4	将二进制文件从主节点复制到 Amazon S3。

每个步骤的详细信息请参阅下面的部分。

连接到集群的主节点

- 按照《Amazon EMR 管理指南》中的使用 SSH 连接到主节点中的说明操作。

将源文件复制到主节点。

1. 将源文件置于 Amazon S3 存储桶中。要了解如何创建存储桶以及如何将数据移至 Amazon S3，请参阅《Amazon Simple Storage Service 用户指南》。

2. 通过输入类似于以下内容的命令，为您的源文件在 Hadoop 集群上创建文件夹：

```
mkdir SourceFiles
```

3. 通过键入类似以下内容的命令，将您的源文件从 Amazon S3 复制到主节点：

```
hadoop fs -get s3://mybucket/SourceFiles SourceFiles
```

使用任何必要的优化构建二进制文件

构建二进制文件的方式取决于多种因素。请按照具体构建工具的说明，设置和配置您的环境。您可以使用 Hadoop 系统规范命令获取集群信息，以确定如何安装您的构建环境。

识别系统规范

- 使用以下命令验证用于构建二进制文件的架构。

  a. 要查看 Debian 版本，请输入以下命令：

  ```
  master$ cat /etc/issue
  ```

  该输出值看上去类似于以下内容。

  ```
  Debian GNU/Linux 5.0
  ```

  b. 要查看公有 DNS 名称和处理器大小，请输入以下命令：

  ```
  master$ uname -a
  ```

  该输出值看上去类似于以下内容。

  ```
  Linux domU-12-31-39-17-29-39.compute-1.internal 2.6.21.7-2.fc8xen #1 SMP Fri
   Feb 15 12:34:28 EST 2008 x86_64 GNU/Linux
  ```

  c. 要查看处理器速度，请输入以下命令：

  ```
  master$ cat /proc/cpuinfo
  ```

  该输出值看上去类似于以下内容。

```
processor : 0
vendor_id : GenuineIntel
model name : Intel(R) Xeon(R) CPU E5430 @ 2.66GHz
flags : fpu tsc msr pae mce cx8 apic mca cmov pat pse36 clflush dts acpi mmx
 fxsr sse sse2 ss ht tm syscall nx lm constant_tsc pni monitor ds_cpl vmx est
 tm2 ssse3 cx16 xtpr cda lahf_lm
...
```

构建了二进制文件后，您就可以将文件复制到 Amazon S3。

将二进制文件从主节点复制到 Amazon S3

- 键入以下命令以将二进制文件复制到 Amazon S3 存储桶：

```
hadoop fs -put BinaryFiles s3://mybucket/BinaryDestination
```

通过流式处理来处理数据

Hadoop Streaming 是 Hadoop 附带的一种实用工具，可让您以非 Java 语言开发 MapReduce 可执行文件。流式处理是以 JAR 文件的形式实现的，这样您就可以像运行标准 JAR 文件一样，从 Amazon EMR API 或命令行运行它。

此部分介绍如何结合使用 Streaming 与 Amazon EMR。

> **Note**
>
> Apache Hadoop Streaming 是一种独立工具。因此，这里并不介绍其所有函数和参数。有关 Hadoop Streaming 的更多信息，请转到 http://hadoop.apache.org/docs/stable/hadoop-streaming/HadoopStreaming.html。

使用 Hadoop Streaming 实用工具

此部分介绍如何使用 Hadoop 的 Streaming 实用工具。

Hadoop 进程

1	以您所选择的编程语言编写映射器和 Reducer 可执行文件。

	按照 Hadoop 的文档中的指示编写流式处理可执行文件。该等程序应从标准输入读取其输入内容，并通过标准输出来输出数据。默认情况下，输入/输出的每一行都代表一条记录，并且每一行中的第一个制表符都用作密钥与值之间的分隔符。
2	在本地测试您的可执行文件，并将它们上载到 Amazon S3。
3	使用 Amazon EMR 命令行界面或 Amazon EMR 控制台可运行您的应用程序。

每个映射器脚本都会以单独进程的形式在集群中启动。每个 Reducer 可执行文件都会通过任务流程将映射器可执行文件的输出转到数据输出中。

大多数 Streaming 应用程序都需要 `input`、`output`、`mapper` 和 `reducer` 参数。下表描述了上述参数和其它可选参数。

参数	描述	必填
-input	输入数据在 Amazon S3 上的位置。 类型：字符串 默认值：无 约束：URI。如果没有指定协议，那么它就可以使用集群的默认文件系统。	是
-output	Amazon S3 上的位置，该位置为 Amazon EMR 上载已处理数据的地方。 类型：字符串 默认值：无 约束：URI 默认值：如果没有指定位置，那么 `input` 会将数据上载至 Amazon EMR 指定的位置。	是
-mapper	映射器可执行文件的名称。	是

参数	描述	必填
	类型：字符串 默认值：无	
-reducer	Reducer 可执行文件的名称。 类型：字符串 默认值：无	是
-cacheFile	一个 Amazon S3 位置，其中包含一些文件可供 Hadoop 复制到本地工作目录（主要目的是提高性能）。 类型：字符串 默认值：无 约束：[URI]#[要在工作目录中创建的符号链接名称]	否
-cacheArchive	提取到工作目录的 JAR 文件。 类型：字符串 默认值：无 约束：[URI]#[要在工作目录中创建的符号链接目录名称]	否
-combiner	合并结果 类型：字符串 默认值：无 约束：Java 类名	否

以下示例代码是写入 Python 的映射器可执行文件。此脚本是 WordCount 示例应用程序的一部分。

```
#!/usr/bin/python
import sys

def main(argv):
  line = sys.stdin.readline()
  try:
    while line:
      line = line.rstrip()
      words = line.split()
      for word in words:
        print "LongValueSum:" + word + "\t" + "1"
      line = sys.stdin.readline()
  except "end of file":
    return None
if __name__ == "__main__":
  main(sys.argv)
```

提交流式处理步骤

本节介绍向集群提交流式处理步骤的基本知识。Streaming 应用程序会从标准输入读取输入内容，然后针对每个输入运行脚本或可执行文件（称为映射器）。每个输入的结果都会保存在本地，通常位于 Hadoop Distributed File System（HDFS）分区上。所有输入经过映射器处理后，第二个脚本或可执行文件 (名为 Reducer) 会处理映射器结果。将 Reducer 的结果发送到标准输出。您可以将一系列 Streaming 步骤串联起来，让一个步骤的输出作为另一个步骤的输入。

映射器和 Reducer 都能够以文件的形式进行引用，或者您也可以提供一个 Java 类。您能够以任一种受支持的语言 (包括 Ruby、Perl、Python、PHP 或 Bash) 来执行映射器和 Reducer。

使用控制台提交流式处理步骤

此示例介绍如何使用 Amazon EMR 控制台向正在运行的集群提交流式处理步骤。

提交流式处理步骤

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 在 Cluster List (集群列表) 中，选择您的集群的名称。
3. 滚动到 Steps (步骤) 部分并展开它，然后选择 Add step (添加步骤)。
4. 在 Add Step (添加步骤) 对话框中：
   - 对于 Step type (步骤类型)，选择 Streaming program (流式程序)。

- 对于 Name (名称)，请接受默认名称（流式程序）或键入新名称。
- 对于映射器，键入或浏览到 Hadoop 中映射器类所在的位置或映射器可执行文件（如 Python 程序）所在的 S3 存储桶。该路径值的形式必须是：*BucketName*/*path*/*MapperExecutable*。
- 对于 Reducer，键入或浏览到 Hadoop 中 Reducer 类所在的位置或 Reducer 可执行文件（如 Python 程序）所在的 S3 存储桶。该路径值的形式必须是：*BucketName*/*path*/*MapperExecutable*。Amazon EMR 支持特殊 aggregate 关键字。有关更多信息，请转到 Hadoop 提供的 Aggregate 库。
- 对于 Input S3 location (输入 S3 位置)，键入或浏览到输入数据的位置。
- 对于 Output S3 location (输出 S3 位置)，键入或浏览到您的 Amazon S3 输出存储桶的名称。
- 对于 Arguments (参数)，将该字段保留为空白。
- 对于 Action on failure (出现故障时的操作)，接受默认选项 Continue (继续)。

5. 选择 Add（添加）。步骤会出现在控制台中，其状态为“Pending”。
6. 步骤的状态会随着步骤的运行从“Pending”变为“Running”，再变为“Completed”。要更新状态，请选择 Actions (操作) 列上方的 Refresh (刷新) 图标。

Amazon CLI

这些示例演示如何使用 Amazon CLI 创建集群并提交流式处理步骤。

使用 Amazon CLI 创建集群并提交 Streaming 步骤

- 若要使用 Amazon CLI 创建集群并提交 Streaming 步骤，请键入以下命令，将 *myKey* 替换为您的 EC2 密钥对的名称。请注意，`--files` 的实际参数应该是指向您脚本位置的 Amazon S3 路径，并且 `-mapper` 和 `-reducer` 的实际参数应该是各自脚本文件的名称。

```
aws emr create-cluster --name "Test cluster" --release-label emr-5.36.1 --
applications Name=Hue Name=Hive Name=Pig --use-default-roles \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge --instance-count 3 \
--steps Type=STREAMING,Name="Streaming Program",ActionOnFailure=CONTINUE,Args=[--
files,pathtoscripts,-mapper,mapperscript,-reducer,reducerscript,aggregate,-
input,pathtoinputdata,-output,pathtooutputbucket]
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\)。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

如果不使用 `--instance-groups` 参数指定实例计数，则将启动单个主节点，其余实例将作为核心节点启动。所有节点都使用该命令中指定的实例类型。

> **Note**
>
> 如果您之前未创建默认 Amazon EMR 服务角色和 EC2 实例配置文件，请先键入 aws emr `create-default-roles` 创建它们，然后再键入 `create-cluster` 子命令。

有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅https://docs.amazonaws.cn/cli/latest/reference/emr。

使用自定义 JAR 处理数据

自定义 JAR 运行您能上载到 Amazon S3 的已编译 Java 程序。您应针对想启动的 Hadoop 版本编译该程序，并将 `CUSTOM_JAR` 步骤提交到 Amazon EMR 集群。有关如何编译 JAR 文件的更多信息，请参阅 使用 Amazon EMR 构建二进制文件。

有关构建 Hadoop MapReduce 应用程序的更多信息，请参阅 Apache Hadoop 文档中的 MapReduce Tutorial。

主题

提交自定义 JAR 步骤

自定义 JAR 运行您能上载到 Amazon S3 的已编译 Java 程序。您应针对想启动的 Hadoop 版本编译该程序，并将 `CUSTOM_JAR` 步骤提交到 Amazon EMR 集群。有关如何编译 JAR 文件的更多信息，请参阅 使用 Amazon EMR 构建二进制文件。

有关构建 Hadoop MapReduce 应用程序的更多信息，请参阅 Apache Hadoop 文档中的 MapReduce Tutorial。

此部分介绍在 Amazon EMR 中提交自定义 JAR 步骤的基础知识。通过提交自定义 JAR 步骤，您可以使用 Java 编程语言编写用于处理数据的脚本。

使用控制台提交自定义 JAR 步骤

此示例介绍如何使用 Amazon EMR 控制台向正在运行的集群提交自定义 JAR 步骤。

使用控制台提交自定义 JAR 步骤

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 在 Cluster List (集群列表) 中，选择您的集群的名称。
3. 滚动到 Steps (步骤) 部分并展开它，然后选择 Add step (添加步骤)。
4. 在 Add step (添加步骤) 对话框中：
   - 对于步骤类型，选择自定义 JAR。
   - 对于 Name (名称)，接受默认名称（自定义 JAR）或键入新名称。
   - 对于 JAR S3 location (JAR S3 位置)，键入或浏览到 JAR 文件的位置。JAR 位置可能是 S3 的路径或类路径中的完全限定的 java 类。
   - 对于参数，以空格分隔的字符串形式键入任何所需参数，或将该字段保留为空白。
   - 对于 Action on failure (出现故障时的操作)，接受默认选项 Continue (继续)。
5. 选择 Add（添加）。步骤会出现在控制台中，其状态为“Pending”。
6. 步骤的状态会随着步骤的运行从“Pending”变为“Running”，再变为“Completed”。要更新状态，请选择 Actions (操作) 列上方的 Refresh (刷新) 图标。

使用 Amazon CLI 启动集群并提交自定义 JAR 步骤

使用 Amazon CLI 启动集群并提交自定义 JAR 步骤

要使用 Amazon CLI 启动集群并提交自定义 JAR 步骤，请键入带 `--steps` 参数的 `create-cluster` 子命令。

- 要启动集群并提交自定义 JAR 步骤，请键入以下命令，并将 *myKey* 替换为您的 EC2 密钥对的名称，将 *mybucket* 替换为您的存储桶名称。

```
aws emr create-cluster --name "Test cluster" --release-label emr-5.36.1 \
--applications Name=Hue Name=Hive Name=Pig --use-default-roles \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge --instance-count 3 \
--steps Type=CUSTOM_JAR,Name="Custom JAR
 Step",ActionOnFailure=CONTINUE,Jar=pathtojarfile,Args=["pathtoinputdata","pathtooutputbuck
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

如果不使用 `--instance-groups` 参数指定实例计数，则将启动单个主节点，其余实例将作为核心节点启动。所有节点都使用您在命令中指定的实例类型。

> **Note**
>
> 如果您之前未创建默认 Amazon EMR 服务角色和 EC2 实例配置文件，请先键入 `aws emr create-default-roles` 创建它们，然后再键入 `create-cluster` 子命令。

有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅https://docs.amazonaws.cn/cli/latest/reference/emr。

第三方依赖项

有时，可能需要包含 MapReduce 类路径 JAR 以与您的程序结合使用。您有两个选项来执行此操作：

- 将 `--libjars s3://`*`URI_to_JAR`* 包含在使用 Amazon CLI 启动集群并提交自定义 JAR 步骤中的过程的步骤选项中。
- 使用 `mapred-site.xml` 中修改过的 `mapreduce.application.classpath` 设置启动集群。使用 `mapred-site` 配置分类。要通过使用 Amazon CLI 的步骤创建集群，内容如下所示：

```
aws emr create-cluster --release-label emr-5.36.1 \
--applications Name=Hue Name=Hive Name=Pig --use-default-roles \
--instance-type m5.xlarge --instance-count 2  --ec2-attributes KeyName=myKey \
--steps Type=CUSTOM_JAR,Name="Custom JAR
 Step",ActionOnFailure=CONTINUE,Jar=pathtojarfile,Args=["pathtoinputdata","pathtooutputbucket
 \
--configurations https://s3.amazonaws.com/mybucket/myfolder/myConfig.json
```

`myConfig.json`:

```
[
```

```
  {
    "Classification": "mapred-site",
    "Properties": {
      "mapreduce.application.classpath": "path1,path2"
    }
  }
]
```

路径的逗号分隔的列表应追加到每个任务的 JVM 的类路径。

为 YARN 容器开启非统一内存访问感知功能

在 Amazon EMR 6.x 及更高版本中，您可以使用非统一内存访问（NUMA）对集群上的数据进行多处理。NUMA 是一种计算机内存设计模式，在这种模式中，处理器访问自身本地内存的速度比访问其他处理器上的内存或处理器之间共享的存储器更快。YARN 容器使用 NUMA 时性能更好，因为这些容器可以绑定到为所有后续内存分配提供服务的特定 NUMA 节点。这可以减少集群访问远程内存的次数。

当 Worker 节点计算机属于多 NUMA 节点时，可以为 YARN 容器启用 NUMA 支持。要确认 Worker 节点是单 NUMA 还是多 NUMA 节点，请运行以下命令。

```
lscpu | grep -i numa
NUMA node(s): 2
```

通常，大于 12x 的实例有两个 NUMA 节点。这不适用于裸机实例。

为 YARN 容器开启 NUMA 感知功能

1. 在您的 Amazon EMR 6.x 集群中使用以下 `yarn-site` 配置。

```
 [
    {
        "classification":"yarn-site",
            "properties":{
                "yarn.nodemanager.linux-container-executor.nonsecure-mode.local-
user":"yarn",
                "yarn.nodemanager.linux-container-executor.group":"yarn",
                "yarn.nodemanager.container-
executor.class":"org.apache.hadoop.yarn.server.nodemanager.LinuxContainerExecutor",
                "yarn.nodemanager.numa-awareness.enabled":"true",
                "yarn.nodemanager.numa-awareness.numactl.cmd":"/usr/bin/numactl",
                "yarn.nodemanager.numa-awareness.read-topology":"true"
```

```
            },
        "configurations":[]
    }
]
```

2. 在集群中提供以下引导操作。

```
#!/bin/bash

sudo yum -y install numactl
echo 1 | sudo tee /proc/sys/kernel/numa_balancing

echo "banned.users=mapred,bin,hdfs" >> /etc/hadoop/conf/container-executor.cfg
rm -rf /var/log/hadoop-yarn/
sudo chown -R yarn:hadoop /var/log/hadoop-yarn/
sudo chmod 755 -R /var/log/hadoop-yarn/

sudo chmod 6050 /etc/hadoop/conf/container-executor.cfg

mkdir /mnt/yarn && sudo chmod 755 -R /mnt/yarn && sudo chown -R yarn:hadoop /mnt/
yarn
mkdir /mnt1/yarn && sudo chmod 755 -R /mnt1/yarn && sudo chown -R yarn:hadoop /
mnt1/yarn
mkdir /mnt2/yarn && sudo chmod 755 -R /mnt2/yarn && sudo chown -R yarn:hadoop /
mnt2/yarn
```

3. 每个容器都必须能够感知 NUMA。您可以使用 NUMA 标志通知每个容器中的 Java 虚拟机（JVM）。例如，要通知 JVM 在 MapReduce 作业中使用 NUMA，请在 `mapred-site.xml` 中添加以下属性。

```
<property>
    <name>mapreduce.reduce.java.opts</name>
    <value>-XX:+UseNUMA</value>
</property>
<property>
    <name>mapreduce.map.java.opts</name>
    <value>-XX:+UseNUMA</value>
</property>
```

4. 要验证您是否已开启 NUMA，请使用以下命令搜索任何 NodeManager 日志文件。

```
grep "NUMA resources allocation is enabled," *
```

要验证 NodeManager 是否已将 NUMA 节点资源分配给容器，请使用以下命令搜索 NodeManager 日志（将 *<container_id>* 替换为您自己的容器 ID）。

```
grep "NUMA node" | grep <container_id>
```

Hadoop 版本历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Hadoop 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Hadoop 版本信息

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.14.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.13.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
		cemanager, hadoop-yarn-timeline-server
emr-6.12.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.11.1	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.11.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.10.1	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.10.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.9.1	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.9.0	3.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.8.1	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.8.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.7.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.36.1	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.36.0	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.6.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.35.0	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.5.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.4.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.3.1	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.3.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.2.1	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.2.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.1.1	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.1.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-6.0.1	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-6.0.0	3.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.34.0	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.33.1	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.33.0	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.32.1	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.32.0	2.10.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.31.1	2.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.31.0	2.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.30.2	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.30.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.30.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.29.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.28.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.28.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.27.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.27.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.26.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.25.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.24.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.24.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.23.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.23.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.22.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.21.2	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.21.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.21.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.20.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.20.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.19.1	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.19.0	2.8.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.18.1	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.18.0	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.17.2	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.17.1	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.17.0	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.16.1	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.16.0	2.8.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.15.1	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.15.0	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.14.2	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.14.1	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.14.0	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.13.1	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.13.0	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.12.3	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.12.2	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.12.1	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.12.0	2.8.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.11.4	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.11.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.11.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.11.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.11.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.10.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.10.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.9.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.9.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.8.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.8.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.8.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.8.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.7.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.7.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.6.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.6.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server
emr-5.5.4	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.5.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.5.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.5.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.5.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.4.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.4.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.3.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.3.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.3.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.2.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.2.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.2.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.2.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.1.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.1.0	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-5.0.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-5.0.0	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.9.6	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.9.5	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.9.4	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.9.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.9.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.9.1	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.8.5	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.8.4	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.8.3	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.8.2	2.7.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.8.0	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.7.4	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.7.2	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.7.1	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.7.0	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.6.0	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.5.0	2.7.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.4.0	2.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.3.0	2.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Amazon EMR 发行版标签	Hadoop 版本	随 Hadoop 安装的组件
emr-4.2.0	2.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.1.0	2.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager
emr-4.0.0	2.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager

Hadoop 发布说明（按版本分类）

Amazon EMR 6.6.0 - Hadoop 发布说明

Amazon EMR 6.6.0 - Hadoop 发布说明

Amazon EMR 6.6.0 - Hadoop 更改

类型	描述
错误	修复了读取 BZip2 文本文件时出现重复记录的问题。
逆向移植	HADOOP-18136：验证 `FileUtils.unTar()` 对缺失的 .tar 文件的处理
逆向移植	HADOOP-17627：逆向移植到 branch-3.2 HADOOP-17371、HADOOP-17621、HADOOP-17625 以将 Jetty 更新到 9.4.39
逆向移植	HADOOP-17655：将 Jetty 升级至 9.4.40
逆向移植	HADOOP-17796：将 jetty 版本升级到 9.4.43
逆向移植	HADOOP-17661：mvn versions:set 无法解析 pom.xml
逆向移植	HADOOP-17236：将 snakeyaml 升级至 1.26 以缓解 CVE-2017-18640
逆向移植	HADOOP-16717：删除 GenericsUtil isLog4jLogger 对 Log4jLoggerAdapter 的依赖性
逆向移植	HADOOP-17633：由于 CVE，将 json-smart 升级至 2.4.2 并将 nimbus-jose-jwt 升级至 9.8
逆向移植	HADOOP-17844：将 JSON smart 升级至 2.4.7
逆向移植	HADOOP-17972：逆向移植 branch-3.2 的 HADOOP-17683（将 commons-io 更新至 2.8.0）
逆向移植	HADOOP-16555：将 commons-compress 更新至 1.19

类型	描述
逆向移植	HADOOP-17370：将 commons-compress 升级到 1.21
逆向移植	HADOOP-17096：修复 ZStandardCompressor 输入缓冲区偏移
逆向移植	HADOOP-17112：通过提交程序将文件保存到 s3a 时，路径中不允许有空格
逆向移植	HADOOP-13500: 同步配置属性对象的迭代
逆向移植	HDFS-14099：在 ZStandardDecompressor 中解压缩多个帧时的未知帧描述符
逆向移植	HDFS-16410: OfflineEditsXmlLoader 中不安全的 Xml 解析
逆向移植	HDFS-14498：LeaseManager 可以在创建失败的文件上永久循环
逆向移植	HDFS-15290：NameNode 启动期间 HttpServer 中的 NPE
逆向移植	HDFS-15293：当接收检查点时放宽接收 fsimage 的条件
逆向移植	HDFS-12979：StandbyNode 应在检查点检查后将 FsImage 上传到 ObserverNode
逆向移植	YARN-10538：将重新调试节点添加到返回到 AM 的更新节点列表中
逆向移植	YARN-10472：将 YARN-10314（YarnClient 仅使用覆盖客户端的 jar 抛出 WebSocketException 的 NoClassDefFoundError）逆向移植到 branch-3.2

类型	描述
逆向移植	YARN-9968：由于 NullPointerException 的原因，Public Localizer 正在退出 NodeManager
逆向移植	YARN-10651：CapacityScheduler 崩溃，AbstractYarnScheduler.updateNodeResource() 中具有 NPE
逆向移植	YARN-9339：将应用程序移至新队列后，应用程序待处理指标不正确
逆向移植	YARN-10438：处理 ClientRMService#getContainerReport() 中的空 containerId
逆向移植	YARN-7266：如果 RollingLevelDb 文件损坏或丢失，ATS 1.5 将无法开启
逆向移植	YARN-9063：如果 RollingLevelDb 文件损坏或丢失，ATS 1.5 将无法开启
逆向移植	YARN-9848：恢复 YARN-4946（当日志聚合未处于终端状态时，RM 不应将应用程序视为 COMPLETED ）。

Apache HBase

HBase 是一种开源、非关系型分布式数据库，它作为 Apache 软件基金会的 Hadoop 项目的一部分开发。HBase 在 Hadoop Distributed File System（HDFS）上运行，为 Hadoop 生态系统提供非关系数据库功能。HBase 包含在 Amazon EMR 发行版 4.6.0 及更高版本中。

HBase 与 Hadoop 无缝协作，共享其文件系统，并充当 MapReduce 框架和执行引擎的直接输入和输出。HBase 还可与 Apache Hive 集成，可通过 HBase 表实现类似 SQL 的查询、与基于 Hive 的表连接以及对 Java 数据库连接（JDBC）的支持。有关 HBase 的更多信息，请参阅 Apache 网站上的 Apache HBase 和 HBase 文档。有关如何将 HBase 用于 Hive 的示例，请参阅Amazon大数据博客文章 Combine NoSQL and massively parallel analytics using Apache HBase and Apache Hive on Amazon EMR。

在 Amazon EMR 上使用 HBase，您还可将 HBase 数据直接备份到 Amazon Simple Storage Service（Amazon S3），并在启动 HBase 集群时从之前创建的备份还原。Amazon EMR 提供与 Amazon S3 集成的其它选项以实现数据持久性和灾难恢复。

- HBase on Amazon S3 – 对于 Amazon EMR 版本 5.2.0 及更高版本，您可使用 HBase on Amazon S3 将集群的 HBase 根目录和元数据直接存储到 Amazon S3。随后，您可以启动新集群，将其指向 Amazon S3 中的根目录位置。一次仅一个集群可使用 Amazon S3 中的 HBase 位置，只读副本集群例外。有关更多信息，请参阅HBase on Amazon S3（Amazon S3 存储模式）。
- HBase 只读副本 – 具有 HBase on Amazon S3 的 Amazon EMR 版本 5.7.0 及更高版本支持只读副本集群。在只读操作中，只读副本集群提供对主集群的存储文件和元数据的只读访问权限。有关更多信息，请参阅使用只读副本集群。
- HBase 快照 – 作为 HBase on Amazon S3 的替代方案，对于 EMR 版本 4.0 及更高版本，您可为直接传输至 Amazon S3 的 HBase 数据创建快照，然后使用快照恢复数据。有关更多信息，请参阅使用 HBase 快照。

> **Important**
>
> 对于 Amazon EMR HBase 集群扩展，不建议对 HBase 集群使用托管扩展或使用自定义策略进行扩展。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 HBase 版本，以及 Amazon EMR 随 HBase 一起安装的组件。

有关此发行版中随 HBase 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 HBase 版本信息

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.14.0	HBase 2.4.17	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

> **Note**
>
> Apache HBase HBCK2 是一个独立的操作工具，用于修复 HBase 区域和系统表。在 Amazon EMR 版本 6.1.0 及更高版本中，主节点 上的 `/usr/lib/hbase-operator-tools/` 中提供 hbase-hbck2.jar。有关如何生成和使用工具的更多信息，请参阅 HBase HBCK2。

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 HBase 版本，以及 Amazon EMR 随 HBase 一起安装的组件。

有关此发行版中随 HBase 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 HBase 版本信息

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.36.1	HBase 1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

主题

- 创建带 HBase 的集群
- HBase on Amazon S3（Amazon S3 存储模式）
- 使用 HBase shell
- 通过 Hive 访问 HBase 表
- 使用 HBase 快照
- 配置 HBase
- 查看 HBase 用户界面
- 查看 HBase 日志文件
- 使用 Ganglia 监控 HBase
- 从早期版本的 HBase 迁移
- HBase 发行版历史记录

创建带 HBase 的集群

此部分中的过程包含使用 Amazon Web Services Management Console和 Amazon CLI 启动集群的基础知识。有关如何计划、配置和启动 EMR 集群的详细信息，请参阅《Amazon EMR 管理指南》中的计划和配置集群。

使用控制台创建带 HBase 的集群

有关使用控制台启动集群的快速步骤，请参阅《Amazon EMR 管理指南》中的Amazon EMR 入门。

使用控制台启动安装了 HBase 的集群

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 依次选择 Create cluster (创建集群) 和 Go to advanced options (转到高级选项)。
3. 在 Software Configuration (软件配置) 中，选择 Amazon Release Version (亚马逊发行版) 4.6.0 或更高版本（建议使用最新版本）。根据需要选择 HBase 和其它应用程序。
4. 对于 Amazon EMR 版本 5.2.0 及更高版本，在 HBase Storage Settings (HBase 存储设置) 下，选择 HDFS 或 S3。有关更多信息，请参阅HBase on Amazon S3（Amazon S3 存储模式）。
5. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

使用 Amazon CLI 创建带 HBase 的集群

使用以下命令创建安装了 HBase 的集群：

```
aws emr create-cluster --name "Test cluster" --release-label emr-5.36.1 \
--applications Name=HBase --use-default-roles --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

如果您使用 HBase on Amazon S3，请指定`--configurations`选项以及对 JSON 配置对象的引用。配置对象必须包含一个 `hbase-site` 分类，此分类使用 `hbase.rootdir` 属性指定 HBase 数据在

Amazon S3 中的存储位置。它还必须包含 hbase 分类，其中使用 hbase.emr.storageMode 属性指定 s3。以下示例演示了具有这些配置设置的 JSON 代码段。

```
[
    {
        "Classification": "hbase-site",
        "Properties": {
            "hbase.rootdir": "s3://MyBucket/MyHBaseStore"
        }
    },
    {
        "Classification": "hbase",
        "Properties": {
            "hbase.emr.storageMode": "s3"
        }
    }
]
```

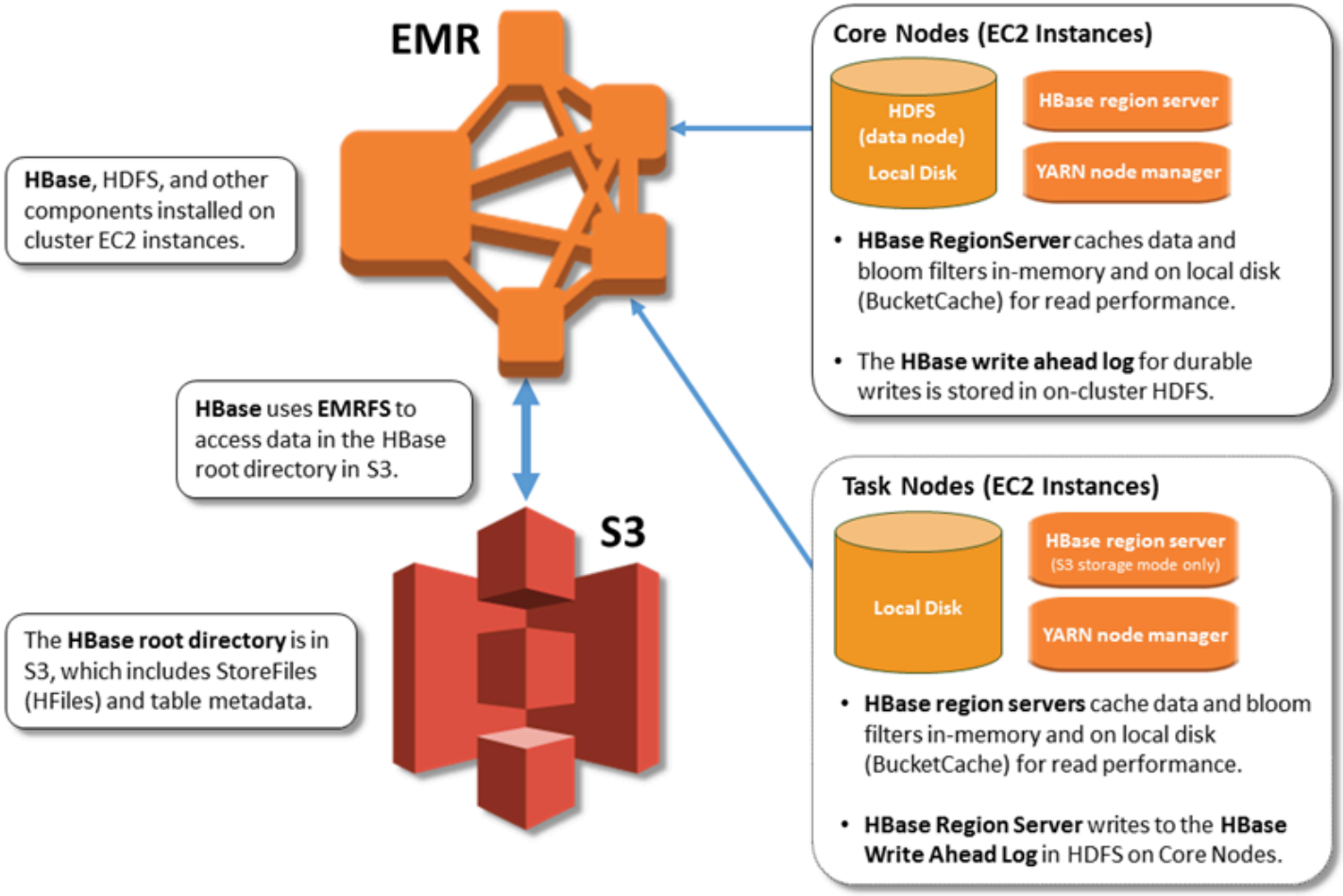
有关 HBase on Amazon S3 的更多信息，请参阅HBase on Amazon S3（Amazon S3 存储模式）。有关分类的更多信息，请参阅配置应用程序。

HBase on Amazon S3（Amazon S3 存储模式）

在 Amazon EMR 版本 5.2.0 及更高版本上运行 HBase 时，您可以启用 HBase on Amazon S3，这将提供以下优势：

- HBase 根目录存储在 Amazon S3 中，包括 HBase 存储文件和表元数据。此数据在集群外部持续存在且可跨 Amazon EC2 可用区访问，您无需使用快照或其它方法进行恢复。
- 对于 Amazon S3 中的存储文件，您可以针对计算要求而非数据要求调整 Amazon EMR 集群的大小（在 HDFS 上为 3 倍复制）。
- 使用 Amazon EMR 版本 5.7.0 及更高版本，您可设置只读副本集群，这允许您将数据的只读副本保留在 Amazon S3 中。如果主集群变得不可用，您可以访问只读副本集群中的数据以同时执行读取操作。
- 在 Amazon EMR 6.2.0 版及更高版本中，持久性 HFile 跟踪使用名为 hbase:storefile 的 HBase 系统表直接跟踪用于读取操作的 HFile 路径。默认情况下，此功能处于启用状态，不需要执行手动迁移。

下图显示与 HBase on Amazon S3 相关的 HBase 组件。



启用 HBase on Amazon S3

您可以通过 Amazon EMR 控制台、Amazon CLI 或 Amazon EMR API 启用 HBase on Amazon S3。该配置是集群创建期间的一个选项。使用控制台时，您可以通过 Advanced options (高级选项) 选择相应设置。在使用 Amazon CLI 时，使用 `--configurations` 选项提供 JSON 配置对象。配置对象的属性指定了 Amazon S3 中的存储模式和根目录位置。您指定的 Amazon S3 位置应位于 Amazon EMR 集群所在的区域内。一次仅一个活动集群可使用 Amazon S3 中的相同 HBase 根目录。有关控制台步骤和使用 Amazon CLI 的详细 create-cluster 示例，请参阅创建带 HBase 的集群。以下 JSON 代码段中显示了示例配置对象。

```
{
  "Classification": "hbase-site",
  "Properties": {
    "hbase.rootdir": "s3://my-bucket/my-hbase-rootdir"}
},
{
```

```
  "Classification": "hbase",
  "Properties": {
  "hbase.emr.storageMode":"s3"
  }
}
```

> **Note**
>
> 如果您使用 Amazon S3 存储桶作为 HBase 的 `rootdir`，您必须在 Amazon S3 URI 的末尾添加斜杠。例如，为了避免出现问题，您必须使用 `"hbase.rootdir: s3://my-bucket/"`，而不是 `"hbase.rootdir: s3://my-bucket"`。

使用只读副本集群

在使用 HBase on Amazon S3 设置主集群后，您可创建和配置只读副本集群，此集群提供对与主集群相同的数据的只读访问权限。当您需要同步访问权限以查询数据或在主集群变得不可用的情况下进行连续访问时，这会很有用。只读副本功能适用于 Amazon EMR 5.7.0 版和更高版本。

主集群和只读副本集群的设置方式相同，但有一个重要差异。两者都指向相同的 `hbase.rootdir` 位置。不过，只读副本集群的 `hbase` 分类包括 `"hbase.emr.readreplica.enabled":"true"` 属性。

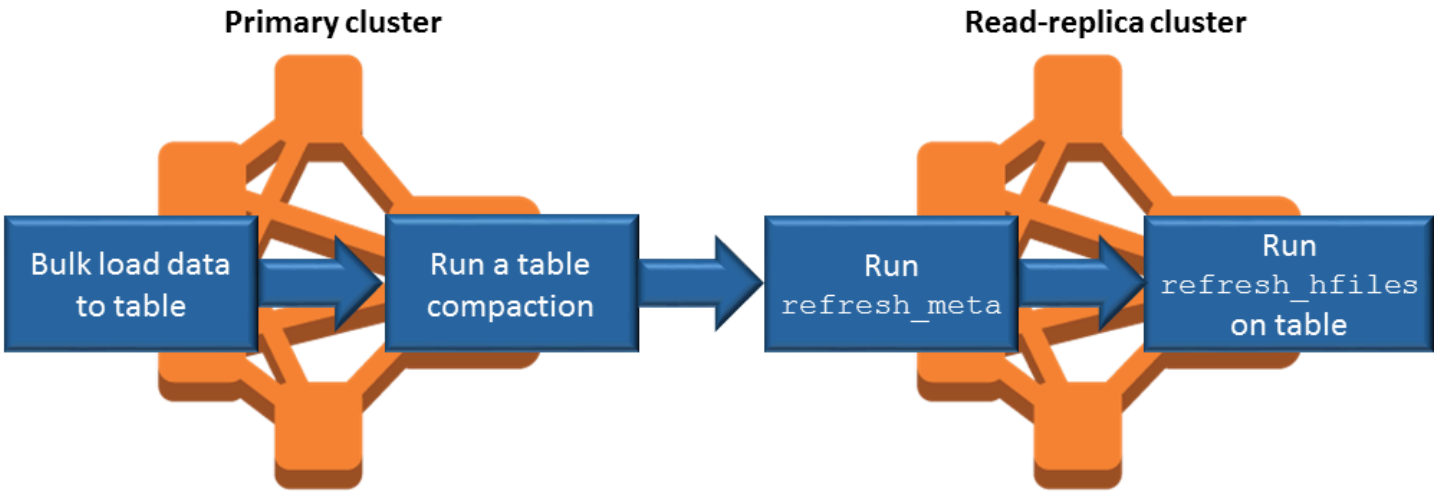
例如，假定主集群的 JSON 分类如主题中前面所示，只读副本集群的配置如下所示：

```
{
  "Classification": "hbase-site",
  "Properties": {
    "hbase.rootdir": "s3://my-bucket/my-hbase-rootdir"}
},
{
  "Classification": "hbase",
  "Properties": {
  "hbase.emr.storageMode":"s3",
  "hbase.emr.readreplica.enabled":"true"
  }
}
```

添加数据时同步只读副本

由于只读副本使用主集群写入 Amazon S3 的 HBase 存储文件和元数据，因此，只读副本仅与 Amazon S3 数据存储一样新。在写入数据时，以下指导信息可帮助您最大程度地缩短主集群和只读副本之间的滞后时间。

- 如果可能，请在主集群上批量加载数据。有关更多信息，请参阅 Apache HBase 文档中的批量加载。
- 将存储文件写入 Amazon S3 的刷新操作应在添加数据后尽快进行。手动刷新或优化刷新设置以最大限度地减少滞后时间。
- 如果压缩可能会自动运行，请运行手动压缩以避免触发压缩时出现不一致。
- 在只读副本集群上，当任何元数据发生变化时（例如，当 HBase 区域拆分或压缩时，或者添加或删除表时），请运行 `refresh_meta` 命令。
- 在只读副本集群上，在表中添加或更改记录后，运行 `refresh_hfiles` 命令。



持久性 HFile 跟踪

持久性 HFile 跟踪使用名为 `hbase:storefile` 的 HBase 系统表直接跟踪用于读取操作的 HFile 路径。将其它数据添加到 HBase 时，新的 HFile 路径将添加到表中。这将删除重命名操作作为关键写入路径 HBase 操作中的提交机制，并通过从 `hbase:storefile` 系统表（而不是文件系统目录列表）读取来缩短打开 HBase 区域时的恢复时间。默认情况下，在 Amazon EMR 6.2.0 版及更高版本上启用此功能，不需要任何手动迁移步骤。

> **Note**
>
> 使用 HBase 存储文件系统表的持久性 HFile 跟踪不支持 HBase 区域复制功能。有关 HBase 区域复制的更多信息，请参阅时间表一致的高可用读取。

禁用持久性 HFile 跟踪

默认情况下，从 EMR 发行版 6.2.0 开始启用持久性 HFile 跟踪。要禁用持久性 HFile 跟踪，请在启动集群时指定以下配置覆盖：

```
{
  "Classification": "hbase-site",
  "Properties": {
    "hbase.storefile.tracking.persist.enabled":"false",

 "hbase.hstore.engine.class":"org.apache.hadoop.hbase.regionserver.DefaultStoreEngine"
  }
}
```

> **Note**
>
> 重新配置 Amazon EMR 集群时，必须更新所有实例组。

手动同步存储文件表

创建新的 HFile 时，存储文件表将保持最新状态。但是，如果存储文件表由于任意原因与数据文件不同步，则可以使用以下命令手动同步数据：

同步线上区域中的存储文件表:

```
hbase org.apache.hadoop.hbase.client.example.RefreshHFilesClient <table>
```

同步离线区域中的存储文件表:

- 删除存储文件表 znode。

```
echo "ls /hbase/storefile/loaded" | sudo -u hbase hbase zkcli
[<tableName>, hbase:namespace]
# The TableName exists in the list
```

```
echo "delete /hbase/storefile/loaded/<tableName>" | sudo -u hbase hbase zkcli
# Delete the Table ZNode
echo "ls /hbase/storefile/loaded" | sudo -u hbase hbase zkcli
[hbase:namespace]
```

- 分配区域（在“hbase shell”中运行）。

```
hbase cli> assign '<region name>'
```

- 如果分配失败。

```
hbase cli> disable '<table name>'
hbase cli> enable '<table name>'
```

扩缩存储文件表

默认情况下，存储文件表可拆分为四个区域。如果存储文件表的写入负载仍然较重，之后可以手动拆分该表。

要拆分特定的热点区域，请使用以下命令（在“hbase shell”中运行）。

```
hbase cli> split '<region name>'
```

要拆分该表，请使用以下命令（在“hbase shell”中运行）。

```
hbase cli> split 'hbase:storefile'
```

操作注意事项

HBase 区域服务器使用 BlockCache 将数据读取存储在内存中，使用 BucketCache 将数据读取存储在本地磁盘上。此外，区域服务器使用 MemStore 将数据写入存储在内存中，并在数据被写入 Amazon S3 中的 HBase 存储文件之前使用预写日志将数据写入存储到 HDFS 中。集群的读取性能与可从内存中或磁盘缓存中读取记录的频率有关。缓存未命中会导致从 Amazon S3 中的存储文件读取记录，与从 HDFS 读取相比，这将产生更大的延迟和标准差。此外，Amazon S3 的最大请求速率低于可从本地缓存中检索内容的速率，因此对于需要进行大量读取操作的工作负载来说，缓存数据可能非常重要。有关 Amazon S3 性能的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的性能优化。

为了提高性能，建议您在 EC2 实例存储中尽可能多地缓存数据集。由于 BucketCache 使用区域服务器的 EC2 实例存储，因此，您可以选择具有足够实例存储的 EC2 实例类型并添加 Amazon EBS 存储来

适应所需的缓存大小。您还可以使用 hbase.bucketcache.size 属性增加附加实例存储和 EBS 卷上的 BucketCache 大小。默认设置为 8192MB。

对于写入，MemStore 刷新的频率以及次要和主要压缩期间存在的存储文件数量可能会显著增加区域服务器响应时间。要获得最佳性能，请考虑增大 MemStore 刷新和 HRegion 块倍数的大小，这将增加主要压缩之间的用时，并增加一致性的滞后 (如果您使用只读副本)。在某些情况下，使用较大的文件块大小（但小于 5GB）触发 EMRFS 中的 Amazon S3 分段上载功能也许能够获得更佳性能。Amazon EMR 的数据块大小默认为 128MB。有关更多信息，请参阅HDFS 配置。在通过刷新和压缩来衡量性能时，我们很少看到有客户的数据块大小超过 1GB。此外，当需要压缩的存储文件较少时，HBase 压缩和区域服务器的性能可能会最佳。

由于需要对大量目录进行重命名，因此从 Amazon S3 中删除表需要花费大量时间。请考虑禁用表而不是删除表。

有一个 HBase 清理器，用于清理旧的 WAL 文件并存储文件。对于 Amazon EMR 版本 5.17.0 及更高版本，全局启用清理器，且以下配置属性可用于控制清理器行为。

配置属性	默认值	描述
hbase.regionserver.hfilecleaner.large.thread.count	1	为清理过期的大型 HFile 而分配的线程数。
hbase.regionserver.hfilecleaner.small.thread.count	1	为清理过期的小型 HFile 而分配的线程数。
hbase.cleaner.scan.dir.concurrent.size	设置为所有可用内核的四分之一。	用于扫描 oldWAL 目录的线程数。
hbase.oldwals.cleaner.thread.size	2	用于清理 oldWAL 目录下的 WAL 的线程数。

对于 Amazon EMR 5.17.0 及更早版本，在运行大量工作负载时，清理器操作会影响查询性能；因此，我们建议您只在非高峰时间启用清理器。清理器拥有以下 HBase shell 命令：

- `cleaner_chore_enabled` 查询是否启用了清理器。
- `cleaner_chore_run` 手动运行清理器来删除文件。
- `cleaner_chore_switch` 启用或禁用清理器并返回清理器的先前状态。例如，`cleaner_chore_switch true` 启用清理器。

用于 HBase on Amazon S3 性能优化的属性

使用 HBase on Amazon S3 时，可调整以下参数来优化工作负载的性能。

配置属性	默认值	描述
`hbase.bucketcache.size`	8192	区域服务器 Amazon EC2 实例存储和 BucketCache 存储的 EBS 卷上预留的磁盘空间量（以 MB 为单位）。此设置适用于所有区域服务器实例。较大的 BucketCache 大小通常对应提高的性能
`hbase.hregion.memstore.flush.size`	134217728	触发对 Amazon S3 的 memstore 刷新的数据限制（以字节为单位）。
`hbase.hregion.memstore.block.multiplier`	4	一个乘数，用于确定阻止更新的 MemStore 上限。如果 MemStore 超过 `hbase.hregion.memstore.flush.size` 乘以此值，则会阻止更新。可能会发生 MemStore 刷新和压缩以取消阻止更新。
`hbase.hstore.blockingStoreFiles`	10	阻止更新之前存储中可以存在的最大存储文件数。

配置属性	默认值	描述
hbase.hregion.max.filesize	10737418240	区域被拆分之前的大小上限。

关闭并恢复集群而不丢失数据

要关闭 Amazon EMR 集群而不丢失尚未写入 Amazon S3 的数据，您需要将 MemStore 缓存刷新到 Amazon S3 以写入新存储文件。首先，您需要禁用所有表格。在向集群添加步骤时，可使用以下步骤配置。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用Amazon CLI和控制台执行步骤。

```
Name="Disable all tables",Jar="command-runner.jar",Args=["/bin/bash","/usr/lib/hbase/bin/disable_all_tables.sh"]
```

或者，您可以直接运行以下 bash 命令。

```
bash /usr/lib/hbase/bin/disable_all_tables.sh
```

禁用所有表格后，使用 HBase shell 和以下命令刷新hbase:meta表。

```
flush 'hbase:meta'
```

然后，您可以运行 Amazon EMR 集群上提供的 shell 脚本来刷新 MemStore 缓存。您可以将它作为步骤添加，也可以使用集群中Amazon CLI直接运行它。此脚本会禁用所有 HBase 表，这会导致每个区域服务器上的 MemStore 刷新到 Amazon S3。如果此脚本成功完成，数据将保留在 Amazon S3 中，并且可以终止集群。

要使用相同 HBase 数据重新启动集群，可在Amazon Web Services Management Console中或使用 hbase.rootdir 配置属性指定与上一个集群相同的 Amazon S3 位置。

使用 HBase shell

在创建 HBase 集群后，下一步是连接到 HBase，以便您可以开始读取和写入数据（只读副本集群不支持数据写入）。您可以使用 HBase shell 来测试命令。

打开 HBase shell

1. 使用 SSH 连接 HBase 集群中的主服务器。有关如何使用 SSH 连接到主节点的信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 运行 `hbase shell`。HBase shell 打开时，系统会显示类似于以下内容的提示符。

```
hbase(main):001:0>
```

您可以根据提示符发出 HBase shell 命令。有关 shell 命令以及如何调用这些命令的详细信息，请在 HBase 提示符处键入 help，然后按 Enter 键。

创建表

通过以下命令可创建一个名为“t1”的表，该表拥有名为“f1”的单列系列。

```
hbase(main):001:0>create 't1', 'f1'
```

设置值

通过以下命令可为表“t1”和列“f1”中的行“r1”设置值“v1”。

```
hbase(main):001:0>put 't1', 'r1', 'f1:col1', 'v1'
```

获取值

通过以下命令可获取表“t1”中的行“r1”的值。

```
hbase(main):001:0>get 't1', 'r1'
```

删除表

以下命令将删除表“t1”。

```
hbase(main):001:0>drop 'ns1:t1',false
```

布尔值对应于您是否要存档表，因此如果要保存，可以将其设置为 `true`。您也可以运行不带布尔值的 `drop 'ns1:t1'` 来存档表。

通过 Hive 访问 HBase 表

HBase 与Apache Hive紧密集成，让您可以在 HBase 中存储的数据上直接运行大规模并行处理的工作负载。要将 Hive 与 HBase 结合使用，您通常可以在同一个集群上启动它们。不过，您可以在单独的集群上启动 Hive 和 HBase。在不同的集群上单独运行 HBase 和 Hive 可以提高性能，因为这可让每个应用程序更高效地利用集群资源。

以下过程说明如何使用 Hive 连接到集群上的 HBase。

> **Note**
>
> 您只能将 Hive 集群连接到单个 HBase 集群。

将 Hive 连接到 HBase

1. 创建安装了 Hive 和 HBase 的单独集群或创建安装了 HBase 和 Hive 的单个集群。
2. 如果您使用单独的集群，请修改您的安全组，以便 HBase 和 Hive 端口在这两个主节点之间是开放的。
3. 使用 SSH 连接到安装了 Hive 的集群的主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
4. 通过以下命令启动 Hive shell。

   ```
   hive
   ```

5. （可选）如果 HBase 和 Hive 位于同一个集群上，则您无需执行此操作。将 Hive 集群上的 HBase 客户端与包含数据的 HBase 集群连接。在以下示例中，使用 HBase 集群的主节点公有 DNS 名称替换 *public-DNS-name*，如：`ec2-50-19-76-67.compute-1.amazonaws.com`。

   ```
   set hbase.zookeeper.quorum=public-DNS-name;
   ```

6. 根据需要继续对 HBase 数据运行 Hive 查询或查看后续步骤。

通过 Hive 访问 HBase 数据

- 在 Hive 和 HBase HBase 集群之间建立连接后 (如上一过程所示)。您可以通过在 Hive 中创建外部表访问存储在 HBase 集群上的数据。

  从主节点上的 Hive 提示符中运行时，以下示例创建了一个外部表，此表引用了存储在名为 `inputTable` 的 HBase 表上的数据。然后，您可以引用 Hive 语句中的 `inputTable`，查询和修改存储在 HBase 集群上的数据。

```
set hbase.zookeeper.quorum=ec2-107-21-163-157.compute-1.amazonaws.com;

create external table inputTable (key string, value string)
     stored by 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
      with serdeproperties ("hbase.columns.mapping" = ":key,f1:col1")
      tblproperties ("hbase.table.name" = "t1");

select count(key) from inputTable ;
```

有关将 HBase 和 Hive 结合使用的更高级的使用案例和示例，请参阅Amazon大数据博客文章 Combine NoSQL and massively parallel analytics using Apache HBase and Apache Hive on Amazon EMR。

使用 HBase 快照

HBase 使用内置快照功能创建表的轻量级备份。在 EMR 集群中，可使用 EMRFS 将这些备份导出到 Amazon S3。您可以使用 HBase shell 在主节点上创建快照。本主题说明如何使用 Shell 或通过将 `command-runner.jar` 与 Amazon CLI 或 Amazon SDK for Java 结合使用的步骤以交互方式运行这些命令。有关 HBase 备份类型的更多信息，请参阅 HBase 文档中的 HBase 备份。

使用表创建快照

```
hbase snapshot create -n snapshotName -t tableName
```

从 Amazon CLI 使用 command-runner.jar：

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps Name="HBase Shell Step",Jar="command-runner.jar",\
Args=[ "hbase", "snapshot", "create","-n","snapshotName","-t","tableName"]
```

Amazon SDK for Java

```
HadoopJarStepConfig hbaseSnapshotConf = new HadoopJarStepConfig()
  .withJar("command-runner.jar")
  .withArgs("hbase","snapshot","create","-n","snapshotName","-t","tableName");
```

> **Note**
>
> 如果您的快照名称不唯一，则创建操作将失败，并返回 -1 或者 255，但您可能看不到说明出现问题的错误消息。要使用相同的快照名称，请先将其删除，然后重新创建。

删除快照

```
hbase shell
>> delete_snapshot 'snapshotName'
```

查看快照信息

```
hbase snapshot info -snapshot snapshotName
```

将快照导出到 Amazon S3

> **Important**
>
> 如果导出快照时没有指定 -mappers 值，HBase 会使用任意计算来确定映射器的数量。此值可能非常大，具体取决于表大小，这会在导出过程中对正在运行的任务产生负面影响。为此，我们建议您指定 -mappers 参数或 -bandwidth 参数（指定每秒使用的带宽，以 MB 为单位），或同时指定这两个参数以限制导出操作使用的集群资源。或者，您可以在低使用率期间运行导出快照操作。

```
hbase snapshot export -snapshot snapshotName \
-copy-to s3://bucketName/folder -mappers 2
```

从 Amazon CLI 使用 command-runner.jar：

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
```

```
--steps Name="HBase Shell Step",Jar="command-runner.jar",\
Args=[ "hbase", "snapshot", "export","-snapshot","snapshotName","-copy-
to","s3://bucketName/folder","-mappers","2","-bandwidth","50"]
```

Amazon SDK for Java:

```
HadoopJarStepConfig hbaseImportSnapshotConf = new HadoopJarStepConfig()
  .withJar("command-runner.jar")
  .withArgs("hbase","snapshot","export",
      "-snapshot","snapshotName","-copy-to",
      "s3://bucketName/folder",
      "-mappers","2","-bandwidth","50");
```

从 Amazon S3 导入快照

虽然这是导入，但这里使用的 HBase 选项仍然是 `export`。

```
sudo -u hbase hbase snapshot export \
-D hbase.rootdir=s3://bucketName/folder \
-snapshot snapshotName \
-copy-to hdfs://masterPublicDNSName:8020/user/hbase \
-mappers 2
```

从 Amazon CLI 使用 `command-runner.jar`：

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps Name="HBase Shell Step",Jar="command-runner.jar", \
Args=["sudo","-u","hbase","hbase snapshot export","-snapshot","snapshotName", \
"-D","hbase.rootdir=s3://bucketName/folder", \
"-copy-to","hdfs://masterPublicDNSName:8020/user/hbase","-mappers","2","-chmod","700"]
```

Amazon SDK for Java:

```
HadoopJarStepConfig hbaseImportSnapshotConf = new HadoopJarStepConfig()
  .withJar("command-runner.jar")
  .withArgs("sudo","-u","hbase","hbase","snapshot","export", "-D","hbase.rootdir=s3://
path/to/snapshot",
      "-snapshot","snapshotName","-copy-to",
      "hdfs://masterPublicDNSName:8020/user/hbase",
      "-mappers","2","-chuser","hbase");
```

通过 HBase shell 中的快照恢复表

```
hbase shell
>> disable tableName
>> restore_snapshot snapshotName
>> enable tableName
```

HBase 目前不支持 HBase shell 中找到的所有快照命令。例如，没有用于还原快照的 HBase 命令行选项，因此您必须在 shell 中还原快照。这意味着 `command-runner.jar` 必须运行 Bash 命令。

> **Note**
>
> 由于此处使用的命令为 `echo`，因此您的 Shell 命令可能仍将失败，即使 Amazon EMR 运行的命令返回退出代码 `0` 也是如此。如果选择将 shell 命令作为步骤运行，请检查步骤日志。

```
echo 'disable tableName; \
restore_snapshot snapshotName; \
enable tableName' | hbase shell
```

以下是使用 Amazon CLI 的步骤。首先，创建以下 `snapshot.json` 文件：

```
[
  {
    "Name": "restore",
    "Args": ["bash", "-c", "echo $'disable \"tableName\"; restore_snapshot
 \"snapshotName\"; enable \"tableName\"' | hbase shell"],
    "Jar": "command-runner.jar",
    "ActionOnFailure": "CONTINUE",
    "Type": "CUSTOM_JAR"
  }
]
```

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps file://./snapshot.json
```

Amazon SDK for Java:

```
HadoopJarStepConfig hbaseRestoreSnapshotConf = new HadoopJarStepConfig()
```

```
  .withJar("command-runner.jar")
  .withArgs("bash","-c","echo $'disable \"tableName\"; restore_snapshot \"snapshotName
\"; enable \"snapshotName\"' | hbase shell");
```

配置 HBase

尽管默认 HBase 设置应当用于大多数应用程序，但是您可以修改 HBase 配置设置。为此，请使用 HBase 配置分类的属性。有关更多信息，请参阅配置应用程序。

以下示例基于存储在 Amazon S3 中的配置文件 `myConfig.json` 创建了一个具有备用 HBase 根目录的集群。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=HBase \
--instance-type m5.xlarge --instance-count 3 --configurations https://s3.amazonaws.com/
mybucket/myfolder/myConfig.json
```

`myConfig.json` 文件指定 `hbase-site` 配置分类的 `hbase.rootdir` 属性，如以下示例中所示。将 *ip-XXX-XX-XX-XXX.ec2.internal* 替换为集群的主节点的内部 DNS 主机名。

```
[
  {
    "Classification":"hbase-site",
    "Properties": {
       "hbase.rootdir": "hdfs://ip-XXX-XX-XX-XXX.ec2.internal:8020/user/
myCustomHBaseDir"
    }
  }
]
```

> **Note**
>
> 对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon

Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

对 YARN 中内存分配的更改

HBase 不是作为 YARN 应用程序运行的，因此有必要重新计算分配给 YARN 及其应用程序的内存，这会导致如果安装 HBase，YARN 可用的总内存会减少。您在规划将 YARN 应用程序和 HBase 共同放置在相同的集群时应考虑这一点。内存小于 64GB 的实例类型有一半的内存可供 NodeManager 使用，然后再分配给 HBase RegionServer。对于内存大于 64GB 的实例类型，HBase RegionServer 内存的上限为 32GB。一般来说，YARN 设置内存是 MapReduce 折叠器任务内存的几倍。

任务配置设置的默认值中的表显示了根据 HBase 所需内存对 YARN 设置的更改。

HBase 端口号

为 HBase 选择的某些端口号与默认端口号不同。以下是适用于 Amazon EMR 上的 HBase 的接口和端口。

HBase 端口

接口	端口	协议
HMaster	16000	TCP
HMaster UI	16010	HTTP
RegionServer	16020	TCP
RegionServer 信息	16030	HTTP
REST 服务器	8070	HTTP
REST UI	8085	HTTP
Thrift 服务器	9090	TCP
Thrift 服务器 UI	9095	HTTP

> **Important**
>
> 在 Amazon EMR 发行版 4.6.0 及更高版本中，`kms-http-port` 为 9700，`kms-admin-port` 为 9701。

要优化的 HBase 站点设置

您可以设置任何或所有的 HBase 站点设置，以便优化 HBase 集群以更好地承载您的应用程序工作负载。建议您将以下设置作为调查的起点。

zookeeper.session.timeout

默认超时时间为 40 秒（40000 毫秒）。如果区域服务器崩溃，则上述值是主服务器注意到区域服务器缺少及开始恢复所需的时长。要帮助更快地恢复主服务器，可以将此值减小到更短的时间。以下示例使用的是 30 秒或 30000 毫秒：

```
[
  {
    "Classification":"hbase-site",
    "Properties": {
       "zookeeper.session.timeout": "30000"
    }
  }
]
```

hbase.regionserver.handler.count

这可定义区域服务器保持打开状态以向表提供请求的线程数。默认值 10 较低，以防止用户在使用具有大量并发客户端的大写缓冲区时终止其区域服务器。经验法则是，当每个请求的负载接近 MB 范围（大放置、使用大缓存扫描）时保持数值较低，而当负载较小（获取、小放置、ICV、删除）时保持数值较高。以下示例将打开的线程数提高到 30：

```
[
  {
    "Classification":"hbase-site",
    "Properties": {
       "hbase.regionserver.handler.count": "30"
    }
```

```
  }
]
```

hbase.hregion.max.filesize

此参数控制单个区域的大小（以字节为单位）。默认情况下，将它设置为 1073741824。如果您向 HBase 集群中写入很多数据并且导致了频繁拆分，您可以增加它的大小以扩大单个区域。这会减少拆分，但需要更多时间来实现从一台服务器到另一台服务器的区域负载均衡。

```
[
  {
    "Classification":"hbase-site",
    "Properties": {
       "hbase.hregion.max.filesize": "1073741824"
    }
  }
]
```

hbase.hregion.memstore.flush.size

此参数控制 Memstore 刷新到磁盘之前的大小上限（以字节为单位）。默认为 134217728。如果您的工作负载由短时间突发的写入操作组成，您可能希望增加此限制，以便所有写入在突发期间都保留在内存中，并在稍后刷新到磁盘。这可以提高突发期间的性能。

```
[
  {
    "Classification":"hbase-site",
    "Properties": {
       "hbase.hregion.memstore.flush.size": "134217728"
    }
  }
]
```

查看 HBase 用户界面

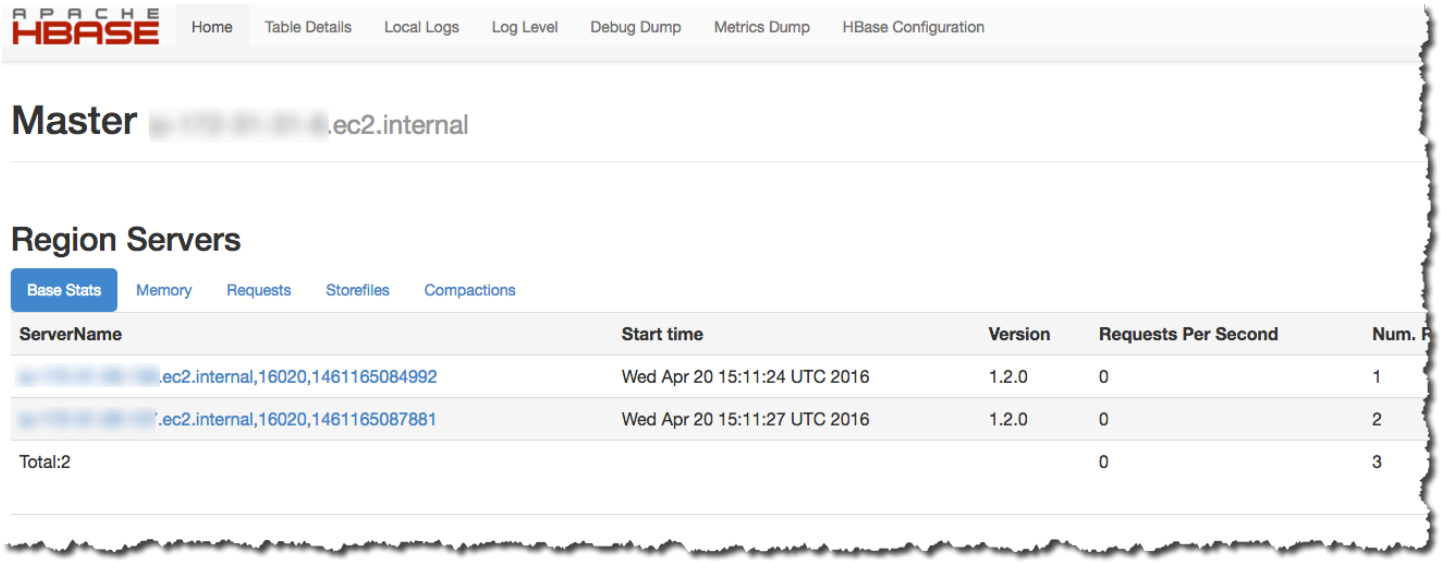
> **Note**
>
> 默认情况下，HBase 用户界面使用的是不安全的 HTTP 连接。要启用安全的 HTTP（HTTPS），请在 HBase 配置中将 `hbase-site` 分类的 `hbase.ssl.enabled` 属性

设置为 `true`。有关在 HBase Web UI 中使用安全的 HTTP（HTTPS）的更多信息，请参阅 Apache HBase 参考指南。

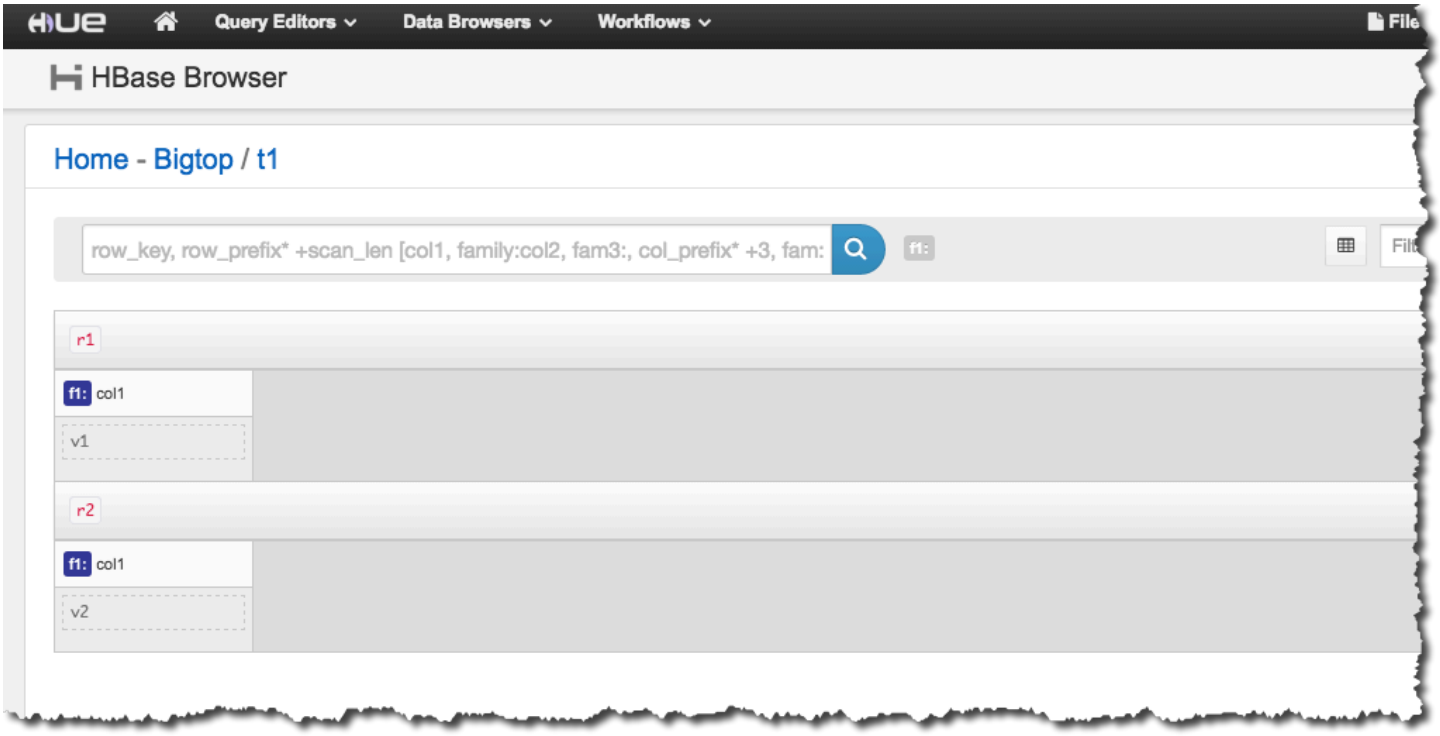
HBase 提供了基于 Web 的用户界面，您可以用它监控您的 HBase 集群。当您在 Amazon EMR 上运行 HBase 时，Web 界面会在主节点上运行，并可以使用端口转发（也称为创建 SSH 隧道）进行查看。

查看 HBase 用户界面

1. 使用 SSH 隧道进入主节点并创建安全连接。有关更多信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 1 部分：使用动态端口转发设置到主节点的 SSH 隧道。
2. 使用代理工具 (如 Firefox 的 FoxyProxy 插件) 安装 Web 浏览器，为 Amazon 域创建 SOCKS 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 2 部分：配置代理设置以查看主节点上托管的网站。
3. 通过设置代理并打开 SSH 连接，您可以通过打开浏览器窗口并访问 http://***master-public-dns-name***:16010/master-status 来查看 HBase UI，其中 *master-public-dns-name* 是集群的主节点的公有 DNS 地址。



您也可以在 Hue 中查看 HBase。例如，下面显示了使用 HBase shell中创建的表 `t1`：



有关 Hue 的更多信息，请参阅Hue。

查看 HBase 日志文件

在其操作过程中，HBase 写入日志文件，其中包含有关配置设置、守护进程操作和异常的详细信息。这些日志文件对于调试 HBase 问题以及跟踪性能非常有用。

如果您配置集群以将日志文件保存到 Amazon S3，您应当知道，日志会每五分钟写入 Amazon S3 一次，因此在最新的日志文件可用之前，可能会出现轻微延迟。

查看主节点上的 HBase 日志

- 您可以使用 SSH 连接主节点，然后导航到 `/var/log/hbase` 目录，从而查看当前的 HBase 日志。除非您在集群启动时启用了针对 Amazon S3 的日志记录，否则这些日志将在集群终止后不再可用。

查看 Amazon S3 上的 HBase 日志

- 要访问 Amazon S3 上的 HBase 日志和其它集群日志，以及要让它们在集群终止后可用，请在创建集群时指定一个 Amazon S3 存储桶以接收这些日志。这可以使用选项 `--log-uri` 实现。有

关为集群启用日志记录的更多信息，请参阅《Amazon EMR 管理指南》中的配置日志记录和调试（可选）。

使用 Ganglia 监控 HBase

Ganglia 开源项目是一个可扩展的分布式系统，旨在监控集群和网格，同时尽量减少对其性能的影响。当您在集群上启用 Ganglia 时，您可以生成报告并查看整个集群的性能，还可以检查单个节点实例的性能。有关 Ganglia 开源项目的更多信息，请参阅 http://ganglia.info/。有关结合 Amazon EMR 集群使用 Ganglia 的更多信息，请参阅Ganglia。

在配置了 Ganglia 的情况下启动集群后，您就可以使用主节点上运行的图形界面来访问 Ganglia 图形和报告。

Ganglia 将日志文件存在主节点上的 `/mnt/var/lib/ganglia/rrds/` 目录中。早期版本的 Amazon EMR 可将日志文件存储在 `/var/log/ganglia/rrds/` 目录中。

使用 Amazon CLI 为 Ganglia 和 HBase 配置集群

- 使用类似于以下内容的 `create-cluster` 命令：

```
aws emr create-cluster --name "Test cluster" --release-label emr-5.36.1 \
--applications Name=HBase Name=Ganglia --use-default-roles \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge \
--instance-count 3
```

> **Note**
>
> 如果默认 Amazon EMR 服务角色和 Amazon EC2 实例配置文件不存在，则会发生错误。请使用 `aws emr create-default-roles` 命令创建它们，然后重试。

有关更多信息，请参阅 Amazon CLI 中的 Amazon EMR 命令。

在 Ganglia Web 界面中查看 HBase 指标

1. 使用 SSH 隧道进入主节点并创建安全连接。有关更多信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 1 部分：使用动态端口转发设置到主节点的 SSH 隧道。

2. 使用代理工具 (如 Firefox 的 FoxyProxy 插件) 安装 Web 浏览器，为 Amazon 域创建 SOCKS 代理。有关更多信息，请参阅《Amazon EMR 管理指南》中的选项 2，第 2 部分：配置代理设置以查看主节点上托管的网站。
3. 通过设置代理和打开 SSH 连接，您可以打开浏览器窗口，通过 http://*master-public-dns-name*/ganglia/ 查看 Ganglia 指标，其中 *master-public-dns-name* 是 HBase 集群中主服务器的公有 DNS 地址。

查看主节点上的 Ganglia 日志文件

- 如果集群仍在运行，您可以使用 SSH 连接主节点，然后导航到 `/mnt/var/lib/ganglia/rrds/` 目录来访问日志文件。对于 EMR 3.x，请导航到 `/var/log/ganglia/rrds` 目录。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。

查看 Amazon S3 上的 Ganglia 日志文件

- 即使您为集群启用日志记录，Ganglia 日志文件也不会自动写入 Amazon S3。要在 Amazon S3 上查看 Ganglia 日志文件，您必须手动将日志从 `/mnt/var/lib/ganglia/rrds/` 推送到 S3 存储桶。

从早期版本的 HBase 迁移

要从之前的 HBase 版本迁移数据，请参阅 Apache HBase 参考指南中的升级和 HBase 版本号和兼容性。您可能需要特别注意从 HBase 1.0 版本之前的版本升级的要求。

HBase 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 HBase 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

HBase 版本信息

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.14.0	2.4.17	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
		client, hadoop-hdfs-datano de, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-ma pred, hadoop-yarn-nodema nager, hadoop-yarn-resour cemanager, hadoop-yarn-timeline-server, hbase-hma ster, hbase-client, hbase-reg ion-server, hbase-rest-server, hbase-thrift-server, hbase-ope rator-tools, zookeeper-client, zookeeper-server
emr-6.13.0	2.4.17	emrfs, emr-ddb, emr-goodi es, emr-kinesis, emr-s3-di st-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datano de, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-ma pred, hadoop-yarn-nodema nager, hadoop-yarn-resour cemanager, hadoop-yarn-timeline-server, hbase-hma ster, hbase-client, hbase-reg ion-server, hbase-rest-server, hbase-thrift-server, hbase-ope rator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.12.0	2.4.17	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.11.1	2.4.15	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.11.0	2.4.15	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.10.1	2.4.15	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.10.0	2.4.15	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-wal-cli, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server
emr-6.9.1	2.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.9.0	2.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server
emr-6.8.1	2.4.12	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.8.0	2.4.12	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server
emr-6.7.0	2.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.36.1	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.36.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.6.0	2.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, hbase-operator-tools, zookeeper-client, zookeeper-server
emr-5.35.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.5.0	2.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-6.4.0	2.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.3.1	2.2.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-6.3.0	2.2.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.2.1	2.2.6-amzn-0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-6.2.0	2.2.6-amzn-0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.1.1	2.2.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-6.1.0	2.2.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-6.0.1	2.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-6.0.0	2.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.34.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.33.1	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.33.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.32.1	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.32.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.31.1	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.31.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.30.2	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.30.1	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.30.0	1.4.13	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.29.0	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.28.1	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.28.0	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.27.1	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.27.0	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.26.0	1.4.10	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.25.0	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.24.1	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.24.0	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.23.1	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.23.0	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.22.0	1.4.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.21.2	1.4.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.21.1	1.4.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.21.0	1.4.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.20.1	1.4.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.20.0	1.4.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.19.1	1.4.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.19.0	1.4.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.18.1	1.4.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.18.0	1.4.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.17.2	1.4.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.17.1	1.4.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.17.0	1.4.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.16.1	1.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.16.0	1.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.15.1	1.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.15.0	1.4.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.14.2	1.4.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.14.1	1.4.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.14.0	1.4.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.13.1	1.4.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.13.0	1.4.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.12.3	1.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.12.2	1.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.12.1	1.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.12.0	1.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.11.4	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.11.3	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.11.2	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.11.1	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.11.0	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.10.1	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.10.0	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.9.1	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.9.0	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.8.3	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.8.2	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.8.1	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.8.0	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.7.1	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.7.0	1.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.6.1	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.6.0	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.5.4	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.5.3	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.5.2	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.5.1	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.5.0	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.4.1	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.4.0	1.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.3.2	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.3.1	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.3.0	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.2.3	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.2.2	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.2.1	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.2.0	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.1.1	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.1.0	1.2.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-5.0.3	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-5.0.0	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.9.6	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.9.5	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.9.4	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.9.3	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.9.2	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.9.1	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.8.5	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.8.4	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.8.3	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.8.2	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.8.0	1.2.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.7.4	1.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.7.2	1.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.7.1	1.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-regionserver, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	HBase 版本	随 HBase 安装的组件
emr-4.7.0	1.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server
emr-4.6.0	1.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, hbase-rest-server, hbase-thrift-server, zookeeper-client, zookeeper-server

Apache HCatalog

HCatalog 是一个工具，通过该工具，您可以访问 Pig、Spark SQL 和/或自定义 MapReduce 应用程序中的 Hive 元存储表。HCatalog 具有 REST 接口和命令行客户端，允许您创建表或执行其它操作。然后，您可以编写应用程序以使用 HCatalog 库访问这些表。有关详细信息，请参阅使用 HCatalog。HCatalog 包含在 Amazon EMR 发行版 4.4.0 及更高版本中。

HCatalog on Amazon EMR 发行版 5.8.0 及更高版本上的 HCatalog 支持使用 Amazon Glue 数据目录作为 Hive 的元存储。有关更多信息，请参阅使用 Amazon Glue 数据目录作为 Hive 的元存储。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 HCatalog 版本，以及 Amazon EMR 随 HCatalog 一起安装的组件。

有关此发行版中随 HCatalog 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 HCatalog 版本信息

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.14.0	HCatalog 3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 HCatalog 版本，以及 Amazon EMR 随 HCatalog 一起安装的组件。

有关此发行版中随 HCatalog 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 HCatalog 版本信息

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.36.1	HCatalog 2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

主题

创建带 HCatalog 的集群

虽然 Hive 项目包含了 HCatalog，但您必须将 HCatalog 作为其自己的应用程序安装。

使用控制台启动安装了 HCatalog 的集群

以下过程创建一个安装了 HCatalog 的集群。有关使用控制台（包括 Advanced Options (高级选项)）创建集群的更多信息，请参阅《Amazon EMR 管理指南》中的计划和配置集群。

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 选择 Create cluster (创建集群) 以使用 Quick Create (快速创建)。

3. 对于 Software Configuration (软件配置) 字段，选择 Amazon Release Version emr-4.4.0 (Amazon 发行版 emr-4.4.0) 或更高版本。
4. 在 Select Applications (选择应用程序) 字段中，选择 All Applications (所有应用程序) 或 HCatalog。
5. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

使用 Amazon CLI 启动安装了 HCatalog 的集群

- 使用下面的命令创建集群：

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --name "Cluster with Hcat" --release-label emr-5.36.1 \
--applications Name=HCatalog --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3 --use-default-roles
```

使用 HCatalog

您可以在使用 Hive 元存储的各种应用程序中使用 HCatalog。此部分中的示例演示如何创建表以及如何在 Pig 和 Spark SQL 上下文中使用该表。

使用 HCatalog HStorer 时禁用直接写入

当应用程序使用 HCatStorer 对存储在 Amazon S3 中的 HCatalog 表进行写入时，禁用 Amazon EMR 的直接写入功能。例如，在使用 Pig STORE 命令或运行将 HCatalog 表写入 Amazon S3 的 Sqoop 作业时，禁用直接写入。您可以通过将 `mapred.output.direct.NativeS3FileSystem` 和 `mapred.output.direct.EmrFileSystem` 配置设置为 `false` 来禁用直接写入功能。以下示例演示如何使用 Java 设置这些配置。

```
Configuration conf = new Configuration();
conf.set("mapred.output.direct.NativeS3FileSystem", "false");
conf.set("mapred.output.direct.EmrFileSystem", "false");
```

使用 HCat CLI 创建表并在 Pig 中使用该数据

在您的集群上创建以下脚本 `impressions.q`：

```
CREATE EXTERNAL TABLE impressions (
    requestBeginTime string, adId string, impressionId string, referrer string,
    userAgent string, userCookie string, ip string
  )
  PARTITIONED BY (dt string)
  ROW FORMAT
    serde 'org.apache.hive.hcatalog.data.JsonSerDe'
    with serdeproperties ( 'paths'='requestBeginTime, adId, impressionId, referrer,
 userAgent, userCookie, ip' )
  LOCATION 's3://[your region].elasticmapreduce/samples/hive-ads/tables/impressions/';
ALTER TABLE impressions ADD PARTITION (dt='2009-04-13-08-05');
```

使用 HCat CLI 执行脚本：

```
% hcat -f impressions.q
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-
log4j.properties
OK
Time taken: 4.001 seconds
OK
Time taken: 0.519 seconds
```

打开 Grunt shell 并访问 `impressions` 中的数据：

```
% pig -useHCatalog -e "A = LOAD 'impressions' USING
 org.apache.hive.hcatalog.pig.HCatLoader();
B = LIMIT A 5;
dump B;"
<snip>
(1239610346000,m9nwdo67Nx6q2kI25qt5On7peICfUM,omkxkaRpNhGPDucAiBErSh1cs0MThC,cartoonnetwork.com
 (compatible; MSIE 7.0; Windows NT 6.0; FunWebProducts; GTB6; SLCC1; .NET CLR
 2.0.50727; Media Center PC
 5.0; .NET,wcVWWTascoPbGt6bdqDbuWTPPHgOPs,69.191.224.234,2009-04-13-08-05)
(1239611000000,NjriQjdODgWBKnkGJUP6GNTbDeK4An,AWtXPkfaWGOaNeL9OOsFU8Hcj6eLHt,cartoonnetwork.com
 (compatible; MSIE 7.0; Windows NT 5.1; GTB6; .NET CLR
 1.1.4322),OaMU1F2gE4CtADVHAbKjjRRks5kIgg,57.34.133.110,2009-04-13-08-05)
```

```
(1239610462000,Irpv3oiu0I5QNQiwSSTIshrLdo9cM1,i1LDq44LRSJF0hbmhB8Gk7k9gMWtBq,cartoonnetwork.com
 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322;
 InfoPath.1),QSb3wkLR4JAIut4Uq6FNFQIR1rCVwU,42.174.193.253,2009-04-13-08-05)
(1239611007000,q2Awfnpe0JAvhInaIp0VGx9KTs0oPO,s3HvTflPB8JIE0IuM6hOEebWWpOtJV,cartoonnetwork.com
 (compatible; MSIE 6.0; Windows NT 5.2; SV1; .NET CLR 1.1.4322;
 InfoPath.1),QSb3wkLR4JAIut4Uq6FNFQIR1rCVwU,42.174.193.253,2009-04-13-08-05)
(1239610398000,c362vpAB0soPKGHRS43cj6TRwNeOGn,jeas5nXbQInGAgFB8jlkhnprN6cMw7,cartoonnetwork.com
 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; GTB6; .NET CLR
 1.1.4322),k96n5PnUmwHKfiUI0TFP0TNMfADgh9,51.131.29.87,2009-04-13-08-05)
7120 [main] INFO  org.apache.pig.Main  - Pig script completed in 7 seconds and 199
 milliseconds (7199 ms)
16/03/08 23:17:10 INFO pig.Main: Pig script completed in 7 seconds and 199 milliseconds
 (7199 ms)
```

使用 Spark SQL 访问表

此示例通过第一个示例中创建的表来创建 Spark DataFrame，并显示前 20 行：

```
% spark-shell --jars /usr/lib/hive-hcatalog/share/hcatalog/hive-hcatalog-core-1.0.0-
amzn-3.jar
<snip>
scala> val hiveContext = new org.apache.spark.sql.hive.HiveContext(sc);
scala> val df = hiveContext.sql("SELECT * FROM impressions")
scala> df.show()
<snip>
16/03/09 17:18:46 INFO DAGScheduler: ResultStage 0 (show at <console>:32) finished in
 10.702 s
16/03/09 17:18:46 INFO DAGScheduler: Job 0 finished: show at <console>:32, took
 10.839905 s
+---------------+-------------------+-------------------+-----------------
+-------------------+-------------------+-------------+---------------+
|requestbegintime|               adid|       impressionid|          referrer|
   useragent|          usercookie|            ip|             dt|
+---------------+-------------------+-------------------+-----------------
+-------------------+-------------------+-------------+---------------+
|   1239610346000|m9nwdo67Nx6q2kI25...|omkxkaRpNhGPDucAi...|cartoonnetwork.com|
Mozilla/4.0 (comp...|wcVWWTascoPbGt6bd...|69.191.224.234|2009-04-13-08-05|
|   1239611000000|NjriQjdODgWBKnkGJ...|AWtXPkfaWGOaNeL9O...|cartoonnetwork.com|
Mozilla/4.0 (comp...|OaMU1F2gE4CtADVHA...| 57.34.133.110|2009-04-13-08-05|
|   1239610462000|Irpv3oiu0I5QNQiwS...|i1LDq44LRSJF0hbmh...|cartoonnetwork.com|
Mozilla/4.0 (comp...|QSb3wkLR4JAIut4Uq...|42.174.193.253|2009-04-13-08-05|
|   1239611007000|q2Awfnpe0JAvhInaI...|s3HvTflPB8JIE0IuM...|cartoonnetwork.com|
Mozilla/4.0 (comp...|QSb3wkLR4JAIut4Uq...|42.174.193.253|2009-04-13-08-05|
```

```
|   1239610398000|c362vpAB0soPKGHRS...|jeas5nXbQInGAgFB8...|cartoonnetwork.com|
Mozilla/4.0 (comp...|k96n5PnUmwHKfiUI0...|  51.131.29.87|2009-04-13-08-05|
|   1239610600000|cjBTpruoaiEtqLuMX...|XwlohBSs8Ipxs1bRa...|cartoonnetwork.com|
Mozilla/4.0 (comp...|k96n5PnUmwHKfiUI0...|  51.131.29.87|2009-04-13-08-05|
|   1239610804000|Ms3eJHNAEItpxvimd...|4SIj4pGmgVLl625BD...|cartoonnetwork.com|
Mozilla/4.0 (comp...|k96n5PnUmwHKfiUI0...|  51.131.29.87|2009-04-13-08-05|
|   1239610872000|h5bccHX6wJReDi1jL...|EFAWIiBdVfnxwAMWP...|cartoonnetwork.com|
Mozilla/4.0 (comp...|k96n5PnUmwHKfiUI0...|  51.131.29.87|2009-04-13-08-05|
|   1239610365000|874NBpGmxNFfxEPKM...|xSvE4XtGbdtXPF2Lb...|cartoonnetwork.com|
Mozilla/5.0 (Maci...|eWDEVVUphlnRa273j...| 22.91.173.232|2009-04-13-08-05|
|   1239610348000|X8gISpUTSqh1A5reS...|TrFblGT99AgE75vuj...|       corriere.it|
Mozilla/4.0 (comp...|tX1sMpnhJUhmAF7AS...|   55.35.44.79|2009-04-13-08-05|
|   1239610743000|kbKreLWB6QVueFrDm...|kVnxx9Ie2i3OLTxFj...|       corriere.it|
Mozilla/4.0 (comp...|tX1sMpnhJUhmAF7AS...|   55.35.44.79|2009-04-13-08-05|
|   1239610812000|9lxOSRpEi3bmEeTCu...|1B2sff99AEIwSuLVV...|       corriere.it|
Mozilla/4.0 (comp...|tX1sMpnhJUhmAF7AS...|   55.35.44.79|2009-04-13-08-05|
|   1239610876000|lijjmCf2kuxfBTnjL...|AjvufgUtakUFcsIM9...|       corriere.it|
Mozilla/4.0 (comp...|tX1sMpnhJUhmAF7AS...|   55.35.44.79|2009-04-13-08-05|
|   1239610941000|t8t8trgjNRPIlmxuD...|agu2u2TCdqWP08rAA...|       corriere.it|
Mozilla/4.0 (comp...|tX1sMpnhJUhmAF7AS...|   55.35.44.79|2009-04-13-08-05|
|   1239610490000|OGRLPVNGxiGgrCmWL...|mJg2raBUpPrC8OlUm...|       corriere.it|
Mozilla/4.0 (comp...|r2k96t1CNjSU9fJKN...|   71.124.66.3|2009-04-13-08-05|
|   1239610556000|OnJID12x0RXKPUgrD...|P7Pm2mPdW6wO8KA3R...|       corriere.it|
Mozilla/4.0 (comp...|r2k96t1CNjSU9fJKN...|   71.124.66.3|2009-04-13-08-05|
|   1239610373000|WflsvKIgOqfIE5KwR...|TJHd1VBspNcua0XPn...|       corriere.it|
Mozilla/5.0 (Maci...|fj2L1ILTFGMfhdrt3...| 75.117.56.155|2009-04-13-08-05|
|   1239610768000|4MJR0XxiVCU1ueXKV...|1OhGWmbvKf8ajoU8a...|       corriere.it|
Mozilla/5.0 (Maci...|fj2L1ILTFGMfhdrt3...| 75.117.56.155|2009-04-13-08-05|
|   1239610832000|gWIrpDiN57i3sHatv...|RNL4C7xPi3tdar2Uc...|       corriere.it|
Mozilla/5.0 (Maci...|fj2L1ILTFGMfhdrt3...| 75.117.56.155|2009-04-13-08-05|
|   1239610789000|pTne9k62kJ14QViXI...|RVxJVIQousjxUVI3r...|        pixnet.net|
Mozilla/5.0 (Maci...|1bGOKiBD2xmui9OkF...| 33.176.101.80|2009-04-13-08-05|
+----------------+--------------------+--------------------+------------------
+--------------------+--------------------+-------------+---------------+
only showing top 20 rows


scala>
```

示例：创建一个 HCatalog 表并使用 Pig 写入该表

您可以创建一个 HCatalog 表并使用 Apache Pig 对此表进行写入，方式是通过使用 Amazon S3 中的数据源的 HCatStorer。HCatalog 要求您禁用直接写入，否则操作将

无提示地失败。要同时将 mapred.output.direct.NativeS3FileSystem 和 mapred.output.direct.EmrFileSystem 配置设置为 false，可以使用 mapred-site 分类，或者通过 Grunt shell 手动操作。以下示例显示一个使用 HCat CLI 创建的表，后跟 Grunt shell 中执行的命令（用于从 Amazon S3 中的示例数据文件填充表）。

要运行此示例，请使用 SSH 连接到主节点。

使用以下内容创建一个 HCatalog 脚本文件 wikicount.q，这样就会创建一个名为 wikicount 的 HCatalog 表。

```
CREATE EXTERNAL TABLE IF NOT EXISTS wikicount(
col1 string,
col2 bigint
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\001'
STORED AS ORC
LOCATION 's3://MyBucket/hcat/wikicount';
```

使用 HCat CLI 命令执行文件中的脚本。

```
hcat -f wikicount.q
```

接下来，使用 -useHCatalog 选项启动 Grunt shell，将配置设置为禁用直接写入，从 S3 位置加载数据，然后将结果写入 wikicount 表。

```
pig -useHCatalog
SET mapred.output.direct.NativeS3FileSystem false;
SET mapred.output.direct.EmrFileSystem false;
A = LOAD 's3://support.elasticmapreduce/training/datasets/wikistats_tiny/' USING
 PigStorage(' ') AS (Site:chararray, page:chararray, views:int, total_bytes:long);
B = GROUP A BY Site;
C = FOREACH B GENERATE group as col1, COUNT(A) as col2;
STORE C INTO 'wikicount' USING org.apache.hive.hcatalog.pig.HCatStorer();
```

HCatalog 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 HCatalog 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

HCatalog 版本信息

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.14.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.13.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.12.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
		yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.11.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.11.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.10.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.10.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.9.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.9.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.8.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.8.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.7.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.36.1	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.36.0	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.6.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.35.0	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.5.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.4.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.3.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.3.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.2.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.2.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.1.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.1.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-6.0.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-6.0.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.34.0	2.3.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.33.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.33.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.32.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.32.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.31.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.31.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.30.2	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.30.1	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.30.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mariadb-server
emr-5.29.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.28.1	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.28.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.27.1	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.27.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.26.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.25.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.24.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.24.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.23.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.23.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.22.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.21.2	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.21.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.21.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.20.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.20.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.19.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.19.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.18.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.18.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.17.2	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.17.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.17.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.16.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.16.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.15.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.15.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.14.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.14.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.14.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.13.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.13.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.12.3	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.12.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.12.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.12.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.11.4	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.11.3	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.11.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.11.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.11.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.10.1	2.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.10.0	2.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.9.1	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.9.0	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.8.3	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.8.2	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.8.1	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.8.0	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.7.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.7.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.6.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.6.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.5.4	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.5.3	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.5.2	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.5.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.5.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.4.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.4.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.3.2	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.3.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.3.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.2.3	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.2.2	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.2.1	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.2.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.1.1	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.1.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-5.0.3	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-5.0.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.9.6	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.9.5	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.9.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.9.3	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.9.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.9.1	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.8.5	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.8.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.8.3	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.8.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.8.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.7.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.7.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.7.1	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server
emr-4.7.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.6.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, hive-metastore-server, mysql-server
emr-4.5.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, hive-metastore-server, mysql-server

Amazon EMR 发行版标签	HCatalog 版本	随 HCatalog 安装的组件
emr-4.4.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hcatalog-client, hcatalog-server, hcatalog-webhcat-server, hive-client, hive-metastore-server, mysql-server

Apache Hive

Hive 是一种开源数据仓库和分析软件程序包，基于 Hadoop 集群运行。Hive 脚本使用称为 Hive QL（查询语言）的类似于 SQL 的语言，对编程模型进行抽象，并支持典型的数据仓库交互。Hive 使您能够避免根据有向无环图（DAG）或 MapReduce 程序以较低级别的计算机语言（例如 Java）编写 Tez 任务的复杂性。

Hive 通过包含序列化格式来扩展 SQL 范例。您也可以通过创建与您的数据匹配的表架构自定义查询处理，而无需接触到数据本身。SQL 仅仅支持原始值类型（如日期、数字和字符串）；与此相反，Hive 表中的值是结构化元素，如 JSON 对象、任何用户定义的数据类型或以 Java 编写的任何函数。

有关 Hive 的更多信息，请参阅 http://hive.apache.org/。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Hive 的版本，以及 Amazon EMR 随 Hive 一起安装的组件。

有关此发行版中随 Hive 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Hive 版本信息

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.14.0	Hive 3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Hive 的版本，以及 Amazon EMR 随 Hive 一起安装的组件。

有关此发行版中随 Hive 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Hive 版本信息

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.36.1	Hive 2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

您可以使用 Amazon EMR 构件存储库构建针对特定 Amazon EMR 发行版（从 Amazon EMR 发行版 5.18.0 开始）附带的准确版本的库和依赖项的任务代码。有关更多信息，请参阅使用 Amazon EMR 项目存储库检查依赖项。

主题

- Amazon EMR 上的 Hive 的区别和注意事项
- 为 Hive 配置外部元存储
- 使用 Hive JDBC 驱动程序
- 改进 Hive 性能
- 使用 Hive Live Long and Process（LLAP）
- Hive 中的加密
- Hive 发行历史记录

Amazon EMR 上的 Hive 的区别和注意事项

Amazon EMR 上的 Apache Hive 和 Apache Hive 之间的区别

本节介绍 Amazon EMR 上的 Hive 和默认 Hive 版本（http://svn.apache.org/viewvc/hive/branches/）之间的区别。

Hive 授权

Amazon EMR 对于 HDFS 支持 Hive 授权，但对于 EMRFS 和 Amazon S3 不支持此授权。默认情况下，Amazon EMR 集群在禁用授权的状态下运行。

Amazon S3 中的 Hive 文件合并操作

如果 `hive.merge.mapfiles` 为 true，那么 Apache Hive 会在仅 map 作业结束时合并小文件，且仅在平均的作业输出大小低于 `hive.merge.smallfiles.avgsize` 设置时才会触发合并。如果最终输出路径位于 HDFS 中，那么 Amazon EMR Hive 的行为将完全相同。如果输出路径位于 Amazon S3 中，将忽略 `hive.merge.smallfiles.avgsize` 参数。在那种情况下，如果 `hive.merge.mapfiles` 设置为 `true`，会始终触发合并任务。

ACID 事务和 Amazon S3

Amazon EMR 6.1.0 及更高版本支持 Hive ACID（原子性、一致性、隔离、持久性）事务，因此它符合数据库的 ACID 属性。借助此功能，您可以使用 Amazon Simple Storage Service（Amazon S3）中的数据在 Hive 托管表中运行 INSERT、UPDATE、DELETE 和 MERGE 操作。

Hive Live Long and Process (LLAP)

默认 Apache Hive 版本 2.0 中添加的 LLAP 功能在 Amazon EMR 5.0 发行版上的 Hive 2.1.0 中不受支持。

Amazon EMR 版本 6.0.0 及更高版本支持 Hive 的 Live Long and Process（LLAP）功能。有关更多信息，请参阅使用 Hive LLAP。

Hive 在 Amazon EMR 发行版 4.x 和 5.x 之间的不同

本节介绍在将 Hive 实现从 Amazon EMR 4.x 发行版上的 Hive 1.0.0 版迁移到 Amazon EMR 5.x 发行版上的 Hive 2.x 之前要考虑的区别。

操作区别和注意事项

- 添加了对 ACID（原子性、一致性、隔离和持久性）事务的支持：Amazon EMR 4.x 上的 Hive 1.0.0 和默认 Apache Hive 之间的这一区别已经消除。
- 已消除对 Amazon S3 的直接写入：Amazon EMR 上的 Hive 1.0.0 和默认 Apache Hive 之间的这一区别已经消除。Amazon EMR 5.x 发行版上的 Hive 2.1.0 现在会创建存储在 Amazon S3 中的临时文件、从这些文件中读取数据以及向其写入数据。因此，要读取和写入同一个表，您不再需要在集群的本地 HDFS 文件系统中创建一个临时表作为解决办法。如果您使用受版本控制的存储桶，请确保如下所述管理这些临时文件。
- 使用 Amazon S3 受版本控制的存储桶时管理临时文件：当您在生成数据的目的地是 Amazon S3 的环境中运行 Hive 查询时，会创建许多临时文件和目录。这是新行为，如上所述。如果您使用受版本控制的 S3 存储桶，那么不删除这些临时文件会使 Amazon S3 凌乱并产生费用。请调整生命周期规则，以便包含 `/_tmp` 前缀的数据在一个短周期 (例如，五天) 后被删除。有关更多信息，请参阅指定生命周期配置。
- Log4j 已更新到 log4j 2：如果您使用 log4j，则可能因为此升级而需要更改您的日志记录配置。有关更多信息，请参阅 Apache log4j 2。

性能区别和注意事项

- 使用 Tez 时的性能区别：对于 Amazon EMR 5.x 发行版，Tez 是 Hive（而非 MapReduce）的默认执行引擎。Tez 为大多数工作流提供了改进的性能。
- 具有多个分区的表：生成大量动态分区的查询可能会失败，并且执行从具有多个分区的表中选择的查询可能需要比预期更长的时间。例如，从 100,000 个分区中进行选择可能需要 10 分钟或更长时间。

Amazon EMR 上的 Hive 的额外功能

Amazon EMR 通过支持 Hive 与其他 Amazon 服务集成的读取和写入 Amazon Simple Storage Service（Amazon S3）和 DynamoDB 等新功能来扩展 Hive。

Hive 中的变量

您可以使用美元符号和大括号在脚本中包括变量。

```
add jar ${LIB}/jsonserde.jar
```

如以下示例所示，您可以在命令行上使用 -d 参数将这些变量的值传递给 Hive：

```
-d LIB=s3://elasticmapreduce/samples/hive-ads/lib
```

您还可以将值传递到执行 Hive 脚本的步骤。

使用控制台将变量值传递到 Hive 步骤

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 选择创建集群。
3. 在步骤部分中，对于 Add Step (添加步骤)，从列表中选择 Hive Program (Hive 程序)，然后选择 Configure and add (配置并添加)。
4. 在 Add Step (添加步骤) 对话框中，参考下表指定参数，然后选择添加。

Field	操作
脚本 S3 位置*	指定 Amazon S3 中脚本存储位置的 URI。该值的形式必须是：*BucketName* /*path*/*ScriptName* 。例如：s3://elasticmapreduce/samples/hive-ads/libs/response-time-stats.q 。
输入 S3 位置	指定 Amazon S3 中输入文件存储位置的 URI（可选）。该值的形式必须是：*BucketName* /*path*/。如果指定完毕，则以名为“INPUT”的参数将其传递到 Hive 脚本。例如：s3://elasticmapreduce/samples/hive-ads/tables/ 。
输出 S3 位置	指定您希望在 Amazon S3 中存储该输出的位置的 URI（可选）。该值的形式必须是：*BucketName* /*path*。如果指定完毕，则以名为“OUTPUT”的参数将其传递到 Hive 脚本。例如：s3://mybucket/hive-ads/output/ 。
Arguments	输入要传递到 Hive 的参数列表（以空格分隔的字符串）。如果您在名为 ${SAMPLE} 的 Hive 脚本中定义了一个路径变量，如： `CREATE EXTERNAL TABLE logs (requestBeginTime STRING, requestEndTime STRING, hostname STRING) PARTITIONED BY (dt STRING) \ ROW FORMAT serde 'com.amazon.elasticmapreduce.JsonSerde'`

Field	操作
	`WITH SERDEPROPERTIES ( 'paths'='requestBeginTime, requestEndTime, hostname' ) LOCATION '${SAMPLE}/tables/impressions';` 要传递该变量的值，请在参数窗口中键入以下内容： `-d SAMPLE=s3://elasticmapreduce/samples/hive-ads/` .
出现故障时的操作	这决定了集群为响应任何错误而执行的操作。此设置的可能值为： • Terminate cluster (终止集群)：如果步骤失败，则终止集群。如果集群启用了终止保护和保持活动状态，则它不会终止。 • Cancel and wait (取消并等待)：如果步骤失败，则取消剩余步骤。如果集群启用了保持活动状态，则集群不会终止。 • 继续：如果该步骤失败，则继续到下一个步骤。

5. 根据需要选择值，然后选择创建集群。

使用 Amazon CLI 将变量值传递到 Hive 步骤

要使用 Amazon CLI 将变量值传递到 Hive 步骤，请使用 `--steps` 参数并包括参数列表。

-

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^) 。

```
aws emr create-cluster --name "Test cluster" --release-label emr-5.36.1 \
--applications Name=Hive Name=Pig --use-default-roles --ec2-attributes
 KeyName=myKey --instance-type m5.xlarge --instance-count 3 \
--steps Type=Hive,Name="Hive Program",ActionOnFailure=CONTINUE,Args=[-f,s3://
elasticmapreduce/samples/hive-ads/libs/response-time-stats.q,-d,INPUT=s3://
elasticmapreduce/samples/hive-ads/tables,-d,OUTPUT=s3://mybucket/hive-ads/output/,-
d,SAMPLE=s3://elasticmapreduce/samples/hive-ads/]
```

有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅https://docs.amazonaws.cn/cli/latest/reference/emr。

使用 Java 开发工具包将变量值传递到 Hive 步骤

- 以下示例演示如何使用开发工具包将变量传递到步骤。有关更多信息，请参阅《Amazon SDK for Java API 参考》中的 Class StepFactory。

```
StepFactory stepFactory = new StepFactory();

   StepConfig runHive = new StepConfig()
     .withName("Run Hive Script")
     .withActionOnFailure("TERMINATE_JOB_FLOW")
     .withHadoopJarStep(stepFactory.newRunHiveScriptStep(“s3://mybucket/script.q”,
      Lists.newArrayList(“-d”,”LIB= s3://elasticmapreduce/samples/hive-ads/lib”));
```

Amazon EMR Hive 查询可适应部分 DynamoDB 架构

在查询 DynamoDB 表时，Amazon EMR Hive 允许您指定一部分列作为数据筛选条件，而不要求您的查询包含所有列，因此可提供最大的灵活性。当采用稀疏数据库架构，并希望根据一些列来筛选记录 (例如根据时间戳筛选) 时，这种部分架构查询技术可以发挥作用。

以下示例显示了如何使用 Hive 查询执行下列操作：

- 创建 DynamoDB 表。
- 选择 DynamoDB 中的一部分项目（行）并进一步将数据范围缩小到特定列。
- 将结果数据复制到 Amazon S3。

```
DROP TABLE dynamodb;
DROP TABLE s3;

CREATE EXTERNAL TABLE dynamodb(hashKey STRING, recordTimeStamp BIGINT, fullColumn
 map<String, String>)
    STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
    TBLPROPERTIES (
     "dynamodb.table.name" = "myTable",
     "dynamodb.throughput.read.percent" = ".1000",
     "dynamodb.column.mapping" = "hashKey:HashKey,recordTimeStamp:RangeKey");

CREATE EXTERNAL TABLE s3(map<String, String>)
     ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
     LOCATION 's3://bucketname/path/subpath/';
```

```
INSERT OVERWRITE TABLE s3 SELECT item fullColumn FROM dynamodb WHERE recordTimeStamp <
 "2012-01-01";
```

下表显示了从 DynamoDB 中选择任意项目组合的查询语法。

查询范例	结果描述
从 *table_name* 中选择 * ;	从指定表选择所有项目 (行) 并包括这些项目对应的所有列的数据。
从 *table_name* 中选择 * , 其中 *field_name* =*value* ;	从指定表选择一些项目 (行) 并包括这些项目对应的所有列的数据。
SELECT *column1_name* , *column2_name* , *column3_name* FROM *table_name* ;	从指定表选择所有项目 (行) 并包括这些项目对应的一些列的数据。
SELECT *column1_name* , *column2_name* , *column3_name* FROM *table_name* WHERE *field_name* =*value*;	从指定表选择一些项目 (行) 并包括这些项目对应的一些列的数据。

在不同 Amazon 区域的 DynamoDB 表之间复制数据

Amazon EMR Hive 提供了可以为每个 DynamoDB 表设置的 dynamodb.region 属性。当两个表的 dynamodb.region 设置不同时，您在两个表之间执行的所有数据复制将自动在指定区域之间发生。

以下示例显示了如何通过用于设置 dynamodb.region 属性的 Hive 脚本创建 DynamoDB 表：

> **Note**
>
> 每个表的 region 属性会覆盖全局 Hive 属性。

```
CREATE EXTERNAL TABLE dynamodb(hashKey STRING, recordTimeStamp BIGINT, map<String,
 String> fullColumn)
    STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
    TBLPROPERTIES (
     "dynamodb.table.name" = "myTable",
```

```
     "dynamodb.region" = "eu-west-1",
     "dynamodb.throughput.read.percent" = ".1000",
     "dynamodb.column.mapping" = "hashKey:HashKey,recordTimeStamp:RangeKey");
```

设置每个表的 DynamoDB 吞吐量值

Amazon EMR Hive 允许您在表定义中设置每个表的 DynamoDB readThroughputPercent 和 writeThroughputPercent 设置。以下 Amazon EMR Hive 脚本显示了如何设置吞吐量值。有关 DynamoDB 吞吐量值的更多信息，请参阅指定表的读取和写入要求。

```
CREATE EXTERNAL TABLE dynamodb(hashKey STRING, recordTimeStamp BIGINT, map<String,
 String> fullColumn)
    STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
    TBLPROPERTIES (
     "dynamodb.table.name" = "myTable",
     "dynamodb.throughput.read.percent" = ".4",
     "dynamodb.throughput.write.percent" = "1.0",
     "dynamodb.column.mapping" = "hashKey:HashKey,recordTimeStamp:RangeKey");
```

为 Hive 配置外部元存储

默认情况下，Hive 会在主节点的文件系统上的 MySQL 数据库中记录元存储信息。元存储包含表以及在其上构建表的基础数据的描述，包括分区名称、数据类型等。集群终止后，所有集群节点都会关闭，包括主节点。当发生此情况时，本地数据会丢失，因为节点文件系统使用的是短暂存储。如果您需要保留元存储，则必须创建一个存在于集群外部的外部元存储。

对于外部元存储，您有两个选项：

- Amazon Glue 数据目录（仅限 Amazon EMR 发行版 5.8.0 或更高版本）。

  有关更多信息，请参阅将 Amazon Glue 数据目录用作 Hive 元存储。。
- Amazon RDS 或 Amazon Aurora。

  有关更多信息，请参阅使用外部 MySQL 数据库或 Amazon Aurora。

> **Note**
>
> 如果您使用 Hive 3 并遇到 Hive 元数据仓的连接太多的问题，请将参数 `datanucleus.connectionPool.maxPoolSize` 配置为一个较小的值或增加数据库服务器

可以处理的连接数。连接数量增加是因 Hive 计算 JDBC 连接的最大数量的方式所致。要计算可确保最佳性能的值，请参阅 Hive 配置属性。

将 Amazon Glue 数据目录用作 Hive 元存储。

使用 Amazon EMR 发行版 5.8.0 或更高版本，您可以将 Hive 配置为使用 Amazon Glue 数据目录作为元存储。当您需要持久的元数据仓或由不同集群、服务、应用程序和 Amazon 账户共享的元数据仓时，我们建议使用此配置。

Amazon Glue 是一项完全托管式提取、转换和加载（ETL）服务，使您能够轻松且经济高效地对数据进行分类、清理和扩充，并在各种数据存储之间可靠地移动数据。Amazon Glue 数据目录跨各种数据源和数据格式提供统一的元数据存储库，从而不仅与 Amazon EMR 集成，还与 Amazon RDS、Amazon Redshift、Redshift Spectrum、Athena 以及任何与 Apache Hive 元存储兼容的应用程序集成。AmazonGlue 爬网程序能够自动从 Amazon S3 源数据推断架构，从而将关联的元数据存储在数据目录中。有关数据目录的更多信息，请参阅《Amazon Glue 开发人员指南》中的填充 Amazon Glue 数据目录。

使用 Amazon Glue 需单独付费。在数据目录中存储和访问数据需按月付费；为 Amazon Glue ETL 作业和爬网程序运行时按小时费率付费（按分计费）；为每个预置的开发端点支付每小时费率（按分计费）。数据目录让您最多可免费存储一百万个对象。如果您存储一百万个以上的对象，将需要为超过一百万的每 100,000 个对象支付 1 美元。数据目录中的对象为表、分区或数据库。有关更多信息，请参阅 Glue 定价。

Important

如果您在 2017 年 8 月 14 日之前使用 Amazon Athena 或 Amazon Redshift Spectrum 创建了表，则数据库和表将存储在 Athena 托管式目录中，该目录与 Amazon Glue 数据目录相互独立。要将 Amazon EMR 与这些表集成，您必须升级到 Amazon Glue 数据目录。有关更多信息，请参阅《Amazon Athena 用户指南》中的升级到 Amazon Glue 数据目录。

指定 Amazon Glue 数据目录作为元存储

您可以使用Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 指定 Amazon Glue 数据目录作为元存储。在使用 CLI 或 API 时，您可以使用 Hive 配置分类指定数据目录。此外，使用 Amazon EMR 5.16.0 及更高版本时，您可以使用配置分类指定其他 Amazon Web Services 账户中的数据目录。在使用控制台时，您可以使用 Advanced Options (高级选项) 或 Quick Options (快速选项) 指定数据目录。

New console

使用新控制台指定 Amazon Glue 数据目录作为 Hive 元存储

1. 登录 Amazon Web Services Management Console 并打开 Amazon EMR 控制台，网址为 https://console.aws.amazon.com/emr。
2. 在左侧导航窗格中的 EMR on EC2 下，选择 Clusters（集群），然后选择 Create cluster（创建集群）。
3. 在 Application bundle（应用程序包）下，选择 Core Hadoop（核心 Hadoop）、HBase 或 Custom（自定义）。如果您自定义集群，请确保选择 Hive 或 HCatalog 作为应用程序之一。
4. 在Amazon Glue 数据目录设置下，选择用于 Hive 表元数据复选框。
5. 选择适用于集群的任何其他选项。
6. 要启动集群，选择 Create cluster（创建集群）。

Old console

使用旧控制台指定 Amazon Glue 数据目录作为 Hive 元存储

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 对于 Release (版本)，选择 emr-5.8.0 或更高版本。
4. 在 Release (版本) 下，选择 Hive 或 HCatalog。
5. 在 Amazon Glue Data Catalog settings (Amazon Glue 数据目录设置) 下，选择 Use for Hive table metadata (用于 Hive 表元数据)。
6. 根据需要为您的集群选择其他选项，选择 Next (下一步)，然后根据需要为您的应用程序配置其他集群选项。

CLI

使用 Amazon CLI 指定 Amazon Glue 数据目录作为 Hive 元存储

有关使用 Amazon CLI 和 EMR API 指定配置分类的更多信息，请参阅配置应用程序。

- 使用 `hive-site` 配置分类指定 `hive.metastore.client.factory.class` 的值，如下例所示：

```
[
  {
    "Classification": "hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
 "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  }
]
```

在 EMR 发行版本 5.28.0、5.28.1、5.29.0 或 6.x 中，如果您要将 Amazon Glue 数据目录用作元数据来创建集群，请将 `hive.metastore.schema.verification` 设置为 `false`。这可以防止 Hive 和 HCatalog 根据 MySQL 验证元数据仓架构。如果没有此配置，主实例组将在 Hive 或 HCatalog 上进行重新配置后暂停。

```
[
  {
    "Classification": "hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
 "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory",
      "hive.metastore.schema.verification": "false"
    }
  }
]
```

如果您在 EMR 发行版 5.28.0、5.28.1 或 5.29.0 上已有集群，则可以使用以下信息将主实例组 `hive.metastore.schema.verification` 设置为 `false`：

```
    Classification = hive-site
    Property       = hive.metastore.schema.verification
    Value          = false
```

要在其他Amazon账户中指定数据目录，请添加 `hive.metastore.glue.catalogid` 属性，如以下示例所示。将 *`acct-id`* 替换为数据目录的Amazon账户。

```
[
  {
    "Classification": "hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
 "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory",
      "hive.metastore.schema.verification": "false",
      "hive.metastore.glue.catalogid": "acct-id"
    }
  }
]
```

IAM 权限

集群的 EC2 实例配置文件必须具有适用于 Amazon Glue 操作的 IAM 权限。此外，如果您为 Amazon Glue 数据目录对象启用加密，还必须允许该角色加密、解密和生成用于加密的Amazon KMS key。

适用于 Amazon Glue 操作的权限

如果使用适用于 Amazon EMR 默认的 EC2 实例配置文件，则无需执行任何操作。附加到 `EMR_EC2_DefaultRole` 的 `AmazonElasticMapReduceforEC2Role` 托管式策略允许所有必要 Amazon Glue 操作。但是，如果您指定自定义 EC2 实例配置文件和权限，则必须配置合适的 Amazon Glue 操作。使用 `AmazonElasticMapReduceforEC2Role` 托管策略作为起点。如需了解更多信息，请参阅《Amazon EMR 管理指南》中的集群 EC2 实例的服务角色（EC2 实例配置文件）。

用于加密和解密 Amazon Glue 数据目录的权限

您的实例配置文件需要使用密钥加密和解密数据的权限。如果以下语句适用，您不必配置这些权限：

- 您使用 Amazon Glue 的托管式密钥启用 Amazon Glue Data Catalog 对象的加密。
- 您使用的是同一Amazon Web Services 账户的集群，其作为 AmazonGlue Data Catalog。

否则，您必须将以下语句添加到附加到 EC2 实例配置文件的权限策略。

```
[
    {
        "Version": "2012-10-17",
        "Statement": [
```

```
            {
                "Effect": "Allow",
                "Action": [
                    "kms:Decrypt",
                    "kms:Encrypt",
                    "kms:GenerateDataKey"
                ],
                "Resource": "arn:aws:kms:region:acct-
id:key/12345678-1234-1234-1234-123456789012"
            }
        ]
    }
]
```

有关 Amazon Glue 数据目录加密的更多信息，请参阅《Amazon Glue 开发人员指南》中的加密您的数据目录。

基于资源的权限

如果您将 Amazon Glue 与 Amazon EMR 中的 Hive、Spark 或 Presto 结合使用，Amazon Glue 支持使用基于资源的策略来控制对数据目录资源的访问权限。这些资源包括数据库、表、连接和用户定义的函数。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的 Amazon Glue 资源策略。

当使用基于资源的策略来限制从 Amazon EMR 中访问 Amazon Glue 时，在权限策略中指定的委托人必须是与创建集群时指定的 EC2 实例配置文件相关联的角色 ARN。例如，对于附加到目录的基于资源的策略，您可以使用以下示例中显示的格式为集群 EC2 实例的默认服务角色指定角色 ARN，将 *EMR_EC2_DefaultRole* 指定为 `Principal`：

```
arn:aws:iam::acct-id:role/EMR_EC2_DefaultRole
```

acct-id 可以与 Amazon Glue 账户 ID 不同。这允许从不同账户中的 EMR 集群进行访问。您可以指定多个委托人，且每个委托人都可以来自不同的账户。

使用 Amazon Glue 数据目录时的注意事项

在将 Amazon Glue 数据目录用作 Hive 的元存储时，请考虑以下项目：

- 不支持使用 Hive Shell 添加辅助 JAR。作为解决方法，请使用 `hive-site` 配置分类来设置 `hive.aux.jars.path` 属性，它会将辅助 JAR 添加到 Hive 类路径中。
- 不支持 Hive 事务。
- 不支持在 Amazon Glue 中重命名表。

- 当您创建 Hive 表而不指定 LOCATION 时，表数据存储在通过 hive.metastore.warehouse.dir 属性指定的位置。默认情况下，这是 HDFS 中的一个位置。如果另一个集群需要访问该表，则它将失败，除非它有足够的权限访问创建该表的集群。此外，由于 HDFS 存储是暂时性的，因此如果集群终止，表数据将丢失，并且必须重新创建该表。建议您在使用 Amazon Glue 创建 Hive 表时，指定 Amazon S3 中的一个 LOCATION。此外，也可以使用 hive-site 配置分类来为 hive.metastore.warehouse.dir 指定 Amazon S3 中的位置，它适用于所有 Hive 表。如果表在 HDFS 位置创建，并且创建该表的集群仍在运行，您可以在 Amazon Glue 中更新 Amazon S3 中表的位置。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的使用 Amazon Glue 控制台上的表。
- 不支持包含引号和撇号的分区值，例如 PARTITION (owner="Doe's").
- emr-5.31.0 及更高版本支持列统计数据。
- 不支持使用 Hive 授权。作为替代方案，考虑使用基于 Amazon Glue 资源的策略。有关更多信息，请参阅将用于 Amazon EMR 访问的基于资源的策略用于 Amazon Glue 数据目录。
- 不支持 Hive 约束。
- 不支持 Hive 中基于成本的优化。
- 不支持设置 hive.metastore.partition.inherit.table.properties。
- 不支持使用以下元存储常量：BUCKET_COUNT, BUCKET_FIELD_NAME, DDL_TIME, FIELD_TO_DIMENSION, FILE_INPUT_FORMAT, FILE_OUTPUT_FORMAT, HIVE_FILTER_FIELD_LAST_ACCESS, HIVE_FILTER_FIELD_OWNER, HIVE_FILTER_FIELD_PARAMS, IS_ARCHIVED, META_TABLE_COLUMNS, META_TABLE_COLUMN_TYPES, META_TABLE_DB, META_TABLE_LOCATION, META_TABLE_NAME, META_TABLE_PARTITION_COLUMNS, META_TABLE_SERDE, META_TABLE_STORAGE, ORIGINAL_LOCATION。
- 使用谓词表达式时，显式值必须位于比较运算符的右侧，否则查询可能会失败。
  - 正确：SELECT * FROM mytable WHERE time > 11
  - 错误：SELECT * FROM mytable WHERE 11 > time
- Amazon EMR 5.32.0 和 6.3.0 及更高版本支持在谓词表达式中使用用户定义的函数 (UDF)。使用早期版本时，查询可能因 Hive 尝试优化查询执行的方式而失败。
- 不支持临时表。
- 建议通过 Amazon EMR 使用应用程序创建表，而不是直接使用 Amazon Glue 创建。通过 Amazon Glue 创建表可能会导致必填字段丢失，并导致查询异常。
- 在 EMR 5.20.0 或更高版本中，当使用 Amazon Glue 数据目录作为元存储时，会自动为 Spark 和 Hive 启用并行分区修剪。此更改通过并行执行多个请求来检索分区，显著缩短查询计划时间。可同时执行的分段总数介于 1 到 10 之间。默认值为 5，这是建议的设置。您可以通过以下方式更改该

值：指定 `hive-site` 配置分类中的属性 `aws.glue.partition.num.segments`。如果发生节流，则可以通过将值更改为 1 来关闭此功能。有关更多信息，请参阅 Amazon Glue 分段结构。

使用外部 MySQL 数据库或 Amazon Aurora

要将外部 MySQL 数据库或 Amazon Aurora 用作您的 Hive 元数据仓，您可在 Hive 中覆盖该元数据仓的原定设置配置值，以指定外部数据库位置 – 要么在 Amazon RDS MySQL 实例上，要么在 Amazon Aurora PostgreSQL 实例上。

> **Note**
>
> Hive 既不支持对元存储表的并发写入访问权限，也不阻止此权限。如果要在两个集群间共享元数据仓信息，您必须确保不会同时写入同一元数据仓表，除非您要写入同一元数据仓表的不同分区。

以下步骤介绍了如何覆盖 Hive 元数据仓位置的默认配置值和使用重新配置的元数据仓位置启动集群。

创建位于 EMR 集群外的元数据仓

1. 创建 MySQL 或 Aurora PostgreSQL 数据库。如果您使用 PostgreSQL，则必须在预置集群之后对其进行配置。创建集群时只支持 MySQL。有关 Aurora MySQL 和 Aurora PostgreSQL 之间的区别的信息，请参阅 Amazon Aurora MySQL 概述和使用 Amazon Aurora PostgreSQL。有关如何创建 Amazon RDS 数据库的一般信息，请参阅 https://aws.amazon.com/rds/。
2. 修改您的安全组，以允许在数据库与 ElasticMapReduce-Master 安全组之间建立 JDBC 连接。有关如何修改安全组以进行访问的信息，请参阅使用 Amazon EMR 托管式安全组。
3. 在 `hive-site.xml` 中设置 JDBC 配置值：

> **Important**
>
> 如果您提供敏感信息（如密码）至 Amazon EMR 配置 API，该信息将仅对拥有充分权限的账户显示。如果您担心此信息可能对其他用户显示，可通过创建以显式方式拒绝 `elasticmapreduce:DescribeCluster` API 密钥许可的角色来使用管理账户创建集群并限制其他用户 (IAM 用户或具有委派凭证的用户) 访问集群服务。

a. 创建一个名为 hiveConfiguration.json 的配置文件，该文件包含对 hive-site.xml 的编辑，如以下示例所示。

以 *hostname* 代替运行数据库的 Amazon RDS 实例的 DNS 地址，以及数据库凭证的*username* 和 *password* 。有关连接到 MySQL 和 Aurora 数据库实例的更多信息，请参阅《Amazon RDS 用户指南》https://docs.amazonaws.cn/AmazonRDS/latest/UserGuide/USER_ConnectToInstance.html中的连接到运行 MySQL 数据库引擎的数据库实例和连接到 Athena 数据库集群。javax.jdo.option.ConnectionURL 是 JDBC 元数据仓的 JDBC 连接字符串。javax.jdo.option.ConnectionDriverName 是 JDBC 元数据仓的驱动程序类名。

MySQL JDBC 驱动程序由 Amazon EMR 进行安装。

值属性不能包含任何空格或回车。所有内容应显示在一行中。

```
[
    {
      "Classification": "hive-site",
      "Properties": {
        "javax.jdo.option.ConnectionURL": "jdbc:mysql://hostname:3306/hive?
createDatabaseIfNotExist=true",
        "javax.jdo.option.ConnectionDriverName": "org.mariadb.jdbc.Driver",
        "javax.jdo.option.ConnectionUserName": "username",
        "javax.jdo.option.ConnectionPassword": "password"
      }
    }
  ]
```

b. 您创建集群时会引用 hiveConfiguration.json 文件，如以下 Amazon CLI 命令中所示。在此命令中，此文件存储在本地，您也可将此文件上传到 Amazon S3 并在此对其进行引用，例如 s3://*DOC-EXAMPLE-BUCKET*/hiveConfiguration.json。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ ) 。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --release-label emr-5.36.1 --instance-type m5.xlarge --
instance-count 2 \
--applications Name=Hive --configurations file://hiveConfiguration.json --use-
default-roles
```

4. 连接到集群的主节点。

   有关如何连接到主节点的信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。

5. 通过输入类似以下内容的命令，创建在 Amazon S3 上指定位置的 Hive 表：

```
CREATE EXTERNAL TABLE IF NOT EXISTS table_name
(
key int,
value int
)
LOCATION s3://DOC-EXAMPLE-BUCKET/hdfs/
```

6. 将 Hive 脚本添加到正在运行的集群。

您的 Hive 集群使用 Amazon RDS 中的元数据运行。通过指定该元数据仓位置，启动共享该元数据仓的所有其他 Hive 集群。

使用 Hive JDBC 驱动程序

您可以将常用商业智能工具（如 Microsoft Excel、MicroStrategy、QlikView 和 Tableau）与 Amazon EMR 结合使用来探索和显示您的数据。许多这类工具都需要 Java 数据库连接（JDBC）驱动程序或开放式数据库连接（ODBC）驱动程序。Amazon EMR 支持 JDBC 和 ODBC 连接。

以下示例演示了如何使用 SQL Workbench/J 作为 SQL 客户端，与 Amazon EMR 中的 Hive 集群连接。有关其他驱动程序的更多信息，请参阅将业务情报工具与 Amazon EMR 结合使用。

在安装和使用 SQL Workbench/J 之前，请下载驱动程序包并安装驱动程序。程序包中包含的驱动程序支持 Amazon EMR 发行版 4.0 及更高版本中提供的 Hive 版本。如需详细的发布说明和文档，请参阅程序包中的 PDF 文档。

- 最新 Hive JDBC 驱动程序包下载

  http://awssupportdatasvcs.com/bootstrap-actions/Simba/latest/

- Hive JDBC 驱动程序的旧版本

  http://awssupportdatasvcs.com/bootstrap-actions/Simba/

安装和配置 SQL Workbench

1. 从 http://www.sql-workbench.net/downloads.html 下载适用于您的操作系统的 SQL Workbench/J 客户端。
2. 安装 SQL Workbench/J。有关更多信息，请参阅 SQL Workbench/J 用户手册中的安装并启动 SQL Workbench/J。
3. Linux、Unix、Mac OS X 用户：在终端会话中，使用下面的命令创建到集群主节点的 SSH 隧道。将 *mmaster-public-dns-name* 替换为主节点 的公有 DNS 名称，将 *path-to-key-file* 替换为您的 Amazon EC2 私有密钥（ `.pem` ）文件的位置和文件名。

   ```
   ssh -o ServerAliveInterval=10 -i path-to-key-file -N -L 10000:localhost:10000
    hadoop@master-public-dns-name
   ```

   Windows 用户：在 PuTTY 会话中，使用 `10000` 作为 Source port (源端口)、使用 *master-public-dns-name*`:10000` 作为 Destination (目标) 来创建到集群主节点 的 SSH 隧道（ 使用本地端口转发 ）。将 *master-public-dns-name* 替换为主节点的公有 DNS 名称。
4. 将 JDBC 驱动程序添加到 SQL Workbench。

   a. 在 Select Connection Profile (选择连接配置文件) 对话框中，单击 Manage Drivers (管理驱动程序)。

   b. 单击 Create a new entry (创建新条目) （ 空白页 ）图标。

   c. 在名称字段中，键入 **`Hive JDBC`**。

   d. 对于 Library (库)，请单击 Select the JAR file(s) (选择 JAR 文件) 图标。

   e. 导航到包含提取的驱动程序的位置。选择您下载的 JDBC 驱动程序包版本中包含的驱动程序，然后单击 Open （ 打开 ）。

      例如，您的 JDBC 驱动程序包可能包括以下 JAR。

      ```
      hive_metastore.jar
      hive_service.jar
      HiveJDBC41.jar
      libfb303-0.9.0.jar
      libthrift-0.9.0.jar
      ```

```
log4j-1.2.14.jar
ql.jar
slf4j-api-1.5.11.jar
slf4j-log4j12-1.5.11.jar
TCLIServiceClient.jar
zookeeper-3.4.6.jar
```

f. 在 Please select one driver (请选择一个驱动程序) 对话框中，选择 `com.amazon.hive.jdbc41.HS2Driver`、确定。

5. 当您返回到 Manage Drivers (管理驱动程序) 对话框时，确认 Classname (类名) 字段已经填写，然后选择确定。
6. 当您返回到 Select Connection Profile (选择连接配置文件) 对话框时，验证驱动程序 字段是否设置为 Hive JDBC，然后在 URL 字段中提供以下 JDBC 连接字符串：`jdbc:hive2://localhost:10000/default`。
7. 选择确定进行连接。连接完成后，连接详细信息将显示在 SQL Workbench/J 窗口顶部。

有关使用 Hive 和 JDBC 界面的更多信息，请参阅 Apache Hive 文档中的 HiveClient 和 HiveJDBCInterface。

改进 Hive 性能

Amazon EMR 提供一些功能，有助于优化使用 Hive 查询、读取和写入保存在 Amazon S3 中的数据的性能。

S3 Select 可通过将处理“向下推送”到 Amazon S3 来提高某些应用程序中 CSV 和 JSON 文件的查询性能。

EMRFS S3 优化提交程序是 OutputCommitter 类的替代，这消除了列表和重命名操作，从而提高使用 EMRFS 编写文件 Amazon S3 时的性能。

主题

- 启用 Hive EMRFS S3 优化提交程序
- 将 S3 Select 与 Hive 结合使用以提高查询性能
- MSCK 优化

启用 Hive EMRFS S3 优化提交程序

Hive EMRFS S3 优化提交器是 EMR Hive 在利用 EMRFS 时为插入查询写入文件所用的另一种方式。提交程序消除了在 Amazon S3 上执行的列表和重命名操作，并提高了应用程序性能。该功能可在 EMR 5.34 和 EMR 6.5 及更高版本中使用。

启用提交程序

如果您想启用 EMR Hive 以使用 `HiveEMRFSOptimizedCommitter` 提交数据作为所有 Hive 托管式表和外部表的默认值，请在 EMR 6.5.0 或 EMR 5.34.0 集群中使用以下 `hive-site` 配置。

```
[
   {
      "classification": "hive-site",
      "properties": {
         "hive.blobstore.use.output-committer": "true"
      }
   }
]
```

> **Note**
>
> 请勿在 `hive.exec.parallel` 设置为 `true` 时启动此功能。

限制

下面是适用于标签的基本限制：

- 不支持启用 Hive 自动合并小文件。即使启用了优化提交程序，也会使用默认的 Hive 提交逻辑。
- 不支持 Hive ACID 表。即使启用了优化提交程序，也会使用默认的 Hive 提交逻辑。
- 写入的文件的文件命名术语从 Hive 的 `<task_id>_<attempt_id>_<copy_n>` 更改为 `<task_id>_<attempt_id>_<copy_n>_<query_id>`。例如，文件命名为 `s3://warehouse/table/partition=1/000000_0` 将更改为 `s3://warehouse/table/partition=1/000000_0-hadoop_20210714130459_ba7c23ec-5695-4947-9d98-8a40ef759222-1`。此 `query_id` 是用户名、时间戳和 UUID 的组合。

- 当自定义分区位于不同的文件系统（HDFS、S3）时，此功能将自动禁用。即使启用，也会使用默认的 Hive 提交逻辑。

将 S3 Select 与 Hive 结合使用以提高查询性能

对于 Amazon EMR 发行版 5.18.0 及更高版本，您可以将 S3 Select 与 Hive on Amazon EMR 搭配使用。S3 Select 可让应用程序仅从对象检索数据子集。对于 Amazon EMR，筛选要处理的大数据集的计算工作从集群“向下推送”到 Amazon S3，这可以在某些应用程序中提高性能和减少 Amazon EMR 与 Amazon S3 之间传输的数据量。

借助基于 CSV 和 JSON 文件的 Hive 表并在 Hive 会话期间将 `s3select.filter` 配置变量设置为 `true`，从而支持 S3 Select。有关更多信息以及示例，请参阅 在代码中指定 S3 Select。

S3 Select 是否适合我的应用程序？

建议您分别在使用和不使用 S3 Select 的情况下测试您的应用程序，以查看 S3 Select 是否适用于您的应用程序。

使用以下准则来确定您的应用程序是否为使用 S3 Select 的候选项：

- 您的查询将筛选掉原始数据集的一半以上的数据。
- 您的查询筛选谓词使用具有 Amazon S3 Select 支持的数据类型的列。有关更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的数据类型。
- 您在 Amazon S3 和 Amazon EMR 集群之间的网络连接具有良好的传输速度和可用带宽。Amazon S3 不压缩 HTTP 响应，因此响应大小可能会根据压缩的输入文件而增大。

注意事项和限制

- 使用客户提供的加密密钥进行的 Amazon S3 服务器端加密（SSE-C）与客户端加密都不受支持。
- 不支持 `AllowQuotedRecordDelimiters` 属性。如果指定该属性，则查询将失败。
- 仅支持采用 UTF-8 格式的 CSV 和 JSON 文件。不支持多行 CSV 和 JSON。
- 仅支持未压缩文件或 gzip、或 bzip2 文件。
- 不支持最后一行中的注释字符。
- 文件末尾的空行不会被处理。
- Hive on Amazon EMR 支持 S3 Select 所支持的基元数据类型。有关更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的数据类型。

在代码中指定 S3 Select

要在 Hive 表中使用 S3 Select，请通过将 `com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat` 指定为 `INPUTFORMAT` 类名称以及使用 `TBLPROPERTIES` 子句为 `s3select.format` 属性指定一个值，来创建表。

默认情况下，S3 Select 在您运行查询时处于禁用状态。通过在您的 Hive 会话中将 `s3select.filter` 设置为 `true` 来启用 S3 Select，如下所示。下面的示例演示了如何在通过基础 CSV 和 JSON 文件创建表时指定 S3 Select，然后使用简单的 select 语句查询表。

Example 基于 CSV 的 CREATE TABLE 语句

```
CREATE TABLE mys3selecttable (
col1 string,
col2 int,
col3 boolean
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS
INPUTFORMAT
  'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://path/to/mycsvfile/'
TBLPROPERTIES (
  "s3select.format" = "csv",
  "s3select.headerInfo" = "ignore"
);
```

Example 基于 JSON 的 CREATE TABLE 语句

```
CREATE TABLE mys3selecttable (
col1 string,
col2 int,
col3 boolean
)
ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe'
STORED AS
INPUTFORMAT
  'com.amazonaws.emr.s3select.hive.S3SelectableTextInputFormat'
OUTPUTFORMAT
```

```
  'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://path/to/json/'
TBLPROPERTIES (
  "s3select.format" = "json"
);
```

Example SELECT TABLE 语句

```
SET s3select.filter=true;
SELECT * FROM mys3selecttable WHERE col2 > 10;
```

MSCK 优化

Hive 在其元数据存储中存储每个表的分区列表。但是，当直接向文件系统添加分区或从文件系统中移除分区时，Hive 元数据存储不会意识到这些变化。对于直接添加到文件系统或从文件系统中移除的分区，MSCK 命令 会更新 Hive 元数据存储中的分区元数据。此命令的语法是：

```
MSCK [REPAIR] TABLE table_name [ADD/DROP/SYNC PARTITIONS];
```

Hive 将按如下方式实现此命令：

1. Hive 从元数据存储中检索表的所有分区。然后根据文件系统中不存在的分区路径列表，创建一个要从元数据存储中移除的分区列表。
2. Hive 收集文件系统中存在的分区路径，将其与元数据存储中的分区列表进行比较，然后生成需要添加到元数据存储的分区列表。
3. Hive 使用 ADD、DROP 或 SYNC 模式更新元数据存储。

> **Note**
>
> 如果元数据存储中有大量分区，检查文件系统中是否存在分区的步骤需要很长时间才能完成运行，因为必须对每个分区执行文件系统的 `exists` API 调用。

在 Amazon EMR 6.5.0 中，Hive 引入了一个名为 `hive.emr.optimize.msck.fs.check` 的标记。启用此标记后，它会让 Hive 检查上面第 2 步中所生成的文件系统分区路径列表中是否存在分区，而不是调用文件系统 API。在 Amazon EMR 6.8.0 中，Hive 默认启用了此优化，无需设置标记 `hive.emr.optimize.msck.fs.check`。

使用 Hive Live Long and Process（LLAP）

Amazon EMR 版本 6.0.0 支持 Hive 的 Live Long and Process（LLAP）功能。与之前的默认 Tez 容器执行模式相比，LLAP 使用具有智能内存中的缓存的持久守护进程来提高查询性能。

Hive LLAP 守护进程作为 YARN 服务进行管理和运行。由于 YARN 服务可以被视为长时间运行的 YARN 应用程序，因此您的部分集群资源专用于 Hive LLAP，不能用于其它工作负载。有关更多信息，请参阅 LLAP 和 YARN Service API。

在 Amazon EMR 上启用 Hive LLAP

要在 Amazon EMR 上启用 Hive LLAP，请在启动集群时提供以下配置。

```
[
  {
    "Classification": "hive",
    "Properties": {
      "hive.llap.enabled": "true"
    }
  }
]
```

有关更多信息，请参阅配置应用程序。

默认情况下，Amazon EMR 将大约 60% 的集群 YARN 资源分配给 Hive LLAP 守护进程。您可以配置分配给 Hive LLAP 的集群 YARN 资源的百分比，以及进行 Hive LLAP 分配时要考虑的任务和核心节点数。

例如，以下配置在三个任务或核心节点上启动具有三个守护进程的 Hive LLAP，并将这三个核心或任务节点的 YARN 资源的 40% 分配给 Hive LLAP 守护进程。

```
[
  {
    "Classification": "hive",
    "Properties": {
      "hive.llap.enabled": "true",
      "hive.llap.percent-allocation": "0.4",
      "hive.llap.num-instances": "3"
    }
  }
```

```
]
```

您可以使用分类 API 中的以下 hive-site 配置来覆盖默认的 LLAP 资源设置。

属性	描述
hive.llap.daemon.yarn.container.mb	LLAP 守护进程容器总大小（以 MB 为单位）
hive.llap.daemon.memory.per.instance.mb	LLAP 守护进程容器中执行程序使用的总内存（以 MB 为单位）
hive.llap.io.memory.size	LLAP 输入/输出的缓存大小
hive.llap.daemon.num.executors	每个 LLAP 守护进程的执行程序数

在集群上手动启动 Hive LLAP

在集群启动过程中，LLAP 使用的所有依赖关系和配置都被打包到 LLAP tar 归档中。如果已使用 "hive.llap.enabled": "true" 启用 LLAP，我们建议您使用 Amazon EMR 重新配置对 LLAP 进行配置更改。

否则，对于 hive-site.xml 的任何手动更改，您必须使用 hive --service llap 命令来重建 LLAP tar 归档，如以下示例所示。

```
# Define how many resources you want to allocate to Hive LLAP

LLAP_INSTANCES=<how many llap daemons to run on cluster>
LLAP_SIZE=<total container size per llap daemon>
LLAP_EXECUTORS=<number of executors per daemon>
LLAP_XMX=<Memory used by executors>
LLAP_CACHE=<Max cache size for IO allocator>

yarn app -enableFastLaunch

hive --service llap \
--instances $LLAP_INSTANCES \
--size ${LLAP_SIZE}m \
--executors $LLAP_EXECUTORS \
```

```
--xmx ${LLAP_XMX}m \
--cache ${LLAP_CACHE}m \
--name llap0 \
--auxhbase=false \
--startImmediately
```

检查 Hive LLAP 的状态

通过 Hive 使用以下命令检查 Hive LLAP 的状态。

```
hive --service llapstatus
```

通过 YARN 使用以下命令检查 Hive LLAP 的状态。

```
yarn app -status (name-of-llap-service)

# example:
yarn app -status llap0 | jq
```

启动或停止 Hive LLAP

由于 Hive LLAP 作为持久 YARN 服务运行，因此您可以通过停止或重新启动 YARN 服务来停止或重新启动 Hive LLAP。以下命令对此进行了演示。

```
yarn app -stop llap0
yarn app -start llap0
```

调整 Hive LLAP 进程守护程序的数量

使用以下命令减少 LLAP 实例的数量。

```
yarn app -flex llap0 -component llap -1
```

有关更多信息，请参阅 Flex a component of a service。

Hive 中的加密

这一部分说明了 Amazon EMR 支持的加密类型。

Hive 中的 Parquet 模块化加密

Parquet 模块化加密提供列级访问控制和加密功能，以增强以 Parquet 文件格式存储的数据的隐私和数据完整性。Amazon EMR Hive 从版本 6.6.0 起提供此功能。

有关以前支持的安全性和完整性解决方案，包括文件加密或存储层加密的信息，详见《Amazon EMR 管理指南》中的 加密选项。这些解决方案可以用于 Parquet 文件，但是利用集成 Parquet 加密机制的新功能可以在列级别实现精细访问控制，并提高性能和安全性。有关此功能的更多信息，请访问 Apache github 页面 Parquet 模块化加密。

用户使用 Hadoop 配置将配置传递给 Parquet 读取器和写入器。有关用户配置读取器和写入器以启用加密以及切换高级功能的详细配置，详见 PARQUET-1854：Parquet 加密管理的属性驱动接口

用法示例

以下示例涉及使用 Amazon KMS 创建加密密钥并写入 Hive 表以进行加密密钥管理。

1. 按照文档 PARQUET-1373：加密密钥管理工具 中的描述，为 Amazon KMS 服务实施一个 KmsClient。以下示例演示了一个实施片段。

```
package org.apache.parquet.crypto.keytools;

import com.amazonaws.AmazonClientException;
import com.amazonaws.AmazonServiceException;
import com.amazonaws.regions.Regions;
import com.amazonaws.services.kms.AWSKMS;
import com.amazonaws.services.kms.AWSKMSClientBuilder;
import com.amazonaws.services.kms.model.DecryptRequest;
import com.amazonaws.services.kms.model.EncryptRequest;
import com.amazonaws.util.Base64;
import org.apache.hadoop.conf.Configuration;
import org.apache.parquet.crypto.KeyAccessDeniedException;
import org.apache.parquet.crypto.ParquetCryptoRuntimeException;
import org.apache.parquet.crypto.keytools.KmsClient;
import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

import java.nio.ByteBuffer;
import java.nio.charset.Charset;
import java.nio.charset.StandardCharsets;

public class AwsKmsClient implements KmsClient {
```

```
   private static final AWSKMS AWSKMS_CLIENT = AWSKMSClientBuilder
           .standard()
           .withRegion(Regions.US_WEST_2)
           .build();
   public static final Logger LOG = LoggerFactory.getLogger(AwsKmsClient.class);

   private String kmsToken;
   private Configuration hadoopConfiguration;

   @Override
   public void initialize(Configuration configuration, String kmsInstanceID, String
kmsInstanceURL, String accessToken) throws KeyAccessDeniedException {
       hadoopConfiguration = configuration;
       kmsToken = accessToken;

   }

   @Override
   public String wrapKey(byte[] keyBytes, String masterKeyIdentifier) throws
KeyAccessDeniedException {
       String value = null;
       try {
           ByteBuffer plaintext = ByteBuffer.wrap(keyBytes);

           EncryptRequest req = new
EncryptRequest().withKeyId(masterKeyIdentifier).withPlaintext(plaintext);
           ByteBuffer ciphertext = AWSKMS_CLIENT.encrypt(req).getCiphertextBlob();

           byte[] base64EncodedValue = Base64.encode(ciphertext.array());
           value = new String(base64EncodedValue, Charset.forName("UTF-8"));
       } catch (AmazonClientException ae) {
           throw new KeyAccessDeniedException(ae.getMessage());
       }
       return value;
   }

   @Override
   public byte[] unwrapKey(String wrappedKey, String masterKeyIdentifier) throws
KeyAccessDeniedException {
       byte[] arr = null;
       try {
           ByteBuffer ciphertext  =
ByteBuffer.wrap(Base64.decode(wrappedKey.getBytes(StandardCharsets.UTF_8)));
```

```
            DecryptRequest request = new
 DecryptRequest().withKeyId(masterKeyIdentifier).withCiphertextBlob(ciphertext);
            ByteBuffer decipheredtext =
 AWSKMS_CLIENT.decrypt(request).getPlaintext();
            arr = new byte[decipheredtext.remaining()];
            decipheredtext.get(arr);
        } catch (AmazonClientException ae) {
            throw new KeyAccessDeniedException(ae.getMessage());
        }
        return arr;
    }
}
```

2. 按照《Amazon Key Management Service 开发者指南》中 创建密钥 部分的描述，使用具有访问权限的 IAM 角色为页脚以及列创建 Amazon KMS 加密密钥。默认 IAM 角色为 EMR_ECS_default。
3. 按照 Apache Hive 资源文档 所述，在 Amazon EMR 集群上的 Hive 应用程序中，使用上述 ADD JAR 语句添加客户端。下面是一个示例语句：

   ```
   ADD JAR 's3://location-to-custom-jar';
   ```

   另一种方法是使用引导操作将 JAR 添加到 Hive 的 auxlib 中。以下是要添加到引导操作的示例行：

   ```
   aws s3 cp 's3://location-to-custom-jar' /usr/lib/hive/auxlib
   ```

4. 设置下列配置值：

   ```
   set
    parquet.crypto.factory.class=org.apache.parquet.crypto.keytools.PropertiesDrivenCryptoFacto
   set
    parquet.encryption.kms.client.class=org.apache.parquet.crypto.keytools.AwsKmsClient;
   ```

5. 创建一个 Parquet 格式的 Hive 表，然后在 SERDEPROPERTIES 中指定 Amazon KMS 密钥，然后在其中插入一些数据：

   ```
   CREATE TABLE my_table(name STRING, credit_card STRING)
   ROW FORMAT SERDE 'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe'
   WITH SERDEPROPERTIES (
     'parquet.encryption.column.key'=<aws-kms-key-id-for-column-1>: credit_card',
     'parquet.encryption.footer.key'='<aws-kms-key-id-for-footer>')
   ```

```
STORED AS parquet
LOCATION “s3://<bucket/<warehouse-location>/my_table”;

INSERT INTO my_table SELECT
java_method ('org.apache.commons.lang.RandomStringUtils','randomAlphabetic',5) as
 name,
java_method ('org.apache.commons.lang.RandomStringUtils','randomAlphabetic',10) as
 credit_card
from (select 1) x lateral view posexplode(split(space(100),' ')) pe as i,x;

select * from my_table;
```

6. 验证当您在同一位置创建无 Amazon KMS 密钥访问权限的外部表时（例如，IAM 角色访问被拒绝时），您将无法读取数据。

```
CREATE EXTERNAL TABLE ext_table (name STRING, credit_card STRING)
ROW FORMAT SERDE 'org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe’
STORED AS parquet
LOCATION “s3://<bucket>/<warehouse-location>/my_table”;

SELECT * FROM ext_table;
```

7. 最后一条语句应触发以下异常：

```
Failed with exception
 java.io.IOException:org.apache.parquet.crypto.KeyAccessDeniedException: Footer key:
 access denied
```

HiveServer2 中的传输中加密

从 Amazon EMR 发行版 6.9.0 开始，HiveServer2（HS2）作为 HiveServer2 中的传输中加密安全配置的一部分，已启用 TLS/SSL。这会影响与在启用传输中加密的 Amazon EMR 集群上运行的 HS2 建立连接的方式。要连接到 HS2，须修改 JDBC URL 中的 TRUSTSTORE_PATH 和 TRUSTSTORE_PASSWORD 参数值。以下 URL 是带有所需参数的 HS2 的 JDBC 连接示例：

```
jdbc:hive2://HOST_NAME:10000/
default;ssl=true;sslTrustStore=TRUSTSTORE_PATH;trustStorePassword=TRUSTSTORE_PASSWORD
```

使用以下相应的集群上或集群外 HiveServer2 加密说明。

On-cluster HS2 access

如果您在 SSH 连接到主节点后使用 Beeline 客户端访问 HiveServer2，请参考 /etc/hadoop/conf/ssl-server.xml 以使用配置 ssl.server.truststore.location 和 ssl.server.truststore.password 来查找 TRUSTSTORE_PATH 和 TRUSTSTORE_PASSWORD 参数值。

以下示例命令可帮助您检索这些配置：

```
TRUSTSTORE_PATH=$(sed -n '/ssl.server.truststore.location/,+2p' /etc/hadoop/conf/
ssl-server.xml | awk -F "[><]" '/value/{print $3}')
TRUSTSTORE_PASSWORD=$(sed -n '/ssl.server.truststore.password/,+2p' /etc/hadoop/
conf/ssl-server.xml | awk -F "[><]" '/value/{print $3}')
```

Off-cluster HS2 access

如果您从 Amazon EMR 集群外部的客户端访问 HiveServer2，可使用以下方法之一来获取 TRUSTSTORE_PATH 和 TRUSTSTORE_PASSWORD：

- 将安全配置期间创建的 PEM 文件转换为 JKS 文件，并在 JDBC 连接 URL 中使用该文件。例如，对于 openssl 和 keytool，请使用以下命令：

  ```
  openssl pkcs12 -export -in trustedCertificates.pem -inkey privateKey.pem -out
   trustedCertificates.p12 -name "certificate"
  keytool -importkeystore -srckeystore trustedCertificates.p12 -srcstoretype pkcs12
   -destkeystore trustedCertificates.jks
  ```

- 或者，参考 /etc/hadoop/conf/ssl-server.xml 以使用配置 ssl.server.truststore.location 和 ssl.server.truststore.password 来查找 TRUSTSTORE_PATH 和 TRUSTSTORE_PASSWORD 参数值。将信任存储库文件下载到客户端计算机，并使用客户端计算机上的路径作为 TRUSTSTORE_PATH。

  有关从 Amazon EMR 集群外部的客户端访问应用程序的更多信息，请参阅使用 Hive JDBC 驱动程序。

Hive 发行历史记录

下表列出了 Amazon EMR 每个发行版本中包含的 Hive 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Hive 版本信息

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.14.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server
emr-6.13.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.12.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server
emr-6.11.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.11.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server
emr-6.10.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.10.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, tez-on-worker, zookeeper-client, zookeeper-server
emr-6.9.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.9.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.8.1	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.8.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.7.0	3.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.36.1	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.36.0	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.6.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-5.35.0	2.3.9	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.5.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.4.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.3.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.3.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.2.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.2.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.1.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.1.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-6.0.1	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server
emr-6.0.0	3.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.34.0	2.3.8	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.33.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.33.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.32.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.32.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.31.1	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.31.0	2.3.7	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.30.2	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.30.1	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn
emr-5.30.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mariadb-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.29.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mysql-server, tez-on-yarn
emr-5.28.1	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.28.0	2.3.6	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, hudi, mysql-server, tez-on-yarn
emr-5.27.1	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.27.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.26.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.25.0	2.3.5	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.24.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.24.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.23.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.23.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.22.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.21.2	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.21.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.21.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.20.1	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.20.0	2.3.4	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.19.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.19.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.18.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.18.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.17.2	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.17.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.17.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, emr-s3-select, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.16.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.16.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.15.1	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.15.0	2.3.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.14.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.14.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.14.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.13.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.13.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.12.3	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.12.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.12.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.12.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.11.4	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.11.3	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.11.2	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.11.1	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.11.0	2.3.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.10.1	2.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.10.0	2.3.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.9.1	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.9.0	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.8.3	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.8.2	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.8.1	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.8.0	2.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.7.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.7.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.6.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.6.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.5.4	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.5.3	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.5.2	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.5.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.5.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.4.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.4.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hive-hbase, hcatalog-server, hive-server2, mysql-server, tez-on-yarn
emr-5.3.2	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.3.1	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn
emr-5.3.0	2.1.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.2.3	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn
emr-5.2.2	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.2.1	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn
emr-5.2.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.1.1	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn
emr-5.1.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-5.0.3	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn
emr-5.0.0	2.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, hive-server, mysql-server, tez-on-yarn

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.9.6	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.9.5	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.9.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.9.3	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.9.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.9.1	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.8.5	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.8.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.8.3	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.8.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.8.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.7.4	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.7.2	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.7.1	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.7.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, hive-server, mysql-server
emr-4.6.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.5.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server
emr-4.4.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server
emr-4.3.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server

Amazon EMR 发行版标签	Hive 版本	随 Hive 安装的组件
emr-4.2.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server
emr-4.1.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server
emr-4.0.0	1.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hive-metastore-server, hive-server, mysql-server

Hive 发布说明（按版本分类）

主题

Amazon EMR 6.14.0 – Hive 发布说明

Amazon EMR 6.14.0 – Hive 更改

类型	描述
改进	HIVE-26762：移除 `HiveFilterSetOpTransposeRule` 中的操作数修剪
错误修复	HIVE-27582：不要在 FetchOperator 中缓存 HBase 表输入格式
错误修复	HIVE-26452：将 JOIN 转换到被多次引用的 MAPJOIN 和 JOIN 列时 NPE
错误修复	HIVE-26416：`AcidUtils.isRawFormatFile()` 对 non-ORC 文件发出 `InvalidProtocolBufferException` 异常
错误修复	HIVE-26105：如果注释列包含特定的中文字符，则显示列会显示额外的值

类型	描述
错误修复	HIVE-25864：Hive 查询优化功能创建错误的计划，以使用窗口函数进行谓词下推
错误修复	HIVE-25224：涉及具有不同 bucketing_versions 的表的多个 INSERT 语句导致错误
错误修复	HIVE-24151：如果字符串包含 non-ASCII 字符，则 MultiDelimitSerDe 会移动数据
错误修复	HIVE-23606：（LLAP）EncodedReaderImpl 的 DirectByteBuffer 清理延迟
错误修复	HIVE-22165：HIVE-14296 在 SessionManager.closeSession 上引入的同步会导致繁忙的 Hive 服务器出现高延迟
错误修复	HIVE-21304：提高存储桶版本使用的稳健性

Amazon EMR 6.13.0 – Hive 发布说明

Amazon EMR 6.13.0 – Hive 更改

类型	描述
改进	升级 Python 脚本以支持 Python3
改进	HIVE-27097：改进 MetaStore 客户端和服务器的重试策略
错误修复	HIVE-21778：CBO：“结构不为空”被评估为可为空，总是会导致查询中筛选条件缺失
错误修复	HIVE-21009：为用户添加了设置绑定用户的功能

类型	描述
错误修复	HIVE-22661：在路径中加载了数据的非存储桶表上的压缩失败
错误修复	HIVE-19718：批量添加分区还会提取每个分区的表
错误修复	HIVE-22173：在编译过程中使用多个横向视图的查询挂起
错误修复	HIVE-27088：合并带有后联接筛选条件的内部和外部联接时结果不正确
错误修复	HIVE-21935：Hive 向量化：使用向量化 UDF 导致性能降低
错误修复	HIVE-25299：对于非 UTC 时区，将时间戳转换为数字数据类型是不正确的
错误修复	HIVE-24626：LLAP：如果所有 IO 电梯线程都忙于排队到其他队列已满的读取器，则读取器线程可能会面临短缺
错误修复	HIVE-27029：Hive 查询失败并显示文件系统关闭错误，HIVE-26352 已完成返工
错误修复	HIVE-26352：Tez 队列访问检查失败，并且 Compaction 上存在 GSS 异常
错误修复	HIVE-24590：操作日志仍然会泄漏 log4j 附加程序
错误修复	HIVE-24552：可能的 HMS 连接泄漏或 loadDynamicPartitions 中的累积
错误修复	HIVE-27069：加入存储桶地图时结果不正确
错误修复	HIVE-27344：在 RecordReaderImpl#close 中添加空值检查

类型	描述
错误修复	HIVE-27439：支持以十进制表示的空格
错误修复	HIVE-27267：使用子查询对十进制分桶列进行存储桶映射联接时结果不正确
错误修复	HIVE-21986：HiveServer Web 用户界面：在默认响应标头中设置 Strict-Transport-Security
错误修复	HIVE-22148：S3A 委托令牌未添加到 Compactor 的任务配置中。
错误修复	HIVE-22622：Hive 允许创建具有重复属性名称的结构
错误修复	HIVE-22008：LIKE 运算符应匹配多行输入
错误修复	HIVE-23144：LLAP：让 QueryTracker 在 ServiceStop 上进行清理
错误修复	HIVE-22391：检查 Hive 查询结果缓存时出现 NPE
错误修复	HIVE-23305：由于争用情况，LlapTaskSchedulerService addNode 中出现 NullPointerException
错误修复	HIVE-22178：Parquet FilterPredicate 在 SchemaEvolution 之后引发了 CastException
错误修复	HIVE-21517：修复 AggregateStatsCache
错误修复	HIVE-21825：改进启用主动/被动 HA 时的客户端错误消息
错误修复	HIVE-23389：FilterMergeRule 可能会导致 AssertionError

类型	描述
错误修复	HIVE-22767：Beeline 无法正确解析注释中的分号
错误修复	HIVE-22996：BasicStats 解析应主动检查 null 或空字符串
错误修复	HIVE-22808：HiveRelFieldTrimmer 无法处理 HiveTableFunctionScan
错误修复	HIVE-22437：LLAP 元数据在锁定元数据时缓存 NPE。
错误修复	HIVE-22606：AvroSerde 在 INFO 级别下记录 avro.schema.literal
错误修复	HIVE-22713：不应对 Join-Fil(*)-RS 结构进行持续传播
错误修复	HIVE-21624：LLAP：线程级别的 Cpu 指标已损坏
错误修复	HIVE-22815：减少在 MROutput 中创建不必要的文件系统对象的次数
错误修复	HIVE-23060：查询失败并显示错误“分组集表达式不在 GROUP BY 键中。在令牌附近遇到错误”
错误修复	HIVE-22236：无法通过选择包含 `NOT IN` 子查询的视图来创建视图
错误修复	HIVE-19886：如果使用 —hiveconf hive.log.file，则日志可能会被定向到 2 个文件
错误修复	HIVE-20620：使用动态分区插入存储桶排序的 MM 表时出现清单冲突
错误修复	HIVE-14557：同时启用 SkewJoin 和 MapJoin 时为空指针

类型	描述
错误修复	HIVE-20471：获取默认数据库路径时出现问题
错误修复	HIVE-20598：修复 HiveAlgorithmsUtil 计算中的错别字
错误修复	HIVE-14737：在 Kerberized Hive Server 2 Web 用户界面中访问 /logs 时出现问题
错误修复	HIVE-20733：GenericUDFOPEqualNS 不得在计划描述中使用 =
错误修复	HIVE-20848：设置 UpdateInputAccessTimeHook 后，查询失败并显示未找到表。
错误修复	HIVE-18929：HiveStringUtils.java 中的 HumanReadableInt 方法存在争用条件。
错误修复	HIVE-20841：LLAP：将动态端口设置为可配置
错误修复	HIVE-20930：筛选模式下的 VectorCoalesce 不生效
错误修复	HIVE-21007：半联接 + Union 可能会导致计划错误
错误修复	HIVE-21074：Hive 分桶表查询修剪不适用于 IS NOT NULL 条件
错误修复	HIVE-21223：当分区不存在时 CachedStore 返回空分区
错误修复	HIVE-19625：Hive#copyFiles 中潜在的 NPE 和隐藏的实际异常
错误修复	HIVE-17020：激进的 RS 重复数据删除可能会错误地删除 OP 树分支

类型	描述
错误修复	HIVE-20168：隐藏了 ReduceSinkOperator 日志记录
错误修复	HIVE-20879：在投影表达式中使用空值会导致 CastException
错误修复	HIVE-20888：TxnHandler：在不可变列表上调用 sort()
错误修复	HIVE-19948：如果字符串内有引号，HiveCLI 无法正确地用分号分隔命令
错误修复	HIVE-20621：在 resultset.next 中调用 GetOperationStatus 会导致增量缓慢
错误修复	HIVE-20854：合理的默认值：Hive 的 Zookeeper 检测信号间隔为 20 分钟，改为 2 分钟
错误修复	HIVE-20330：HCatLoader 无法为具有多个输入的作业处理多个 InputJobInfo 对象
错误修复	HIVE-20787：MapJoinBytesTableContainer dummyRow 案例无法处理重用
错误修复	HIVE-20331：使用 union all、横向视图和 Join 进行查询失败，并显示“无法在子运算符中找到父项”
错误修复	HIVE-19968：未引发 UDF 异常
错误修复	HIVE-20410：中止在事务表上插入覆盖会导致“没有足够的历史记录可用于……”错误
错误修复	HIVE-20059：Hive 流式传输应在例外情况下无条件尝试阴影前缀
错误修复	HIVE-19424：MetaDataFormatters 中的 NPE

类型	描述
错误修复	HIVE-20355：HiveConnection.setSchema 的清理参数
错误修复	HIVE-20858：未使用 Utilities.createEmptyBuckets 中的配置正确初始化序列化器
错误修复	HIVE-20424：schematool 不得污染 beeline 历史记录
错误修复	HIVE-20338：LLAP：对具有 HDFS 协议隐含和 POSIX 突变语义的文件系统强制使用合成文件 ID
错误修复	HIVE-11708：逻辑运算符使用空值引起 ClassCastExceptions
错误修复	HIVE-21082：在 HPL/SQL 中，声明语句不支持字符类型的变量
错误修复	HIVE-16690：根据 LLAP 集群大小配置 Tez 笛卡尔乘积边缘
错误修复	HIVE-21296：删除 varchar 分区引发异常
错误修复	HIVE-14516：OrcInputFormat.SplitGenerator.callInternal
错误修复	HIVE-20981：流式传输/AbstractRecordWriter 泄漏了 HeapMemoryMonitor
错误修复	HIVE-20043：HiveServer2：SessionState 在 AtomicBoolean 周围有一个静态同步块
错误修复	HIVE-20191：如果补丁为空，则 PreCommit 补丁应用程序不会失败
错误修复	HIVE-20400：创建表应始终使用完全限定路径，以避免潜在的 FS 歧义

类型	描述
错误修复	在访问偏斜列之前，添加对 skewedInfo 的空值检查

Amazon EMR 6.12.0 – Hive 发布说明

Amazon EMR 6.12.0 – Hive 更改

类型	描述
改进	添加了对 JDK 11 和 JDK 17 运行时系统的支持
改进	添加了使用 S3 Select 时对查询区分大小写和保留关键字列名的支持。要使用它，请以“s3select.column.mapping”= "*column1*:*fieldName1* , *column2*:*fieldName2* ,..." 的格式定义表格属性
改进	HIVE-23133：不同硬件架构的数值运算可能有不同的结果
改进	HIVE-27145：对剩余的数学函数使用 StrictMath 作为 HIVE-23133 的后续函数
错误修复	修复在 EMR Hive 6.4.0 中移植 HIVE-22900 而导致的 get_partitions_by_filter 和 get_num_partitions_by_filter HMS API 中的通配符不兼容问题
错误修复	HIVE-26736：带有 WITH 子句的嵌套视图授权失败
错误修复	HIVE-22416：启用并行执行后，与 MR 相关的操作日志丢失
错误修复	HIVE-19653：带有分组集的 groupby 的谓词下推不正确

类型	描述
错误修复	HIVE-22094：使用 ClassCastException 查询失败：hive.ql.exec.vector.DecimalColumnVector 无法转换到 hive.ql.exec.vector.Decimal64ColumnVector
错误修复	HIVE-26340：如果查询具有大写窗口功能，则向量化 PTF 运算符会失败
错误修复	HIVE-26184：当某些键高度偏斜时，使用 GROUP BY 的 COLLECT_SET 速度非常慢
错误修复	HIVE-26373：从包含 Avro 数据的 HBase 表中读取时间戳时出现 ClassCastException
错误修复	HIVE-26388：当 CTAS 查询的源表中存在非字符串类型的列时出现 ClassCastException 升级 HIVE-26172：Hive – 由于 CVE-2021-36373 和 CVE-2021-36374，将 Ant 升级到 1.10.11
错误修复	HIVE-26114：使用带前缀空格的 dfs 命令修复 jdbc 连接 hiveserver2 会导致异常
错误修复	HIVE-26396：trunc 函数在精度截取方面存在问题，且结果中有很多 0
错误修复	HIVE-26446：HiveProtoLoggingHook 无法填充分区表的 TablesWritten 字段。
错误修复	HIVE-26639：ConstantVectorExpression 和 ExplainTask 不应依赖默认的字符集
错误修复	HIVE-22670：使用向量化读取器读取 parquet 文件时出现 ArrayIndexOutOfBoundsException
错误修复	HIVE-23607：权限问题：在另一个视图上创建视图成功，但更改视图失败

类型	描述
错误修复	HIVE-25498：包含超过 31 个计数的不同函数的查询返回错误结果
错误修复	HIVE-25780：DistinctExpansion 创建了 64 个以上的分组集 II
错误修复	HIVE-23868：窗口化函数规范：支持 0 在之前/之后
错误修复	HIVE-24539：OrcInputFormat 架构生成应遵循列分隔符
错误修复	HIVE-23476：LLAP：也要为 mmap case 预先分配竞技场
错误修复	HIVE-25806：LlapCacheAwareFs 中的可能泄漏 – Parquet、LLAP IO
错误修复	HIVE-23498：在 ThriftHttpCliService 上禁用 HTTP Trace 方法
错误修复	HIVE-25729：完全启动后应通知 ThriftUni onObjectInspector
错误修复	HIVE-23846：避免对位向量进行不必要的序列化和反序列化
错误修复	HIVE-24233：except 子查询在禁用 cbo 的情况下引发空指针
错误修复	HIVE-24276：HiveServer2 loggerconf jsp 跨站脚本攻击（XSS）漏洞
错误修复	HIVE-25721：外部联接结果错误
错误修复	HIVE-25223：带限制的选择不会在非本地表上返回任何行

类型	描述
错误修复	HIVE-25794：CombineHiveRecordReader：循环中的日志语句会导致内存压力
错误修复	HIVE-23602：使用 Java Concurrent Package 作为操作句柄集
错误修复	HIVE-24045：没有与创建默认数据库的时间相关的日志
错误修复	HIVE-24305：如果将值括在引号中，avro 十进制架构无法正确填充比例/精度
错误修复	HIVE-25844：异常反序列化错误可能会导致 beeline 立即终止
错误修复	HIVE-25040：删除数据库级联无法删除永久函数
错误修复	HIVE-23501：将复杂类型转换为基本类型时 VectorDeserializeRow 中出现 AOOB
错误修复	HIVE-23704：Thrift HTTP 服务器无法正确处理身份验证处理
错误修复	HIVE-23529：使用 row_deserialize 时 uniontype 的 CTAS 被破坏
错误修复	HIVE-24144：HiveDatabaseMetaData 中的 getIdentifierQuoteString 返回不正确的值
错误修复	HIVE-23850：当主题不是包含分组集的列时，允许 PPD
错误修复	HIVE-24036：序列化用于 getSplits UDF 调用的计划时出现 Kryo 异常
错误修复	HIVE-25919：在 HBaseStorageHandler 中推送布尔列谓词时出现 ClassCastException

类型	描述
错误修复	HIVE-25261：RetryingHMSHandler 应该用对目标的简短描述来封装 MetaException
错误修复	HIVE-24792：操作中可能出现线程泄漏
错误修复	HIVE-23409：如果 TezSession 应用程序因时间线服务关闭而重新打开失败，则 SessionPool 中的默认 TezSession 将在重试后关闭
错误修复	HIVE-23615：不要在 Beeline Commands 类中使用空指针
错误修复	HIVE-24849：当位置有大量文件时创建外部表套接字超时（影响 3.1.2）
错误修复	HIVE-24193：在重命名的 hive acid 表上选择查询不会产生任何输出
错误修复	HIVE-25209：使用 SUM 函数的 SELECT 查询生成意外结果
错误修复	HIVE-23666：当 groupby 运算符没有设置分组时，会跳过 checkHashModeEfficiency
错误修复	HIVE-23873：当 CBO 关闭时，使用 NPE 查询 Hive JDBCStorageHandler 表失败
错误修复	HIVE-24149：HiveStreamingConnection 不关闭 HMS 连接
错误修复	HIVE-25561：被终止的任务不应提交文件。（影响 2.x 和 3.x 版本）
错误修复	HIVE-25683：在 AcidUtils.isRawFormatFile 中关闭读取器
错误修复	HIVE-24294：TezSessionPool 会话可能会引发 AssertionError 错误

类型	描述
错误修复	HIVE-24182：永久 UDF 存在 Ranger 授权问题
错误修复	HIVE-22805：使用条件数组或映射的向量化未实现并引发错误
错误修复	HIVE-22828：Decimal64：NVL 和 CASE 语句隐式地将 decimal64 转换为 128
错误修复	HIVE-21398：不应将包含估计统计数据的列视为唯一键
错误修复	HIVE-22490：添加路径中包含特殊字符的 jar 引发错误
错误修复	HIVE-22700：未经授权的压缩可能会泄漏内存
错误修复	HIVE-22053：创建函数时函数名称未标准化
错误修复	HIVE-22595：在带有外部架构的 Avro 表上进行动态分区插入失败
错误修复	HIVE-21795：在分区表上正在进行 mapjoin 时，Rollup 摘要行可能会丢失
错误修复	HIVE-22987：当 DataTypePhysicalVariation 为空时 VectorCoalesce 中出现 ClassCast Exception
错误修复	HIVE-22814：向量化中的 getDataTypePhysicalVariation 中的 ArrayIndexOutOfBound
错误修复	HIVE-22523：如果 LlapRecordReader 中的错误处理程序队列已满，则该错误处理程序可能会被阻塞
错误修复	HIVE-21796：ArrayWritableObjectInspector.equals 可能需要 O(2^nesting_depth) 时间

类型	描述
错误修复	HIVE-22929：性能：带引号的标识符解析通过 String.replaceAll() 使用一次性正则表达式
错误修复	HIVE-21641：与 beeline 相比，Llap 外部客户端返回精度/小数位数不同的十进制列
错误修复	HIVE-22207：Tez：当集群上的“dfs.block.size”为“128m”时，SplitGenerator 会引发 NumberFormatException
错误修复	HIVE-22114：当所有存储桶都为空时，仅限分区插入表的插入查询失败
错误修复	HIVE-22165：HIVE-14296 在 SessionManager.closeSession 上引入的同步会导致繁忙的 Hive 服务器出现高延迟
错误修复	HIVE-22744：具有多个出边的顶点的 TezTask 应具有成比例的排序内存
错误修复	HIVE-22072：更改表格以进行列更改不会更新约束引用
错误修复	HIVE-22075：修复 HIVE-14200 中的 max-reducers=1 回归
错误修复	HIVE-22527：Hive on Tez：合并小文件的作业将提交到另一个队列（默认队列）
错误修复	HIVE-22816：QueryCache：在 CTE 扩展后，使用视图的查询可以对其进行缓存
错误修复	HIVE-22733：在 Hive 中禁用操作日志属性后，HS2 仍在保存操作日志
错误修复	HIVE-22699：遮罩 UDF 应遮蔽数值 0

类型	描述
错误修复	HIVE-23356：使用分组集表达式处理查询时，哈希聚合始终处于禁用状态。
错误修复	HIVE-21568：HiveRelOptUtil.isRowFilteringPlan 应该跳过 Project
错误修复	HIVE-21760：对于 SMB 联接，应绕过 Sharedwork 优化
错误修复	HIVE-22712：无论用户定义的队列如何，ReExec 驱动程序都会在默认队列中执行提交查询
错误修复	HIVE-21397：适用于 Hive 托管的 BloomFilter [ACID] 表未按预期工作
错误修复	HIVE-23011：在比较联接时，共享工作优化程序应检查剩余谓词
错误修复	HIVE-21412：PostExecOrcFileDump 不适用于 ACID 表
错误修复	HIVE-22201：如果没有选择大表，ConvertJoinMapJoin#checkShuffleSizeForLargeTable 会引发 ArrayIndexOutOfBoundsException
错误修复	HIVE-21971：由于“ReflectionUtils::CONSTRUCTOR_CACHE”与临时函数 + GenericUDF，HS2 泄露了类加载程序
错误修复	HIVE-21368：向量化：不必要的 Decimal64 ->HiveDecimal 转换
错误修复	HIVE-25416：Hive 元存储内存泄漏是因为 datanucleus-api-jdo 错误

类型	描述
错误修复	HIVE-22219：关闭节点管理器会阻止 LLAP 服务的重启
错误修复	HIVE-21793：即使 hive.stats.fetch.column.stats 设置为 false，CBO 也会检索列统计信息
错误修复	HIVE-22163：CBO：启用 CBO 会开启统计数据估计，即使禁用估算功能也是如此
错误修复	HIVE-18735：像丢失交易属性一样创建表
错误修复	HIVE-22433：Hive JDBC 存储处理器：从 BOOLEAN 和从 JDBC 数据来源获取的 TIMESTAMP DataType 结果不正确
错误修复	HIVE-19430：大量待处理事件上存在 ObjectStore.cleanNotificationEvents OutOfMemory
错误修复	HIVE-20785：JDBC DatabaseMetaData.getPrimaryKeys 方法中存在密钥名称错误
错误修复	HIVE-16116：当 beeline.properties 中存在 beeline.hiveconfvariables={} 时 beeline 会引发 NPE
错误修复	HIVE-20066：将 hive.load.data.owner 与完整主体进行比较
错误修复	HIVE-20489：解释查询挂起的计划
错误修复	HIVE-21033：忘记关闭操作会切断更多的 HiveServer2 输出
错误修复	HIVE-19888：来自 SessionState 的误导性“MMETASTORE_FILTER_HOOK will be ignored”警告

类型	描述
错误修复	HIVE-20303：INSERT OVERWRITE TABLE db.table PARTITION (...) IF NOT EXISTS 引发 InvalidTableException
错误修复	HIVE-16144：CompactionInfo 没有 equals/HashCode 但在 Set 中使用
错误修复	HIVE-20818：使用 WHERE 子查询创建的视图会将子查询中引用的视图视为直接输入
错误修复	HIVE-21005：LLAP：每次拆分读取更多条带会泄露 ZlibCodecs
错误修复	HIVE-20771：LazyBinarySerDe 在空结构上失败。
错误修复	HIVE-18852：变更表验证中出现误导性错误消息
错误修复	HIVE-21124：HPL/SQL 不支持 CREATE TABLE LIKE 语句
错误修复	HIVE-20935：在 EC2 中上传 llap 包压缩包 tarball 失败导致 LLAP 服务启动失败
错误修复	HIVE-20409：Hive ACID：更新/删除/合并无法清理 hdfs 暂存目录
错误修复	HIVE-20570：包含 hive.optimize.union.remove=true 的 Union ALL 计划不正确
错误修复	HIVE-20421：hive-default.xml.template 中存在非法字符实体“\b”
错误修复	HIVE-19133：HS2 WebUI 分阶段性能指标显示不正确

类型	描述
错误修复	HIVE-18977：使用 JDO 和直接 SQL 列出分区会返回不同的结果
错误修复	HIVE-20034：回滚 MetaStore 异常处理更改以实现向后兼容
错误修复	HIVE-20672：LlapTaskSchedulerService 中的日志线程应每隔固定时间间隔报告一次
错误修复	HIVE-12812：默认启用 mapred.input.dir.recursive 以支持与聚合函数的合并
错误修复	HIVE-20147：Hive 流式传输提取满足于同步日志记录
错误修复	HIVE-19203：HiveMetaStore 中的线程安全问题
错误修复	HIVE-20091：Tez：为 FileSinkOperator 输出添加安全凭证
错误修复	HIVE-16906：在连接 ATS 之前，Hive ATSHook 应检查 yarn.timeline-service.enabled
错误修复	HIVE-20714：显示单个属性的 tblproperties 将返回名称列中的值
错误修复	HIVE-24730：Shims 类以静默方式覆盖来自 hive-site.xml 和 tez-site.xml 的值
错误修复	HIVE-22055：从文本文件加载数据后，选择计数给出的结果不正确

Amazon EMR 6.11.0 – Hive 发布说明

Amazon EMR 6.11.0 – Hive 更改

类型	描述
改进	增加了对多线程删除分区的支持，以提高删除分区的性能
改进	支持读取编码的 Hive 查询文件
改进	默认情况下为 Hive on Tez 作业启用 Tez Shuffle Handler
错误	添加了一个选项，以允许在启用 hive.groupby.skewindata 时对 Reducer 启用确定性密钥分配，以修复错误结果（在 HIVE-20220 中报告）
错误	修复了配置默认分区名称时统计数据计算失败的问题
错误	遵守在启用了传输中加密的集群中以开箱即用的方式为 HiveServer2 配置了 SSL 时传递的任何自定义 SSL 分类参数
逆向移植	HIVE-23617：修复了存储 api FindBug 问题
逆向移植	HIVE-26408：向量化：修复暂存列的取消分配，不要重复使用子级 ConstantVectorExpression 作为输出
逆向移植	HIVE-23614：始终将 HiveConfig 传递给 removeTempOrDuplicateFiles
逆向移植	HIVE-23354：从 compareTempOrDuplicateFiles 中删除文件大小完整性检查
逆向移植	HIVE-20344：为引发 AccessControlException 的 SBA 修复了 PrivileGeSynchronizer。还引入

类型	描述
	了 hive.privilege.synchronizer 属性来禁用权限同步器
逆向移植	HIVE-15826：支持为所有 SerDes 配置“serialization.encoding”
逆向移植	HIVE-18284：修复使用 dynpart 排序优化插入带有“distribute by”子句的数据时出现的 NPE
逆向移植	HIVE-24930：在向量化代码路径中不使用来自子操作的 operator.setDone() 短路（如果 childSize == 1）
逆向移植	HIVE-24523：LazySimpleSerde 的向量化读取路径不支持时间戳的 SERDEPROPERTIES
逆向移植	HIVE-23265：在设置了限制和偏移的情况下返回重复的 rowset
逆向移植	HIVE-21492：VectorizedParquetRecordReader 无法读取使用 thrift/自定义工具生成的 parquet 文件
逆向移植	HIVE-22540：向量化：Decimal64 列不适用于 VectorizedBatchUtil.makeLikeColumnVector()
逆向移植	HIVE-22588：切换向量 groupby 模式时，刷新其余分组集的剩余行
逆向移植	HIVE-22551：BytesColumnVector initBuffer 应该一致地清除向量和长度
逆向移植	HIVE-22448：CBO：使用分组按键扩展不同的多个计数
逆向移植	HIVE-22248：修复持续存在的统计问题

类型	描述
逆向移植	HIVE-22210：向量化可以重复使用筛选中涉及的计算输出列
逆向移植	HIVE-21531：向量化：所有空哈希码都不是使用 Murmur3 计算的
逆向移植	HIVE-20419：向量化：防止在 hashmap 键中使用 VectorPartitionDesc 后发生更改
逆向移植	HIVE-19388：VectorMapJoinCommonOperator 初始化期间的 ClassCastException
逆向移植	HIVE-21584：Java 11 准备工作：系统类加载程序不是 URLClassLoader
逆向移植	HIVE-25107：类路径日志记录应处于 DEBUG 级别（#2271）
逆向移植	HIVE-22097：不兼容适用于 java 11 的 java.util.ArrayList
逆向移植	HIVE-23938：LLAP：JDK11 – 某些 GC 日志文件轮换相关的 jvm 参数无法再使用
逆向移植	HIVE-26226：在 upgrade-acid 中将 jdk.tools dep 从 hive-metastore 中排除
逆向移植	HIVE-17879：升级 Datanucleus Maven 插件
逆向移植	HIVE-27004：DateTimeFormatterBuilder#appendZoneText 无法在高于 8 的 Java 版本中解析“UTC+”
逆向移植	HIVE-16812：VectorizedOrcAcidRowBatchReader 无法过滤删除事件
逆向移植	HIVE-17917：VectorizedOrcAcidRowBatchReader.computeOffsetAndBucket 优化

类型	描述
逆向移植	HIVE-19985：ACID：跳过解码只读查询的 ROW__ID 部分
逆向移植	HIVE-20635：VectorizedOrcAcidRowBatchReader 不会过滤原始文件的删除事件
升级	将 Javadoc 升级到 3.3.1
升级	将 Javassist 升级到 3.24.1-GA
升级	将 apache-directory-server 更新到 2.0.0-M14

新配置

名称	分类	描述
hive.metastore.fs.drop.partition.threads	hive-site	删除分区线程池中的核心线程数。
hive.metastore.fs.drop.partition.keepalive.time	hive-site	空闲的丢弃分区异步线程（来自线程池）在终止之前等待新任务到达的时间（以秒为单位）。
hive.metastore.fs.drop.partition.threadpool.max.queue.size	hive-site	线程池中用于从文件系统中删除分区的最大队列大小。
hive.groupby.enable.deterministic.distribution	hive-site	启用向 Reducer 的密钥分配确定性。它将在调用用于随机分区的 rand 函数时传递一个恒定的种子值。
hive.privilege.synchronizer	hive-site	是否在 HiveServer2 中定期同步来自外部授权者的权限。

名称	分类	描述
hive.cli.query.file.encoding	hive-site	cli 参数中提供的所有类型的查询文件（查询文件、init 查询文件、rc 文件等）的文件编码。
hive.emr.tez.shuffle.enabled	hive-site	Hive on Tez 作业现在默认使用 tez_shuffle 而不是 mapreduce_shuffle 作为默认 Shuffle 处理程序。

已弃用的配置

由于 HIVE-23354 的原因，以下配置属性已弃用并且在 Amazon EMR 发行版 6.11.0 及更高版本中不再支持。

名称	默认值
`hive.mapred.reduce.tasks.speculative.execution`	`false`
`tez.am.speculation.enabled`	`false`

Amazon EMR 6.10.0 – Hive 发布说明

Amazon EMR 6.10.0 – Hive 更改

类型	描述
特征	通过 IAM PassThrough（HiveCLI/Steps API）为 Apache Hive 查询（写入）启用基于 Amazon Lake Formation 的访问控制。
改进	默认情况下禁用配置 `hive.log.explain.output` 以减小日志大小

类型	描述
逆向移植	HIVE-26408：向量化：修复暂存列的取消分配，不要重复使用子级 ConstantVectorExpression 作为输出
逆向移植	HIVE-22269：修复由于 HIVE-20703 导致的统计数据丢失导致的动态分区插入查询中错误的 Reducer 计数。
逆向移植	HIVE-22891：在非 LLAP 执行模式下跳过 CombineHiveRecord 中的 PartitionDesc 提取
逆向移植	HIVE-23804：在 Hive 元存储架构中为列统计数据特定表添加默认数据库，使其向后兼容
逆向移植	HIVE-25277：使用昂贵的 ListFiles 的 Cloud 对象存储缓慢删除 Hive 分区
逆向移植	HIVE-19202：由于 HiveAggregate.isBucketedInput() 中的 NullPointerException，CBO 失败
逆向移植	HIVE-19048：修复 beeline Initscript 错误被忽略的问题
逆向移植	HIVE-21085：实体化视图注册表启动非外部 tez 会话
逆向移植	HIVE-21675：如果视图已经存在，CREATE VIEW IF NOT EXISTS 将返回错误而不是“确定”。这是 Hive 2 的回归。
逆向移植	HIVE-21646：Tez：防止 TezTasks 转义线程日志上下文
逆向移植	HIVE-22054：避免使用递归列出检查目录是否为空

类型	描述
逆向移植	HIVE-16587：插入带有嵌套空值的复杂类型时为 NPE
逆向移植	HIVE-22647：默认启用会话池
逆向移植	HIVE-13288：DagUtils.localizeResource 中存在令人混淆的异常消息
逆向移植	HIVE-23870：在 WritableHiveCharObjectInspector.getPrimitiveJavaObject / HiveCharWritable 中优化多个文本转换
逆向移植	HIVE-21498：将 Thrift 升级到 0.13.0
逆向移植	HIVE-24378：在转换十进制之前，不会删除前导空格和尾随空格
逆向移植	HIVE-21341：合理的默认值：hive.server2.idle.operation.timeout 和 hive.server2.idle.session.timeout 过高
逆向移植	HIVE-22465：在 TezConfigurationFactory 中添加 ssl conf
逆向移植	HIVE-24710：优化 count(*) 的 PTF 迭代以降低 CPU 和 IO 成本
逆向移植	HIVE-15406：考虑对新的“trunc”函数进行向量化
逆向移植	HIVE-21541：修复 HIVE-15406 中缺少的 asf 标题
逆向移植	HIVE-24808：缓存已解析的日期
逆向移植	HIVE-24746：PTF：TimestampValueBoundaryScanner 可以在范围计算期间进行优化

类型	描述
逆向移植	HIVE-25059：在复制过程中，更改事件被转换为重命名
逆向移植	HIVE-25142：在映射联接快速哈希表中重新哈希会导致大密钥损坏
逆向移植	HIVE-23756：向 package.jdo 文件添加了更多限制
逆向移植	HIVE-25150：在进行十进制转换之前不会移除制表符，类似于作为 HIVE-24378 一部分修复的空格字符
逆向移植	HIVE-25093：date_format() UDF 仅以 UTC 时区返回输出
逆向移植	HIVE-25268：如果本地时区不是世界标准时间，则对于 1900 年之前的日期，date_format udf 将返回错误的结果
逆向移植	HIVE-25338：如果输入为空，则在 conv UDF 中出现 AIOBE
逆向移植	HIVE-22400：带时间的 UDF 分钟返回空值
逆向移植	HIVE-25058：PTF: TimestampValueBoundaryScanner 可以在范围计算 pt2 - isDistanceGreater 期间进行优化
逆向移植	HIVE-25449：datediff() 在某些非 UTC 时区的 tez 任务中运行时给出了错误的输出
逆向移植	HIVE-23688：向量化：IndexArrayOutOfBoundsException 适用于包含空值的地图类型列

类型	描述
逆向移植	HIVE-22247：当分区的任务输出为空时 HiveHFileOutputFormat 会引发 FileNotFoundException
逆向移植	HIVE-25570：Hive 应发送完整的 URL 路径以获得命令插入覆盖位置的授权
逆向移植	HIVE-22903：如果分区子句中有常量表达式，向量化的 row_number() 会在一批之后重置行号
逆向移植	HIVE-25549：在 PARTITION BY 或 ORDER BY 子句中带有表达式的窗口函数的结果错误
逆向移植	HIVE-25579：LOAD 覆盖会附加而不是覆盖
逆向移植	HIVE-25659：应根据 SQL 数据库允许的最大参数来拆分带有 IN/(NOT IN) 的元存储直接 sql 查询
逆向移植	HIVE-20502：修复使用列统计数据时运行 skewjoin_mapjoin10.q 时出现的 NPE 问题。
逆向移植	HIVE-25765：当文件大小较大时，skip.header.line.count 属性会跳过 FetchOperator 中每个块的行
错误	在 `hive.stats.column.autogather` 和 `hive.groupby.skewindata` 都启用的特定情况下，在插入时修复 NPE
错误	在未设置 `mapred.tasktracker.expiry.interval` 值时修复 NPE

Amazon EMR 6.9.0 – Hive 发布说明

Amazon EMR 6.9.0 – Hive 更改

类型	描述
升级	将 Jetty 升级到 9.4.48.v20220622
升级	对于 Hadoop 3.3.3 的支持
特征	Amazon EMR Hive 使用 GCSC API 与 Lake Formation 集成，以实现交互式工作负载。
特征	Amazon EMR Hive 与 Iceberg 的集成。
改进	使用 Amazon EMR 安全配置启用传输中加密后，在 HiveServer2 中启用 SSL。
改进	默认情况下启用 Hive EMRFS Amazon S3 优化提交程序。有关更多信息，请参阅 启用 Hive EMRFS S3 优化提交程序。
改进	添加仅继承 InputFormat mapred 版本的 `HiveHBaseTableInputFormatV2` 以修复 SPARK-34210。将 `hive.hbase.inputformat.v2` 设置为 `true` 以使用它。
改进	等待 TezaM 在后台使用 hive.cli.tez.session.async 启动，而不是终止它后立即启动新版本。使用 `hive.emr.cli.tez.session.open.timeout` 以秒为单位设置此超时。
改进	添加选项 hive.conf.restricted.list.append，以将逗号分隔的配置附加到现有的受限配置列表 `hive.conf.restricted.list` 中。

类型	描述
改进	由于未为数据库定义位置而导致 Hive 查询失败时，会出现更清晰的错误消息。
逆向移植	HIVE-24484：将 Hadoop 升级到 3.3.1，并将 Tez 升级到 0.10.2
逆向移植	HIVE-22398：通过 SHIMLoader 移除 YARN 队列管理。
逆向移植	HIVE-23190：LLAP：修改 IndexCache 以将文件系统对象传递给 TezSpillRecord。
逆向移植	HIVE-22185：HADOOP-15832 会导致使用 MiniYarn 集群时出现测试问题。
逆向移植	HIVE-21670：将 `mockito-all` 替换为 `mockito-core` 依赖项。
逆向移植	HIVE-24542：准备 Guava 以进行升级。
逆向移植	HIVE-23751：QTest：覆盖 `ProxyFileSystem` 中的 `#mkdirs()` 方法以在 HADOOP-16582 之后对齐。
逆向移植	HIVE-21603：准备 Java 11：更新 powermock 版本。
逆向移植	HIVE-24083：Hadoop 3.3.0 中出现 `hcatalog` 错误：需要身份验证类型。
逆向移植	HIVE-24282：除非明确提及，否则显示列不得对输出列进行排序。
逆向移植	HIVE-20656：合理的默认值：映射聚合内存配置过于激进。

类型	描述
逆向移植	HIVE-25443：如果值超过 1024 个，Arrow SerDe 无法对复杂数据类型进行序列化/反序列化。
逆向移植	HIVE-19792：将 orc 升级到 1.5.2 并启用 decimal_64 架构发展测试。
逆向移植	HIVE-20437：处理从浮点数、双精度浮点数和十进制数转换的架构发展。
逆向移植	HIVE-21987：Hive 无法读取使用十进制注释的 Parquet int32。
逆向移植	HIVE-20038：对非分桶表和未分区表的更新查询会引发 NPE。

Amazon EMR 6.9.0 – Hive 已知问题

- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。

Amazon EMR 6.8.0 – Hive 发布说明

Amazon EMR 6.8.0 – Hive 更改

类型	描述
改进	减少 msck 命令中的文件系统调用。性能改进（在 10k 以上的分区上提高约 15-20 倍）

类型	描述
逆向移植	HIVE-20678：HivehBaseTableOutputFormat 应实现 HiveOutputFormat 以确保兼容性
逆向移植	HIVE-21040：msck 在目录树的最后一级列出不必要的文件
逆向移植	HIVE-21460：加载数据后再进行 select * 查询会导致结果不正确
逆向移植	HIVE-21660：当使用 union all 和 later with explode 时结果错误
逆向移植	HIVE-22505：classCastException 由错误的向量化运算符选择导致
逆向移植	HIVE-22513：过滤器运算中强制转换列的持续传播可能会导致不正确的结果
逆向移植	HIVE-23435：完整的外部联接结果缺少行
逆向移植	HIVE-24209：启用向量化时，NOT BETWEEN 运算的搜索参数转换不正确
逆向移植	HIVE-24934：GenericUDFSQCountCheck 中不需要 VectorizedExpressions 注释
逆向移植	HIVE-25278：HiveProjectJoinTransposeRule 可能会使用窗口表达式进行无效转换
逆向移植	HIVE-25505：如果第一行为空，则 header.skip.header.line.count 的结果不正确
逆向移植	HIVE-26080：将 accumulo-core 升级到 1.10.1
逆向移植	HIVE-26235：二进制列上的 OR 条件返回空结果

类型	描述
错误	修复启动期间 stderr 中的多个 SLF4J 绑定警告日志
错误	修复当分区和表位于不同文件系统时 SHOW TABLE EXTENDED 查询失败并出现 Wrong FS 错误的问题。

Amazon EMR 6.8.0 – Hive 已知问题

- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。

Amazon EMR 6.7.0 – Hive 发布说明

Amazon EMR 6.7.0 – Hive 更改

类型	描述
特征	Amazon EMR Hive 与 LakeFormation 的集成
特征	适用于 Hive EMRFS Amazon S3 优化提交程序的其他审核日志记录。Hive config：`hive.blobstore.output-committer.logging` ，默认值：`false`
特征	如果在插入覆盖时删除目标目录，选择结果为空，会导致未分区的表/静态分区的行为与 Hive 2.x 类似。Hive config：`hive.emr.iow.clean.target.dir` ，默认值：`false`

类型	描述
错误	修复了在将 Hive EMRFS Amazon S3 优化提交程序与分区存储桶排序结合使用时出现间歇性查询失败的问题。
升级	已将 Hive 升级到版本 3.1.3。请参阅 Apache Hive 3.1.3 发布说明以了解更多详细信息。
升级	已将 Parquet 升级到 1.12.2。
逆向移植	HIVE-20065：元存储不应依赖 jackson 1.x
逆向移植	HIVE-20071：迁移到 jackson 2.x 并阻止使用
逆向移植	HIVE-20607：TxnHandler 应使用 PreparedStatement 来执行直接 SQL 查询
逆向移植	HIVE-20740：移除 ObjectStore.setConf 方法中的全局锁定
逆向移植	HIVE-20961：停用 NVL 实施
逆向移植	HIVE-22059：hive-exec jar 不包含（fasterxml）jackson 库
逆向移植	HIVE-22351：修复 TestObjectStore 中线程化 ObjectStore 的错误用法
逆向移植	HIVE-23534：在捕获 MetaException 时 RetryingMetaStoreClient#invoke 中出现 NPE，但无消息
逆向移植	HIVE-24048：将 Jackson 组件统一到版本 2.10.最新版 – Hive
逆向移植	HIVE-24768：在所有地方均使用 jackson-bom 进行版本替换

类型	描述
逆向移植	HIVE-24816：由于 CVE-2020-25649 的原因，将 jackson 升级到 2.10.5.1 或 2.11.0+
逆向移植	HIVE-25971：Tez 任务关闭因缓存线程池未关闭而延迟
逆向移植	HIVE-26036：ObjectStore 中的 getMTable() 导致 NPE

Amazon EMR 6.7.0 – Hive 已知问题

- 如果使用窗口函数对与交集函数相同的列进行查询，则可能会导致如 HIVE-25278 中报告的转换无效问题，并导致查询结果不正确或查询失败。解决方法是在查询级别为此类查询禁用 CBO。修复程序将在 6.7.0 之后的 Amazon EMR 发行版中提供。有关更多信息，请联系 Amazon Support。
- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。

Amazon EMR 6.6.0 – Hive 发布说明

Amazon EMR 6.6.0 – Hive 更改

类型	描述
升级	将 Parquet 升级到 1.12.1。
升级	将 jetty jars 版本升级到 9.4.43.v20210629
错误	修复了在 Hive 集群上启用 LLAP 时导致在所有任务/核心节点上安装 Hive 的问题。

类型	描述
逆向移植	HIVE-25942：由于 CVE-2021-29425 而将 commons-io 升级到 2.8.0
逆向移植	HIVE-25726：由于 CVE-2020-13936 而将速度升级到 2.3
逆向移植	HIVE-25680：授权 #get_table_meta HiveMetastore 服务器 API 使用任何 HiveMetastore 授权模型。
逆向移植	HIVE-25554：将 arrow 版本升级到 0.15
逆向移植	HIVE-25242：使用 vectorized.adaptor = chosen 时，查询执行速度极慢
逆向移植	HIVE-25085：MetaStore 客户端不再在会话之间共享。
逆向移植	HIVE-24827：Hive 聚合查询为非文本文件返回错误结果。
逆向移植	HIVE-24683：如果路径不存在，Hadoop23Shims getFileId 容易出现 NPE
逆向移植	HIVE-24656：对于 map 和数组类型为 null 的查询，出现 CBO 失败
逆向移植	HIVE-24556：针对没有孙子的情况优化 DefaultGraphWalker
逆向移植	HIVE-24408：将 Parquet 升级到 1.11.1
逆向移植	HIVE-24391: 修复 branch-3.1 中的 TestOrcFile 失败
逆向移植	HIVE-24362：对于具有大量节点的树而言，AST 树处理非最优

类型	描述
逆向移植	HIVE-24316: 在 branch-3.1 中将 ORC 从 1.5.6 升级到 1.5.8
逆向移植	HIVE-24307：带属性文件和 -e 参数的 Beeline 失败
逆向移植	HIVE-24245：带有计数和不同分区的向量化 PTF 会产生错误结果。
逆向移植	HIVE-24224：修复在压缩文件上跳过 Hive on Tez 的标头/脚注
逆向移植	HIVE-24157：严格模式在 CAST 时间戳 ↔ 数值上失败
逆向移植	HIVE-24113：GenericUDFToUnixTimeStamp 中出现 NPE
逆向移植	HIVE-23987：将 arrow 版本升级到 0.11.0
逆向移植	HIVE-23972：将外部客户端 ID 添加到 LLAP 外部客户端
逆向移植	HIVE-23806：避免在扩展架构的情况下清除所有分区中的列统计数据状态。这提高了 alter table add columns 语句的运行时间。
逆向移植	HIVE-23779: BasicStatsTask 信息无法在 beeline 控制台中打印
逆向移植	HIVE-23306：如果 System.getProperty 设置了配置，则 RESET 命令不起作用
逆向移植	HIVE-23164：由于非进程守护程序线程，服务器未正确终止
逆向移植	HIVE-22967：支持 Hive on Tez 的 hive.reloadable.aux.jars.path

类型	描述
逆向移植	HIVE-22934：Hive 服务器交互式日志计数器用于错误流
逆向移植	HIVE-22901: 变量替换可能导致循环引用中出现 OOM
逆向移植	HIVE-22769: 压缩文本文件拆分生成期间出现查询结果不正确和查询失败
逆向移植	HIVE-22716: ByteBuffer 的读取操作在 ParquetFooterInputFromCache 处中断
逆向移植	HIVE-22648：将 Parquet 升级到 1.11.0
逆向移植	HIVE-22640：Decimal64ColumnVector：当分区列类型为十进制时，出现ClassCastException
逆向移植	HIVE-22621: 不稳定的测试用例：TestLlapSignerImpl.testSigning
逆向移植	HIVE-22533：修复可能出现的 LLAP 进程守护程序 Web UI 漏洞
逆向移植	HIVE-22532：PTFPPD 可能会通过 Rank/DenseRank 函数错误地推送限制
逆向移植	HIVE-22514: HiveProtoLoggingHook 可能会消耗大量内存
逆向移植	HIVE-22476：当 hive.fetch.task.conversion 设置为 none 时，hive datediff 函数所提供的结果不一致
逆向移植	HIVE-22429：在 hive 3 上通过 bucketing_version 1 迁移的集群表使用 bucketing_version 2 进行插入
逆向移植	HIVE-22412: StatsUtils 在解释期间抛出 NPE

类型	描述
逆向移植	HIVE-22360：如果加载的文件的列多于表架构中的列，MultiDelimitSerDe 会在最后一列返回错误结果
逆向移植	HIVE-22332：自 ORC-540 以来，Hive 会确保有效的架构发展设置
逆向移植	HIVE-22331：不带参数的 unix_timestamp 以毫秒为单位（而不是秒）返回时间戳
逆向移植	HIVE-22275: OperationManager.queryIdOperation 并未正确清理多个 queryId
逆向移植	HIVE-22273：删除临时目录时，访问检查失败
逆向移植	HIVE-22270：将 commons-io 升级到 2.6
逆向移植	HIVE-22241：实现 UDF 以使用其内部表示和 Gregorian-Julian 混合日历来解释日期/时间戳
逆向移植	HIVE-22241：实现 UDF 以使用其内部表示和 Gregorian-Julian 混合日历来解释日期/时间戳
逆向移植	HIVE-22232：当 hive.order.columnalignment 设置为 false 时，出现 NPE
逆向移植	HIVE-22231：通过 knox 进行大量 hive 查询失败，并显示“Broken pipe Write failed（断开链接回写失败）”
逆向移植	HIVE-22221：Llap 外部客户端 – 需要减少 LlapBaseInputFormat#getSplits
逆向移植	HIVE-22208：当查询（包括在具有掩码列的表上的联接）被重写时，具有保留关键字的列名无法转义

类型	描述
逆向移植	HIVE-22197：Common Merge join 抛出类强制转换异常。
逆向移植	HIVE-22170：from_unixtime 和 unix_timestamp 会使用用户会话时区
逆向移植	HIVE-22169：Tez：SplitGenerator 尝试查找对 Tez 而言不存在的计划文件
逆向移植	HIVE-22168：从 llap 缓存热路径中删除极为昂贵的日志记录
逆向移植	HIVE-22161：UDF：FunctionRegistry 在 org.apache.hadoop.hive.ql.udf.UDFType 类上同步
逆向移植	HIVE-22120：修复在特定边界条件下 map 端左外连接中出现的错误结果/ArrayOutOfBound 异常
逆向移植	HIVE-22115：如果属性设置为 false，会阻止创建查询 routing appender
逆向移植	HIVE-22113：防止在 AMReporter 相关的 RuntimeException 上关闭 LLAP
逆向移植	HIVE-22106：删除 partition-eval 的 cross-query 同步
逆向移植	HIVE-22099：自 HIVE-20007 以来，几个与日期相关的 UDF 无法正确处理 Julian 日期
逆向移植	HIVE-22037：HS2 因 OOM 而关闭时，会记录日志
逆向移植	HIVE-21976：在 Calcite HiveSortLimit 中，偏移量为 null 而不是零

类型	描述
逆向移植	HIVE-21924：即使存在标头/脚注，也可拆分文本文件
逆向移植	HIVE-21913：GenericUDTFGetSplits 会以与 LLAP 相同的方式处理用户名
逆向移植	HIVE-21905: 围绕 FetchOperator 类的泛型改进
逆向移植	HIVE-21902：HiveServer2 UI：jetty 响应标头需要 X-Frame-Options
逆向移植	HIVE-21888：将 hive.parquet.timestamp.skip.conversion 默认设置为 true
逆向移植	HIVE-21868: 矢量化 CAST...FORMAT
逆向移植	HIVE-21864：LlapBaseInputFormat#closeAll
逆向移植	HIVE-21863：改进 WHEN 表达式的 Vectorizer 类型转换
逆向移植	HIVE-21862：ORC ppd 产生带有时间戳的错误结果
逆向移植	HIVE-21846：在 TezAM 中创建一个定期获取 LlapDaemon 指标的线程
逆向移植	HIVE-21837：当选定的列完全具有 null 值时，MapJoin 会引发异常
逆向移植	HIVE-21834：避免不必要的调用以简化筛选条件
逆向移植	HIVE-21832：获取平均队列/服务/响应时间的新指标
逆向移植	HIVE-21827：SemanticAnalyzer 中的多个调用不通过 getTableObjectByName 方法

类型	描述
逆向移植	HIVE-21822：通过新的 API 方法公开 LlapDaemon 指标
逆向移植	HIVE-21818: CBO：TableRelOptHiveTable 的复制包含元存储流量
逆向移植	HIVE-21815：ORC 文件中的统计数据被解析两次
逆向移植	HIVE-21805：HiveServer2：使用快速 ShutdownHookManager API
逆向移植	HIVE-21799：当连接键位于聚合列时，NullPointerException 在 DynamicPartitionPruningOptimization 中
逆向移植	HIVE-21794：将具体化视图参数添加到 sqlStdAuthSafeVarNameRegexes
逆向移植	HIVE-21768：JDBC：删除没有括起来的 UNION 查询的默认联合前缀
逆向移植	HIVE-21746：动态分区散列连接期间的 ArrayIndexOutOfBoundsException，禁用 CBO
逆向移植	HIVE-21717：移动任务中的目录重命名失败。
逆向移植	HIVE-21685：包含多个 IN 子句的查询出现简化错误
逆向移植	HIVE-21681：Describe formatted 显示了多个主键的错误信息
逆向移植	HIVE-21651：将 protobuf serde 移动到 hive-exec 中。
逆向移植	HIVE-21619：在 SQL 解释扩展中输出的时间戳类型缺乏精度

类型	描述
逆向移植	HIVE-21592：如果表达式包含 CONCAT，则不会显示 OptimizedSql
逆向移植	HIVE-21576: 引入 CAST...FORMAT 和 SQL:2016 日期时间格式的限制列表
逆向移植	HIVE-21573：如果身份验证设置为 delegatio nToken，二进制传输会忽略主体
逆向移植	HIVE-21550: TestObjectStore 测试不稳定 - 无法在请求的时间内获得锁定
逆向移植	HIVE-21544：常量传播会破坏折叠期间的 coalesce/case/when 表达式
逆向移植	HIVE-21539：如果 GroupBy + where 子句在同一列，会导致错误的查询重写
逆向移植	HIVE-21538：Beeline：尽管控制台读取器并未传递到连接参数，但会提供密码来源
逆向移植	HIVE-21509：LLAP 可能会缓存损坏的列向量并返回错误的查询结果
逆向移植	HIVE-21499：如果创建命令失败并显示 AlreadyExistsException，则不会从注册表中删除该函数
逆向移植	HIVE-21496: 自动调整无序缓冲区大小可能会溢出
逆向移植	HIVE-21468: JDBC 存储处理程序的标识符名称区分大小写
逆向移植	HIVE-21467: 移除已弃用的 junit.framework.As sert 导入

类型	描述
逆向移植	HIVE-21435：LlapBaseInputFormat 会在构建 SubmitWorkRequestProto 时从 TASK_ATTEMPT_ID conf 获取任务号（如果存在）
逆向移植	HIVE-21389：HIVE-21247 后，Hive 分发缺少 javax.ws.rs-api.jar
逆向移植	HIVE-21385: 允许禁止向 JDBC 源下推不可拆分的计算
逆向移植	HIVE-21383：JDBC 存储处理程序：通过使用目录和架构检索表（如果已指定）
逆向移植	HIVE-21382: 按密钥减少优化分组 - 在 query23 中未减少密钥
逆向移植	HIVE-21362：添加输入格式和 serde 以从 protobuf 文件中读取。
逆向移植	HIVE-21340: CBO：修剪馈送至 SemiJoin 的非密钥列
逆向移植	HIVE-21332: 清除非锁定缓冲区，而不清除锁定的缓冲区
逆向移植	HIVE-21329：根据运算符管道，自定义 Tez 运行时无序输出缓冲区的大小
逆向移植	HIVE-21295：StorageHandler 会按照 Hive 约定将日期转换为字符串
逆向移植	HIVE-21294：向量化：1-reducer Shuffle 可以跳过对象哈希函数
逆向移植	HIVE-21255：删除 JdbcStorageHandler 中的 QueryConditionBuilder

类型	描述
逆向移植	HIVE-21253：支持 JDBC StorageHandler 中的 DB2
逆向移植	HIVE-21232：LLAP：添加友好的缓存未命中拆分关联提供程序
逆向移植	HIVE-21214：MoveTask：使用 attemptId 而不是文件大小来删除文件 compareTempOrDuplicateFiles 的重复数据
逆向移植	HIVE-21184: 添加解释内容并解释带有成本信息的格式化 CBO 计划
逆向移植	HIVE-21182：在计划期间跳过设置 hive scratch dir
逆向移植	HIVE-21171：如果 RPC 处于启用状态，请跳过为 tez 创建暂存目录
逆向移植	HIVE-21126: 允许在 LlapBaseInputFormat#getSplit 中进行会话级别查询
逆向移植	HIVE-21107：动态分区哈希连接期间，出现“Cannot find field（无法找到字段）”错误
逆向移植	HIVE-21061：CTAS 查询失败，并显示空源的 IllegalStateException
逆向移植	HIVE-21041：从逻辑计划中获取架构出现 NPE，ParseException
逆向移植	HIVE-21013：JdbcStorageHandler 无法在 Oracle 中找到分区列
逆向移植	HIVE-21006：扩展 SharedWorkOptimizer，以便出现再利用机会时删除半连接

类型	描述
逆向移植	HIVE-20992：将配置 hive.metastore.dbaccess.ssl.properties 拆分为更有意义的配置
逆向移植	HIVE-20989：JDBC - The GetOperationStatus + 日志可以通过睡眠阻止查询进度
逆向移植	HIVE-20988：在多列上使用主键进行 group by 查询时，结果错误
逆向移植	HIVE-20985：如果选择运算符输入是临时列，向量化可能会复用其中一部分作为输出
逆向移植	HIVE-20978：会将“hive.jdbc.*”添加到 sqlStdAuthSafeVarNameRegexes
逆向移植	HIVE-20953：如果在创建函数时无法将其添加到元存储，则将其从函数注册表中删除。
逆向移植	HIVE-20952: 清理 VectorizationContext.java
逆向移植	HIVE-20951：LLAP：始终将 Xms 设置为 50%
逆向移植	HIVE-20949：改进物理计划中的 PKFK 基数估算
逆向移植	HIVE-20944：查询编译期间不验证统计数据
逆向移植	HIVE-20940: 桥接 Calcite 类型解析比 Hive 更严格的案例。
逆向移植	HIVE-20937：Postgres jdbc 查询失败，并显示“LIMIT must not be negative（LIMIT 不得为负数）”
逆向移植	HIVE-20926：当 bloom 筛选条件条目较高或没有统计数据时，半联接减少提示失败

类型	描述
逆向移植	HIVE-20920：使用 SQL 约束来改进连接重新排序算法
逆向移植	HIVE-20918: 用于启用/禁用将计算从 Calcite 下推到 JDBC 连接的标记
逆向移植	HIVE-20915: 为 HoS 和 MR 提供动态排序分区优化
逆向移植	HIVE-20910：由于动态分区排序优化，在分桶表中插入失败
逆向移植	HIVE-20899：LLAP YARN 服务的 Keytab URI 仅限于支持 HDFS
逆向移植	HIVE-20898：对于时间相关函数，参数不会转换为非空型
逆向移植	HIVE-20881：常量传播导致投影过度简化
逆向移植	HIVE-20880：更新 hive.stats.filter.in.min.ratio 的默认值
逆向移植	HIVE-20873：对 VectorHashKeyWrapperTwoLong 使用 Murmur 哈希值，以减少哈希冲突
逆向移植	HIVE-20868：如果 TezDummyOperator 在 MapRecordProcessor 中的 getFinalOp 中具有子操作，SMB Join 会间歇性失败
逆向移植	HIVE-20853：在 llap 进程守护程序 API 中公开 shuffleHandler.regerDag
逆向移植	HIVE-20850: 如果可能，将案例条件从投影推送到维度表

类型	描述
逆向移植	HIVE-20842：修复 HIVE-20660 中引入的逻辑，以估算 group by 的统计数据
逆向移植	HIVE-20839：动态分区哈希连接期间，出现“Cannot find field（无法找到字段）”错误
逆向移植	HIVE-20835: 约束与 MV 重写之间的交互可能会在 Calcite 计划程序中创建循环
逆向移植	HIVE-20834：Hive QueryResultCache 条目保持从缓存查询中引用 SemanticAnalyzer
逆向移植	HIVE-20830：在某些情况下，JdbcStorageHandler 范围查询断言失败
逆向移植	HIVE-20829：JdbcStorageHandler 范围拆分会引发 NPE
逆向移植	HIVE-20827：空数组的结果不一致
逆向移植	HIVE-20826: 增强 HiveSemiJoin 规则，将左侧的联接 + 分组依据转换为左半联接
逆向移植	HIVE-20821：将 SUM0 重写为 SUM + COALESCE 组合
逆向移植	HIVE-20815：JdbcRecordReader.next 不会吃掉异常
逆向移植	HIVE-20813：udf to_epoch_milli 也需要支持无时区的时间戳。
逆向移植	HIVE-20804：通过约束实现 group by 深度优化
逆向移植	HIVE-20792：插入带区域的时间戳会截断数据
逆向移植	HIVE-20788：创建筛选条件时，扩展的 SJ 缩减可能导致列错误回溯

类型	描述
逆向移植	HIVE-20778: 如果计划中的所有联接都是通过去相关逻辑创建的，则可能无法触发联接重新排序
逆向移植	HIVE-20772：在 LLAP 中记录每个任务的 CPU 计数器
逆向移植	HIVE-20768：添加滚动窗口 UDF
逆向移植	HIVE-20767：连接运算符之间的多个项目可能会对使用约束的连接重新排序造成影响
逆向移植	HIVE-20762：NOTIFICATION_LOG 清理间隔会被硬编码为 60s，间隔过小
逆向移植	HIVE-20761：选择对 notification_sequence 表进行更新具有重试时间间隔，且重试计数过小
逆向移植	HIVE-20751：将 arrow 版本升级到 0.10.0
逆向移植	HIVE-20746：HiveProtoHookLogger 无法在一天结束时关闭文件。
逆向移植	HIVE-20744：使用 SQL 约束来改进连接重新排序算法
逆向移植	HIVE-20740：删除 ObjectStore.setConf 方法中的全局锁定。此 cherrypick 会将适用于 Hive 3.2 和 4.x 的 HIVE-20740 反向移植到 3.1.x
逆向移植	HIVE-20734：Beeline：如果 beeline-site.xml 存在并且 hive CLI 重定向到 beeline，它会使用系统用户名/虚拟密码而不会提示输入一个 beeline
逆向移植	HIVE-20731：JdbcStorageHandler 中的密钥库文件应获授权

类型	描述
逆向移植	HIVE-20720：将分区列选项添加到 JDBC 处理程序
逆向移植	HIVE-20719: 在 hive.optimize.sort.dynamic.partition 优化和向量化开启的情况下，SELECT 语句在 UPDATE 后失败
逆向移植	HIVE-20718: 添加带有约束条件的 perf cli 驱动程序
逆向移植	HIVE-20716：将 hive.cbo.stats.correlated.multi.key.joins 的默认值设置为 true
逆向移植	HIVE-20712：HivePointLookupOptimizer 会提取深层案例
逆向移植	HIVE-20710：常量折叠可能不会创建没有类型的 null 常量
逆向移植	HIVE-20706: external_jdbc_table2.q 间歇性失败
逆向移植	HIVE-20704：扩展 HivePreFilteringRule 以支持其他功能
逆向移植	HIVE-20703：将动态排序分区优化置于基于成本的决策之下
逆向移植	HIVE-20702：在 mapjoin 选择期间，考虑数据结构感知估算的负载
逆向移植	HIVE-20692：启用 NOT x IS (NOT) [TRUE|FALSE] 表达式的折叠
逆向移植	HIVE-20691：修复 org.apache.hadoop.hive.cli.TestMiniLlapCliDriver.testCliDriver[cttl]
逆向移植	HIVE-20682：如果主线程关闭了共享 sessionHive，则异步查询执行可能会失败

类型	描述
逆向移植	HIVE-20676：HiveServer2：PrivilegeSynchronizer 未设置为进程守护程序状态
逆向移植	HIVE-20660：可通过将总行数绑定到源表，改进 group by 统计数据估算
逆向移植	HIVE-20652：JdbcStorageHandler 将两个不同数据源的连接推送到 jdbc 驱动程序
逆向移植	HIVE-20651：JdbcStorageHandler 密码会进行加密
逆向移植	HIVE-20649：LLAP 感知内存管理器可用于 Orc 写入器
逆向移植	HIVE-20648：LLAP：按运算符的向量组会使用每个执行程序的内存
逆向移植	HIVE-20646：如果分区筛选条件具有 IS NOT NULL，则不会下推到元存储查询
逆向移植	HIVE-20644：避免经由 Hive Runtime 异常泄露敏感信息
逆向移植	HIVE-20636：对外部连接后的 null 值估算数目进行改进
逆向移植	HIVE-20632：如果在查询表上创建具体化视图，则使用 get_splits UDF 查询会失败
逆向移植	HIVE-20627: 并发异步查询间歇性地失败并抛出 LockException，同时导致内存泄漏
逆向移植	HIVE-20623：共享工作：扩展 LLAP 中 map-join 缓存条目的共享
逆向移植	HIVE-20619: 默认情况下，在 HiveServer2 中包含 MultiDelimitSerDe

类型	描述
逆向移植	HIVE-20618：在连接选择期间，可能会为非分桶表选择 BucketMapJoin
逆向移植	HIVE-20617：修复 IN 表达式中的常量类型，使其具有正确类型
逆向移植	HIVE-20612: 为 CBO 创建新的联接多密钥关联标记
逆向移植	HIVE-20603：更改表位置文件系统后，插入分区时出现“Wrong FS（FS 错误）”错误
逆向移植	HIVE-20601：DbNotificationListener 中 ALTER_PARTITION 事件中的 EnvironmentContext 为 null
逆向移植	HIVE-20583：仅对 HiveConnection 中的 kerberos 身份验证使用规范主机名
逆向移植	HIVE-20582：hive proto 日志记录中的 hflush 可配置
逆向移植	HIVE-20563：向量化：如果 THEN/ELSE 类型和结果类型不同，CASE WHEN 表达式会失败
逆向移植	HIVE-20558：将 hive.hashtable.key.count.adjustment 的默认值更改为 0.99
逆向移植	HIVE-20552: 更快地从 LogicalPlan 获取架构
逆向移植	HIVE-20550：切换 WebHCat 以使用 beeline 提交 Hive 查询
逆向移植	HIVE-20537：使用与 CBO 和 Hive 中不同的不相关列，进行多列连接估算

类型	描述
逆向移植	HIVE-20524：在从 Hive 版本 2 升级到版本 3 以从 ALTER TABLE VARCHAR 转到 DECIMAL 的过程中，架构发展检查中断
逆向移植	HIVE-20522：由于字段可为 null 值，HiveFilterSetOpTransposeRule 可能引发断言错误
逆向移植	HIVE-20521：HS2 doAs=true 与 hadoop.tmp.dir、MR 和 S3A 文件系统存在权限问题
逆向移植	HIVE-20515：使用不同文件系统中的结果缓存、查询临时目录和结果缓存目录时，查询结果为空
逆向移植	HIVE-20508：Hive 不支持“user@realm”类型的用户名
逆向移植	HIVE-20507：Beeline：添加实用程序命令以从 beeline-site.xml 中检索所有 uri
逆向移植	HIVE-20505：将 org.openjdk.jmh:jmh-core 升级到 1.21
逆向移植	HIVE-20503：在 mapjoin 选择期间，使用数据结构感知估算
逆向移植	HIVE-20498：支持列统计数据自动收集的日期类型
逆向移植	HIVE-20496：向量化：向量化 PTF IllegalStateException
逆向移植	HIVE-20494：GenericUDFRestrictInformationSchema 在 HIVE-19440 之后被破坏
逆向移植	HIVE-20477：如果表达式包含 IN，则不会显示 OptimizedSql

类型	描述
逆向移植	HIVE-20467: 创建/删除资源计划时允许 IF NOT EXISTS/IF EXISTS
逆向移植	HIVE-20462：如果视图已存在，则“CREATE VIEW IF NOT EXISTS（如不存在则创建视图）”失败
逆向移植	HIVE-20455：security.authorization.PrivilegeSynchonizer.run 出现 日志过多的问题
逆向移植	HIVE-20439：在 llap 的连接选择期间，限制膨胀内存
逆向移植	HIVE-20433：字符串隐式转换为时间戳过慢
逆向移植	HIVE-20432：将 BETWEEN 重写为 IN，以便统计估算的整数类型
逆向移植	HIVE-20423: 将 NULLS LAST 设置为默认空顺序
逆向移植	HIVE-20418：如果查询未选择列，LLAP IO 可能无法处理正确禁用行索引的 ORC 文件
逆向移植	HIVE-20412：HiveMetaHook 中出现 NPE
逆向移植	HIVE-20406：Nested Coalesce 给出错误结果
逆向移植	HIVE-20399：对于 MM 表而言，不完全限定的 CTAS w/a 自定义表位置失败
逆向移植	HIVE-20393：Semijoin Reduction：markSemiJoinForDPP 行为不一致
逆向移植	HIVE-20391: HiveAggregateReduceFunctionsRule 在分解聚合函数时可能会推断出错误的返回类型

类型	描述
逆向移植	HIVE-20383：hive proto 事件挂钩出现无效队列名称和同步问题。
逆向移植	HIVE-20367：向量化：支持 PTF AVG、MAX、MIN、SUM 进行流式传输
逆向移植	HIVE-20366: 因为筛选条件为空，TPC-DS query78 统计数据估计值已关闭
逆向移植	HIVE-20364：更新 hive.map.aggr.hash .min.reduction 的默认值
逆向移植	HIVE-20352：向量化：支持分组功能
逆向移植	HIVE-20347: hive.optimize.sort.dynamic. partition 应该适用于分区的 CTAS 和 MV
逆向移植	HIVE-20345：如果从不同的调用中删除表，则删除数据库可能会挂起
逆向移植	HIVE-20343: Hive 3：CTAS 未以 transacti onal_properties 为准
逆向移植	HIVE-20340: 当时间戳函数的输出用作 Strin 时，Druid 需要从 Timestamp 显式 CAST 为 STRING
逆向移植	HIVE-20339：向量化：取消不需要的限制导致某些 RANK 和 PTF 无法向量化
逆向移植	HIVE-20337：CachedStore：getPartitionsB yExpr 未正确填充分区列表
逆向移植	HIVE-20336：屏蔽和筛选具体化视图的策略
逆向移植	HIVE-20326：创建约束时，将 RELY 作为默认值而不是 NO RELY

类型	描述
逆向移植	HIVE-20321：向量化：将 1 col VectorHashKeyWrapper 的内存大小减少至 <1 Cache Line
逆向移植	HIVE-20320: 开启 hive.optimize.remove.sq_count_check 标记
逆向移植	HIVE-20315：向量化：修复更多的 NULL/错误结果问题，并避免不必要的强制转换/转换
逆向移植	HIVE-20314：具体化视图重写包括分区修剪
逆向移植	HIVE-20312：允许 arrow 客户端将其 BufferAllocator 与 LlapOutputFormatService 结合使用
逆向移植	HIVE-20302：LLAP：IO 中的非向量化执行忽略了 ROW__ID 等虚拟列
逆向移植	HIVE-20300：VectorFileSinkArrowOperator
逆向移植	HIVE-20299：LLAP 签名者单元测试中存在潜在争用
逆向移植	HIVE-20296：改进 HivePointLookupOptimizerRule，便于从更复杂的上下文中提取
逆向移植	HIVE-20294：向量化：修复 COALESCE / ELT 中的 NULL/错误结果问题
逆向移植	HIVE-20292: 已定义主要约束的 tpcds query93 中的联接排序出错
逆向移植	HIVE-20290：延迟初始化 ArrowColumnarBatchSerDe，因而它不会在 GetSplits 期间分配缓冲区
逆向移植	HIVE-20281：SharedWorkOptimizer 失败并显示“operator cache contents and actual plan differ（运算符缓存内容与实际计划不一致）”

类型	描述
逆向移植	HIVE-20277：向量化：FILTER 不支持 case 表达式返回布尔值
逆向移植	HIVE-20267：扩展 WebUI 以包含表单，便于动态配置日志级别
逆向移植	HIVE-20263：HiveReduceExpressionsWithStatsRule 中的拼写错误多变
逆向移植	HIVE-20260：如果另一列的筛选条件更改行数，不会扩展列的 NDV
逆向移植	HIVE-20252：Semijoin Reduction：如果小表端在上游有映射联接，则可能无法检测到半连接分支引起的循环。
逆向移植	HIVE-20245：向量化：修复 BETWEEN/IN 中的 NULL/错误结果问题
逆向移植	HIVE-20241: 支持 CTAS 语句中的分区规范
逆向移植	HIVE-20240: Semijoin Reduction：使用本地变量检查外部表的条件
逆向移植	HIVE-20226：如果请求 maxEvents 超过表的 max_rows，HMS getNextNotification 会引发异常
逆向移植	HIVE-20225: SerDe 支持 Teradata 二进制格式
逆向移植	HIVE-20213: 将 Calcite 升级到 1.17.0
逆向移植	HIVE-20212：http 模式下的 Hiveserver2 错误地发出指标 default.General.open_connections
逆向移植	HIVE-20210：当筛选非分区列且转换最少时，Simple Fetch 优化程序会导致 mapReduce

类型	描述
逆向移植	HIVE-20209: 首次尝试在 repl 转储中连接元存储失败
逆向移植	HIVE-20207：向量化：修复 Filter/Compare 中的 NULL/错误结果问题
逆向移植	HIVE-20204：IN 期间进行类型转换
逆向移植	HIVE-20203：Arrow SerDe 泄露 DirectByteBuffer
逆向移植	HIVE-20197: 矢量化：添加 DECIMAL_64 测试，添加日期/间隔/时间戳算法，并添加更多 GROUP BY 聚合
逆向移植	HIVE-20193: cboInfo 不在解释计划 json 中
逆向移植	HIVE-20192：HS2 与嵌入式元存储配合使用时，泄露 JdOperistencManager 对象
逆向移植	HIVE-20183：如果源表包含空存储桶，则从分桶表插入可能导致数据丢失
逆向移植	HIVE-20177: 矢量化：减少 GroupBy 流式处理模式下的 KeyWrapper 分配
逆向移植	HIVE-20174: 矢量化：修复 GROUPBY 聚合函数中的 NULLLLL/错误结果问题
逆向移植	HIVE-20172：StatsUpdater 在尝试连接到远程元存储时失败，并显示 GSS 异常
逆向移植	HIVE-20153: Count 和 Sum UDF 在 Hive 2+ 中消耗更多内存
逆向移植	HIVE-20152：如果 repl 转储失败，重置数据库状态可以重命名表

类型	描述
逆向移植	HIVE-20149：TestHiveCli 失败/超时
逆向移植	HIVE-20130：实现更好的信息架构同步器日志记录
逆向移植	HIVE-20129：恢复到 orc 表的基于位置的架构发展
逆向移植	HIVE-20118：SessionStateUserAuthenticator.getGroupNames
逆向移植	HIVE-20116：TezTask 正在使用父记录器
逆向移植	HIVE-20115: ACID 表不应使用页脚扫描进行分析
逆向移植	HIVE-20103：WM：如果至少使用了一个，则仅有 Aggregate DAG 计数器
逆向移植	HIVE-20101：BloomkFilter：避免完全使用本地字节 [] 数组
逆向移植	HIVE-20100：OpTraits：如果检测到不匹配，Select Optraits 会停止
逆向移植	HIVE-20098：统计数据：获取日期列分区统计信息时，出现 NPE
逆向移植	HIVE-20095：修复将计算推送到 jdbc 外部表的功能
逆向移植	HIVE-20093: LlapOutputFomatService：使用 ArrowBuf 和 Netty 进行会计工作
逆向移植	HIVE-20090：对 semijoin reduction 筛选条件创建进行扩展，以便发现新机会
逆向移植	HIVE-20088：Beeline 配置位置路径组装不正确

类型	描述
逆向移植	HIVE-20082：HiveDecimal 转换为字符串无法正确格式化十进制数字
逆向移植	HIVE-20069：在 DPP 和 Semijoin 优化的情况下，修复重优化
逆向移植	HIVE-20051：跳过对临时表的授权
逆向移植	HIVE-20044：Arrow Serde 会填充字符值并正确处理空字符串
逆向移植	HIVE-20028：元存储客户端缓存配置使用不正确
逆向移植	HIVE-20025: 清理由 HiveProtoLoggingHook 创建的事件文件
逆向移植	HIVE-20020: Hive contrib jar 不应该在 lib 中
逆向移植	HIVE-20013: 为 to_date 函数添加隐式日期类型转换
逆向移植	HIVE-20011：从 proto 日志记录挂钩中的追加模式迁移
逆向移植	HIVE-20005: acid_table_stats、acid_no_buckets 等 - 分支的查询结果变更
逆向移植	HIVE-20004: ConvertDecimal64ToDecimal 使用的错误比例会导致结果不正确
逆向移植	HIVE-19995: ACID 表的聚合行流量
逆向移植	HIVE-19993：无法使用同样显示为列名的表别名

类型	描述
逆向移植	HIVE-19992：向量化：HIVE-19951 的后续 —> 添加对 SchemaEvolution.isOnlyImplicitConversion 的调用，以便仅在不是数据类型隐式转换的情况下禁用 ORC 的编码 LLAP I/O
逆向移植	HIVE-19989：元存储对 HADOOP2 指标使用错误的应用程序名称
逆向移植	HIVE-19981：由 HiveStrictManagedMigration 实用程序转换为外部表的托管表会设置为在删除表时删除数据
逆向移植	HIVE-19967：SMB 联接：需要适用于 PTFOperator ala GBY Op 的 Optraits
逆向移植	HIVE-19935：Hive WM 会话已终止：无法更新 LLAP 任务计数
逆向移植	HIVE-19924: 标记 Repl 负载运行的 distcp 任务
逆向移植	HIVE-19891：使用自定义分区目录插入外部表，可能导致数据丢失
逆向移植	HIVE-19850：Tez 中的动态分区修剪导致“No work found for tablescan（并未找到表扫描的任何工作）”错误
逆向移植	HIVE-19806: 对 qtests 输出进行排序以避免测试结果出现不稳定
逆向移植	HIVE-19770: 支持 CBO 以用于在 select 语句中包含多个相同列的查询
逆向移植	HIVE-19769: 为数据库和表名创建专用对象
逆向移植	HIVE-19765: 向 BlobstoreCliDriver 添加 Parquet 特定的测试

类型	描述
逆向移植	HIVE-19759：不稳定的测试：TestRpc#testServerPort
逆向移植	HIVE-19711：重构 Hive Schema Tool
逆向移植	HIVE-19701：getDelegationTokenFromMetaStore 无需同步
逆向移植	HIVE-19694：若要创建具体化视图语句，应在运行 MV 的 SQL 语句之前检查 MV 名称是否冲突。
逆向移植	HIVE-19674: 按十进制常量分组下推到 Druid 表
逆向移植	HIVE-19668: 重复的 org.antlr.runtime.CommonToken 和重复的字符串浪费了超过 30% 的堆
逆向移植	HIVE-19663: 重构 LLAP IO 报告生成
逆向移植	HIVE-19661：切换 Hive UDF 以使用 Re2J 正则表达式引擎
逆向移植	HIVE-19628：LLAP testSigning 中可能出现 NPE
逆向移植	HIVE-19568：主动/被动 HS2 HA：禁止直接连接到被动 HS2 实例
逆向移植	HIVE-19564：向量化：修复 Arithmetic 中的 NULL/错误结果问题
逆向移植	HIVE-19552：启用 TestMiniDruidKafkaCliDriver#druidkafkamini_basic.q
逆向移植	HIVE-19432：如果 hive 的数据库和表过多，GetTablesOperation 过慢

类型	描述
逆向移植	HIVE-19360: CBO：在 QueryPlan 对象中添加 "optimizedSQL"
逆向移植	HIVE-19326: 统计数据自动收集：UNION 查询期间聚合不正确
逆向移植	HIVE-19313：TestJdbcWithDBTokenStoreNoDoAs 测试失败
逆向移植	HIVE-19285: 将日志添加到 MetaDataOperation 的子类中
逆向移植	HIVE-19235: 更新 Minimr 测试的黄金文件
逆向移植	HIVE-19104：当测试 MetaStore 以重试启动时，实例是独立的
逆向移植	HIVE-18986：如果表包含大量列，表重命名会在 dataNucleus 中运行 java.lang.StackOverflowError
逆向移植	HIVE-18920：CBO：在初次查询之前初始化 Janino 提供程序
逆向移植	HIVE-18873：在 HiveInputFormat 处以静默方式跳过 MR 的谓词下推，可能导致存储处理程序产生错误结果
逆向移植	HIVE-18871：由于将 hive.aux.jars.path 设置为 hdfs:// 而导致 hive on tez 执行错误
逆向移植	HIVE-18725: 如果存在错误的列引用，则改进子查询的错误处理
逆向移植	HIVE-18696: 在 HiveMetaStore.add_partitions_core 方法中可能无法正确清理分区文件夹（如果存在）

类型	描述
逆向移植	HIVE-18453: ACID：添加 "CREATE TRANSACTIONAL TABLE" 语法以统一 ACID ORC 和 Parquet 支持
逆向移植	HIVE-18201: 为 sq_count_chec 禁用 XPROD_EDGE
逆向移植	HIVE-18140: 在基本统计数据混合大小写中，分区表的统计数据可能会出错
逆向移植	HIVE-17921: 在 LLAP 中使用 struct 进行聚合会产生错误的结果
逆向移植	HIVE-17896: TopNKey：创建独立的可矢量化的 TopNKey 运算符
逆向移植	HIVE-17840：如果 transactionalListeners.notifyEvent 失败，HiveMetaStore 会存入异常
逆向移植	HIVE-17043：如果日后不引用，则密钥会从分组中删除非唯一列
逆向移植	HIVE-17040：在具有 FK 关系的情况下，实现连接消除
逆向移植	HIVE-16839：如果同时更改同一个分区，openTransaction/commitTransaction 调用会失衡
逆向移植	HIVE-16100：动态排序分区优化器丢失同级运算符
逆向移植	HIVE-15956：删除大量分区时出现 StackOverflowError

类型	描述
逆向移植	HIVE-15177：当 kerberos 身份验证类型设置为 fromSubject 并且主体包含 _HOST 时，使用 hive 进行身份验证失败
逆向移植	HIVE-14898：若出现空授权标头错误，HS2 不会记录调用堆栈
逆向移植	HIVE-14493: 具体化视图的分区支持
逆向移植	HIVE-14431: 将 COALESCE 识别为 CASE
逆向移植	HIVE-13457：创建 HS2 REST API 端点以监控信息
逆向移植	HIVE-12342: 将 hive.optimize.index.filter 的默认值设置为 true
逆向移植	HIVE-10296：当 hive 在元存储上运行多连接查询时，发现强制转换异常
逆向移植	HIVE-6980：使用 direct sql 删除表

Amazon EMR 6.6.0 – Hive 配置更改

- 作为 OSS 更改 HIVE-20703 的一部分，用于对动态分区进行排序的属性 `hive.optimize.sort.dynamic.partition` 已替换为 `hive.optimize.sort.dynamic.partition.threshold`。

  `hive.optimize.sort.dynamic.partition.threshold` 配置具有以下潜在值：

Value	描述
`0`（默认值）	使用 ORC 文件时，对动态分区进行排序的优化将作为基于成本的决策。INSERT 查询中允许的最大写入器数根据（执行程序/容器内存）*（ORC 占用的内存百分比）除以单个写入器占用的最大内存（条带大小）计算得出。

Value	描述
-1	禁用优化以对动态分区进行完全排序。
1	对动态分区启用全局排序。这样可以使 Reducer 中的每个分区值只打开一个记录写入器，从而减小 Reducer 的内存压力。
2 （或更大的整数）	告知 Hive 使用指定的整数作为最大写入器数的阈值。

Amazon EMR 6.6.0 – Hive 已知问题

- 如果使用窗口函数对与交集函数相同的列进行查询，则可能会导致如 HIVE-25278 中报告的转换无效问题，并导致查询结果不正确或查询失败。解决方法是在查询级别为此类查询禁用 CBO。如需进一步的信息，请联系 Amazon Support。
- Amazon EMR 6.6.0 包含 Hive 软件版本 3.1.2。Hive 3.1.2 引入了一项功能，如果文本文件包含页眉和页脚，则可将其拆分（HIVE-21924）。Apache Tez App Master 读取您的每个文件以确定数据范围内的偏移点。如果您的查询读取大量小型文本文件，这些行为综合起来可能会影响性能。解决方法是使用 `CombineHiveInputFormat` 并通过配置以下属性调整最大拆分大小：

```
SET hive.tez.input.format=org.apache.hadoop.hive.ql.io.CombineHiveInputFormat;
SET mapreduce.input.fileinputformat.split.maxsize=16777216;
```

- 在 6.6.0 到 6.9.x 版 Amazon EMR 中，带有动态分区和 ORDER BY 或 SORT BY 子句的 INSERT 查询将始终具有两个 Reducer。此问题是由于 OSS 更改 HIVE-20703 所致，该更改将动态排序分区优化置于基于成本的决策之下。如果您的工作负载不需要对动态分区进行排序，建议将 `hive.optimize.sort.dynamic.partition.threshold` 属性设置为 `-1` 以禁用新功能并获得计算正确的 Reducer 数量。此问题已作为 HIVE-22269 的一部分在 OSS Hive 中修复，并已在 Amazon EMR 6.10.0 中修复。

Hudi

Apache Hudi 是一种开源数据管理框架，用于通过提供记录级插入、更新、更新插入和删除功能来简化增量数据处理和数据管道开发工作。更新插入指的是将记录插入到现有数据集中（如果它们不存在）或对数据集进行更新（如果它们存在）的功能。通过高效地管理数据在 Amazon S3 中的布局方式，Hudi 允许近乎实时地摄取和更新数据。Hudi 仔细维护对数据集执行的操作的元数据，以帮助确保操作是原子级且是一致的。

Hudi 集成了 Apache Spark、Apache Hive 和 Presto。在 Amazon EMR 发行版 6.1.0 及更高版本中，Hudi 还与 Trino（PrestoSQL）集成。

在 Amazon EMR 5.28.0 版本及更高版本中，EMR 默认情况下会在安装 Spark、Hive、Presto 或 Flink 时安装 Hudi 组件。您可以使用 Spark 或 Hudi DeltaStreamer 实用程序来创建或更新 Hudi 数据集。您可以使用 Hive、Spark、Presto 或 Flink 以交互方式查询 Hudi 数据集，或使用增量拉取功能构建数据处理管道。增量拉取是指仅拉取两个操作之间更改的数据的功能。

这些功能使得 Hudi 适用于以下使用案例：

- 处理来自传感器和其它需要特定数据插入和更新事件的物联网 (IoT) 设备的流数据。
- 在用户可能会选择被忘记或修改其对数据使用方式的同意的应用程序中，遵守数据隐私法规。
- 实施更改数据捕获 (CDC) 系统，该系统允许您随着时间的推移将更改应用于数据集。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Hudi 版本，以及 Amazon EMR 随 Hudi 一起安装的组件。

有关此发行版中随 Hudi 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Hudi 版本信息

Amazon EMR 发行版标签	Hudi 版本	随 Hudi 安装的组件
emr-6.14.0	Hudi 0.13.1-amzn-2	Not available.

> **Note**
>
> Amazon EMR 发行版 6.8.0 随附 Apache Hudi 0.11.1；但是，Amazon EMR 6.8.0 集群也与 Hudi 0.12.0 中的开源 `hudi-spark3.3-bundle_2.12` 兼容。

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Hudi 版本，以及 Amazon EMR 随 Hudi 一起安装的组件。

有关此发行版中随 Hudi 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Hudi 版本信息

Amazon EMR 发行版标签	Hudi 版本	随 Hudi 安装的组件
emr-5.36.1	Hudi 0.10.1-amzn-1	Not available.

主题

Hudi 的工作原理

当将 Hudi 与 Amazon EMR 搭配使用时，您可以使用 Spark Data Source API 或 Hudi DeltaStreamer 实用程序将数据写入数据集。Hudi 将数据集整理到 *basepath* 下一个分区的目录结构中，类似于传统的 Hive 表。如何将数据布局为这些目录中的文件的具体细节取决于您选择的数据集类型。您可以选择“写入时复制 (CoW)”或“读取时合并 (MOM)”。

无论数据集类型如何，数据集中的每个分区都由其相对于 basepath 的 partitionpath 唯一标识。在每个分区中，记录分布到多个数据文件中。有关更多信息，请参阅 Apache Hudi 文档中的文件管理。

Hudi 中的每个操作都有一个相应的提交，由一个单调递增的时间戳标识，称为 Instant。Hudi 将对数据集执行的一系列操作保留为时间轴。Hudi 依靠此时间轴提供读取器和写入器之间的快照隔离，并支持回滚到前一个时间点。有关 Hudi 记录的操作和操作状态的更多信息，请参阅 Apache Hudi 文档中的 Timeline。

了解数据集存储类型：写入时复制与读取时合并

创建 Hudi 数据集时，可以指定数据集在写入时复制或读取时合并。

- 写入时复制（CoW）– 数据以列状格式存储（Parquet），并且每次更新都会在写入过程中创建一个新版本的文件。CoW 是默认存储类型。
- 读取时合并（MOR）– 数据使用列式（Parquet）和基于行（Avro）的格式的组合进行存储。更新记录到基于行的增量文件中，并根据需要进行压缩以创建新版本的列式文件。

对于 CoW 数据集，每次更新记录时，包含该记录的文件都会使用更新后的值进行重写。对于 MoR 数据集，每次进行更新时，Hudi 仅写入已更改记录对应的行。MoR 更适合写入或更改繁重而读取量较少的工作负载。CoW 更适合更改频率较低但读取量繁重的工作负载。

Hudi 为访问数据提供三个逻辑视图：

- 读取优化视图 – 提供来自 CoW 表的最新提交数据集和来自 MOR 表的最新压缩数据集。
- 增量视图 – 提供 CoW 数据集中两个操作之间的更改流，以馈送给下游作业和提取、转换、加载（ETL）工作流。
- 实时视图 – 通过内联合并列式和基于行的文件，从 MOR 表中提供最新提交的数据。

当您查询读取优化的视图时，查询将返回所有压缩数据，但不包括最新的增量提交。查询此数据可提供良好的读取性能，但忽略最新的数据。当您查询实时视图时，Hudi 会在读取时将压缩的数据与增量提交合并。最新的数据可用于查询，但合并的计算开销使查询性能降低。通过查询压缩数据或实时数据的功能，您可以在查询时在性能和灵活性之间进行选择。

有关在存储类型之间权衡的更多信息，请参阅 Apache Hudi 文档中的存储类型和视图。

在 MoR 的 Hive 元数据仓中创建两个表：一个具有您指定的名称的表（即读取优化视图）和一个附加了 `_rt` 的同名表（即实时视图）。您可以查询这两个表。

将 Hudi 数据集注册到您的元数据仓

当您向 Hive 元数据仓注册 Hudi 表时，您可以像对待任何其它表一样，使用 Hive、Spark SQL 或 Presto 查询 Hudi 表。此外，您可以通过将 Hive 和 Spark 配置为使用 Amazon Glue 数据目录作为元数据仓来将 Hudi 与 Amazon Glue 进行集成。对于 MoR 表，Hudi 将数据集注册为元数据仓中的两个表：一个具有您指定的名称的表（即读取优化视图）和一个附加了 `_rt` 的同名表（即实时视图）。

当您使用 Spark 创建 Hudi 数据集时，您可以通过将 HIVE_SYNC_ENABLED_OPT_KEY 选项设置为 "true" 并提供其它必需的属性来向 Hive 元数据仓注册 Hudi 表。有关更多信息，请参阅使用 Hudi 数据集。此外，您可以使用 hive_sync_tool 命令行实用程序将 Hudi 数据集单独注册为元数据仓中的表。

在 Amazon EMR 上使用 Hudi 的注意事项和限制

- 记录键字段不能为 Null 或空 – 您指定为记录键字段的字段不能具有 null 或空值。
- 默认情况下在更新插入和插入时更新架构 – Hudi 提供一个接口 HoodieRecordPayload，用于确定如何合并输入 DataFrame 和现有 Hudi 数据集以生成新的更新数据集。Hudi 提供类 OverwriteWithLatestAvroPayload 的默认实现，它会覆盖现有记录并更新在输入 DataFrame 中指定的架构。要自定义此逻辑以实现合并和部分更新，您可以使用 DataSourceWriteOptions.PAYLOAD_CLASS_OPT_KEY 参数提供 HoodieRecordPayload 接口的实现。
- 删除需要架构 – 删除时，必须指定记录键、分区键和预组合键字段。其它列可以成为 null 或空，但需要完整的架构。
- MoR 表限制 – MoR 表不支持保存点。您可以使用来自 Spark SQL、Presto 或 Hive 的读取优化视图或实时视图（*tableName*_rt）。使用读取优化视图仅公开基本文件数据，不会公开基本数据和日志数据的合并视图。
- Hive
  - 要在 Hive 元数据仓中注册表，Hudi 需要 Hive Thrift 服务器在默认端口 10000 上运行。如果使用自定义端口覆盖此端口，请传递 HIVE_URL_OPT_KEY 选项，如以下示例所示。

    ```
    .option(DataSourceWriteOptions.HIVE_URL_OPT_KEY, "jdbc:hive2://localhost:override-port-number
    ```

  - Spark 中的 timestamp 数据类型在 Hive 中注册为 long 数据类型，而不注册为 Hive 的 timestamp 类型。
- Presto
  - 在版本低于 0.6.0 的 Hudi 中，Presto 不支持读取 MoR 实时表。
  - Presto 仅支持快照查询。
  - 要使 Presto 正确解释 Hudi 数据集列，请将 hive.parquet_use_column_names 值设置为 true。
    - 要设置会话的值，请在 Presto shell 中运行以下命令：

      ```
      set session hive.parquet_use_column_names=true
      ```

- 要在集群级别设置值，请使用 `presto-connector-hive` 配置分类将 `hive.parquet.use_column_names` 设置为 `true`，如以下示例所示。有关更多信息，请参阅配置应用程序。

```
[
  {
    "Classification": "presto-connector-hive",
    "Properties": {
      "hive.parquet.use-column-names": "true"
    }
  }
]
```

- HBase 索引
  - 用于构建 Hudi 的 HBase 版本可能与 EMR 发行指南中列出的内容有所不同。要为 Spark 会话提取正确的依赖项，请运行以下命令。

```
spark-shell \
--jars /usr/lib/spark/external/lib/spark-avro.jar,/usr/lib/hudi/cli/lib/*.jar \
--conf "spark.serializer=org.apache.spark.serializer.KryoSerializer" \
--conf "spark.sql.hive.convertMetastoreParquet=false"
```

创建安装了 Hudi 的集群

在 Amazon EMR 版本 5.28.0 及更高版本中， Amazon EMR 默认情况下会在安装 Spark、Hive 或 Presto 时安装 Hudi 组件。要在 Amazon EMR 上使用 Hudi，请在安装了以下一个或多个应用程序后创建集群：

- Hadoop
- Hive
- Spark
- Presto
- Flink

您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 创建集群。

使用 Amazon Web Services Management Console 创建包含 Hudi 的集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 在 Software Configuration（软件配置）下，对于 Release（发行版），选择 emr-5.28.0 或更高版本，然后选择 Hadoop、Hive、Spark、Presto、Tez 以及集群需要的其他应用程序。
4. 根据应用程序的需要配置其它选项，然后选择下一步。
5. 根据需要配置 Hardware (硬件) 和 General cluster settings (常规集群设置) 的选项。
6. 对于 Security Options (安全选项)，我们建议您选择一个 EC2 key pair (EC2 密钥对)，您可以使用它通过 SSH 连接到主节点命令行。这允许您运行本指南中描述的 Spark shell 命令、Hive CLI 命令和 Hudi CLI 命令。
7. 根据需要选择其它安全选项，然后选择 Create cluster (创建集群)。

使用 Hudi 数据集

Hudi 支持通过 Spark 在 Hudi 数据集中插入、更新和删除数据。有关更多信息，请参阅 Apache Hudi 文档中的写入 Hudi 表格。

以下示例演示如何启动交互式 Spark Shell、使用 Spark 提交，或如何使用 Amazon EMR Notebooks 在 Amazon EMR 上使用 Hudi。您也可以使用 Hudi DeltaStreamer 实用程序或其它工具来写入数据集。在本节中，示例演示使用 Spark shell 处理数据集，同时使用 SSH 作为默认 `hadoop` 用户连接到主节点。

使用 Amazon EMR 6.7 及更高版本启动 Spark Shell

运行 `spark-shell`、`spark-submit` 或 `spark-sql` 使用 Amazon EMR 6.7.0 或更高版本时，传递以下命令。

> **Note**
>
> Amazon EMR 6.7.0 使用 Apache Hudi 0.11.0-amzn-0，相比于之前的 Hudi 版本有明显改进。有关更多信息，请参阅 Apache Hudi 0.11.0 Migration Guide（《Apache Hudi 0.11.0 迁移指南》）。此选项卡上的示例反映了这些更改。

在主节点上打开 Spark Shell

1. 使用 SSH 连接到主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以启动 Spark shell。要使用 PySpark Shell，请将 *spark-shell* 替换为 *pyspark*。

```
spark-shell --jars /usr/lib/hudi/hudi-spark-bundle.jar \
--conf "spark.serializer=org.apache.spark.serializer.KryoSerializer" \
--conf
 "spark.sql.catalog.spark_catalog=org.apache.spark.sql.hudi.catalog.HoodieCatalog"
 \
--conf "spark.sql.extensions=org.apache.spark.sql.hudi.HoodieSparkSessionExtension"
```

使用 Amazon EMR 6.6 及更早版本启动 Spark Shell

运行 `spark-shell`、`spark-submit` 或 `spark-sql` 使用 Amazon EMR 6.6.x 或更早版本时，传递以下命令。

> **Note**
>
> - Amazon EMR 6.2 和 5.31 及更高版本（Hudi 0.6.x 及更高版本）可以在配置中省略 `spark-avro.jar`。
> - Amazon EMR 6.5 和 5.35 及更高版本（Hudi 0.9.x 及更高版本）可以从配置中省略 `spark.sql.hive.convertMetastoreParquet=false`。
> - Amazon EMR 6.6 和 5.36 及更高版本（Hudi 0.10.x 及更高版本）必须包含 Version: 0.10.0 Spark Guide（《版本：0.10.0 Spark 指南》）中所述的 `HoodieSparkSessionExtension` 配置：
>
> ```
> --conf
>  "spark.sql.extensions=org.apache.spark.sql.hudi.HoodieSparkSessionExtension"
>  \
> ```

在主节点上打开 Spark Shell

1. 使用 SSH 连接到主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以启动 Spark shell。要使用 PySpark Shell，请将 *spark-shell* 替换为 *pyspark*。

```
spark-shell \
--conf "spark.serializer=org.apache.spark.serializer.KryoSerializer" \
--conf "spark.sql.hive.convertMetastoreParquet=false" \
--jars /usr/lib/hudi/hudi-spark-bundle.jar,/usr/lib/spark/external/lib/spark-
avro.jar
```

使用 Amazon EMR 6.7 及更高版本将 Hudi 与 Amazon EMR Notebooks 结合使用

要将 Hudi 与 Amazon EMR Notebooks 结合使用，您必须首先将 Hudi jar 文件从本地文件系统复制到笔记本集群的主节点 上的 HDFS。然后，您可以使用笔记本编辑器来配置 EMR Notebook 以使用 Hudi。

将 Hudi 与 Amazon EMR Notebooks 搭配使用

1. 为 Amazon EMR Notebooks 创建并启动集群。有关更多信息，请参阅《Amazon EMR 管理指南》中的为笔记本创建 Amazon EMR 集群。
2. 使用 SSH 连接到集群的主节点，然后将 jar 文件从本地文件系统复制到 HDFS，如以下示例所示。在此示例中，我们在 HDFS 中创建了一个目录，以便清晰地管理文件。如果需要，您可以在 HDFS 中选择自己的目的地。

```
hdfs dfs -mkdir -p /apps/hudi/lib
```

```
hdfs dfs -copyFromLocal /usr/lib/hudi/hudi-spark-bundle.jar /apps/hudi/lib/hudi-
spark-bundle.jar
```

3. 打开笔记本编辑器，输入以下示例中的代码，然后运行它。

```
%%configure
{ "conf": {
            "spark.jars":"hdfs:///apps/hudi/lib/hudi-spark-bundle.jar",
            "spark.serializer":"org.apache.spark.serializer.KryoSerializer",
```

```
            "spark.sql.catalog.spark_catalog":
 "org.apache.spark.sql.hudi.catalog.HoodieCatalog",

 "spark.sql.extensions":"org.apache.spark.sql.hudi.HoodieSparkSessionExtension"
          }}
```

使用 Amazon EMR 6.6 及更早版本将 Hudi 与 Amazon EMR Notebooks 结合使用

要将 Hudi 与 Amazon EMR Notebooks 结合使用，您必须首先将 Hudi jar 文件从本地文件系统复制到笔记本集群的主节点 上的 HDFS。然后，您可以使用笔记本编辑器来配置 EMR Notebook 以使用 Hudi。

将 Hudi 与 Amazon EMR Notebooks 搭配使用

1. 为 Amazon EMR Notebooks 创建并启动集群。有关更多信息，请参阅《Amazon EMR 管理指南》中的为笔记本创建 Amazon EMR 集群。
2. 使用 SSH 连接到集群的主节点，然后将 jar 文件从本地文件系统复制到 HDFS，如以下示例所示。在此示例中，我们在 HDFS 中创建了一个目录，以便清晰地管理文件。如果需要，您可以在 HDFS 中选择自己的目的地。

   ```
   hdfs dfs -mkdir -p /apps/hudi/lib
   ```

   ```
   hdfs dfs -copyFromLocal /usr/lib/hudi/hudi-spark-bundle.jar /apps/hudi/lib/hudi-
   spark-bundle.jar
   ```

   ```
   hdfs dfs -copyFromLocal /usr/lib/spark/external/lib/spark-avro.jar /apps/hudi/lib/
   spark-avro.jar
   ```

3. 打开笔记本编辑器，输入以下示例中的代码，然后运行它。

   ```
   { "conf": {
              "spark.jars":"hdfs:///apps/hudi/lib/hudi-spark-bundle.jar,hdfs:///apps/
   hudi/lib/spark-avro.jar",
              "spark.serializer":"org.apache.spark.serializer.KryoSerializer",
              "spark.sql.hive.convertMetastoreParquet":"false"
            }}
   ```

初始化 Hudi 的 Spark 会话

使用 Scala 时，您必须在 Spark 会话中导入以下类。这需要在每个 Spark 会话中完成一次。

```
import org.apache.spark.sql.SaveMode
import org.apache.spark.sql.functions._
import org.apache.hudi.DataSourceWriteOptions
import org.apache.hudi.DataSourceReadOptions
import org.apache.hudi.config.HoodieWriteConfig
import org.apache.hudi.hive.MultiPartKeysValueExtractor
import org.apache.hudi.hive.HiveSyncConfig
import org.apache.hudi.sync.common.HoodieSyncConfig
```

写入 Hudi 数据集

以下示例演示如何创建 DataFrame 并将其作为 Hudi 数据集写入。

> **Note**
>
> 要将代码示例粘贴到 Spark shell 中，请在提示符处键入 **`:paste`**，粘贴示例，然后按 **`CTRL + D`**。

每次向 Hudi 数据集写入 DataFrame 时，都必须指定 `DataSourceWriteOptions`。这些选项中的许多选项在写入操作之间可能是相同的。以下示例使用 *`hudiOptions`* 变量指定常用选项，随后的示例使用这些选项。

使用 Amazon EMR 6.7 及更高版本的 Scala 进行写入

> **Note**
>
> Amazon EMR 6.7.0 使用 Apache Hudi 0.11.0-amzn-0，相比于之前的 Hudi 版本有明显改进。有关更多信息，请参阅 Apache Hudi 0.11.0 Migration Guide（《Apache Hudi 0.11.0 迁移指南》）。此选项卡上的示例反映了这些更改。

```
// Create a DataFrame
val inputDF = Seq(
 ("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
```

```
 ("101", "2015-01-01", "2015-01-01T12:14:58.597216Z"),
 ("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
 ("103", "2015-01-01", "2015-01-01T13:51:40.519832Z"),
 ("104", "2015-01-02", "2015-01-01T12:15:00.512679Z"),
 ("105", "2015-01-02", "2015-01-01T13:51:42.248818Z")
 ).toDF("id", "creation_date", "last_update_time")

//Specify common DataSourceWriteOptions in the single hudiOptions variable
val hudiOptions = Map[String,String](
  HoodieWriteConfig.TBL_NAME.key -> "tableName",
  DataSourceWriteOptions.TABLE_TYPE.key -> "COPY_ON_WRITE",
  DataSourceWriteOptions.RECORDKEY_FIELD_OPT_KEY -> "id",
  DataSourceWriteOptions.PARTITIONPATH_FIELD_OPT_KEY -> "creation_date",
  DataSourceWriteOptions.PRECOMBINE_FIELD_OPT_KEY -> "last_update_time",
  DataSourceWriteOptions.HIVE_SYNC_ENABLED_OPT_KEY -> "true",
  DataSourceWriteOptions.HIVE_TABLE_OPT_KEY -> "tableName",
  DataSourceWriteOptions.HIVE_PARTITION_FIELDS_OPT_KEY -> "creation_date",
  HoodieSyncConfig.META_SYNC_PARTITION_EXTRACTOR_CLASS.key ->
 "org.apache.hudi.hive.MultiPartKeysValueExtractor",
  HoodieSyncConfig.META_SYNC_ENABLED.key -> "true",
  HiveSyncConfig.HIVE_SYNC_MODE.key -> "hms",
  HoodieSyncConfig.META_SYNC_TABLE_NAME.key -> "tableName",
  HoodieSyncConfig.META_SYNC_PARTITION_FIELDS.key -> "creation_date"
)

// Write the DataFrame as a Hudi dataset
(inputDF.write
    .format("hudi")
    .options(hudiOptions)
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY,"insert")
    .mode(SaveMode.Overwrite)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 Amazon EMR 6.6 及更早版本的 Scala 进行写入

```
// Create a DataFrame
val inputDF = Seq(
 ("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
 ("101", "2015-01-01", "2015-01-01T12:14:58.597216Z"),
 ("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
 ("103", "2015-01-01", "2015-01-01T13:51:40.519832Z"),
 ("104", "2015-01-02", "2015-01-01T12:15:00.512679Z"),
 ("105", "2015-01-02", "2015-01-01T13:51:42.248818Z")
```

```
 ).toDF("id", "creation_date", "last_update_time")

//Specify common DataSourceWriteOptions in the single hudiOptions variable
val hudiOptions = Map[String,String](
  HoodieWriteConfig.TABLE_NAME -> "tableName",
  DataSourceWriteOptions.TABLE_TYPE_OPT_KEY -> "COPY_ON_WRITE",
  DataSourceWriteOptions.RECORDKEY_FIELD_OPT_KEY -> "id",
  DataSourceWriteOptions.PARTITIONPATH_FIELD_OPT_KEY -> "creation_date",
  DataSourceWriteOptions.PRECOMBINE_FIELD_OPT_KEY -> "last_update_time",
  DataSourceWriteOptions.HIVE_SYNC_ENABLED_OPT_KEY -> "true",
  DataSourceWriteOptions.HIVE_TABLE_OPT_KEY -> "tableName",
  DataSourceWriteOptions.HIVE_PARTITION_FIELDS_OPT_KEY -> "creation_date",
  DataSourceWriteOptions.HIVE_PARTITION_EXTRACTOR_CLASS_OPT_KEY ->
 classOf[MultiPartKeysValueExtractor].getName
)

// Write the DataFrame as a Hudi dataset
(inputDF.write
    .format("org.apache.hudi")
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY,
 DataSourceWriteOptions.INSERT_OPERATION_OPT_VAL)
    .options(hudiOptions)
    .mode(SaveMode.Overwrite)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 PySpark 进行写入

```
# Create a DataFrame
inputDF = spark.createDataFrame(
    [
        ("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
        ("101", "2015-01-01", "2015-01-01T12:14:58.597216Z"),
        ("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
        ("103", "2015-01-01", "2015-01-01T13:51:40.519832Z"),
        ("104", "2015-01-02", "2015-01-01T12:15:00.512679Z"),
        ("105", "2015-01-02", "2015-01-01T13:51:42.248818Z"),
    ],
    ["id", "creation_date", "last_update_time"]
)

# Specify common DataSourceWriteOptions in the single hudiOptions variable
hudiOptions = {
'hoodie.table.name': 'tableName',
```

```
'hoodie.datasource.write.recordkey.field': 'id',
'hoodie.datasource.write.partitionpath.field': 'creation_date',
'hoodie.datasource.write.precombine.field': 'last_update_time',
'hoodie.datasource.hive_sync.enable': 'true',
'hoodie.datasource.hive_sync.table': 'tableName',
'hoodie.datasource.hive_sync.partition_fields': 'creation_date',
'hoodie.datasource.hive_sync.partition_extractor_class':
 'org.apache.hudi.hive.MultiPartKeysValueExtractor'
}

# Write a DataFrame as a Hudi dataset
inputDF.write \
.format('org.apache.hudi') \
.option('hoodie.datasource.write.operation', 'insert') \
.options(**hudiOptions) \
.mode('overwrite') \
.save('s3://DOC-EXAMPLE-BUCKET/myhudidataset/')
```

> **Note**
>
> 您可能会在代码示例和通知中看到“hoodie”而不是 Hudi。Hudi 代码库广泛使用旧的“hoodie”拼写。

Hudi 的 DataSourceWriteOptions 引用

选项	描述
TABLE_NAME	要在其中注册数据集的表名称。
TABLE_TYPE_OPT_KEY	可选。指定数据集是创建为 "COPY_ON_WRITE" 还是 "MERGE_ON_READ" 。默认为 "COPY_ON_WRITE" 。
RECORDKEY_FIELD_OPT_KEY	其值将用作 HoodieKey 的 recordKey 组件的记录键字段。实际值将通过对字段值调用 .toString() 来获得。可使用点表示法指定嵌套字段，例如 a.b.c。

选项	描述
PARTITIONPATH_FIELD_OPT_KEY	其值将用作 HoodieKey 的 partition Path 组件的分区路径字段。实际值将通过对字段值调用 .toString() 来获得。
PRECOMBINE_FIELD_OPT_KEY	在实际写入之前在预合并中使用的字段。如果两个记录具有相同的键值，Hudi 为预合并选择字段值最大的记录（由 Object.compareTo(..) 确定）。

仅在元数据仓中注册 Hudi 数据集表时才需要以下选项。如果您未将 Hudi 数据集注册为 Hive 元数据仓中的表，则不需要这些选项。

Hive 的 DataSourceWriteOptions 引用

选项	描述
HIVE_DATABASE_OPT_KEY	要同步到的 Hive 数据库。默认为 "default"。
HIVE_PARTITION_EXTRACTOR_CLASS_OPT_KEY	用于将分区字段值提取到 Hive 分区列中的类。
HIVE_PARTITION_FIELDS_OPT_KEY	数据集中用于确定 Hive 分区列的字段。
HIVE_SYNC_ENABLED_OPT_KEY	设置为 "true" 时，将向 Apache Hive 元数据仓注册数据集。默认为 "false"。
HIVE_TABLE_OPT_KEY	必需。Hive 中要同步到的表的名称。例如，"my_hudi_table_cow"。
HIVE_USER_OPT_KEY	可选。同步时要使用的 Hive 用户名。例如，"hadoop"。
HIVE_PASS_OPT_KEY	可选。由 HIVE_USER_OPT_KEY 指定的用户的 Hive 密码。
HIVE_URL_OPT_KEY	Hive 元数据仓 URL。

更新插入数据

以下示例演示如何通过编写 DataFrame 来更新插入数据。与之前的插入示例不同，`OPERATION_OPT_KEY` 值设置为 `UPSERT_OPERATION_OPT_VAL`。此外，还指定 `.mode(SaveMode.Append)` 以指示应追加记录。

使用 Amazon EMR 6.7 及更高版本的 Scala 进行更新插入

> **Note**
>
> Amazon EMR 6.7.0 使用 Apache Hudi 0.11.0-amzn-0，相比于之前的 Hudi 版本有明显改进。有关更多信息，请参阅 Apache Hudi 0.11.0 Migration Guide（《Apache Hudi 0.11.0 迁移指南》）。此选项卡上的示例反映了这些更改。

```
// Create a new DataFrame from the first row of inputDF with a different creation_date
 value
val updateDF = inputDF.limit(1).withColumn("creation_date", lit("new_value"))

(updateDF.write
    .format("hudi")
    .options(hudiOptions)
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY, "upsert")
    .mode(SaveMode.Append)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 Amazon EMR 6.6 及更早版本的 Scala 进行更新插入

```
// Create a new DataFrame from the first row of inputDF with a different creation_date
 value
val updateDF = inputDF.limit(1).withColumn("creation_date", lit("new_value"))

(updateDF.write
    .format("org.apache.hudi")
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY,
 DataSourceWriteOptions.UPSERT_OPERATION_OPT_VAL)
    .options(hudiOptions)
    .mode(SaveMode.Append)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 PySpark 进行更新插入

```
from pyspark.sql.functions import lit

# Create a new DataFrame from the first row of inputDF with a different creation_date
 value
updateDF = inputDF.limit(1).withColumn('creation_date', lit('new_value'))

updateDF.write \
    .format('org.apache.hudi') \
    .option('hoodie.datasource.write.operation', 'upsert') \
    .options(**hudiOptions) \
    .mode('append') \
    .save('s3://DOC-EXAMPLE-BUCKET/myhudidataset/')
```

删除记录

要硬删除记录，您可以更新插入一个空的负载。在这种情况下，PAYLOAD_CLASS_OPT_KEY 选项指定 EmptyHoodieRecordPayload 类。该示例使用更新插入示例中使用的相同 DataFrame updateDF，以便指定相同的记录。

使用 Amazon EMR 6.7 及更高版本的 Scala 进行删除

> **Note**
>
> Amazon EMR 6.7.0 使用 Apache Hudi 0.11.0-amzn-0，相比于之前的 Hudi 版本有明显改进。有关更多信息，请参阅 Apache Hudi 0.11.0 Migration Guide（《Apache Hudi 0.11.0 迁移指南》）。此选项卡上的示例反映了这些更改。

```
(updateDF.write
    .format("hudi")
    .options(hudiOptions)
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY, "delete")
    .mode(SaveMode.Append)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 Amazon EMR 6.6 及更早版本的 Scala 进行删除

```
(updateDF.write
```

```
    .format("org.apache.hudi")
    .option(DataSourceWriteOptions.OPERATION_OPT_KEY,
 DataSourceWriteOptions.UPSERT_OPERATION_OPT_VAL)
    .option(DataSourceWriteOptions.PAYLOAD_CLASS_OPT_KEY,
 "org.apache.hudi.common.model.EmptyHoodieRecordPayload")
    .mode(SaveMode.Append)
    .save("s3://DOC-EXAMPLE-BUCKET/myhudidataset/"))
```

使用 PySpark 进行删除

```
updateDF.write \
    .format('org.apache.hudi') \
    .option('hoodie.datasource.write.operation', 'upsert') \
    .option('hoodie.datasource.write.payload.class',
 'org.apache.hudi.common.model.EmptyHoodieRecordPayload') \
    .options(**hudiOptions) \
    .mode('append') \
    .save('s3://DOC-EXAMPLE-BUCKET/myhudidataset/')
```

您还可以通过以下方式硬删除数据：将 OPERATION_OPT_KEY 设置为 DELETE_OPERATION_OPT_VAL 来删除您提交的数据集中的所有记录。有关执行软删除的说明，以及有关删除 Hudi 表中存储的数据的详细信息，请参阅 Apache Hudi 文档中的 Deletes。

从 Hudi 数据集读取

要在当前时间点检索数据，Hudi 默认情况下执行快照查询。以下是查询在 写入 Hudi 数据集 中写入 S3 的数据集的示例。将 *s3://DOC-EXAMPLE-BUCKET/myhudidataset* 替换为您的表的路径，并为每个分区级别添加通配符星号，外加一个额外的星号。在此示例中，有一个分区级别，因此我们添加了两个通配符号。

使用 Amazon EMR 6.7 及更高版本的 Scala 进行读取

> **Note**
>
> Amazon EMR 6.7.0 使用 Apache Hudi 0.11.0-amzn-0，相比于之前的 Hudi 版本有明显改进。有关更多信息，请参阅 Apache Hudi 0.11.0 Migration Guide（《Apache Hudi 0.11.0 迁移指南》）。此选项卡上的示例反映了这些更改。

```
(val snapshotQueryDF = spark.read
```

```
    .format("hudi")
    .load(s3://DOC-EXAMPLE-BUCKET/myhudidataset)
    .show()
```

使用 Amazon EMR 6.6 及更早版本的 Scala 进行读取

```
(val snapshotQueryDF = spark.read
    .format("org.apache.hudi")
    .load("s3://DOC-EXAMPLE-BUCKET/myhudidataset" + "/*/*"))

snapshotQueryDF.show()
```

使用 PySpark 进行读取

```
snapshotQueryDF = spark.read \
    .format('org.apache.hudi') \
    .load('s3://DOC-EXAMPLE-BUCKET/myhudidataset' + '/*/*')

snapshotQueryDF.show()
```

递增查询

您还可以使用 Hudi 执行增量查询，以获取自给定提交时间戳以来已更改的记录流。为此，请将 QUERY_TYPE_OPT_KEY 字段设置为 QUERY_TYPE_INCREMENTAL_OPT_VAL。然后，为 BEGIN_INSTANTTIME_OPT_KEY 添加一个值，以获取自指定时间以来写入的所有记录。递增查询的效率通常是批处理查询的十倍，因为它们只处理更改的记录。

执行增量查询时，请使用根（基）表路径，而不需要用于快照查询的通配符星号。

> **Note**
>
> Presto 不支持递增查询。

使用 Scala 进行增量查询

```
(val incQueryDF = spark.read
    .format("org.apache.hudi")
    .option(DataSourceReadOptions.QUERY_TYPE_OPT_KEY,
 DataSourceReadOptions.QUERY_TYPE_INCREMENTAL_OPT_VAL)
```

```
    .option(DataSourceReadOptions.BEGIN_INSTANTTIME_OPT_KEY, <beginInstantTime>)
    .load("s3://DOC-EXAMPLE-BUCKET/myhudidataset" ))

incQueryDF.show()
```

使用 PySpark 进行增量查询

```
readOptions = {
  'hoodie.datasource.query.type': 'incremental',
  'hoodie.datasource.read.begin.instanttime': <beginInstantTime>,
}

incQueryDF = spark.read \
    .format('org.apache.hudi') \
    .options(**readOptions) \
    .load('s3://DOC-EXAMPLE-BUCKET/myhudidataset')

incQueryDF.show()
```

有关从 Hudi 数据集读取的更多信息，请参阅 Apache Hudi 文档中的 查询 Hudi 表。

使用 Hudi CLI

您可以使用 Hudi CLI 管理 Hudi 数据集，以查看有关提交、文件系统、统计信息等的信息。还可以使用 CLI 手动执行压缩、计划压缩或取消计划的压缩。有关更多信息，请参阅 Apache Hudi 文档中的 CLI 互动。

启动 Hudi CLI 并连接到数据集

1. 使用 SSH 连接主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 在命令行中，键入 `/usr/lib/hudi/cli/bin/hudi-cli.sh`。

   命令提示符更改为 `hudi->`。
3. 键入以下代码以连接到数据集。将 *s3://DOC-EXAMPLE-BUCKET/myhudidataset* 替换为您想要使用的数据集的路径。我们使用的值与前面示例中建立的值相同。

   ```
   connect --path s3://DOC-EXAMPLE-BUCKET/myhudidataset
   ```

   命令提示符将更改以包括您连接到的数据集，如以下示例所示。

```
hudi:myhudidataset->
```

Hudi 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Hudi 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Hudi 版本信息。

Amazon EMR 发行版标签	Hudi 版本	随 Hudi 安装的组件
emr-6.14.0	0.13.1-amzn-2	Not available.
emr-6.13.0	0.13.1-amzn-1	Not available.
emr-6.12.0	0.13.1-amzn-0	Not available.
emr-6.11.1	0.13.0-amzn-0	Not available.
emr-6.11.0	0.13.0-amzn-0	Not available.
emr-6.10.1	0.12.2-amzn-0	Not available.
emr-6.10.0	0.12.2-amzn-0	Not available.
emr-6.9.1	0.12.1-amzn-0	Not available.
emr-6.9.0	0.12.1-amzn-0	Not available.
emr-6.8.1	0.11.1-amzn-0	Not available.
emr-6.8.0	0.11.1-amzn-0	Not available.
emr-6.7.0	0.11.0-amzn-0	Not available.
emr-5.36.1	0.10.1-amzn-1	Not available.
emr-5.36.0	0.10.1-amzn-1	Not available.

Amazon EMR 发行版标签	Hudi 版本	随 Hudi 安装的组件
emr-6.6.0	0.10.1-amzn-0	Not available.
emr-5.35.0	0.9.0-amzn-2	Not available.
emr-6.5.0	0.9.0-amzn-1	Not available.
emr-6.4.0	0.8.0-amzn-0	Not available.
emr-6.3.1	0.7.0-amzn-0	Not available.
emr-6.3.0	0.7.0-amzn-0	Not available.
emr-6.2.1	0.6.0-amzn-1	Not available.
emr-6.2.0	0.6.0-amzn-1	Not available.
emr-6.1.1	0.5.2-incubating-amzn-2	Not available.
emr-6.1.0	0.5.2-incubating-amzn-2	Not available.
emr-6.0.1	0.5.0-incubating-amzn-1	Not available.
emr-6.0.0	0.5.0-incubating-amzn-1	Not available.
emr-5.34.0	0.9.0-amzn-0	Not available.
emr-5.33.1	0.7.0-amzn-1	Not available.
emr-5.33.0	0.7.0-amzn-1	Not available.
emr-5.32.1	0.6.0-amzn-0	Not available.
emr-5.32.0	0.6.0-amzn-0	Not available.
emr-5.31.1	0.6.0-amzn-0	Not available.
emr-5.31.0	0.6.0-amzn-0	Not available.
emr-5.30.2	0.5.2-incubating	Not available.

Amazon EMR 发行版标签	Hudi 版本	随 Hudi 安装的组件
emr-5.30.1	0.5.2-incubating	Not available.
emr-5.30.0	0.5.2-incubating	Not available.
emr-5.29.0	0.5.0-incubating	Not available.
emr-5.28.1	0.5.0-incubating	Not available.
emr-5.28.0	0.5.0-incubating	Not available.

Hue

Hue（Hadoop 用户体验）是基于 Web 的开源图形用户界面，可用于 Amazon EMR 和 Apache Hadoop。Hue 将多个不同的 Hadoop 生态系统项目组合在一起，形成一个可配置界面。Amazon EMR 还在 Amazon EMR 中添加了特定于 Hue 的自定义项。Hue 充当在您的集群上运行的应用程序的前端，使您可以使用可能更加熟悉或对用户更友好的界面与应用程序进行交互。通过 Hue 中的应用程序 (如 Hive 和 Pig 编辑器)，无需登录集群即可使用各应用程序相应的 Shell 交互式运行脚本。在集群启动后，您可以使用 Hue 或类似界面与应用程序进行完全交互。有关 Hue 的更多信息，请参阅 http://gethue.com。

默认情况下，Hue 是您使用 Amazon EMR 控制台启动集群时安装的。您可通过以下方式选择不安装 Hue：在启动集群时使用 Amazon EMR 控制台中的 Advanced options (高级选项)；或在通过 Amazon CLI 使用 `create-cluster` 时显式指定 `--applications` 选项并忽略 Hue。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Hue 版本，以及 Amazon EMR 随 Hue 一起安装的组件。

有关此发行版中随 Hue 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Hue 版本信息

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.14.0	Hue 4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Hue 版本，以及 Amazon EMR 随 Hue 一起安装的组件。

有关此发行版中随 Hue 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Hue 版本信息

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.36.1	Hue 4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

主题

Hue on Amazon EMR 支持和不支持的功能

- Amazon S3 和 Hadoop 文件系统（HDFS）浏览器
  - 通过适当的权限，您可以在临时 HDFS 存储和属于您账户的 S3 存储桶之间浏览和移动数据。
  - 默认情况下，Hue 中的超级用户可以访问允许 Amazon EMR IAM 角色访问的所有文件。新建用户不会自动拥有对 Amazon S3 filebrowser 的访问权限，并且必须为其组启用 `filebrowser.s3_access` 权限。
- Hive – 对数据运行交互式查询。此外，这也是编程或批处理查询原型的一种有用方法。
- Pig – 对数据运行脚本或发出交互式命令。

- Oozie – 创建并监控 Oozie 工作流。
- 元存储管理器 – 可用于查看和操作 Hive 元存储的内容（导入/创建、删除等）。
- 任务浏览器 – 查看您提交的 Hadoop 任务的状态。
- 用户管理 – 管理 Hue 账户并将 LDAP 用户与 Hue 集成。
- Amazon示例 – 有多个“ready-to-run”示例可使用 Hue 中的应用程序处理来自各种Amazon服务的示例数据。登录 Hue 后，您将转到 Hue 应用程序主页，其中预安装了示例。
- 仅 Amazon EMR 5.9.0 版或更高版本支持 Livy Server。
- 要使用 Hue Notebook for Spark，您必须在 Hue 中安装 Livy 和 Spark。
- 不支持 Hue 控制面板。
- 不支持 PostgreSQL。

连接到 Hue Web 用户界面

连接到 Hue Web 用户界面的过程与连接到集群主节点上托管的任何 HTTP 接口相同。以下过程将介绍如何访问 Hue 用户界面。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看 EMR 集群上托管的 Web 界面。

查看 Hue Web 用户界面

1. 按照《Amazon EMR 管理指南》中的使用动态端口转发设置到主节点的 SSH 隧道的说明操作。
2. 在浏览器中键入以下地址以打开 Hue Web 界面：`http://`*`master public DNS`*`:8888`，其中 *`master public dns`* 是您的集群主节点的公有 DNS 名称，例如，`ec2-11-22-333-44.compute-1.amazonaws.com`。
3. 在 Hue 登录屏幕上，如果您是首次登录的管理员，请输入用户名和密码以创建您的 Hue 超级用户账户，然后选择创建账户。否则，请键入您的用户名和密码，然后选择 Create account (创建账户)，或输入管理员提供的凭证。

将 Hue 与 Amazon RDS 中的远程数据库结合使用

默认情况下，Hue 用户信息和查询历史记录存储在主节点上的本地 MySQL 数据库中。或者，您可以使用 Amazon S3 中存储的配置和 Amazon Relational Database Service（Amazon RDS）中的 MySQL 数据库创建一个或多个启用了 Hue 的集群。这样，无需使 Amazon EMR 集群保持运行，您就可以保存 Hue 创建的用户信息和查询历史记录。我们建议使用 Amazon S3 服务器端加密来存储配置文件。

首先为 Hue 创建远程数据库。

创建外部 MySQL 数据库

1. 通过以下网址打开 Amazon RDS 控制台：https://console.aws.amazon.com/rds/。
2. 点击 Launch a DB Instance (启动数据库实例)。
3. 选择 MySQL，然后单击 Select (选择)。
4. 保留默认选择 Multi-AZ Deployment and Provisioned IOPS Storage (多可用区部署和预置 IOPS 存储)，并单击 Next (下一步)。
5. 保留 Instance Specifications (实例规格) 的默认值，指定设置，然后单击 Next (下一步)。
6. 在 Configure Advanced Settings (配置高级设置) 页面上，选择相应的安全组和数据库名称。您使用的安全组必须至少允许从集群主节点中对端口 3306 进行入口 TCP 访问。如果此时您尚未创建集群，则可以允许所有主机连接到端口 3306 并在启动集群之后调整安全组。单击 Launch DB Instance (启动数据库实例)。
7. 在 RDS 控制面板中，选择 Instances (实例)，然后选择您刚刚创建的实例。当您的数据库可用时，记下数据库名称、用户名、密码和 RDS 实例主机名。您将在创建和配置集群时用到此信息。

使用 Amazon CLI 在启动集群时为 Hue 指定外部 MySQL 数据库

要在使用 Amazon CLI 启动集群时为 Hue 指定外部 MySQL 数据库，请在创建 RDS 实例时使用所记录的信息，以使用配置对象配置 `hue.ini`

> **Note**
>
> 您可以创建使用同一个外部数据库的多个集群，但是每个集群将共享查询历史记录和用户信息。

- 使用 Amazon CLI 创建安装了 Hue 的集群，使用您创建的外部数据库并使用指定数据库属性的 Hue 配置分类引用配置文件。以下示例创建一个安装了 Hue 的集群，引用了 Amazon S3 中的配置文件 `myConfig.json`，该文件指定数据库配置。

  > **Note**
  >
  > 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Hue
 Name=Spark Name=Hive \
--instance-type m5.xlarge --instance-count 3 \
--configurations https://s3.amazonaws.com/mybucket/myfolder/myConfig.json --use-
default-roles
```

下面显示的是 myConfig.json 文件的内容示例。讲 *dbname*、*username*、*password* 和 *RDS instance hostname* 替换为您之前记下的 RDS 控制面板中的值。

```
[{
  "Classification": "hue-ini",
  "Properties": {},
  "Configurations": [
    {
      "Classification": "desktop",
      "Properties": {},
      "Configurations": [
        {
          "Classification": "database",
          "Properties": {
            "name": "dbname",
            "user": "username",
            "password": "password",
            "host": "RDS instance hostname",
            "port": "3306",
            "engine": "mysql"
          },
          "Configurations": []
        }
      ]
    }
  ]
}]
```

故障排除

发生 Amazon RDS 故障转移时

由于 Hue 数据库实例无响应或正在进行故障转移，因此用户可能会在运行查询时遇到延迟。以下是有关此问题的一些事实和准则：

- 如果登录 Amazon RDS 控制台，则可以搜索故障转移事件。例如，要查看是否正在进行或已发生故障转移，请查找诸如“多可用区实例故障转移已启动”和“多可用区实例故障转移已完成”之类的事件。
- RDS 实例完成故障转移大约需要 30 秒。
- 如果在 Hue 中遇到比正常查询更长的响应，请尝试重新执行查询。

Hue 的高级配置

本节包括以下主题。

主题

为 LDAP 用户配置 Hue

通过与 LDAP 集成，用户可以使用存储在 LDAP 目录中的现有凭证登录 Hue。将 Hue 与 LDAP 集成时，不需要在 Hue 中独立管理用户信息。以下信息演示了 Hue 与 Microsoft Active Directory 的集成，但配置选项类似于任何 LDAP 目录。

LDAP 身份验证首先需要绑定到服务器并建立连接。然后，建立的连接将用于任何后续查询，从而搜索 LDAP 用户信息。除非您的 Active Directory 服务器允许匿名连接，否则需要使用绑定可分辨名称和密码建立连接。绑定可分辨名称（DN）由 `bind_dn` 配置设置定义。绑定密码由 `bind_password` 配置设置定义。Hue 有两种绑定 LDAP 请求的方法：搜索绑定和直接绑定。将 Hue 与 Amazon EMR 一起使用的首选方法是搜索绑定。

在 Active Directory 中使用搜索绑定时，Hue 会通过用户名属性（由 `user_name_attr config` 定义）来查找需要从基本可分辨名称（或 DN）中检索的属性。当 Hue 用户不知道完整 DN 时，搜索绑定非常有用。

例如，您可能已将 `user_name_attr config` 设置为使用通用名称（CN）。在这种情况下，Active Directory 服务器使用登录期间提供的 Hue 用户名在目录树中搜索匹配的通用名称，从基本可分辨名称开始。如果找到 Hue 用户的通用名称，则服务器返回用户的可分辨名称。然后，Hue 构造一个可分辨名称，用于通过执行绑定操作对用户进行身份验证。

> **Note**
>
> 搜索绑定可搜索所有目录子树中的用户名，从基本可分辨名称开始。在 Hue LDAP 配置中指定的基本可分辨名称应该是用户名的最近父级，否则 LDAP 身份验证性能可能会受到影响。

在 Active Directory 中使用直接绑定时，必须使用精确的 `nt_domain` 或 `ldap_username_pattern` 进行身份验证。当使用直接绑定时，如果 nt 域（由 `nt_domain` 配置设置定义）属性已定义，则使用以下形式创建用户可分辨名称模板：`<login username>@nt_domain`。此模板用于搜索以基本可分辨名称开始的所有目录子树。如果未配置 nt 域，Hue 会为用户搜索精确的可分辨名称模式（由 `ldap_username_pattern` 配置设置定义）。在这种情况下，服务器在所有目录子树中搜索匹配的 `ldap_username_pattern` 值，从基本可分辨名称开始。

使用 Amazon CLI 启动带有针对 Hue 的 LDAP 属性的集群

- 要为 `hue-ini` 指定 LDAP 属性，请创建一个安装了 Hue 的集群并引用包含 LDAP 的配置属性的 json 文件。下面显示了一个示例命令，此命令引用 Amazon S3 中存储的配置文件 `myConfig.json`。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Hue
 Name=Spark Name=Hive \
--instance-type m5.xlarge --instance-count 3 --configurations https://
s3.amazonaws.com/mybucket/myfolder/myConfig.json.
```

  下面显示的是 `myConfig.json` 的内容示例。

```
[
    {
        "Classification": "hue-ini",
        "Properties": {},
        "Configurations": [
            {
                "Classification": "desktop",
                "Properties": {},
                "Configurations": [
                    {
                        "Classification": "ldap",
                        "Properties": {},
                        "Configurations": [
                            {
```

```
                                "Classification": "ldap_servers",
                                "Properties": {},
                                "Configurations": [
                                    {
                                        "Classification": "yourcompany",
                                        "Properties": {
                                            "base_dn":
 "DC=yourcompany,DC=hue,DC=com",
                                            "ldap_url": "ldap://ldapurl",
                                            "search_bind_authentication": "true",
                                            "bind_dn":
 "CN=hue,CN=users,DC=yourcompany,DC=hue,DC=com",
                                            "bind_password": "password"
                                        },
                                        "Configurations": []
                                    }
                                ]
                            }
                        ]
                    },
                    {
                        "Classification": "auth",
                        "Properties": {
                         "backend": "desktop.auth.backend.LdapBackend"
                              }
                    }
                ]
            }

        ]
    }
]
```

> **Note**
>
> 对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

在 Hue 中查看 LDAP 设置

1. 验证您是否有到 Amazon EMR 集群主节点的活动 VPN 连接或 SSH 隧道。然后，在浏览器中键入 *master-public-dns*:8888 以打开 Hue Web 界面。
2. 使用 Hue 管理员凭证登录。如果 Did you know? (您知道吗？) 窗口打开，单击 Got it, prof! (明白了，教授！) 可关闭该窗口。
3. 在工具栏中，单击 Hue 图标。
4. 在 About Hue (关于 Hue) 页面上，单击 Configuration (配置)。
5. 在 Configuration Sections and Variables (配置部分和变量) 部分，单击 Desktop (桌面)。
6. 滚动到 ldap 部分以查看您的设置。

Hue 发行版历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Hue 的版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Hue 版本信息

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.14.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.13.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
		hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.12.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.11.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.11.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.10.1	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.10.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.9.1	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.9.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.8.1	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.8.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.7.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.36.1	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.36.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.6.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.35.0	4.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.5.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.4.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.3.1	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.3.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.2.1	4.8.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.2.0	4.8.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.1.1	4.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.1.0	4.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-6.0.1	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-6.0.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.34.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.33.1	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.33.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.32.1	4.8.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.32.0	4.8.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.31.1	4.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.31.0	4.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.30.2	4.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.30.1	4.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.30.0	4.6.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mariadb-server, oozie-client, oozie-server
emr-5.29.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.28.1	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.28.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.27.1	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.27.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.26.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.25.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.24.1	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.24.0	4.4.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.23.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.23.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.22.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.21.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.21.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.21.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.20.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.20.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.19.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.19.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.18.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.18.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.17.2	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.17.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.17.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.16.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.16.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.15.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.15.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.14.2	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.14.1	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.14.0	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.13.1	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.13.0	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.12.3	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.12.2	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.12.1	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.12.0	4.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.11.4	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.11.3	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.11.2	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.11.1	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.11.0	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.10.1	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.10.0	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.9.1	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.9.0	4.0.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.8.3	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.8.2	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.8.1	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.8.0	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.7.1	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.7.0	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.6.1	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.6.0	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hue-server, mysql-server, oozie-client, oozie-server
emr-5.5.4	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.5.3	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.5.2	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.5.1	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.5.0	3.12.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.4.1	3.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.4.0	3.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.3.2	3.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.3.1	3.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.3.0	3.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.2.3	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.2.2	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.2.1	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.2.0	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.1.1	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-5.1.0	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-5.0.3	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-5.0.0	3.10.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.9.6	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.9.5	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.9.4	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.9.3	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.9.2	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.9.1	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.8.5	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.8.4	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.8.3	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.8.2	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.8.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.7.4	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server
emr-4.7.2	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-client, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.7.1	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.7.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.6.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.5.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.4.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.3.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server
emr-4.2.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server

Amazon EMR 发行版标签	Hue 版本	随 Hue 安装的组件
emr-4.1.0	3.7.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hue-server, mysql-server, oozie-server

Iceberg

Apache Iceberg 是 Amazon Simple Storage Service（Amazon S3）中适用于大型数据集的开放表格式。它提供快速的大型表查询性能、原子提交、并发写入和 SQL 兼容表演进等功能。从 Amazon EMR 6.5.0 开始，您可以在使用 Iceberg 表格式的 Amazon EMR 集群上使用 Apache Spark 3。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Iceberg 版本，以及 Amazon EMR 随 Iceberg 一起安装的组件。

有关此发行版中随 Iceberg 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Iceberg 版本信息

Amazon EMR 发行版标签	Iceberg 版本	随 Iceberg 安装的组件
emr-6.14.0	Iceberg 1.3.1-amzn-0	Not available.

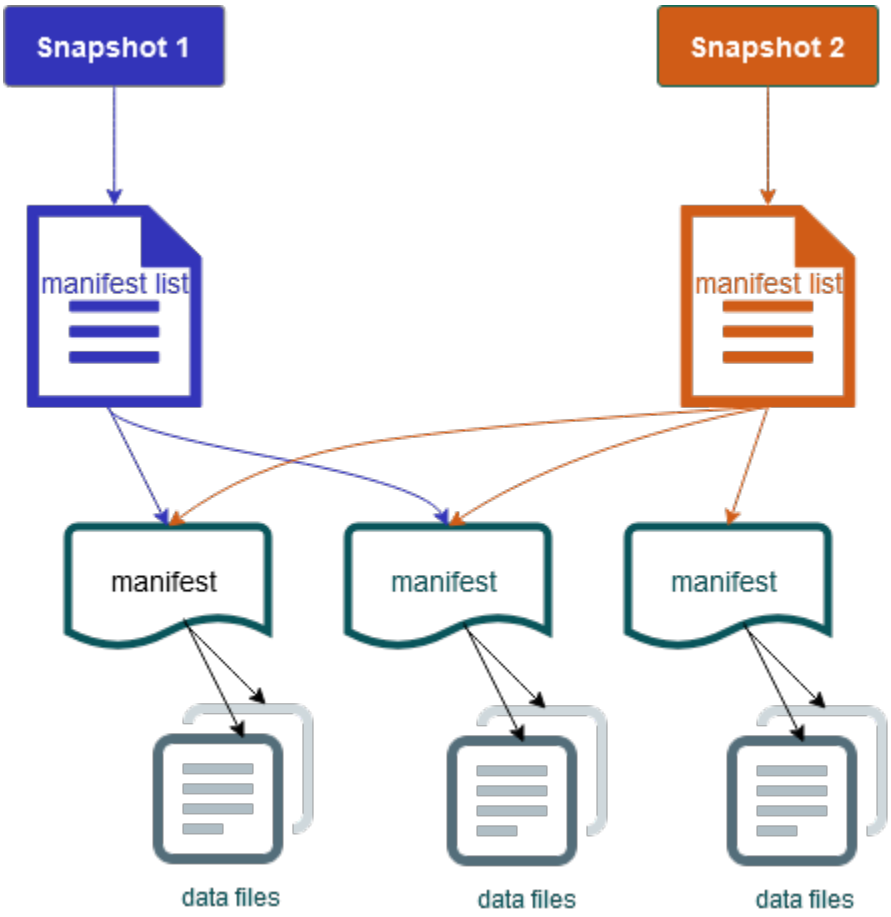
主题

Iceberg 的工作原理

Iceberg 跟踪表中而非目录中的单个数据文件。这样，写入器便可在原位置创建数据文件（文件不会移动或更改）。此外，写入器只能在显式提交时将文件添加到表中。表状态在元数据文件中维护。对表状态的所有更改都会创建一个新的元数据文件，该文件会原子式替换旧的元数据。表元数据文件跟踪表架构、分区配置和其他属性。

其还包括表内容的快照。每个快照都是表中数据文件在某个时间点的完整集合。快照在元数据文件中列出，但快照中的文件存储在单独的清单文件中。通过在表元数据文件之间的原子转换来实现快照隔离。读取器使用加载表元数据时最新的快照。读取器在刷新并选取新的元数据位置之前不会受到更改的影响。快照中的数据文件存储在一个或多个清单文件中，其中包含表中每个数据文件、分区数据及其指标一行。快照是清单中所有文件的并集。清单文件还可以在快照之间共享，以避免重写不常更改的元数据。

Iceberg 快照图



Iceberg 提供以下功能：

- 支持 Simple Storage Service (Amazon S3) 数据湖中的 ACID 事务处理和时间行程。
- 提交重试次数受益于乐观并发的性能优势。
- 由于解决了文件级冲突问题，因此具有高并发能力。
- 通过元数据中每列的最小最大统计数据，您可以跳过文件，从而提高选择性查询的性能。
- 您可以通过分区发展将表整理为灵活的分区布局，以便能够更新分区架构。然后，查询和数据量可以在不依赖物理目录的情况下进行更改。
- 支持架构发展和强制执行。
- Iceberg 表充当幂等性数据汇和可重放的源。其能够通过一次精确的管道支持流式处理和批处理。幂等性数据汇会跟踪过去成功的写入操作。因此，数据汇可以在失败时再次请求数据，并可在数据已多次发送时丢弃数据。
- 查看历史记录和谱系，包括表发展、操作历史记录和每次提交的统计数据。

- 从现有数据集迁移时，支持数据格式（Parquet、ORC、Avro）和分析引擎（Spark、Trino、PrestoDB、Flink、Hive）选择。

使用安装有 Iceberg 的集群

本节包含将 Iceberg 与 Spark、Trino、Flink 和 Hive 结合使用的信息。

将 Iceberg 集群与 Spark 结合使用

从 Amazon EMR 版本 6.5.0 开始，您可以将 Iceberg 用于您的 Spark 集群，无需包含引导操作。对于 Amazon EMR 版本 6.4.0 及更早版本，您可以使用引导操作来预装所有需要的依赖项。

在本教程中，您将通过 Amazon CLI 在 Amazon EMR Spark 集群上使用 Iceberg。要使用控制台创建安装了 Iceberg 的集群，请按照使用 Amazon Athena、Amazon EMR 和 Amazon Glue 构建 Apache Iceberg 数据湖中的步骤操作。

创建 Iceberg 集群

您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 创建安装了 Iceberg 的集群。在本教程中，您将通过 Amazon CLI 在 Amazon EMR 集群上使用 Iceberg。要使用控制台创建安装了 Iceberg 的集群，请按照使用 Amazon Athena、Amazon EMR 和 Amazon Glue 构建 Apache Iceberg 数据湖中的步骤操作。

要在 Amazon EMR 上将 Iceberg 与 Amazon CLI 结合使用，请首先按照以下步骤创建一个集群。有关使用 Amazon CLI 指定 Iceberg 分类的信息，请参阅在创建集群时使用 Amazon CLI 提供配置或在创建集群时，使用 Java SDK 提供配置。

1. 创建 `configurations.json` 文件并输入以下内容：

```
[{
    "Classification":"iceberg-defaults",
    "Properties":{"iceberg.enabled":"true"}
}]
```

2. 接下来，使用以下配置创建集群。将实例 Amazon S3 桶路径和子网 ID 替换为您自己的值。

```
aws emr create-cluster --release-label emr-6.5.0 \
--applications Name=Spark \
--configurations file://iceberg_configurations.json \
```

```
--region us-east-1 \
--name My_Spark_Iceberg_Cluster \
--log-uri s3://DOC-EXAMPLE-BUCKET/ \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole_V2 \
--ec2-attributes
 InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef0
```

您还可以创建一个包含 Spark 应用程序的 Amazon EMR 集群，并且将文件 `/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar` 作为 Spark 任务中的 JAR 依赖关系包含在内。有关更多信息，请参阅提交应用程序。

要将 jar 作为 Spark 作业中的依赖项包含在内，请将以下配置属性添加到 Spark 应用程序中：

```
--conf "spark.jars=/usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar"
```

有关 Spark 作业依赖项的更多信息，请参阅 Apache Spark 文档 Running Spark on Kubernetes（在 Kubernetes 上运行 Spark）中的 Dependency Management（依赖项管理）。

为 Iceberg 初始化 Spark 会话

以下示例演示如何启动交互式 Spark Shell、使用 Spark 提交，或如何使用 Amazon EMR Notebooks 在 Amazon EMR 上使用 Iceberg。

spark-shell

1. 使用 SSH 连接主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以启动 Spark shell。要使用 PySpark Shell，请使用 `pyspark` 替换 `spark-shell`。

   ```
   spark-shell \
   --conf
    "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"
    \
   --conf "spark.sql.catalog.dev=org.apache.iceberg.spark.SparkCatalog" \
   --conf "spark.sql.catalog.dev.type=hadoop" \
   --conf "spark.sql.catalog.dev.warehouse=s3://DOC-EXAMPLE-BUCKET/example-prefix/"
   ```

spark-submit

1. 使用 SSH 连接主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 输入以下命令以为 Iceberg 启动 Spark 会话。

```
spark-submit \
--conf
 "spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions"
 \
--conf "spark.sql.catalog.dev=org.apache.iceberg.spark.SparkCatalog" \
--conf "spark.sql.catalog.dev.type=hadoop" \
--conf "spark.sql.catalog.dev.warehouse=s3://DOC-EXAMPLE-BUCKET/example-prefix/"
```

EMR Studio notebooks

要使用 EMR Studio notebooks 初始化 Spark 会话，请使用 Amazon EMR notebook 中的 `%%configure` 魔法命令配置 Spark 会话，如以下示例所示。有关更多信息，请参阅 Amazon EMR 管理指南中的使用 EMR Notebooks 魔法命令。

```
%%configure -f
{
"conf":{

 "spark.sql.extensions":"org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions",
    "spark.sql.catalog.dev":"org.apache.iceberg.spark.SparkCatalog",
    "spark.sql.catalog.dev.type":"hadoop",
    "spark.sql.catalog.dev.warehouse":"s3://DOC-EXAMPLE-BUCKET/example-prefix/"
    }
}
```

写入 Iceberg 表

以下示例演示如何创建 DataFrame 并将其作为 Iceberg 数据集写入。这些示例演示使用 Spark Shell 处理数据集，同时使用 SSH 作为原定设置将 hadoop 用户连接到主节点（master node）。

> **Note**
>
> 要将代码示例粘贴到 Spark Shell 中，请在提示符处键入 `:paste`，粘贴示例，然后按 CTRL +D。

PySpark

Spark 包含一个基于 Python 的 Shell `pyspark`，您可以用它来设计以 Python 编写的 Spark 程序的原型。在主节点上调用 `pyspark`。

```
## Create a DataFrame.
data = spark.createDataFrame([
 ("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
 ("101", "2015-01-01", "2015-01-01T12:14:58.597216Z"),
 ("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
 ("103", "2015-01-01", "2015-01-01T13:51:40.519832Z")
],["id", "creation_date", "last_update_time"])

## Write a DataFrame as a Iceberg dataset to the Amazon S3 location.
spark.sql("""CREATE TABLE IF NOT EXISTS dev.db.iceberg_table (id string,
creation_date string,
last_update_time string)
USING iceberg
location 's3://DOC-EXAMPLE-BUCKET/example-prefix/db/iceberg_table'""")

data.writeTo("dev.db.iceberg_table").append()
```

Scala

```
import org.apache.spark.sql.SaveMode
import org.apache.spark.sql.functions._

// Create a DataFrame.
val data = Seq(
("100", "2015-01-01", "2015-01-01T13:51:39.340396Z"),
("101", "2015-01-01", "2015-01-01T12:14:58.597216Z"),
("102", "2015-01-01", "2015-01-01T13:51:40.417052Z"),
("103", "2015-01-01", "2015-01-01T13:51:40.519832Z")
).toDF("id", "creation_date", "last_update_time")

// Write a DataFrame as a Iceberg dataset to the Amazon S3 location.
```

```
spark.sql("""CREATE TABLE IF NOT EXISTS dev.db.iceberg_table (id string,
creation_date string,
last_update_time string)
USING iceberg
location 's3://DOC-EXAMPLE-BUCKET/example-prefix/db/iceberg_table'""")

data.writeTo("dev.db.iceberg_table").append()
```

从 Iceberg 表读取

PySpark

```
df = spark.read.format("iceberg").load("dev.db.iceberg_table")
df.show()
```

Scala

```
val df = spark.read.format("iceberg").load("dev.db.iceberg_table")
df.show()
```

Spark SQL

```
SELECT * from dev.db.iceberg_table LIMIT 10
```

配置 Spark 属性以使用 Amazon Glue 数据目录作为 Iceberg 表元数据仓

要将 Amazon Glue 数据目录作为 Iceberg 表的元数据仓，请按如下方式设置 Spark 配置属性:

```
spark-submit \
    --conf spark.sql.catalog.my_catalog=org.apache.iceberg.spark.SparkCatalog \
    --conf spark.sql.catalog.my_catalog.warehouse=s3://<bucket>/<prefix> \
    --conf spark.sql.catalog.my_catalog.catalog-
impl=org.apache.iceberg.aws.glue.GlueCatalog \
    --conf spark.sql.catalog.my_catalog.io-impl=org.apache.iceberg.aws.s3.S3FileIO \
    --conf spark.sql.catalog.my_catalog.lock-
impl=org.apache.iceberg.aws.dynamodb.DynamoDbLockManager \
    --conf spark.sql.catalog.my_catalog.lock.table=myGlueLockTable
```

将 Iceberg 集群与 Trino 结合使用

从 Amazon EMR 版本 6.6.0 开始，您可以将 Iceberg 用于您的 Trino 集群。

在本教程中，您将通过 Amazon CLI 在 Amazon EMR Trino 集群上使用 Iceberg。要使用控制台创建安装了 Iceberg 的集群，请按照使用 Amazon Athena、Amazon EMR 和 Amazon Glue 构建 Apache Iceberg 数据湖中的步骤操作。

创建 Iceberg 集群

要在 Amazon EMR 上将 Iceberg 与 Amazon CLI 结合使用，请首先按照以下步骤创建一个集群。有关使用 Amazon CLI 指定 Iceberg 分类的信息，请参阅在创建集群时使用 Amazon CLI 提供配置或在创建集群时，使用 Java SDK 提供配置。

1. 创建 `iceberg.properties` 文件，然后为您选择的目录设置一个值。例如，假设您想将 Hive 元存储作为目录使用，则您的文件应包含以下内容。

   ```
   connector.name=iceberg
   hive.metastore.uri=thrift://localhost:9083
   ```

   如果您想将 Amazon Glue Data Catalog 作为目录使用，则您的文件应包含以下内容。

   ```
   connector.name=iceberg
   iceberg.catalog.type=glue
   ```

2. 创建一个会将 `iceberg.properties` 从 Amazon S3 复制到 `/etc/trino/conf/catalog/iceberg.properties` 的引导操作，如下例所示。有关引导操作的信息，请参阅 创建引导操作以安装其他软件。

   ```
   set -ex
   sudo aws s3 cp s3://DOC-EXAMPLE-BUCKET/iceberg.properties /etc/trino/conf/catalog/
   iceberg.properties
   ```

3. 使用以下配置创建一个集群，将示例引导操作脚本路径和密钥名称替换为您自己的值。

   ```
   aws emr create-cluster --release-label emr-6.7.0 \
   --applications Name=Trino \
   --region us-east-1 \
   --name My_Trino_Iceberg_Cluster \
   --bootstrap-actions '[{"Path":"s3://DOC-EXAMPLE-BUCKET","Name":"Add
    iceberg.properties"}]' \
   ```

```
--instance-groups InstanceGroupType=MASTER,InstanceCount=1,InstanceType=c3.4xlarge
 InstanceGroupType=CORE,InstanceCount=3,InstanceType=c3.4xlarge \
--use-default-roles \
--ec2-attributes KeyName=<key-name>
```

为 Iceberg 初始化 Trino 会话

要初始化 Trino 会话，请运行以下命令。

```
trino-cli --catalog iceberg
```

写入 Iceberg 表

使用以下 SQL 命令创建并写入您的表。

```
trino> SHOW SCHEMAS;
trino> CREATE TABLE default.iceberg_table (
            id int,
            data varchar,
            category varchar)
       WITH (
            format = 'PARQUET',
            partitioning = ARRAY['category', 'bucket(id, 16)'],
            location = 's3://DOC-EXAMPLE-BUCKET/<prefix>')

trino> INSERT INTO default.iceberg_table VALUES (1,'a','c1'), (2,'b','c2'),
 (3,'c','c3');
```

从 Iceberg 表读取

要从 Iceberg 表读取，请运行以下命令。

```
trino> SELECT * from default.iceberg_table;
```

将 Iceberg 集群与 Flink 结合使用

从 Amazon EMR 版本 6.9.0 开始，您可以将 Iceberg 与 Flink 集群结合使用，而无需使用开源 Iceberg Flink 集成时所需的设置步骤。

创建 Iceberg 集群

您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 创建安装了 Iceberg 的集群。在本教程中，您将通过 Amazon CLI 在 Amazon EMR 集群上使用 Iceberg。要使用控制台创建安装了 Iceberg 的集群，请按照使用 Amazon Athena、Amazon EMR 和 Amazon Glue 构建 Apache Iceberg 数据湖中的步骤操作。

要在 Amazon EMR 上将 Iceberg 与 Amazon CLI 结合使用，请首先按照以下步骤创建一个集群。有关使用 Amazon CLI 指定 Iceberg 分类的信息，请参阅在创建集群时使用 Amazon CLI 提供配置或在创建集群时，使用 Java SDK 提供配置。使用以下内容创建名为 `configurations.json` 的文件：

```
[{
"Classification":"iceberg-defaults",
    "Properties":{"iceberg.enabled":"true"}
}]
```

接下来，使用以下配置创建集群，将示例 Amazon S3 桶路径和子网 ID 替换为您自己的值：

```
aws emr create-cluster --release-label emr-6.9.0 \
--applications Name=Flink \
--configurations file://iceberg_configurations.json \
--region us-east-1 \
--name My_flink_Iceberg_Cluster \
--log-uri s3://DOC-EXAMPLE-BUCKET/ \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole \
--ec2-attributes InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef
```

您还可以创建一个其中包含 Flink 应用程序的 Amazon EMR 6.9.0 集群，并且将文件 `/usr/share/aws/iceberg/lib/iceberg-flink-runtime.jar` 用作 Flink 作业中的 JAR 依赖项。

使用 Flink SQL 客户端

SQL 客户端脚本位于 `/usr/lib/flink/bin` 下。您可以使用以下命令运行脚本：

```
flink-yarn-session -d # starting the Flink YARN Session in detached mode
./sql-client.sh
```

这将启动 Flink SQL Shell。

Flink 示例

创建 Iceberg 表

Flink SQL

```
CREATE CATALOG glue_catalog WITH (
   'type'='iceberg',
   'warehouse'='<WAREHOUSE>',
   'catalog-impl'='org.apache.iceberg.aws.glue.GlueCatalog',
   'io-impl'='org.apache.iceberg.aws.s3.S3FileIO',
   'lock-impl'='org.apache.iceberg.aws.glue.DynamoLockManager',
   'lock.table'='myGlueLockTable'
 );

USE CATALOG  glue_catalog;

CREATE DATABASE IF NOT EXISTS <DB>;

USE <DB>;

CREATE TABLE IF NOT EXISTS `glue_catalog`.`<DB>`.`sample` (id int, data string);
```

表 API

```
EnvironmentSettings settings =
                EnvironmentSettings.newInstance().inBatchMode().build();

TableEnvironment tEnv = TableEnvironment.create(settings);

String warehouse = "<WAREHOUSE>";
String db = "<DB>";

tEnv.executeSql(
                "CREATE CATALOG glue_catalog WITH (\n"
                        + "   'type'='iceberg',\n"
                        + "   'warehouse'='"
                        + warehouse
                        + "',\n"
                        + "   'catalog-impl'='org.apache.iceberg.aws.glue.GlueCatalog',
\n"
                        + "   'io-impl'='org.apache.iceberg.aws.s3.S3FileIO'\n"
                        + " );");
```

```
tEnv.executeSql("USE CATALOG  glue_catalog;");
tEnv.executeSql("CREATE DATABASE IF NOT EXISTS " + db + ";");
tEnv.executeSql("USE " + db + ";");
tEnv.executeSql(
        "CREATE TABLE `glue_catalog`.`" + db + "`.`sample` (id bigint, data string);");
```

写入 Iceberg 表

Flink SQL

```
INSERT INTO `glue_catalog`.`<DB>`.`sample` values (1, 'a'),(2,'b'),(3,'c');
```

表 API

```
tEnv.executeSql(
        "INSERT INTO `glue_catalog`.`"
                + db
                + "`.`sample` values (1, 'a'),(2,'b'),(3,'c');");
```

数据流 API

```
final StreamExecutionEnvironment env =
 StreamExecutionEnvironment.getExecutionEnvironment();

final StreamTableEnvironment tableEnv = StreamTableEnvironment.create(env);

String db = "<DB Name>";

String warehouse = "<Warehouse Path>";

GenericRowData rowData1 = new GenericRowData(2);
rowData1.setField(0, 1L);
rowData1.setField(1, StringData.fromString("a"));

DataStream<RowData> input = env.fromElements(rowData1);

Map<String, String> props = new HashMap<();
props.put("type", "iceberg");
props.put("warehouse", warehouse);
props.put("io-impl", "org.apache.iceberg.aws.s3.S3FileIO");
```

```
CatalogLoader glueCatlogLoader =
        CatalogLoader.custom(
                "glue",
                props,
                new Configuration(),
                "org.apache.iceberg.aws.glue.GlueCatalog");

TableLoader tableLoader =
        TableLoader.fromCatalog(glueCatlogLoader, TableIdentifier.of(db, "sample"));

DataStreamSink<Void> dataStreamSink =
        FlinkSink.forRowData(input).tableLoader(tableLoader).append();

env.execute("Datastream Write");
```

从 Iceberg 表读取

Flink SQL

```
SELECT * FROM `glue_catalog`.`<DB>`.`sample`;
```

表 API

```
Table result = tEnv.sqlQuery("select * from `glue_catalog`.`" + db + "`.`sample`;");
```

数据流 API

```
final StreamExecutionEnvironment env =
 StreamExecutionEnvironment.getExecutionEnvironment();

final StreamTableEnvironment tableEnv = StreamTableEnvironment.create(env);

String db = "<DB Name>";

String warehouse = "<Warehouse Path>";

Map<String, String> props = new HashMap<>();
props.put("type", "iceberg");
props.put("warehouse", warehouse);
props.put("io-impl", "org.apache.iceberg.aws.s3.S3FileIO");

CatalogLoader glueCatlogLoader =
        CatalogLoader.custom(
```

```
                "glue",
                props,
                new Configuration(),
                "org.apache.iceberg.aws.glue.GlueCatalog");

TableLoader tableLoader =
        TableLoader.fromCatalog(glueCatlogLoader, TableIdentifier.of(db, "sample"));

DataStream<RowData> batch =

 FlinkSource.forRowData().env(env).tableLoader(tableLoader).streaming(false).build();

batch.print().name("print-sink");
```

使用 Hive 目录

确保如 使用 Hive 元存储和 Glue 目录配置 Flink 中所述解析 Flink 和 Hive 依赖项。

运行 Flink 作业

向 Flink 提交作业的一种方法是使用每个作业的 Flink YARN 会话。这可以通过以下命令启动：

```
sudo flink run -m yarn-cluster -p 4 -yjm 1024m -ytm 4096m $JAR_FILE_NAME
```

将 Iceberg 集群与 Hive 结合使用

在 Amazon EMR 发行版 6.9.0 及更高版本中，您可以将 Iceberg 与 Hive 集群结合使用，而无需执行开源 Iceberg Hive 集成所需的设置步骤。对于 Amazon EMR 版本 6.8.0 及更早版本，您可以使用引导操作安装 `iceberg-hive-runtime` jar 来配置 Hive for Iceberg 支持。

Amazon EMR 6.9.0 包括 Hive 3.1.3 与 Iceberg 0.14.1 集成的所有功能，还包括 Amazon EMR 增加的功能，例如在运行时自动选择支持的执行引擎（EKS 6.9.0 上的 Amazon EMR）。

创建 Iceberg 集群

您可以使用 Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 创建安装了 Iceberg 的集群。在本教程中，您将通过 Amazon CLI 在 Amazon EMR 集群上使用 Iceberg。要使用控制台创建安装了 Iceberg 的集群，请按照使用 Amazon Athena、Amazon EMR 和 Amazon Glue 构建 Iceberg 数据湖中的步骤操作。

要在 Amazon EMR 上将 Iceberg 与 Amazon CLI 结合使用，请首先使用以下步骤创建一个集群。有关使用 Amazon CLI 或 Java SDK 指定 Iceberg 分类的信息，请参阅 在创建集群时使用

Amazon CLI 提供配置 或 在创建集群时，使用 Java SDK 提供配置。使用以下内容创建名为 configurations.json 的文件：

```
[{
    "Classification":"iceberg-defaults",
    "Properties":{"iceberg.enabled":"true"}
}]
```

接下来，使用以下配置创建集群，将示例 Amazon S3 桶路径和子网 ID 替换为您自己的值：

```
aws emr create-cluster --release-label emr-6.9.0 \
--applications Name=Hive \
--configurations file://iceberg_configurations.json \
--region us-east-1 \
--name My_hive_Iceberg_Cluster \
--log-uri s3://DOC-EXAMPLE-BUCKET/ \
--instance-type m5.xlarge \
--instance-count 2 \
--service-role EMR_DefaultRole \
--ec2-attributes InstanceProfile=EMR_EC2_DefaultRole,SubnetId=subnet-1234567890abcdef
```

Hive Iceberg 集群执行以下操作：

- 在 Hive 中加载 Iceberg Hive 运行时 jar 并为 Hive 引擎启用 Iceberg 相关配置。
- 启用 Amazon EMR Hive 的动态执行引擎选择，以防止用户设置支持的执行引擎实现 Iceberg 兼容性。

> **Note**
>
> Hive Iceberg 集群目前不支持 Amazon Glue Data Catalog。默认 Iceberg 目录为 HiveCatalog，它对应于为 Hive 环境配置的元存储。有关目录管理的更多信息，请参阅 Apache Hive documentation（Apache Hive 文档）中的 Catalog Management（目录管理）。

功能支持

Amazon EMR 6.9.0 支持 Hive 3.1.3 和 Iceberg 0.14.1。该功能支持仅限于 Hive 3.1.2 和 3.1.3 的 Iceberg 兼容功能。支持以下命令：

- 在 Amazon EMR 发行版 6.9.0 到 6.12.x 版本中，您必须将 `libfb303` jar 包含在 Hive `auxlib` 目录中。使用以下命令将其包含在内：

```
sudo /usr/bin/ln -sf /usr/lib/hive/lib/libfb303-*.jar /usr/lib/hive/auxlib/
libfb303.jar
```

  在 Amazon EMR 6.13 及更高版本中，`libfb303` jar 会自动符号链接到 Hive `auxlib` 目录。

- 创建表
  - 非分区表 - 可以通过提供存储处理程序在 Hive 中创建外部表，如下所示：

```
CREATE EXTERNAL TABLE x (i int) STORED BY
 'org.apache.iceberg.mr.hive.HiveIcebergStorageHandler'
```

  - 分区表 - 可以在 Hive 中创建外部分区表，如下所示：

```
CREATE EXTERNAL TABLE x (i int) PARTITIONED BY (j int) STORED BY
 'org.apache.iceberg.mr.hive.HiveIcebergStorageHandler'
```

> **Note**
>
> Hive 3 不支持 ORC/AVRO/PARQUET 的 `STORED AS` 文件格式。默认且唯一的选项是 Parquet。

- 删除表 - `DROP TABLE` 命令用于删除表，如以下示例中所示：

```
DROP TABLE [IF EXISTS] table_name [PURGE];
```

- 读取表 - `SELECT` 语句可用于读取 Hive 中的 Iceberg 表，如以下示例中所示。支持的执行引擎为 MR 和 Tez。

```
SELECT * FROM table_name
```

  有关 Hive 选择语法的信息，请参阅 LanguageManual Select。有关在 Hive 中使用 Iceberg 表的选择语句的信息，请参阅 Apache Iceberg Select。

- 插入到表中 - HiveQL 的 `INSERT INTO` 语句仅适用于支持 Map Reduce 执行引擎的 Iceberg 表。Amazon EMR 用户无需显式设置执行引擎，因为 Amazon EMR Hive 会在运行时为 Iceberg 表选择引擎。

- 单表插入 - 例如：

```
INSERT INTO table_name VALUES ('a', 1);
INSERT INTO table_name SELECT...;
```

- 多表插入 - 支持在语句中插入非原子多表。示例：

```
FROM source
 INSERT INTO table_1 SELECT a, b
 INSERT INTO table_2 SELECT c,d;
```

在 Amazon EMR 上使用 Iceberg 的注意事项和限制

本节包含将 Iceberg 与 Spark、Trino、Flink 和 Hive 结合使用的注意事项和限制。

将 Iceberg 与 Spark 结合使用的注意事项

- 原定设置下，Amazon EMR 6.5.0 不支持 Iceberg 在 Amazon EMR on EKS 上运行。Amazon EMR 6.5.0 自定义映像可供您传递 `--jars local:///usr/share/aws/iceberg/lib/iceberg-spark3-runtime.jar` 作为 `spark-submit` 参数，用于在 Amazon EMR on EKS 上创建 Iceberg 表。有关更多信息，请参阅《Amazon EMR on EKS 开发指南》中的使用自定义映像在 Amazon EMR 中提交 Spark 工作负载。您也可以联系 Amazon Web Services Support 获取帮助。从 Amazon EMR 6.6.0 开始，Amazon EMR on EKS 支持 Iceberg。
- 使用 Amazon Glue 作为 Iceberg 的目录时，请确保您在其中创建表的数据库存在于 Amazon Glue 中。如果您使用的是类似 Amazon Lake Formation 的服务并且无法加载目录，请确保您有访问该服务的适当权限来执行命令。

将 Iceberg 与 Trino 结合使用的注意事项

- Amazon EMR 6.5 不提供对 Iceberg 的原生 Trino Iceberg Catalog 支持。Trino 需要使用 Iceberg v0.11，因此我们建议为 Trino 启动独立于 Spark 集群的 Amazon EMR 集群，并在该集群上包括 Iceberg v0.11。
- 使用 Amazon Glue 作为 Iceberg 的目录时，请确保您在其中创建表的数据库存在于 Amazon Glue 中。如果您使用的是类似 Amazon Lake Formation 的服务并且无法加载目录，请确保您有访问该服务的适当权限来执行命令。

将 Iceberg 与 Flink 结合使用的注意事项

使用 Amazon Glue 作为 Iceberg 的目录时，请确保您在其中创建表的数据库存在于 Amazon Glue 中。如果您使用的是类似 Amazon Lake Formation 的服务并且无法加载目录，请确保您有访问该服务的适当权限来执行命令。

将 Iceberg 与 Hive 结合使用的注意事项

- Iceberg 支持以下查询类型：
  - 创建表
  - 删除表
  - 插入到表中
  - 读取表
- DML（数据操作语言）操作仅支持 MR（MapReduce）执行引擎，而 MR 在 Hive 3.1.3 中已弃用。
- 带有 Hive 的 Iceberg 目前不支持 Amazon Glue Data Catalog。
- 错误处理不够强大。在配置错误的情况下，插入查询可能会成功完成。但是，无法更新元数据可能会导致数据丢失。

Iceberg 发布历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Iceberg 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Iceberg 版本信息

Amazon EMR 发行版标签	Iceberg 版本	随 Iceberg 安装的组件
emr-6.14.0	1.3.1-amzn-0	Not available.
emr-6.13.0	1.3.0-amzn-1	Not available.
emr-6.12.0	1.3.0-amzn-0	Not available.

Amazon EMR 发行版标签	Iceberg 版本	随 Iceberg 安装的组件
emr-6.11.1	1.2.0-amzn-0	Not available.
emr-6.11.0	1.2.0-amzn-0	Not available.
emr-6.10.1	1.1.0-amzn-0	Not available.
emr-6.10.0	1.1.0-amzn-0	Not available.
emr-6.9.1	0.14.1-amzn-0	Not available.
emr-6.9.0	0.14.1-amzn-0	Not available.
emr-6.8.1	0.14.0-amzn-0	Not available.
emr-6.8.0	0.14.0-amzn-0	Not available.
emr-6.7.0	0.13.1-amzn-0	Not available.
emr-6.6.0	0.13.1	Not available.
emr-6.5.0	0.12.0	Not available.

Iceberg 发布说明（按版本分类）

- Amazon EMR 6.9.0 – Iceberg 发布说明

Amazon EMR 6.9.0 – Iceberg 发布说明

Amazon EMR 6.9.0 – Iceberg 更改

类型	描述
特征	Amazon EMR Flink 与 Iceberg 的集成。
特征	Amazon EMR Hive 与 Iceberg 的集成。

类型	描述
特征	支持在 Amazon FSx for Lustre 上缓存 Iceberg 元数据文件，以改进查询计划时间。
逆向移植	PR 5050：Flink 1.15：支持内联插入 SQL 注释中的写入选项。
逆向移植	PR 5282：Amazon：通过打开新的数据文件流，修复 PUT 重试失败。
逆向移植	PR 5318：Flink 1.15：缩小 FlinkSource 和 IcebergSource（FLIP-27）之间的差距，并新增了一个在 Flink SQL 中使用 FLIP-27 源的选择配置。
逆向移植	PR 5344：Flink 1.14：缩小 FlinkSource 和 IcebergSource（FLIP-27）之间的差距，并新增了一个在 Flink SQL 中使用 FLIP-27 源的选择配置。
逆向移植	PR 5393：Flink 1.14、Flink 1.15：避免在 FLIP-27 源读取器中将 Iceberg MetricContext 转换为 Flink 指标。
逆向移植	PR 5401：Flink 1.14、Flink1.15：PR #5393 中缺失 FLIP-27 源读取器指标的 IcebergSourceReader 组。
逆向移植	PR 5679：Spark 3.2、Spark 3.3：修复 MergeRows 节点的可空性传播。
逆向移植	PR 5860：Spark 3.3：修复在 Date 分区表上运行 RewriteManifestProcedure 时的 QueryFailure 问题。
逆向移植	PR 5880：Spark 3.3：修复读取时合并投影中的可空性。

类型	描述
逆向移植	PR 5917：Spark 3.2：修复读取时合并投影中的可空性。

Amazon EMR 上的 Jupyter 笔记本

Jupyter 笔记本是一个开源 Web 应用程序，可用于创建和共享包含实时代码、公式、可视化效果和叙述性文本的文档。Amazon EMR 为您提供了三个使用 Jupyter 笔记本的选项：

主题

- EMR Studio
- 基于 Jupyter 笔记本的 Amazon EMR Notebook
- JupyterHub

EMR Studio

Amazon EMR Studio 是一个基于 Web 的集成开发环境 (IDE)，适用于依托 Amazon EMR 集群运行的完全托管式 Jupyter 笔记本。您可以为团队设置 EMR Studio，以开发、可视化和调试用 R、Python、Scala 和 PySpark 编写的应用程序。

我们建议在 Amazon EMR 上使用 Jupyter 笔记本时使用 EMR Studio。详情请参见《Amazon EMR 管理指南》中的 EMR Studio。

基于 Jupyter 笔记本的 Amazon EMR Notebook

EMR Notebooks 是 Amazon EMR 控制台中内置的一个 Jupyter 笔记本环境，您可以在该环境中快速创建 Jupyter 笔记本，将它们连接到 Spark 集群，然后在控制台中打开 Jupyter 笔记本编辑器，以便远程运行查询和代码。EMR 笔记本独立于集群保存在 Amazon S3 中，可实现持久性存储、快速访问和灵活度。您可以打开多个笔记本、将多个笔记本连接到单个集群，以及在不同集群上重新使用笔记本。

详情请参见《Amazon EMR 管理指南》中的 EMR Notebooks。

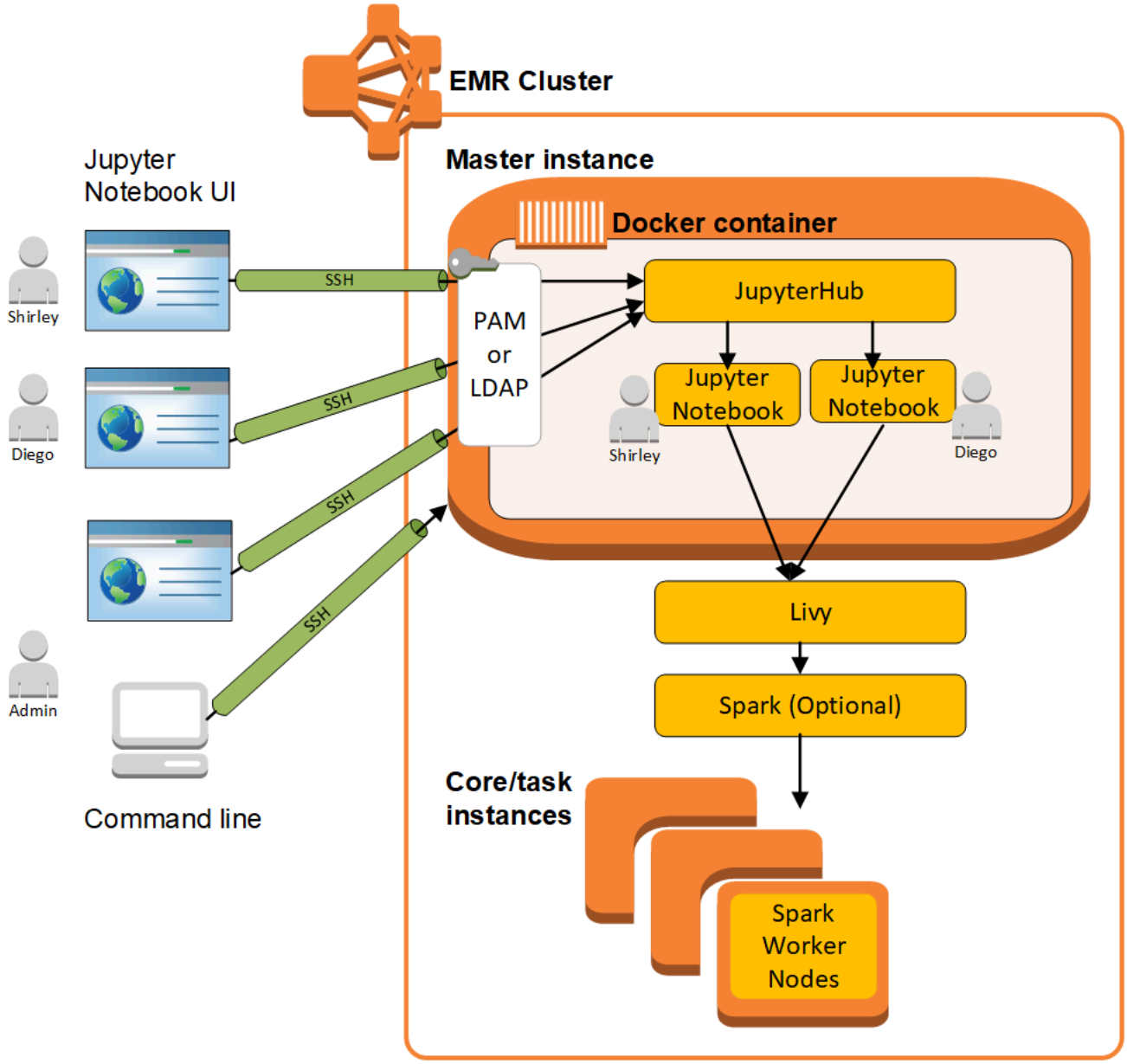
JupyterHub

Jupyter notebook 是一个开源 Web 应用程序，可用于创建和共享包含实时代码、公式、可视化效果和叙述性文本的文档。JupyterHub 允许托管单用户 Jupyter notebook 服务器的多个实例。使用 JupyterHub、Amazon EMR 创建集群时，将在集群的主节点上创建一个 Docker 容器。JupyterHub、Jupyter 需要的所有组件和 Sparkmagic 将在此容器内运行。

Sparkmagic 是内核库，内核允许 Jupyter notebook 通过 Apache Livy（适用于 Spark 的 REST 服务器）与在 Amazon EMR 上运行的 Apache Spark 通信。使用 JupyterHub 创建集群时，将自动安装

Spark 和 Apache Livy。适用于 Jupyter 的默认 Python 3 内核与可与 Sparkmagic 一起使用的 PySpark 3、PySpark 和 Spark 内核一起提供。通过使用 Python 和 Scala，可以使用这些内核运行临时 Spark 代码和交互式 SQL 查询。可以在 Docker 容器内手动安装其它内核。有关更多信息，请参阅安装其它内核和库。

下图描述了 JupyterHub on Amazon EMR 的组件以及笔记本用户和管理员对应的身份验证方法。有关更多信息，请参阅添加 Jupyter notebook 用户和管理员。



下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 JupyterHub 版本，以及 Amazon EMR 随 JupyterHub 一起安装的组件。

有关此发行版中随 JupyterHub 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 JupyterHub 版本信息

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.14.0	JupyterHub 1.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 JupyterHub 版本，以及 Amazon EMR 随 JupyterHub 一起安装的组件。

有关此发行版中随 JupyterHub 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 JupyterHub 版本信息

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.36.1	JupyterHub 1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server,

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
		spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

JupyterHub on Amazon EMR 附带的 Python 3 内核为 3.6.4。

在 Amazon EMR 版本和 Amazon EC2 AMI 版本之间，`jupyterhub` 容器内安装的库可能不同。

使用 **`conda`** 列出已安装的库

- 在主节点命令行上运行以下命令：

```
sudo docker exec jupyterhub bash -c "conda list"
```

使用 **`pip`** 列出已安装的库

- 在主节点命令行上运行以下命令：

```
sudo docker exec jupyterhub bash -c "pip freeze"
```

主题

- 使用 JupyterHub 创建集群
- 在 Amazon EMR 上使用 JupyterHub 时的注意事项
- 配置 JupyterHub
- 在 Amazon S3 中配置笔记本的持久性
- 连接到主节点和笔记本服务器
- JupyterHub 配置和管理
- 添加 Jupyter notebook 用户和管理员
- 安装其它内核和库
- JupyterHub 发行版历史记录

使用 JupyterHub 创建集群

您可以使用Amazon Web Services Management Console、Amazon Command Line Interface 或 Amazon EMR API 通过 JupyterHub 创建 Amazon EMR 集群。确保不使用在完成步骤后自动终止的选项（Amazon CLI 中的 `--auto-terminate` 选项）创建此集群。此外，确保管理员和笔记本用户可以访问创建集群时使用的密钥对。有关更多信息，请参阅《Amazon EMR 管理指南》中的对 SSH 凭证使用密钥对。

使用控制台创建已安装 JupyterHub 的集群

执行以下过程以使用 Amazon EMR 控制台中的 Advanced Options (高级选项) 创建已安装 JupyterHub 的集群。

使用 Amazon EMR 控制台创建已安装 JupyterHub 的 Amazon EMR 集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 在 Software Configuration (软件配置) 下：
   - 对于版本，选择 emr-5.36.1，然后选择 JupyterHub。
   - 如果您使用 Spark，要使用 Amazon Glue 数据目录作为 Spark SQL 的元存储，请选择 Use for Spark table metadata (用于 Spark 表元数据)。有关更多信息，请参阅使用 Amazon Glue 数据目录作为 Spark SQL 的元存储。
   - 对于 Edit software settings (编辑软件设置)，请选择 Enter configuration (输入配置) 并指定值，或选择 Load JSON from S3 (从 S3 加载 JSON) 并指定 JSON 配置文件。有关更多信息，请参阅配置 JupyterHub。
4. 在 Add steps (添加步骤) (可选) 下，配置创建集群后要运行的步骤，确保 Auto-terminate cluster after the last step is completed (完成最后的步骤后，自动终止集群) 未选中，然后选择 Next (下一步)。
5. 选择 Hardware Configuration (硬件配置) 选项、Next (下一步)。有关更多信息，请参阅《Amazon EMR 管理指南》中的配置集群硬件和联网。
6. 选择 General Cluster Settings (常规集群设置) 和 Next (下一步) 选项。
7. 选择 Security Options (安全选项) 以指定密钥对，然后选择 Create Cluster (创建集群)。

使用 Amazon CLI 创建已安装 JupyterHub 的集群

要启动已安装 JupyterHub 的集群，请使用 `aws emr create-cluster` 命令，对于 `--applications` 选项，请指定 `Name=JupyterHub`。以下示例启动 Amazon EMR 上包含两个 EC2 实例（一个是主实例，另一个是核心实例）的 JupyterHub 集群。此外，已启用调试，日志存储在 `--log-uri` 所指定的 Amazon S3 位置中。指定密钥对提供对集群中 Amazon EC2 实例的访问权限。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name="MyJupyterHubCluster" --release-label emr-5.36.1 \
--applications Name=JupyterHub --log-uri s3://MyBucket/MyJupyterClusterLogs \
--use-default-roles --instance-type m5.xlarge --instance-count 2 --ec2-attributes
 KeyName=MyKeyPair
```

在 Amazon EMR 上使用 JupyterHub 时的注意事项

使用 JupyterHub on Amazon EMR 时，请注意以下事项。

- > **Warning**
  >
  > 用户笔记本和文件将保存到主节点上的文件系统中。这是短暂存储，在集群终止后将不复存在。集群终止后，此数据如果未备份的化将丢失。建议使用 `cron` 作业或其它适用于应用程序的方式安排定期备份。
  >
  > 此外，如果容器重启，在容器内进行的配置更改可能不复存在。建议为容器配置编写脚本或以其它方式实现容器配置的自动化，以便可以更轻松地重现自定义。
- 不支持使用 Amazon EMR 安全配置设置的 Kerberos 身份验证。
- 不支持 OAuthenticator。

配置 JupyterHub

通过连接到集群主节点并编辑配置文件，可以自定义 JupyterHub on Amazon EMR 和独立用户笔记本的配置。在更改值之后，重启 `jupyterhub` 容器。

修改以下文件中的属性以配置 JupyterHub 和独立 Jupyter notebook：

- `jupyterhub_config.py` – 默认情况下，此文件保存在主节点上的 `/etc/jupyter/conf/` 目录中。有关更多信息，请参阅 JupyterHub 文档中的配置基础知识。
- `jupyter_notebook_config.py` – 默认情况下，该文件保存在 `/etc/jupyter/` 目录中，并作为默认值复制到 `jupyterhub` 容器中。有关更多信息，请参阅 Jupyter notebook 文档中的配置文件和命令行选项。

您也可以使用 `jupyter-sparkmagic-conf` 配置分类自定义 Sparkmagic，这会更新 Sparkmagic 的 `config.json` 文件中的值。有关可用设置的更多信息，请参阅 GitHub 上的 example_config.json。有关在 Amazon EMR 中对应用程序使用配置分类的更多信息，请参阅配置应用程序。

以下示例使用 Amazon CLI 启动集群，以引用 Sparkmagic 配置分类设置的文件 `MyJupyterConfig.json`。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --use-default-roles --release-label emr-5.14.0 \
--applications Name=Jupyter --instance-type m4.xlarge --instance-count 3 \
--ec2-attributes KeyName=MyKey,SubnetId=subnet-1234a5b6 --configurations file://
MyJupyterConfig.json
```

`MyJupyterConfig.json` 的示例内容如下所示：

```
[
    {
    "Classification":"jupyter-sparkmagic-conf",
    "Properties": {
      "kernel_python_credentials" : "{\"username\":\"diego\",\"base64_password\":
\"mypass\",\"url\":\"http:\/\/localhost:8998\",\"auth\":\"None\"}"
      }
    }
]
```

> **Note**
>
> 对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

在 Amazon S3 中配置笔记本的持久性

您可以在 Amazon EMR 中配置 JupyterHub 集群，以便用户保存的笔记本保存在 Amazon S3（集群 EC2 实例上的短暂存储之外）中。

在创建集群时使用 `jupyter-s3-conf` 配置分类指定 Amazon S3 持久性。有关更多信息，请参阅配置应用程序。

除了使用 `s3.persistence.enabled` 属性启用 Amazon S3 持久性之外，还请使用 `s3.persistence.bucket` 属性在保存笔记本的 Amazon S3 中指定存储桶。每个用户的笔记本将保存到指定存储桶中的 `jupyter/`*`jupyterhub-user-name`* 文件夹。该存储桶必须已存在于 Amazon S3 中，并且您在创建集群时指定的 EC2 实例配置文件的角色必须对此存储桶具有权限（默认情况下，此角色为 `EMR_EC2_DefaultRole`）。有关更多信息，请参阅为Amazon服务的 Amazon EMR 权限配置 IAM 角色。

当您使用相同的配置分类属性启动新集群时，用户可以打开内容来自已保存位置的笔记本。

请注意，当您启用了 Amazon S3 后，如果将文件作为模块导入到笔记本中，这会导致文件上载到 Amazon S3。导入文件而不启用 Amazon S3 持久性时，这些文件会上载到 JupyterHub 容器。

以下示例启用 Amazon S3 持久性。用户保存的笔记本保存在每个用户的 `s3://MyJupyterBackups/jupyter/`*`jupyterhub-user-name`* 文件夹中，其中 *`jupyterhub-user-name`* 是一个用户名（如 `diego`）。

```
[
    {
        "Classification": "jupyter-s3-conf",
        "Properties": {
            "s3.persistence.enabled": "true",
            "s3.persistence.bucket": "MyJupyterBackups"
        }
    }
```

```
]
```

连接到主节点和笔记本服务器

JupyterHub 管理员和笔记本用户必须使用 SSH 隧道连接到集群主节点，然后连接到主节点上 JupyterHub 服务的 Web 接口。有关配置 SSH 隧道和使用此隧道代理 Web 连接的更多信息，请参阅《Amazon EMR 管理指南》中的连接到集群。

默认情况下，JupyterHub on Amazon EMR 是通过主节点上的端口 9443 提供的。内部 JupyterHub 代理也通过端口 9443 为笔记本实例服务。JupyterHub 和 Jupyter Web 接口可通过以下模式的 URL 进行访问：

https://*MasterNodeDNS*:9443

可以使用 `c.JupyterHub.port` 文件中的 `jupyterhub_config.py` 属性指定不同的端口。有关更多信息，请参阅 JupyterHub 文档中的联网基础知识。

默认情况下，JupyterHub on Amazon EMR 对使用 HTTPS 的 SSL 加密使用自签名凭证。用户连接时，系统将提示用户信任自签名凭证。可以使用自己的受信任凭证和密钥。将主节点上 `server.crt` 目录中的默认凭证文件 `server.key` 和密钥文件 `/etc/jupyter/conf/` 更换为自己的凭证和密钥文件。使用 `c.JupyterHub.ssl_key` 文件中的 `c.JupyterHub.ssl_cert` 和 `jupyterhub_config.py` 属性指定 SSL 材料。有关更多信息，请参阅 JupyterHub 文档中的安全性设置。在更新 `jupyterhub_config.py` 之后，重启容器。

JupyterHub 配置和管理

JupyterHub 和相关组件在名为 `jupyterhub`、运行 Ubuntu 操作系统的 Docker 容器内运行。有多种方法可用于管理此容器内运行的组件。

> **Warning**
>
> 在此容器内执行的自定义将在此容器重启后不复存在。建议为容器配置编写脚本或以其它方式实现容器配置的自动化，以便可以更轻松地重现自定义。

使用命令行管理

当使用 SSH 连接到主节点后，可以通过使用 Docker 命令行界面 (CLI) 并按名称 (`jupyterhub`) 或 ID 指定容器来发出命令。例如，`sudo docker exec jupyterhub` *`command`* 将运行容器内运行的操作系统或应用程序识别的命令。可以使用此方法将用户添加到操作系统和在 Docker 容器内安装其它应

用程序和库。例如，默认容器映像包括用于安装软件包的 Conda，因此可能在主节点命令行上运行以下命令以在容器内安装应用程序 Keras：

```
sudo docker exec jupyterhub conda install keras
```

通过提交步骤管理

步骤是将工作提交到集群的一种方式。可以在启动集群时提交步骤，也可以将步骤提交给正在运行的集群。可以使用 `command-runner.jar` 将在命令行上运行的命令作为步骤提交。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 CLI 和控制台执行步骤，以及在 Amazon EMR 集群上运行命令和脚本。

例如，可以在本地计算机上使用以下 Amazon CLI 命令安装 Keras，方式与之前示例中通过主节点的命令行执行的方式一样：

```
aws emr add-steps --cluster-id MyClusterID --steps Name="Command
 Runner",Jar="command-runner.jar",Args="/usr/bin/sudo","/usr/bin/
docker","exec","jupyterhub","conda","install","keras"
```

此外，可以为步骤的序列编写脚本，并将脚本上载到 Amazon S3，然后在创建集群或将脚本作为步骤添加时使用 `script-runner.jar` 运行脚本。有关更多信息，请参阅在 Amazon EMR 集群上运行命令和脚本。有关示例，请参阅 the section called “示例：用于添加多个用户的清除脚本”。

使用 REST API 管理

适用于 JupyterHub 的 Jupyter、JupyterHub 和 HTTP 代理提供可用于发送请求的 REST API。要将请求发送到 JupyterHub，必须随请求传递一个 API 令牌。可以从主节点命令行使用 `curl` 命令执行 REST 命令。有关更多信息，请参阅以下资源：

- JupyterHub 文档中的使用 JupyterHub 的 REST API（包括 API 令牌生成说明）
- GitHub 上的 Jupyter notebook 服务器 API
- GitHub 上的 configurable-http-proxy

以下示例演示了使用 JupyterHub 的 REST API 获取用户列表。此命令传递以前生成的管理员令牌并使用 JupyterHub 的默认端口 9443，以将输出传输到 jq 来方便查看：

```
curl -XGET -s -k https://$HOST:9443/hub/api/users \
-H "Authorization: token $admin_token" | jq .
```

添加 Jupyter notebook 用户和管理员

可以使用两种方式之一为至 JupyterHub 的用户执行身份验证，以便它们可创建笔记本以及（可选）管理 JupyterHub。最简单的方法是，使用 JupyterHub 的可插入验证模块 (PAM)。此外，JupyterHub on Amazon EMR 支持适用于 JupyterHub 的 LDAP Authenticator 插件，用于从 LDAP 服务器（如 Microsoft Active Directory Server）获取用户身份。此部分提供了通过每种身份验证方法添加用户的说明和示例。

JupyterHub on Amazon EMR 包含一个具有管理员权限的默认用户。此用户名为 `jovyan`，密码为 `jupyter`。强烈建议将此用户替换为另一个具有管理员权限的用户。您可以在创建集群时使用一个步骤以执行该操作，也可以在集群运行时连接到主节点以执行该操作。

主题

- 使用 PAM 身份验证
- 使用 LDAP 身份验证
- 用户模拟

使用 PAM 身份验证

在 JupyterHub on Amazon EMR 中创建 PAM 用户是一个两步过程。第一步是，将用户添加到在主节点上的 `jupyterhub` 容器中运行的操作系统，以及为每个用户添加一个相应的用户主目录。第二步是，将这些操作系统用户作为 JupyterHub 用户添加到 JupyterHub 中，这一流程称为加入白名单。添加 JupyterHub 用户之后，这些用户可连接到 JupyterHub URL 并提供其操作系统凭证以便访问。

用户登录后，JupyterHub 将为用户打开保存在主节点上用户主目录 `/var/lib/jupyter/home/`*`username`* 中的笔记本服务器实例。如果笔记本服务器实例不存在，则 JupyterHub 将在用户的主目录中生成一个笔记本实例。以下部分演示如何将用户分别添加到操作系统和 JupyterHub，后接用于添加多个用户的基本清除脚本。

将操作系统用户添加到容器

以下示例先在容器内使用 useradd 命令添加单个用户 diego 并为该用户创建一个主目录。第二个命令使用 chpasswd 为此用户设置密码 diego。在使用 SSH 连接时，命令将在主节点命令行上运行。还可以使用步骤运行这些命令，如之前的通过提交步骤管理中所述。

```
sudo docker exec jupyterhub useradd -m -s /bin/bash -N diego
sudo docker exec jupyterhub bash -c "echo diego:diego | chpasswd"
```

添加 JupyterHub 用户

可以使用 JupyterHub 中的 Admin (管理员) 面板或 REST API 添加用户和管理员，或仅添加用户。

使用 JupyterHub 中的“Admin (管理员)”面板添加用户和管理员

1. 使用 SSH 连接到主节点并使用具有管理员权限的身份登录 https://*MasterNodeDNS*:9443。
2. 选择 Control Panel (控制面板)、Admin (管理员)。
3. 选择 User (用户)、Add Users (添加用户)，或选择 Admin (管理员)、Add Admins (添加管理员)。

使用 REST API 添加用户

1. 使用 SSH 连接到主节点并在主节点上使用以下命令，或将此命令作为步骤运行。
2. 获取管理员令牌以发出 API 请求并将以下步骤中的 *AdminToken* 替换为管理员令牌。
3. 使用以下命令，以将 *UserName* 替换为容器内已创建的操作系统用户。

```
curl -XPOST -H "Authorization: token AdminToken" "https://$(hostname):9443/hub/api/
users/UserName
```

> **Note**
>
> 首次登录 JupyterHub Web 界面时，系统会自动将您添加为 JupyterHub 非管理员用户。

示例：用于添加多个用户的清除脚本

以下示例清除脚本将与此部分中的前述步骤配合使用，以创建多个 JupyterHub 用户。此脚本可以直接在主节点上运行，也可上载到 Amazon S3 并在之后作为步骤运行。

此脚本先建立一组用户名，并使用 `jupyterhub token` 命令为默认管理员 jovyan 创建一个 API 令牌。然后，它在 `jupyterhub` 容器中为每个用户创建一个操作系统用户，以为每个用户分配一个与其用户名相同的初始密码。最后，它调用 REST API 操作以在 JupyterHub 中创建每个用户。它在脚本中传递之前生成的令牌并将 REST 响应传输到 `jq` 以方便查看。

```
# Bulk add users to container and JupyterHub with temp password of username
set -x
USERS=(shirley diego ana richard li john mary anaya)
TOKEN=$(sudo docker exec jupyterhub /opt/conda/bin/jupyterhub token jovyan | tail -1)
```

```
for i in "${USERS[@]}";
do
   sudo docker exec jupyterhub useradd -m -s /bin/bash -N $i
   sudo docker exec jupyterhub bash -c "echo $i:$i | chpasswd"
   curl -XPOST --silent -k https://$(hostname):9443/hub/api/users/$i \
 -H "Authorization: token $TOKEN" | jq
done
```

将此脚本保存到 Amazon S3 中的位置（如 `s3://mybucket/createjupyterusers.sh`）。然后，可以使用 `script-runner.jar` 将此脚本作为步骤运行。

示例：创建集群时运行脚本（Amazon CLI）

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name="MyJupyterHubCluster" --release-label emr-5.36.1 \
--applications Name=JupyterHub --log-uri s3://MyBucket/MyJupyterClusterLogs \
--use-default-roles --instance-type m5.xlarge --instance-count 2 --ec2-attributes
 KeyName=MyKeyPair \
--steps Type=CUSTOM_JAR,Name=CustomJAR,ActionOnFailure=CONTINUE,\
Jar=s3://region.elasticmapreduce/libs/script-runner/script-runner.jar,Args=["s3://
mybucket/createjupyterusers.sh"]
```

在现有集群上运行脚本（Amazon CLI）

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr add-steps --cluster-id j-XXXXXXXX --steps Type=CUSTOM_JAR,\
Name=CustomJAR,ActionOnFailure=CONTINUE,\
Jar=s3://region.elasticmapreduce/libs/script-runner/script-runner.jar,Args=["s3://
mybucket/createjupyterusers.sh"]
```

使用 LDAP 身份验证

轻型目录访问协议 (LDAP) 是一种应用程序协议，用于查询和修改与 LDAP 兼容目录服务提供程序（如 Active Directory 或 OpenLDAP server）中存储的资源（如用户和计算机）对应的对象。可以结合使用适用于 JupyterHub 的 LDAP Authenticator 插件和 JupyterHub on Amazon EMR 来使用 LDAP 执行用户身份验证。此插件处理 LDAP 用户的登录会话并为 Jupyter 提供用户信息。这使用户能够通过使用存储在 LDAP 兼容服务器中的身份的凭证连接到 JupyterHub 和笔记本。

此部分中的步骤演练以下使用适用于 JupyterHub 的 LDAP Authenticator 插件设置和启用 LDAP 的步骤。在连接到主节点命令行时执行这些步骤。有关更多信息，请参阅连接到主节点和笔记本服务器。

1. 创建一个包含 LDAP 服务器相关信息（如主机 IP 地址、端口、绑定名称等）的 LDAP 配置文件。
2. 修改 `/etc/jupyter/conf/jupyterhub_config.py` 以启用适用于 JupyterHub 的 LDAP Authenticator 插件。
3. 创建并运行在 `jupyterhub` 容器内配置 LDAP 的脚本。
4. 在 LDAP 中查询用户，然后在容器内为所有用户创建主目录。JupyterHub 需要主目录托管笔记本。
5. 运行重启 JupyterHub 的脚本

> **Important**
>
> 在设置 LDAP 之前，测试网络基础设施以确保 LDAP 服务器和集群主节点可以根据需要进行通信。TLS 一般通过普通 TCP 连接使用端口 389。如果 LDAP 连接使用 SSL，那么显而易见，适用于 SSL 的 TCP 端口为 636。

创建 LDAP 配置文件

下方的示例使用以下占位符配置值。将这些值替换为与您的实施匹配的参数。

- 正在运行的是 LDAP 服务器版本 3，通过端口 389 提供。这是适用于 LDAP 的标准非 SSL 端口。
- 基本可分辨名称 (DN) 为 `dc=example, dc=org`。

使用文本编辑器创建内容与下类似的 ldap.conf 文件。使用适用于 LDAP 实施的值。将 *`host`* 替换为 LDAP 服务器的 IP 地址或可解析主机名。

```
base dc=example,dc=org
uri ldap://host
```

```
ldap_version 3
binddn cn=admin,dc=example,dc=org
bindpw admin
```

启用适用于 JupyterHub 的 LDAP Authenticator 插件

使用文本编辑器修改 /etc/jupyter/conf/jupyterhub_config.py 文件并添加与下类似的 ldapauthenticator 属性。将 *host* 替换为 LDAP 服务器的 IP 地址或可解析主机名。此示例假定用户对象位于名为 *people* 的组织部门内，并使用之前使用 ldap.conf 建立的可分辨名称组件。

```
c.JupyterHub.authenticator_class = 'ldapauthenticator.LDAPAuthenticator'
c.LDAPAuthenticator.use_ssl = False
c.LDAPAuthenticator.server_address = 'host'
c.LDAPAuthenticator.bind_dn_template = 'cn={username},ou=people,dc=example,dc=org'
```

在容器内配置 LDAP

使用文本编辑器创建包含以下内容的清除脚本：

```
#!/bin/bash

# Uncomment the following lines to install LDAP client libraries only if
# using Amazon EMR release version 5.14.0. Later versions install libraries by default.
# sudo docker exec jupyterhub bash -c "sudo apt-get update"
# sudo docker exec jupyterhub bash -c "sudo apt-get -y install libnss-ldap libpam-ldap
 ldap-utils nscd"

# Copy ldap.conf
sudo docker cp ldap.conf jupyterhub:/etc/ldap/
sudo docker exec jupyterhub bash -c "cat /etc/ldap/ldap.conf"

# configure nss switch
sudo docker exec jupyterhub bash -c "sed -i 's/\(^passwd.*\)/\1 ldap/g' /etc/
nsswitch.conf"
sudo docker exec jupyterhub bash -c "sed -i 's/\(^group.*\)/\1 ldap/g' /etc/
nsswitch.conf"
sudo docker exec jupyterhub bash -c "sed -i 's/\(^shadow.*\)/\1 ldap/g' /etc/
nsswitch.conf"
sudo docker exec jupyterhub bash -c "cat /etc/nsswitch.conf"

# configure PAM to create home directories
sudo docker exec jupyterhub bash -c "echo 'session required        pam_mkhomedir.so
 skel=/etc/skel umask=077' >> /etc/pam.d/common-session"
```

```
sudo docker exec jupyterhub bash -c "cat /etc/pam.d/common-session"

# restart nscd service
sudo docker exec jupyterhub bash -c "sudo service nscd restart"

# Test
sudo docker exec jupyterhub bash -c "getent passwd"

# Install ldap plugin
sudo docker exec jupyterhub bash -c "pip install jupyterhub-ldapauthenticator"
```

将脚本保存到主节点，然后从主节点命令行运行它。例如，对于另存为 `configure_ldap_client.sh` 的脚本，使此文件成为可执行文件：

```
chmod +x configure_ldap_client.sh
```

并运行此脚本：

```
./configure_ldap_client.sh
```

将属性添加到 Active Directory

要查找每个用户并在数据库中创建相应的条目，JupyterHub docker 容器需要 Active Directory 中相应用户对象的以下 UNIX 属性。有关更多信息，请参阅文章 Clarification regarding the status of identity management for Unix (IDMU) and NIS server role in Windows Server 2016 technical preview and beyond 中的 How do I continue to edit the GID/UID RFC 2307 attributes now that the Unix Attributes Plug-in is no longer available for the Active Directory Users and Computers MMC snap-in? 部分。

- `homeDirectory`

  这是用户主目录的位置，通常是 /home/*username*。

- `gidNumber`

  这是一个大于 60000 的值，尚未被其它用户使用。检查 `etc/passwd` 文件中是否有正在使用的 GID。

- `uidNumber`

  这是一个大于 60000 的值，尚未被其它组使用。检查 `etc/group` 文件中是否有正在使用的 UID。

- `uid`

这与 *username* 相同。

创建用户主目录

JupyterHub 需要容器内的主目录才能验证 LDAP 用户的身份和存储实例数据。以下示例演示了 LDAP 目录中的两个用户 shirley 和 diego。

第一步是在 LDAP 服务器中使用 ldapsearch 查询每个用户的用户 ID 和组 ID 信息（如以下示例中所示），以将 *host* 替换为 LDAP 服务器的 IP 地址和可解析的主机名：

```
ldapsearch -x -H ldap://host \
 -D "cn=admin,dc=example,dc=org" \
 -w admin \
 -b "ou=people,dc=example,dc=org" \
 -s sub \
 "(objectclass=*)" uidNumber gidNumber
```

此 `ldapsearch` 命令将为 shirley 和 diego 用户返回看上去与下类似的 LDIF 格式的响应。

```
# extended LDIF

# LDAPv3
# base <ou=people,dc=example,dc=org> with scope subtree
# filter: (objectclass=*)
# requesting: uidNumber gidNumber sn

# people, example.org
dn: ou=people,dc=example,dc=org

# diego, people, example.org
dn: cn=diego,ou=people,dc=example,dc=org
sn: B
uidNumber: 1001
gidNumber: 100

# shirley, people, example.org
dn: cn=shirley,ou=people,dc=example,dc=org
sn: A
uidNumber: 1002
gidNumber: 100
```

```
# search result
search: 2
result: 0 Success

# numResponses: 4
# numEntries: 3
```

通过使用响应中的信息，在容器内运行命令以为每个用户公用名 (cn) 创建一个主目录。使用 `uidNumber` 和 `gidNumber` 确定用户对主目录的所有权。以下示例命令将为用户 *`shirley`* 执行此操作。

```
sudo docker container exec jupyterhub bash -c "mkdir /home/shirley"
sudo docker container exec jupyterhub bash -c "chown -R $uidNumber /home/shirley"
sudo docker container exec jupyterhub bash -c "sudo chgrp -R $gidNumber /home/shirley"
```

> **Note**
>
> JupyterHub 的 LDAP 身份验证器不支持本地用户创建。有关更多信息，请参阅关于本地用户创建的 LDAP 身份验证器配置说明。
> 要手动创建本地用户，请使用以下命令。
>
> ```
> sudo docker exec jupyterhub bash -c "echo 'shirley:x:$uidNumber:$gidNumber::/
> home/shirley:/bin/bash' >> /etc/passwd"
> ```

重新启动 Jupyterhub 容器

运行以下命令重新启动 `jupyterhub` 容器：

```
sudo docker stop jupyterhub
sudo docker start jupyterhub
```

用户模拟

在 Amazon EMR 上执行期间，在 Jupyter notebook 中运行的 Spark 作业将访问多个应用程序。例如，Sparkmagic 接收用户在 Jupyter 中运行的 PySpark3 代码，Sparkmagic 使用 HTTP POST 请求将其提交到 Livy，然后 Livy 创建一个 Spark 作业以使用 YARN 在集群上执行。

默认情况下，以这种方式提交的 YARN 作业以 `livy` 用户身份运行，而不管启动该作业的用户如何。通过设置用户模拟，您也可以将笔记本用户的用户 ID 作为与 YARN 作业关联的用户。每个用户启动的

作业分别与 shirley 和 diego 相关联，而不是由与 livy 用户关联的 shirley 和 diego 同时启动作业。这有助于审核 Jupyter 使用情况以及在组织中管理应用程序。

只有在从 Sparkmagic 到 Livy 的调用未进行身份验证时，才支持该配置。不支持在 Hadoop 应用程序和 Livy 之间提供身份验证或代理层的应用程序（如 Apache Knox Gateway）。在本节中配置用户模拟的步骤假定 JupyterHub 和 Livy 在同一主节点上运行。如果您的应用程序具有单独的集群，则需要修改步骤 3：为用户创建 HDFS 主目录，以便在 Livy 主节点上创建 HDFS 目录。

配置用户模拟的步骤

- 步骤 1：配置 Livy
- 步骤 2：添加用户
- 步骤 3：为用户创建 HDFS 主目录

步骤 1：配置 Livy

在创建集群时，您可以使用 livy-conf 和 core-site 配置分类启用 Livy 用户模拟，如以下示例所示。将配置分类保存为 JSON，然后在创建集群时引用该分类，或者指定内联的配置分类。有关更多信息，请参阅配置应用程序。

```
[
  {
    "Classification": "livy-conf",
    "Properties": {
      "livy.impersonation.enabled": "true"
    }
  },
  {
    "Classification": "core-site",
    "Properties": {
      "hadoop.proxyuser.livy.groups": "*",
      "hadoop.proxyuser.livy.hosts": "*"
    }
  }
]
```

步骤 2：添加用户

使用 PAM 或 LDAP 添加 JupyterHub 用户。有关更多信息，请参阅使用 PAM 身份验证和使用 LDAP 身份验证。

步骤 3：为用户创建 HDFS 主目录

您已连接到主节点以创建用户。在仍连接到主节点时，复制以下内容并将其保存到脚本文件中。该脚本为主节点上的每个 JupyterHub 用户创建 HDFS 主目录。该脚本假定您使用默认管理员用户 ID *jovyan*。

```
#!/bin/bash

CURL="curl --silent -k"
HOST=$(curl -s http://169.254.169.254/latest/meta-data/local-hostname)

admin_token() {
    local user=jovyan
    local pwd=jupyter
    local token=$($CURL https://$HOST:9443/hub/api/authorizations/token \
        -d "{\"username\":\"$user\", \"password\":\"$pwd\"}" | jq ".token")
    if [[ $token != null ]]; then
        token=$(echo $token | sed 's/"//g')
    else
        echo "Unable to get Jupyter API Token."
        exit 1
    fi
    echo $token
}

# Get Jupyter Admin token
token=$(admin_token)

# Get list of Jupyter users
users=$(curl -XGET -s -k https://$HOST:9443/hub/api/users \
 -H "Authorization: token $token" | jq '.[].name' | sed 's/"//g')

# Create HDFS home dir
for user in ${users[@]};
do
 echo "Create hdfs home dir for $user"
 hadoop fs -mkdir /user/$user
 hadoop fs -chmod 777 /user/$user
done
```

安装其它内核和库

当创建包含 JupyterHub on Amazon EMR 的集群时，将在 Docker 容器上安装适用于 Jupyter 的默认 Python 3 内核和适用于 Sparkmagic 的 PySpark 和 Spark 内核。可以安装其它内核。还可以安装其它库和软件包，然后将它们导入相应的 shell。

安装内核

内核安装在 Docker 容器中。安装内核最简单的方式是，创建包含安装命令的清除脚本，将脚本保存到主节点，然后使用 `sudo docker exec jupyterhub` *`script_name`* 命令以在 `jupyterhub` 容器内运行脚本。以下示例脚本安装内核，然后在主节点上安装内核的一些库，以便之后在 Jupyter 中使用内核时可以导出库。

```
#!/bin/bash

# Install Python 2 kernel
conda create -n py27 python=2.7 anaconda
source /opt/conda/envs/py27/bin/activate
apt-get update
apt-get install -y gcc
/opt/conda/envs/py27/bin/python -m pip install --upgrade ipykernel
/opt/conda/envs/py27/bin/python -m ipykernel install

# Install libraries for Python 2
/opt/conda/envs/py27/bin/pip install paramiko nltk scipy numpy scikit-learn pandas
```

要在容器内安装内核和库，请打开至主节点的终端连接，将脚本保存到 `/etc/jupyter/install_kernels.sh`，然后在主节点命令行上运行以下命令：

```
sudo docker exec jupyterhub bash /etc/jupyter/install_kernels.sh
```

使用库和安装其它库

JupyterHub on Amazon EMR 上预安装有适用于 Python 3 的一组核心的机器学习和数据科学库。可以使用 `sudo docker exec jupyterhub bash -c "conda list"` 和 `sudo docker exec jupyterhub bash -c "pip freeze"`。

如果 Spark 作业需要 Worker 节点上的库，建议使用引导操作运行脚本以在创建集群时安装库。集群创建过程中，引导操作将在所有集群节点上运行，这将简化安装。如果于集群运行后在核心/Worker 节点上安装库，则操作更复杂。我们在此部分中提供了示例 Python 程序以演示如何安装这些库。

此部分中演示的引导操作和 Python 程序示例都使用保存到 Amazon S3 的清除脚本在所有节点上安装库。

以下示例中引用的脚本将通过 pip 安装适用于 Python 3 内核的 paramiko、nltk、scipy、scikit-learn 和 pandas：

```
#!/bin/bash

sudo python3 -m pip install boto3 paramiko nltk scipy scikit-learn pandas
```

创建脚本后，将其上载到 Amazon S3 中的位置（例如，s3://mybucket/install-my-jupyter-libraries.sh）。有关更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的上传对象，以便可以在引导操作或 Python 程序中使用此操作。

指定将在使用 Amazon CLI 创建集群时在所有节点上安装库的引导操作

1. 创建与之前的示例类似的脚本并将脚本保存在 Amazon S3 中的位置。我们将使用示例 s3://mybucket/install-my-jupyter-libraries.sh。
2. 创建已安装 JupyterHub 的集群并使用 --bootstrap-actions 选项的 Path 参数指定脚本位置，如以下示例所示：

   > **Note**
   >
   > 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

   ```
   aws emr create-cluster --name="MyJupyterHubCluster" --release-label emr-5.36.1 \
   --applications Name=JupyterHub --log-uri s3://MyBucket/MyJupyterClusterLogs \
   --use-default-roles --instance-type m5.xlarge --instance-count 2 --ec2-attributes
    KeyName=MyKeyPair \
   --bootstrap-actions Path=s3://mybucket/install-my-jupyter-
   libraries.sh,Name=InstallJupyterLibs
   ```

指定将在使用控制台创建集群时在所有节点上安装库的引导操作

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。

2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 根据应用程序的情况，指定 Software and Steps (软件和步骤) 和 Hardware (硬件) 的设置。
4. 在 General Cluster Settings (常规集群设置) 屏幕上，展开 Bootstrap Actions (引导操作)。
5. 对于 Add bootstrap action (添加引导操作)，选择 Custom action (自定义操作)、Configure and add (配置和添加)。
6. 对于 Name (名称)，输入一个易于理解的名称。对于 Script location (脚本位置)，输入 Amazon S3 中脚本的位置（我们使用的示例为 s3://mybucket/install-my-jupyter-libraries.sh）。保留 Optional arguments (可选参数) 为空，然后选择 Add (添加)。
7. 指定集群的其它设置，然后选择 Next (下一步)。
8. 指定安全设置，然后选择 Create cluster (创建集群)。

Example 在运行集群的核心节点上安装库

在 Jupyter 内的主节点上安装库之后，可以通过不同的方式将库安装在运行的核心节点上。以下示例显示了编写为在本地计算机上运行的 Python 程序。当在本地运行 Python 程序时，此程序将使用 Amazon Systems Manager 的 `AWS-RunShellScript` 运行此节中之前所示的示例脚本，从而在集群的核心节点上安装库。

```
import argparse
import time
import boto3


def install_libraries_on_core_nodes(cluster_id, script_path, emr_client, ssm_client):
    """
    Copies and runs a shell script on the core nodes in the cluster.

    :param cluster_id: The ID of the cluster.
    :param script_path: The path to the script, typically an Amazon S3 object URL.
    :param emr_client: The Boto3 Amazon EMR client.
    :param ssm_client: The Boto3 AWS Systems Manager client.
    """
    core_nodes = emr_client.list_instances(
        ClusterId=cluster_id, InstanceGroupTypes=["CORE"]
    )["Instances"]
    core_instance_ids = [node["Ec2InstanceId"] for node in core_nodes]
    print(f"Found core instances: {core_instance_ids}.")

    commands = [
```

```
        # Copy the shell script from Amazon S3 to each node instance.
        f"aws s3 cp {script_path} /home/hadoop",
        # Run the shell script to install libraries on each node instance.
        "bash /home/hadoop/install_libraries.sh",
    ]
    for command in commands:
        print(f"Sending '{command}' to core instances...")
        command_id = ssm_client.send_command(
            InstanceIds=core_instance_ids,
            DocumentName="AWS-RunShellScript",
            Parameters={"commands": [command]},
            TimeoutSeconds=3600,
        )["Command"]["CommandId"]
        while True:
            # Verify the previous step succeeded before running the next step.
            cmd_result = ssm_client.list_commands(CommandId=command_id)["Commands"][0]
            if cmd_result["StatusDetails"] == "Success":
                print(f"Command succeeded.")
                break
            elif cmd_result["StatusDetails"] in ["Pending", "InProgress"]:
                print(f"Command status is {cmd_result['StatusDetails']}, waiting...")
                time.sleep(10)
            else:
                print(f"Command status is {cmd_result['StatusDetails']}, quitting.")
                raise RuntimeError(
                    f"Command {command} failed to run. "
                    f"Details: {cmd_result['StatusDetails']}"
                )


def main():
    parser = argparse.ArgumentParser()
    parser.add_argument("cluster_id", help="The ID of the cluster.")
    parser.add_argument("script_path", help="The path to the script in Amazon S3.")
    args = parser.parse_args()

    emr_client = boto3.client("emr")
    ssm_client = boto3.client("ssm")

    install_libraries_on_core_nodes(
        args.cluster_id, args.script_path, emr_client, ssm_client
    )
```

```
if __name__ == "__main__":
    main()
```

JupyterHub 发行版历史记录

下表列出了 Amazon EMR 每个发行版本中包含的 JupyterHub 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

JupyterHub 版本信息

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.14.0	1.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.13.0	1.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server,

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
		spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.12.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.11.1	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.11.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.10.1	1.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.10.0	1.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.9.1	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.9.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.8.1	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.8.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.7.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.36.1	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.36.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.6.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.35.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.5.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.4.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.3.1	1.2.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.3.0	1.2.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.2.1	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.2.0	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.1.1	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.1.0	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-6.0.1	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-6.0.0	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.34.0	1.4.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.33.1	1.2.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.33.0	1.2.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.32.1	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.32.0	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.31.1	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.31.0	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.30.2	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.30.1	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.30.0	1.1.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.29.0	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.28.1	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.28.0	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.27.1	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.27.0	1.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.26.0	0.9.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.25.0	0.9.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.24.1	0.9.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.24.0	0.9.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.23.1	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.23.0	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.22.0	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.21.2	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.21.1	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.21.0	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.20.1	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.20.0	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.19.1	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.19.0	0.9.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.18.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.18.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.17.2	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.17.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.17.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.16.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.16.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.15.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.15.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.14.2	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub
emr-5.14.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Amazon EMR 发行版标签	JupyterHub 版本	随 JupyterHub 安装的组件
emr-5.14.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, jupyterhub

Apache Livy

Livy 可通过 REST 接口实现与运行 Spark 的 EMR 集群的交互。您可以使用 REST 接口或 RPC 客户端库提交 Spark 任务或 Spark 代码段、同步或异步检索结果并管理 Spark 上下文。有关更多信息，请参阅 Apache Livy 网站。Livy 包含在 Amazon EMR 发行版本 5.9.0 及更高版本中。

要访问 Livy Web 界面，请设置连接到主节点 的 SSH 隧道和代理连接。有关更多信息，请参阅查看 EMR 集群上托管的 Web 界面。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Livy 的版本，以及 Amazon EMR 随 Livy 一起安装的组件。

有关此发行版中随 Livy 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Livy 版本信息

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.14.0	Livy 0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Livy 的版本，以及 Amazon EMR 随 Livy 一起安装的组件。

有关此发行版中随 Livy 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Livy 版本信息

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.36.1	Livy 0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

主题

使用 Apache Livy 启用 HTTPS

1. 在启用了传输加密的情况下预置 Amazon EMR 集群。要了解有关加密的更多信息，请参阅加密静态数据和传输中的数据。
2. 创建以下内容的名为 `livy_ssh.sh` 的文件。

```
#!/bin/bash

KEYSTORE_FILE=`awk '/ssl.server.keystore.location/{getline; print}' /etc/hadoop/
conf/ssl-server.xml | sed -e 's/<[^>]*>//g' | tr -d ' \t\n\r\f'`
KEYSTORE_PASS=`awk '/ssl.server.keystore.password/{getline; print}' /etc/hadoop/
conf/ssl-server.xml | sed -e 's/<[^>]*>//g' | tr -d ' \t\n\r\f'`
KEY_PASS=`awk '/ssl.server.keystore.keypassword/{getline; print}' /etc/hadoop/conf/
ssl-server.xml | sed -e 's/<[^>]*>//g' | tr -d ' \t\n\r\f'`
```

```
echo "livy.keystore $KEYSTORE_FILE
livy.keystore.password $KEYSTORE_PASS
livy.key-password $KEY_PASS" | sudo tee -a /etc/livy/conf/livy.conf >/dev/null

sudo systemctl restart livy-server.service
```

3. 作为 Amazon EMR 步骤运行以下脚本。此脚本将修改 `/etc/livy/conf/livy.conf` 以激活 SSL。

```
--steps '[{"Args":["s3://DOC-EXAMPLE-BUCKET/
livy_ssl.sh"],"Type":"CUSTOM_JAR","ActionOnFailure":"CONTINUE","Jar":"s3://
us-east-1.elasticmapreduce/libs/script-runner/script-
runner.jar","Properties":"","Name":"Custom JAR"}]'
```

4. 重新启动 Apache Livy 服务，以使更改生效。要重新启动 Apache Livy，请参阅停止和重新启动进程。
5. 测试客户端现在是否可以使用 HTTPS 进行通信。例如，要提交任务，请运行以下代码。

```
curl -k -X POST --data '{"file": "local:///usr/lib/spark/examples/jars/spark-
examples.jar",
"className": "org.apache.spark.examples.SparkPi"}' \
-H "Content-Type: application/json" \
https://EMR_Master_Node_Host:8998/batches
```

如果您已成功启用 HTTPS，Livy 会发送一个响应，指示该命令已被接受且批处理任务已提交。

```
{"id":1,"name":null,"owner":null,"proxyUser":null,"state":"starting","appId":null,"appInfo"
{"driverLogUrl":null,"sparkUiUrl":null},"log":["stdout: ","\nstderr: ","\nYARN
 Diagnostics: "]}
```

Livy 发行历史记录

下表列出了 Amazon EMR 的每个发行版本中所包含的 Livy 的版本，以及在安装应用程序时一同安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Livy 版本信息

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.14.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.13.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.12.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
		yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.11.1	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.11.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.10.1	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.10.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.9.1	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.9.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.8.1	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.8.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.7.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.36.1	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.36.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.6.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.35.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.5.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.4.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.3.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.3.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.2.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.2.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.1.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.1.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-6.0.1	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-6.0.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.34.0	0.7.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.33.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.33.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.32.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.32.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.31.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.31.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.30.2	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.30.1	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.30.0	0.7.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-notebook-env, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.29.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.28.1	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.28.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.27.1	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.27.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.26.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.25.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.24.1	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.24.0	0.6.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.23.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.23.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.22.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.21.2	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.21.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.21.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.20.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.20.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.19.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.19.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.18.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx
emr-5.18.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server, nginx

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.17.2	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.17.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.17.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.16.1	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.16.0	0.5.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.15.1	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.15.0	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.14.2	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.14.1	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.14.0	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.13.1	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.13.0	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.12.3	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.12.2	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.12.1	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.12.0	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.11.4	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.11.3	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.11.2	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.11.1	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.11.0	0.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.10.1	0.4.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Amazon EMR 发行版标签	Livy 版本	随 Livy 安装的组件
emr-5.10.0	0.4.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.9.1	0.4.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server
emr-5.9.0	0.4.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, livy-server

Apache MXNet

Apache MXNet 是专为构建神经网络和其他深度学习应用程序设计的加速库。MXNet 自动执行常见工作流程并优化数值计算。MXNet 可帮助您设计神经网络架构，而不必专注于实施低级计算，如线性代数运算。MXNet 包含在 Amazon EMR 发行版本 5.10.0 及更高版本中。

有关更多信息，请参阅 Apache MXNet Web 站点。

下表列出了 Amazon EMR 6.x 系列的最新发行版本附带的 MXNet 的版本，以及 Amazon EMR 随 MXNet 一起安装的组件。

有关此发行版中随 MXNet 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 MXNet 版本信息

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.14.0	MXNet 1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

下表列出了 Amazon EMR 5.x 系列的最新发行版本附带的 MXNet 的版本，以及 Amazon EMR 随 MXNet 一起安装的组件。

有关此发行版中随 MXNet 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 MXNet 版本信息

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.36.1	MXNet 1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
		hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

MXNet 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 MxNet 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

MXNet 版本信息

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.14.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.13.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
		cemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.12.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.11.1	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.11.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.10.1	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.10.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.9.1	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.9.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.8.1	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.8.0	1.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.7.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.36.1	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.36.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.6.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.35.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.5.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.4.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.3.1	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.3.0	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.2.1	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.2.0	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.1.1	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-6.1.0	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.0.1	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-6.0.0	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.34.0	1.8.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.33.1	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.33.0	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.32.1	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.32.0	1.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.31.1	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.31.0	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.30.2	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.30.1	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.30.0	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.29.0	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.28.1	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.28.0	1.5.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.27.1	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.27.0	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.26.0	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.25.0	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.24.1	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.24.0	1.4.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.23.1	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.23.0	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.22.0	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.21.2	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.21.1	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.21.0	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.20.1	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.20.0	1.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.19.1	1.3.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.19.0	1.3.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.18.1	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.18.0	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.17.2	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.17.1	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.17.0	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.16.1	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.16.0	1.2.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.15.1	1.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.15.0	1.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.14.2	1.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.14.1	1.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.14.0	1.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet, opencv
emr-5.13.1	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.13.0	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.12.3	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.12.2	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.12.1	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.12.0	1.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.11.4	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.11.3	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.11.2	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.11.1	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet

Amazon EMR 发行版标签	MXNet 版本	随 MXNet 安装的组件
emr-5.11.0	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.10.1	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet
emr-5.10.0	0.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mxnet

Apache Oozie

使用 Apache Oozie 工作流调度器管理和协调 Hadoop 任务。有关详细信息，请参阅 http://oozie.apache.org/。

Amazon EMR 上不支持 Oozie 本机 Web 界面。如要使用 Oozie 的前端界面，请尝试 Hue Oozie 应用程序。有关更多信息，请参阅Hue。Oozie 包含在 Amazon EMR 发行版 5.0.0 及更高版本中。Oozie 作为沙盒应用程序包含在早期版本中。有关更多信息，请参阅Amazon EMR 4.x 发行版。

如果您使用基于 Amazon Linux AMI（创建日期为 2018-08-11）的自定义 Amazon Linux AMI，则 Oozie 服务器无法启动。如果您使用 Oozie，请根据具有不同创建日期的 Amazon Linux AMI ID 创建自定义 AMI。您可以使用以下 Amazon CLI 命令返回所有 2018.03 版本的 HVM Amazon Linux AMI 的镜像 ID 列表以及发布日期，以便您可以根据需要选择合适的 Amazon Linux AMI。将 MyRegion 替换为您的区域标识符，如 us-west-2。

```
aws ec2 --region MyRegion describe-images --owner amazon --query 'Images[?
Name!=`null`]|[?starts_with(Name, `amzn-ami-hvm-2018.03`) == `true`].
[CreationDate,ImageId,Name]' --output text | sort -rk1
```

下表列出了 Amazon EMR 6.x 系列的最新发行版本附带的 Oozie 的版本，以及 Amazon EMR 随 Oozie 一起安装的组件。

有关此发行版中随 Oozie 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Oozie 版本信息

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.14.0	Oozie 5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client,

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
		oozie-server, tez-on-yarn, tez-on-worker

下表列出了 Amazon EMR 5.x 系列的最新发行版本附带的 Oozie 的版本，以及 Amazon EMR 随 Oozie 一起安装的组件。

有关此发行版中随 Oozie 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Oozie 版本信息

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.36.1	Oozie 5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

主题

- 将 Oozie 与 Amazon RDS 中的远程数据库结合使用
- 为 Oozie 配置 Java 版本
- Oozie 发行历史记录

将 Oozie 与 Amazon RDS 中的远程数据库结合使用

默认情况下，Oozie 用户信息和查询历史记录存储在主节点上的本地 MySQL 数据库中。或者，您可以使用 Amazon S3 中存储的配置和 Amazon Relational Database Service（Amazon RDS）中的

MySQL 数据库创建一个或多个启用了 Oozie 的集群。这样，无需使 Amazon EMR 集群保持运行，您就可以保存 Oozie 创建的用户信息和查询历史记录。我们建议使用 Amazon S3 服务器端加密来存储配置文件。

首先为 Oozie 创建远程数据库。

创建外部 MySQL 数据库

1. 通过以下网址打开 Amazon RDS 控制台：https://console.aws.amazon.com/rds/。
2. 选择 Launch a DB Instance (启动数据库实例)。
3. 选择 MySQL，然后选择 Select（选择）。
4. 保留默认选择 Multi-AZ Deployment and Provisioned IOPS Storage（多可用区部署和预调配 IOPS 存储)，并选择 Next（下一步）。
5. 保留 Instance Specifications（实例规格）的默认值，指定 Settings（设置），然后选择 Next（下一步）。
6. 在 Configure Advanced Settings (配置高级设置) 页面上，选择相应的安全组和数据库名称。您使用的安全组必须至少允许从集群主节点通过端口 3306 进行入站 TCP 访问。如果此时您尚未创建集群，则可以允许所有主机连接到端口 3306 并在启动集群之后调整安全组。选择 Launch DB Instance (启动数据库实例)。
7. 在 RDS 控制面板中，选择 Instances (实例)，然后选择您刚刚创建的实例。当您的数据库可用时，记下数据库名称、用户名、密码和 RDS 实例主机名。您将在创建和配置集群时用到此信息。

使用 Amazon CLI 在启动集群时为 Oozie 指定外部 MySQL 数据库

要在使用 Amazon CLI 启动集群时为 Oozie 指定外部 MySQL 数据库，请在创建 RDS 实例时使用所记录的信息，以使用配置对象配置 `oozie-site`。

> **Note**
>
> 您可以创建使用同一个外部数据库的多个集群，但是每个集群将共享查询历史记录和用户信息。

- 使用 Amazon CLI 创建安装了 Oozie 的集群，使用您创建的外部数据库，并引用包含 Oozie 配置分类的配置文件，该文件指定数据库属性。以下示例创建一个安装了 Oozie 的集群，引用了 Amazon S3 中的配置文件 `myConfig.json`，该文件指定数据库配置。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Oozie
 Name=Spark Name=Hive \
--instance-type m5.xlarge --instance-count 3 \
--configurations https://s3.amazonaws.com/mybucket/myfolder/myConfig.json --use-
default-roles
```

下面显示的是 myConfig.json 文件的内容示例。将 *JDBC URL*、*username* 和 *password* 替换为您的 RDS 实例的 JDBC URL、用户名和密码。

> **Important**
>
> JDBC URL 必须包含数据库名称作为后缀。例如 jdbc:mysql://oozie-external-db.xxxxxxxxxx.us-east-1.rds.amazonaws.com:3306/dbname。

```
[{
  "Classification": "oozie-site",
    "Properties": {
        "oozie.service.JPAService.jdbc.driver": "org.mariadb.jdbc.Driver",
        "oozie.service.JPAService.jdbc.url": "JDBC URL",

        "oozie.service.JPAService.jdbc.username": "username",
        "oozie.service.JPAService.jdbc.password": "password"
    },
    "Configurations": []
}]
```

为 Oozie 配置 Java 版本

Oozie 运行多个 Java 虚拟机（JVM）进程。本页说明如何为每个流程配置 Java 版本。

- Oozie Server：在 `oozie-env` 分类中设置 `JAVA_HOME`，以更新 `EmbeddedOozieServer` 的 Java 版本。
- Oozie Launcher AM：Oozie Launcher AM 是一项单映射器 MR 作业，它调用相应的应用程序客户端库，例如 Hadoop 和 Hive。除非另有配置，否则 Oozie Launcher AM 的运行时系统版本与 EMR 集群中 Hadoop 的 Java 运行时相同。要为 Oozie Launcher AM 配置 Java 运行时系统，请在作业的 `workflow.xml` 中设置以下属性：

  ```
  <property>
       <name>mapred.child.env</name>
       <value>JAVA_HOME=/path/to/JAVA_HOME</value>
   </property>
  ```

  此属性可确保 Ooozie 作业的 Oozie Launcher AM 在您指定的 Java 版本上运行，而不是在 Hadoop 中设置的 Java 版本上运行。
- 应用程序客户端可执行文件：由于 Oozie Launcher AM 默认调用应用程序客户端，因此客户端可执行文件的 Java 运行时系统与 Oozie Launcher AM 相同。
- 由 Ooozie 作业启动的应用程序：除非另有说明，否则由 Oozie 作业启动的实际应用程序 JVM 的运行时系统版本与 EMR 集群中 Hadoop 的 Java 运行时系统相同。根据用于在 Ooozie 作业中启动应用程序的 Oozie 工作流程操作的类型（Spark 或 Hive 操作），您可以更新 Oozie 作业的 `workflow.xml` 中更新实际应用程序 JVM 的默认 Java 运行时系统。

Oozie 发行历史记录

下表列出了 Amazon EMR 的每个发行版本中所包含的 Oozie 的版本，以及在安装应用程序时一同安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Oozie 版本信息

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.14.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
		httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker
emr-6.13.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker
emr-6.12.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.11.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker
emr-6.11.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.10.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker
emr-6.10.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.9.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.9.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.8.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.8.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.7.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.36.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.36.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.6.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.35.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.5.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.4.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.3.1	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.3.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.2.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.2.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.1.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.1.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-6.0.1	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-6.0.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.34.0	5.2.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.33.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.33.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.32.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.32.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.31.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.31.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.30.2	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.30.1	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.30.0	5.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.29.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.28.1	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.28.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.27.1	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.27.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.26.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.25.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.24.1	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.24.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.23.1	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.23.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.22.0	5.1.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.21.2	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.21.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.21.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.20.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.20.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.19.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.19.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.18.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.18.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.17.2	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.17.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.17.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.16.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.16.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.15.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.15.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.14.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.14.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.14.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.13.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.13.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.12.3	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.12.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.12.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.12.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.11.4	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.11.3	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.11.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.11.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.11.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.10.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.10.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.9.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.9.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.8.3	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.8.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.8.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.8.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.7.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.7.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.6.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.6.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.5.4	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.5.3	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.5.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.5.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.5.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.4.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.4.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.3.2	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.3.1	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.3.0	4.3.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.2.3	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.2.2	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.2.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.2.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.1.1	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.1.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn
emr-5.0.3	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Amazon EMR 发行版标签	Oozie 版本	随 Oozie 安装的组件
emr-5.0.0	4.2.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, oozie-client, oozie-server, tez-on-yarn

Apache Phoenix

Apache Phoenix 用于 OLTP 和操作分析，让您可以使用标准 SQL 查询和 JDBC API 来处理 Apache HBase 底层存储。有关更多信息，请参阅 Phoenix in 15 minutes or less。Phoenix 包含在 Amazon EMR 发行版 4.7.0 及更高版本中。

如果您从早期版本的 Amazon EMR 升级到 Amazon EMR 发行版 5.4.0 或更高版本并使用二级索引，请按 Apache Phoenix 文档中所述升级本地索引。Amazon EMR 将从 `hbase-site` 分类中删除所需配置，但索引需要重新填充。支持在线和离线升级索引。在线升级为默认值，这意味着，在从版本 4.8.0 或更高版本的 Phoenix 客户端初始化时重新填充索引。要指定离线升级，请在 `phoenix.client.localIndexUpgrade` 分类中将 `phoenix-site` 配置设置为 false，然后将 SSH 设置为主节点以运行 `psql [zookeeper] -1`。

下表列出了 Amazon EMR 6.x 系列的最新发行版本附带的 Phoenix 的版本，以及 Amazon EMR 随 Phoenix 一起安装的组件。

有关此发行版中随 Phoenix 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Phoenix 版本信息

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.14.0	Phoenix 5.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
		server, zookeeper-client, zookeeper-server

下表列出了 Amazon EMR 5.x 系列的最新发行版本附带的 Phoenix 的版本，以及 Amazon EMR 随 Phoenix 一起安装的组件。

有关此发行版中随 Phoenix 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Phoenix 版本信息

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.36.1	Phoenix 4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

主题

- 使用 Phoenix 创建集群
- Phoenix 客户端
- Phoenix 发行历史记录

使用 Phoenix 创建集群

您可以通过在控制台中或使用 Amazon CLI 创建集群时选择 Phoenix 来安装此应用程序。以下过程和示例说明如何使用 Phoenix 和 HBase 创建集群。有关使用控制台（包括 Advanced Options (高级选项)）创建集群的更多信息，请参阅《Amazon EMR 管理指南》https://docs.amazonaws.cn/emr/latest/ManagementGuide/emr-plan.html中的计划和配置集群。

在控制台中通过使用用来创建集群的 Quick Options 安装的 Phoenix 启动集群

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 选择 Create cluster (创建集群) 以使用 Quick Create (快速创建)。
3. 在 Software Configuration (软件配置) 下，选择与您的应用程序对应的最新版本。Phoenix 只有在选择了 Amazon 发行版 emr-4.7.0 或更高版本时才会作为选项显示。
4. 在 Applications (应用程序) 下，选择第二个选项：HBase: HBase ***ver*** with Ganglia ***ver***, Hadoop ***ver***, Hive ***ver***, Hue ***ver***, Phoenix ***ver***, and ZooKeeper ***ver***。
5. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

以下示例通过使用默认配置设置安装的 Phoenix 启动集群。

使用 Amazon CLI 启动安装了 Phoenix 和 HBase 的集群

- 使用下面的命令创建集群：

```
aws emr create-cluster --name "Cluster with Phoenix" --release-label emr-5.36.1 \
--applications Name=Phoenix Name=HBase --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3 --use-default-roles
```

自定义 Phoenix 配

在创建集群时，您可使用 `hbase-site.xml` 配置分类在 `hbase-site` 中设置值来配置 Phoenix。

有关更多信息，请参阅 Phoenix 文档中的配置和优化。

以下示例说明如何使用存储在 Amazon S3 中的 JSON 文件来为 `phoenix.schema.dropMetaData` 属性指定 `false` 的值。可以为单个分类指定多个属性。有关更多信息，请参阅配置应用程序。随后，`create-cluster` 命令会将 JSON 文件引用为 `--configurations` 参数。

保存到 /mybucket/myfolder/myconfig.json 的 JSON 文件的内容如下所示。

```
[
    {
      "Classification": "hbase-site",
      "Properties": {
        "phoenix.schema.dropMetaData": "false"
      }
    }
  ]
```

引用 JSON 文件的 `create cluster` 命令如以下示例所示。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Phoenix \
Name=HBase --instance-type m5.xlarge --instance-count 2 \
--configurations https://s3.amazonaws.com/mybucket/myfolder/myconfig.json
```

> **Note**
>
> 仅 Amazon EMR 5.23.0 和更高版本支持任何 Phoenix 配置分类的重新配置请求，Amazon EMR 5.21.0 或 5.22.0 版本不支持该请求。有关更多信息，请参阅为运行的集群中的实例组提供配置

Phoenix 客户端

您可以通过以下方式连接到 Phoenix：使用内置了完全依赖项的 JDBC 客户端，或使用已采用 Phoenix Query Server 且只能在集群的主节点上运行的“瘦客户端”（例如，通过使用 SQL 客户端、步骤、命令行、SSH 端口转发等）。使用“胖”JDBC 客户端时，仍需能够访问集群的所有节点，因为它将直接连接到 HBase 服务。“瘦”Phoenix 客户端只需要通过默认端口 8765 访问 Phoenix 查询服务器。Phoenix 中有几个脚本使用这些客户端。

使用 Amazon EMR 步骤通过 Phoenix 进行查询

以下过程从 HBase 还原快照，并使用该数据运行 Phoenix 查询。您可以扩展此示例或创建一个新脚本，以利用 Phoenix 的客户端来满足您的需求。

1. 使用以下命令创建安装了 Phoenix 的集群：

```
aws emr create-cluster --name "Cluster with Phoenix" --log-uri s3://myBucket/
myLogFolder --release-label emr-5.36.1 \
--applications Name=Phoenix Name=HBase --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3 --use-default-roles
```

2. 创建后，将以下文件上传到 Amazon S3：

   copySnapshot.sh

```
sudo su hbase -s /bin/sh -c 'hbase snapshot export \
 -D hbase.rootdir=s3://us-east-1.elasticmapreduce.samples/hbase-demo-customer-data/
snapshot/ \
-snapshot customer_snapshot1 \
-copy-to hdfs://masterDNSName:8020/user/hbase \
-mappers 2 -chuser hbase -chmod 700'
```

   runQuery.sh

```
aws s3 cp s3://myBucket/phoenixQuery.sql /home/hadoop/
/usr/lib/phoenix/bin/sqlline-thin.py http://localhost:8765 /home/hadoop/
phoenixQuery.sql
```

   phoenixQuery.sql

> **Note**
>
> 在使用 Amazon EMR 5.26.0 及更高版本时，只需在以下示例中添加 COLUMN_ENCODED_BYTES=0。

```
CREATE VIEW "customer" (
pk VARCHAR PRIMARY KEY,
"address"."state" VARCHAR,
```

```
"address"."street" VARCHAR,
"address"."city" VARCHAR,
"address"."zip" VARCHAR,
"cc"."number" VARCHAR,
"cc"."expire" VARCHAR,
"cc"."type" VARCHAR,
"contact"."phone" VARCHAR)
COLUMN_ENCODED_BYTES=0;

CREATE INDEX my_index ON "customer" ("customer"."state") INCLUDE("PK",
 "customer"."city", "customer"."expire", "customer"."type");

SELECT "customer"."type" AS credit_card_type, count(*) AS num_customers FROM
 "customer" WHERE "customer"."state" = 'CA' GROUP BY "customer"."type";
```

使用 Amazon CLI 将文件提交到 S3 存储桶：

```
aws s3 cp copySnapshot.sh s3://myBucket/
aws s3 cp runQuery.sh s3://myBucket/
aws s3 cp phoenixQuery.sql s3://myBucket/
```

3. 使用以下步骤创建提交给您在步骤 1 中创建的集群的表：

   createTable.json

```
[
  {
    "Name": "Create HBase Table",
    "Args": ["bash", "-c", "echo $'create \"customer\",\"address\",\"cc\",\"contact
\"' | hbase shell"],
    "Jar": "command-runner.jar",
    "ActionOnFailure": "CONTINUE",
    "Type": "CUSTOM_JAR"
  }
]
```

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps file://./createTable.json
```

4. 使用 `script-runner.jar` 运行您之前上传到 S3 存储桶的 `copySnapshot.sh` 脚本：

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
```

```
--steps Type=CUSTOM_JAR,Name="HBase Copy Snapshot",ActionOnFailure=CONTINUE,\
Jar=s3://region.elasticmapreduce/libs/script-runner/script-
runner.jar,Args=["s3://myBucket/copySnapshot.sh"]
```

这将运行 MapReduce 作业以将快照数据复制到集群 HDFS。

5. 使用以下步骤还原复制到集群的快照：

   restoreSnapshot.json

```
[
  {
    "Name": "restore",
    "Args": ["bash", "-c", "echo $'disable \"customer\"; restore_snapshot
 \"customer_snapshot1\"; enable \"customer\"' | hbase shell"],
    "Jar": "command-runner.jar",
    "ActionOnFailure": "CONTINUE",
    "Type": "CUSTOM_JAR"
  }
]
```

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps file://./restoreSnapshot.json
```

6. 使用 `script-runner.jar` 运行您之前上传到 S3 存储桶的 `runQuery.sh` 脚本：

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF \
--steps Type=CUSTOM_JAR,Name="Phoenix Run Query",ActionOnFailure=CONTINUE,\
Jar=s3://region.elasticmapreduce/libs/script-runner/script-
runner.jar,Args=["s3://myBucket/runQuery.sh"]
```

查询运行并将结果返回到步骤的 `stdout`。完成此步骤可能需要几分钟时间。

7. 通过您在步骤 1 中创建集群时使用的日志 URI 检查步骤的 `stdout` 的结果。结果应该类似以下内容：

```
+-------------------------------------------+-----------------------------------+
|             CREDIT_CARD_TYPE              |             NUM_CUSTOMERS         |
+-------------------------------------------+-----------------------------------+
| american_express                          | 5728                              |
| dankort                                   | 5782                              |
| diners_club                               | 5795                              |
```

```
| discover                                  | 5715                                 |
| forbrugsforeningen                        | 5691                                 |
| jcb                                       | 5762                                 |
| laser                                     | 5769                                 |
| maestro                                   | 5816                                 |
| mastercard                                | 5697                                 |
| solo                                      | 5586                                 |
| switch                                    | 5781                                 |
| visa                                      | 5659                                 |
+-------------------------------------------+--------------------------------------+
```

Phoenix 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Phoenix 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Phoenix 版本信息

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.14.0	5.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.13.0	5.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.12.0	5.1.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.11.1	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.11.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.10.1	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.10.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.9.1	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.9.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.8.1	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.8.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.7.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.36.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.36.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.6.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, hbase-operator-tools, phoenix-library, phoenix-connectors, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.35.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.5.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.4.0	5.1.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.3.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.3.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.2.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.2.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.1.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.1.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-6.0.1	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-6.0.0	5.0.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.34.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.33.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.33.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.32.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.32.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.31.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.31.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.30.2	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.30.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.30.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.29.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.28.1	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.28.0	4.14.3	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.27.1	4.14.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.27.0	4.14.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.26.0	4.14.2	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.25.0	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.24.1	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.24.0	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.23.1	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.23.0	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.22.0	4.14.1	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.21.2	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.21.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.21.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.20.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.20.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.19.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.19.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.18.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.18.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.17.2	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.17.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.17.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.16.1	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.16.0	4.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.15.1	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.15.0	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.14.2	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.14.1	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.14.0	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.13.1	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.13.0	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.12.3	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.12.2	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.12.1	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.12.0	4.13.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.11.4	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.11.3	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.11.2	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.11.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.11.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.10.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.10.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.9.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.9.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.8.3	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.8.2	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.8.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.8.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.7.1	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.7.0	4.11.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.6.1	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.6.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.5.4	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.5.3	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.5.2	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.5.1	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.5.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.4.1	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.4.0	4.9.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.3.2	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.3.1	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.3.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.2.3	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.2.2	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.2.1	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.2.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.1.1	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.1.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-5.0.3	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-5.0.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.9.6	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.9.5	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.9.4	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.9.3	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.9.2	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.9.1	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.8.5	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.8.4	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.8.3	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.8.2	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.8.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.7.4	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.7.2	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	Phoenix 版本	随 Phoenix 安装的组件
emr-4.7.1	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server
emr-4.7.0	4.7.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-mapred, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hbase-hmaster, hbase-client, hbase-region-server, phoenix-library, phoenix-query-server, zookeeper-client, zookeeper-server

Apache Pig

Apache Pig 是一个开源 Apache 库，它在 Hadoop 的顶层运行，提供一种脚本语言，您可以用来转换大型数据集，而不必用底层计算机语言 (例如 Java) 编写复杂的代码。该库使用一种叫做 Pig Latin 的语言编写的、类似 SQL 的命令，并基于有向无环图 (DAG) 或 MapReduce 程序将这些命令转换为 Tez 任务。Pig 可与各种格式的结构化和非结构化数据结合使用。有关 Pig 的详细信息，请参阅 http://pig.apache.org/。

您可以通过交互方式或批处理方式执行 Pig 命令。要以交互方式使用 Pig，请创建到主节点的 SSH 连接，并使用 Grunt Shell 提交命令。要以批处理方式使用 Pig，请编写 Pig 脚本，将脚本上传到 Amazon S3，并作为集群步骤提交。有关向集群提交工作的更多信息，请参阅《Amazon EMR 管理指南》中的向集群提交工作。

当您使用 Pig 将输出写入 Amazon S3 中的 HCatalog 表时，请通过以下方式禁用 Amazon EMR 直接写入：将 `mapred.output.direct.NativeS3FileSystem` 和 `mapred.output.direct.EmrFileSystem` 属性设置为 `false`。有关更多信息，请参阅使用 HCatalog。在 Pig 脚本中，可使用 `SET mapred.output.direct.NativeS3FileSystem false` 和 `SET mapred.output.direct.EmrFileSystem false` 命令。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Pig 的版本，以及 Amazon EMR 随 Pig 一起安装的组件。

有关此发行版中随 Pig 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Pig 版本信息

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.14.0	Pig 0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
		timeline-server, pig-client, tez-on-yarn, tez-on-worker

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Pig 的版本，以及 Amazon EMR 随 Pig 一起安装的组件。

有关此发行版中随 Pig 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Pig 版本信息

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.36.1	Pig 0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-dist-cp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

主题

提交 Pig 工作

此部分演示如何向 Amazon EMR 集群提交 Pig 工作。后面的示例会生成一个报告，其中包含已传输的字节总数、前 50 个 IP 地址的列表、前 50 个外部引用站点的列表以及前 50 个使用 Bing 和 Google

的搜索词。Pig 脚本位于 Amazon S3 存储桶 `s3://elasticmapreduce/samples/pig-apache/do-reports2.pig` 中。输入数据位于 Amazon S3 存储桶 `s3://elasticmapreduce/samples/pig-apache/input` 中。输出保存到 Amazon S3 存储桶。

使用 Amazon EMR 控制台提交 Pig 工作

此示例介绍如何使用 Amazon EMR 控制台向集群添加 Pig 步骤。

提交 Pig 步骤

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 选择创建集群以创建安装有 Pig 的集群。有关如何创建群集的步骤，请参阅规划和配置 Amazon EMR 集群。
3. 按照使用 SSH 连接主节点中概述的步骤，打开终端并 SSH 连接到集群的主节点。完成此操作后，请运行以下步骤。

   ```
   sudo mkdir -p /home/hadoop/lib/pig/
   sudo aws s3 cp s3://elasticmapreduce/libs/pig/0.3/piggybank-0.3-amzn.jar /home/hadoop/lib/pig/piggybank.jar
   ```

4. 在控制台中，点击 Cluster List (集群列表)，并选择您创建的集群。
5. 滚动到 Steps (步骤) 部分并展开它，然后选择 Add step (添加步骤)。
6. 在 Add step (添加步骤) 对话框中：
   - 对于 Step type (步骤类型)，选择 Pig program (Pig 程序)。
   - 对于 Name (名称)，接受默认名称 (Pig program) 或键入新名称。
   - 对于 Script S3 location (脚本 S3 位置)，键入 Pig 脚本的位置。例如：**`s3://elasticmapreduce/samples/pig-apache/do-reports2.pig`**。
   - 对于 Input S3 location (输入 S3 位置)，键入输入数据的位置。例如：**`s3://elasticmapreduce/samples/pig-apache/input`**。
   - 对于 Output S3 location (输出 S3 位置)，键入或浏览到您的 Amazon S3 输出存储桶的名称。
   - 对于 Arguments (参数)，将该字段保留为空白。
   - 对于 Action on failure (出现故障时的操作)，接受默认选项 Continue (继续)。
7. 选择 Add（添加）。步骤会出现在控制台中，其状态为“Pending”。

8. 步骤的状态会随着步骤的运行从“Pending”变为“Running”，再变为“Completed”。要更新状态，请选择 Actions (操作) 列上方的 Refresh (刷新) 图标。步骤完成后，请检查 Amazon S3 存储桶以确认 Pig 步骤的输出文件存在。

使用 Amazon CLI 提交 Pig 工作

使用 Amazon CLI 提交 Pig 步骤

使用 Amazon CLI 启动集群时，请使用 `--applications` 参数安装 Pig。要提交 Pig 步骤，请使用 `--steps` 参数。

1. 要启动安装有 Pig 的群集，请键入以下命令，用您的 EC2 key pair 和 Amazon S3 存储桶的名称替换 *myKey* 和 *DOC-EXAMPLE-BUCKET/*。

```
aws emr create-cluster \
--name "Test cluster" \
--log-uri s3://DOC-EXAMPLE-BUCKET/ \
--release-label emr-5.36.1 \
--applications Name=Pig \
--use-default-roles \
--ec2-attributes KeyName=myKey \
--instance-type m5.xlarge \
--instance-count 3
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\)。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

如果不使用 `--instance-groups` 参数指定实例计数，则将启动单个主节点，其余实例将作为核心节点启动。所有节点都使用该命令中指定的实例类型。

> **Note**
>
> 如果您之前未创建默认 EMR 服务角色和 EC2 实例配置文件，请先键入 `aws emr create-default-roles` 创建它们，然后再键入 `create-cluster` 子命令。

2. 要提交 Pig 步骤，请输入以下命令，使用您的群集 ID 和 Amazon S3 存储桶的名称替换 *myClusterId* 和 *DOC-EXAMPLE-BUCKET*。

```
aws emr add-steps \
--cluster-id myClusterId \
--steps Type=PIG,Name="Pig Program",ActionOnFailure=CONTINUE,Args=[-f,s3://
elasticmapreduce/samples/pig-apache/do-reports2.pig,-p,INPUT=s3://elasticmapreduce/
samples/pig-apache/input,-p,OUTPUT=s3://DOC-EXAMPLE-BUCKET/pig-apache/output]
```

   此命令将返回一个步骤 ID，您可以用它来检查您的步骤的 `State`。

3. 使用带有 `describe-step` 命令的步骤，查询步骤的状态。

```
aws emr describe-step --cluster-id myClusterId --step-id s-1XXXXXXXXXXA
```

   随着步骤的运行，步骤的 `State` 从 `PENDING` 变为 `RUNNING` 再变为 `COMPLETED`。步骤完成后，请检查 Amazon S3 存储桶以确认 Pig 步骤的输出文件存在。

有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅 Amazon CLI 命令参考。

从 Pig 调用由用户定义的函数

Pig 提供从 Pig 脚本内部调用用户定义的函数 (UDF) 的功能。您可以完成此操作，以便实施自定义处理并在 Pig 脚本使用。当前受支持的语言为 Java、Python/Jython 和 JavaScript (不过，对于 JavaScript 的支持仍然是试验性的。)

以下部分描述了如何通过 Pig 注册函数，以便从 Pig Shell 或者从 Pig 脚本内部调用它们。有关将 Pig 与 UDF 一起使用的更多信息，请参阅适用于您的 Pig 版本的 Pig 文档。

从 Pig 中调用 JAR 文件

您可以在 Pig 脚本中使用 `REGISTER` 命令，从而通过 Pig 使用自定义 JAR 文件。JAR 文件在本地或者在远程文件系统（如 Amazon S3）上。当 Pig 脚本运行时，Amazon EMR 会自动下载 JAR 文件到主节点上，然后将 JAR 文件上传到 Hadoop 分布式缓存。通过这种方法，集群中的所有实例可自动地根据需要使用该 JAR 文件。

与 Pig 一起使用 JAR 文件

1. 将自定义 JAR 文件上传到 Amazon S3 中。

2. 使用 Pig 脚本中的 REGISTER 命令，在自定义 JAR 文件的 Amazon S3 上指定存储桶。

```
REGISTER s3://mybucket/path/mycustomjar.jar;
```

从 Pig 调用 Python/Jython 脚本

您可以通过 Pig 注册 Python 脚本，然后，从 Pig Shell 或者在 Pig 脚本中调用这些脚本中的函数。您可以通过 `register` 关键字指定该脚本的位置，从而完成此操作。

因为 Pig 是以 Java 编写的，所以它使用 Jython 脚本引擎解析 Python 脚本。有关 Jython 的详细信息，请转到 http://www.jython.org/。

从 Pig 调用 Python/Jython 脚本

1. 编写 Python 脚本并将其上传到 Amazon S3 中的位置。它应该是创建该 Pig 集群的同一账户所拥有的存储桶，或者该位置拥有相关权限，使得创建该集群的账户可以进行访问。在此示例中，脚本上传到 s3://mybucket/pig/python。
2. 启动 Pig 集群。如果您正在从 Grunt Shell 访问 Pig，请运行交互式集群。如果您正在从脚本运行 Pig 命令，请启动已编写脚本的 Pig 集群。此示例将启动交互式集群。有关如何创建 Pig 集群的详细信息，请参阅 提交 Pig 工作。
3. 有关交互式集群，请使用 SSH 连接到主节点并运行 Grunt Shell。有关更多信息，请参阅通过 SSH 登录主节点。
4. 通过在命令行键入 `pig` 的方式运行 Pig 的 Grunt Shell：

```
pig
```

5. 在 Grunt 命令提示符中使用 `register` 关键字，通过 Pig 注册 Jython 库和 Python 脚本（如以下命令所示），您可以其中指定脚本在 Amazon S3 中的位置：

```
grunt> register 'lib/jython.jar';
grunt> register 's3://mybucket/pig/python/myscript.py' using jython as myfunctions;
```

6. 加载输入数据。以下示例从 Amazon S3 位置加载输入：

```
grunt> input = load 's3://mybucket/input/data.txt' using TextLoader as
 (line:chararray);
```

7. 现在，您可以通过 `myfunctions` 引用脚本中的函数的方式，从 Pig 内部调用这些函数：

```
grunt> output=foreach input generate myfunctions.myfunction($1);
```

Pig 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Pig 的版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Pig 版本信息

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.14.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker
emr-6.13.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
		timeline-server, pig-client, tez-on-yarn, tez-on-worker
emr-6.12.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker
emr-6.11.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.11.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker
emr-6.10.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.10.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn, tez-on-worker
emr-6.9.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.9.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.8.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.8.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.7.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.36.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.36.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.6.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.35.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.5.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.4.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.3.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.3.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.2.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.2.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-6.1.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-6.1.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.34.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.33.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.33.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.32.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.32.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.31.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.31.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.30.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.30.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.30.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.29.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.28.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.28.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.27.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.27.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.26.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.25.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.24.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.24.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.23.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.23.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.22.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.21.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.21.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.21.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.20.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.20.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.19.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.19.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.18.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.18.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.17.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.17.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.17.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.16.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.16.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.15.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.15.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.14.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.14.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.14.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.13.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.13.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.12.3	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.12.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.12.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.12.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.11.4	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.11.3	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.11.2	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.11.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.11.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.10.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.10.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.9.1	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.9.0	0.17.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.8.3	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.8.2	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.8.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.8.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.7.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.7.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.6.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.6.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.5.4	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.5.3	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.5.2	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.5.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.5.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.4.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.4.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.3.2	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.3.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.3.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.2.3	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.2.2	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.2.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.2.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.1.1	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.1.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-5.0.3	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn
emr-5.0.0	0.16.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, pig-client, tez-on-yarn

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.9.6	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.9.5	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.9.4	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.9.3	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.9.2	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.9.1	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.8.5	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.8.4	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.8.3	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.8.2	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.8.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.7.4	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.7.2	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.7.1	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.7.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.6.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.5.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.4.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.3.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.2.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client
emr-4.1.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Amazon EMR 发行版标签	Pig 版本	随 Pig 安装的组件
emr-4.0.0	0.14.0	emrfs, emr-ddb, emr-goodies, emr-kinesis, emr-s3-distcp, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, pig-client

Presto 和 Trino

> **Note**
>
> 2020 年 12 月，PrestoSQL 更名为 Trino。Amazon EMR 6.4.0 及更高版本使用名称 Trino，而早期版本使用名称 PrestoSQL。

Presto 是一种快速 SQL 查询引擎，专为对来自多个源的大型数据集进行交互式分析查询而设计。有关更多信息，请参阅 Presto 网站。Amazon EMR 发行版 5.0.0 及更高版本包含 Presto。早期发行版包含 Presto，将其用作沙盒应用程序。有关更多信息，请参阅Amazon EMR 4.x 发行版。Amazon EMR 发行版 6.1.0 及更高版本支持 Presto 之外的 Trino（PrestoSQL）。有关更多信息，请参阅PrestoDB 和 Trino 安装。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Presto 的版本，以及 Amazon EMR 随 Presto 一起安装的组件。

有关此发行版中随 Presto 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Presto 版本信息

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.14.0	Presto 0.281	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Presto 的版本，以及 Amazon EMR 随 Presto 一起安装的组件。

有关此发行版中随 Presto 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Presto 版本信息

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.36.1	Presto 0.267	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Trino (PrestoSQL) 的版本，以及 Amazon EMR 随 Trino (PrestoSQL) 一起安装的组件。

有关此发行版中随 Trino（PrestoSQL）安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Trino（PrestoSQL）版本信息

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.14.0	Trino (PrestoSQL) 422	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
		cemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

主题

将 Presto 与 Amazon Glue 数据目录配合使用

使用 Amazon EMR 发行版 5.10.0 及更高版本，您可以指定 Amazon Glue 数据目录作为 Presto 的默认 Hive 元存储。当您需要持久性的元数据存储或由不同集群、服务、应用程序和 Amazon Web Services 账户共享的元数据存储时，我们建议使用此配置。

Amazon Glue 是一项完全托管式提取、转换和加载（ETL）服务，使您能够轻松且经济高效地对数据进行分类、清理和扩充，并在各种数据存储之间可靠地移动数据。Amazon Glue 数据目录跨各种数据源和数据格式提供统一的元数据存储库，从而不仅与 Amazon EMR 集成，还与 Amazon RDS、Amazon Redshift、Redshift Spectrum、Athena 以及任何与 Apache Hive 元存储兼容的应用程序集成。AmazonGlue 爬网程序能够自动从 Amazon S3 源数据推断架构，从而将关联的元数据存储在数据目录中。有关数据目录的更多信息，请参阅《Amazon Glue 开发人员指南》中的填充 Amazon Glue 数据目录。

使用 Amazon Glue 需单独付费。在数据目录中存储和访问数据需按月付费；为 Amazon Glue ETL 作业和爬网程序运行时按小时费率付费（按分计费）；为每个预置的开发端点支付每小时费率（按分计费）。数据目录让您最多可免费存储一百万个对象。如果您存储一百万个以上的对象，将需要为超过一百万的每 100,000 个对象支付 1 美元。数据目录中的对象为表、分区或数据库。有关更多信息，请参阅 Glue 定价。

> **Important**
>
> 如果您在 2017 年 8 月 14 日之前使用 Amazon Athena 或 Amazon Redshift Spectrum 创建了表，则数据库和表将存储在 Athena 托管式目录中，该目录与 Amazon Glue 数据目录相互独立。要将 Amazon EMR 与这些表集成，您必须升级到 Amazon Glue 数据目录。有关更多信息，请参阅《Amazon Athena 用户指南》中的升级到 Amazon Glue 数据目录。

指定 Amazon Glue 数据目录作为元存储

您可以使用Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 指定 Amazon Glue 数据目录作为元存储。在使用 CLI 或 API 创建集群时，您可以使用 Presto 配置分类指定数据目录。此外，使用 Amazon EMR 5.16.0 及更高版本时，您可以使用配置分类指定其他 Amazon Web Services 账户中的数据目录。在使用控制台时，您可以使用 Advanced Options (高级选项) 或 Quick Options (快速选项) 指定数据目录。

New console

使用新控制台指定 Amazon Glue 数据目录作为 Hive 元存储

1. 登录 Amazon Web Services Management Console 并打开 Amazon EMR 控制台，网址为 https://console.aws.amazon.com/emr。
2. 在左侧导航窗格中的 EMR on EC2 下，选择 Clusters（集群），然后选择 Create cluster（创建集群）。
3. 在 Application bundle（应用程序包）下，选择 Presto。
4. 在Amazon Glue Data Catalog 设置下，选择用于 Presto 表元数据复选框。
5. 选择适用于集群的任何其他选项。
6. 要启动集群，选择 Create cluster（创建集群）。

Old console

使用旧控制台指定 Amazon Glue 数据目录作为默认 Presto 元存储

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 在 Software Configuration 下，选择版本 emr-5.10-0 或更高版本并选择 Presto。
4. 选择 Use for Presto table metadata，选择 Next，然后完成适合您的应用程序的集群的其他设置。

CLI

使用 Amazon CLI 指定 Amazon Glue 数据目录作为默认 Hive 元存储

有关如何在创建集群时指定以下配置分类的示例，请参阅配置应用程序。

Amazon EMR 5.16.0 及更高版本

- 将 `hive.metastore` 属性设置为 `glue`，如以下 JSON 示例所示。

```
[
  {
    "Classification": "presto-connector-hive",
    "Properties": {
      "hive.metastore": "glue"
    }
  }
]
```

要指定其他 Amazon Web Services 账户的数据目录，请添加 `hive.metastore.glue.catalogid` 属性，如以下 JSON 示例所示。将 *`acct-id`* 替换为数据目录的Amazon账户。Amazon EMR 版本 5.15.0 及更早版本不支持使用其他 Amazon Web Services 账户中的数据目录。

```
[
  {
    "Classification": "presto-connector-hive",
    "Properties": {
      "hive.metastore": "glue",
```

```
      "hive.metastore.glue.catalogid": "acct-id"
    }
  }
]
```

Amazon EMR 5.10.0 至 5.15.0

将 `hive.metastore.glue.datacatalog.enabled` 属性设置为 `true`，如以下 JSON 示例所示：

```
[
  {
    "Classification": "presto-connector-hive",
    "Properties": {
      "hive.metastore.glue.datacatalog.enabled": "true"
    }
  }
]
```

Amazon EMR 6.1.0 及更高版本使用 PrestoSQL (Trino)

从 EMR 版本 6.1.0 起，PrestoSQL 还支持 Glue 作为默认配置 Hive 元存储。为此，请使用 `prestosql-connector-hive` 配置分类将 `hive.metastore` 属性设置为 `glue`，如以下 JSON 示例所示。

Amazon EMR 6.4.0 及更高版本使用新名称 Trino 而不是 PrestoSQL。如果您使用 Trino，请在以下配置分类中将 *`prestosql-connector-hive`* 替换为 `trino-connector-hive`。

```
[
  {
    "Classification": "prestosql-connector-hive",
    "Properties": {
      "hive.metastore": "glue"
    }
  }
]
```

要在长时间运行的集群上切换元存储，您可以连接到主节点，直接编辑 `/etc/presto/conf/catalog/hive.properties` 文件中的属性值并重新启动 Presto 服务器（`sudo restart presto-server`），以便为您的发行版相应地手动设置这些值。如果将此方法与 Amazon EMR

5.15.0 及更高版本结合使用，请确保将 `hive.table-statistics-enabled` 设置为 `false`。在使用发行版 5.16.0 和更高版本时，不需要使用该设置；但不支持表和分区统计信息。

IAM 权限

集群的 EC2 实例配置文件必须具有适用于 Amazon Glue 操作的 IAM 权限。此外，如果您为 Amazon Glue 数据目录对象启用加密，还必须允许该角色加密、解密和生成用于加密的Amazon KMS key。

适用于 Amazon Glue 操作的权限

如果使用适用于 Amazon EMR 默认的 EC2 实例配置文件，则无需执行任何操作。附加到 `EMR_EC2_DefaultRole` 的 `AmazonElasticMapReduceforEC2Role` 托管式策略允许所有必要 Amazon Glue 操作。但是，如果您指定自定义 EC2 实例配置文件和权限，则必须配置合适的 Amazon Glue 操作。使用 `AmazonElasticMapReduceforEC2Role` 托管策略作为起点。如需了解更多信息，请参阅《Amazon EMR 管理指南》中的集群 EC2 实例的服务角色（EC2 实例配置文件）。

用于加密和解密 Amazon Glue 数据目录的权限

您的实例配置文件需要使用密钥加密和解密数据的权限。如果以下语句适用，您不必配置这些权限：

- 您使用 Amazon Glue 的托管式密钥启用 Amazon Glue Data Catalog 对象的加密。
- 您使用的是同一Amazon Web Services 账户的集群，其作为 AmazonGlue Data Catalog。

否则，您必须将以下语句添加到附加到 EC2 实例配置文件的权限策略。

```
[
    {
        "Version": "2012-10-17",
        "Statement": [
            {
                "Effect": "Allow",
                "Action": [
                    "kms:Decrypt",
                    "kms:Encrypt",
                    "kms:GenerateDataKey"
                ],
                "Resource": "arn:aws:kms:region:acct-
id:key/12345678-1234-1234-1234-123456789012"
            }
        ]
```

```
    }
]
```

有关 Amazon Glue 数据目录加密的更多信息，请参阅《Amazon Glue 开发人员指南》中的加密您的数据目录。

基于资源的权限

如果您将 Amazon Glue 与 Amazon EMR 中的 Hive、Spark 或 Presto 结合使用，Amazon Glue 支持使用基于资源的策略来控制对数据目录资源的访问权限。这些资源包括数据库、表、连接和用户定义的函数。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的 Amazon Glue 资源策略。

当使用基于资源的策略来限制从 Amazon EMR 中访问 Amazon Glue 时，在权限策略中指定的委托人必须是与创建集群时指定的 EC2 实例配置文件相关联的角色 ARN。例如，对于附加到目录的基于资源的策略，您可以使用以下示例中显示的格式为集群 EC2 实例的默认服务角色指定角色 ARN，将 *EMR_EC2_DefaultRole* 指定为 `Principal`：

```
arn:aws:iam::acct-id:role/EMR_EC2_DefaultRole
```

acct-id 可以与 Amazon Glue 账户 ID 不同。这允许从不同账户中的 EMR 集群进行访问。您可以指定多个委托人，且每个委托人都可以来自不同的账户。

使用 Amazon Glue 数据目录时的注意事项

在使用 Amazon Glue 数据目录作为 Presto 的元存储时，请考虑以下项目：

- 不支持在 Amazon Glue 中重命名表。
- 当您创建 Hive 表而不指定 `LOCATION` 时，表数据存储在通过 `hive.metastore.warehouse.dir` 属性指定的位置。默认情况下，这是 HDFS 中的一个位置。如果另一个集群需要访问该表，则它将失败，除非它有足够的权限访问创建该表的集群。此外，由于 HDFS 存储是暂时性的，因此如果集群终止，表数据将丢失，并且必须重新创建该表。建议您在使用 Amazon Glue 创建 Hive 表时，指定 Amazon S3 中的一个 `LOCATION`。此外，也可以使用 `hive-site` 配置分类来为 `hive.metastore.warehouse.dir` 指定 Amazon S3 中的位置，它适用于所有 Hive 表。如果表在 HDFS 位置创建，并且创建该表的集群仍在运行，您可以在 Amazon Glue 中更新 Amazon S3 中表的位置。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的使用 Amazon Glue 控制台上的表。
- 不支持包含引号和撇号的分区值，例如 `PARTITION (owner="Doe's")`.
- emr-5.31.0 及更高版本支持列统计数据。

- 不支持使用 Hive 授权。作为替代方案，考虑使用基于 Amazon Glue 资源的策略。有关更多信息，请参阅将用于 Amazon EMR 访问的基于资源的策略用于 Amazon Glue 数据目录。

使用 S3 Select Pushdown 搭配 Presto 提高性能

使用 Amazon EMR 发行版 5.18.0 及更高版本，您可以将 S3 Select Pushdown 与 Presto on Amazon EMR 搭配使用。此功能允许 Presto 将投影操作（例如，`SELECT`）和谓词操作（例如，`WHERE`）的计算工作“下推”至 Amazon S3。这允许查询仅从 Amazon S3 中检索所需数据，从而可以提高性能并减少某些应用程序在 Amazon EMR 和 Amazon S3 之间传输的数据量。

S3 Select Pushdown 是否适合我的应用程序？

建议您分别在使用和不使用 S3 Select Pushdown 的情况下对您的应用程序进行基准检验，以查看其是否适用于您的应用程序。

使用以下准则来确定您的应用程序是否为使用 S3 Select 的候选项：

- 您的查询将筛选掉原始数据集的一半以上的数据。
- 您的查询筛选谓词使用具有 Presto 和 S3 Select 支持的数据类型的列。S3 Select Pushdown 不支持时间戳、实数和双精度数据类型。建议对数值数据使用十进制数据类型。有关 S3 Select 支持的数据类型的更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的数据类型。
- 您在 Amazon S3 和 Amazon EMR 集群之间的网络连接具有良好的传输速度和可用带宽。Amazon S3 不压缩 HTTP 响应，因此响应大小可能会根据压缩的输入文件而增大。

注意事项和限制

- 仅支持以 CSV 格式存储的对象。对象可以解压缩，也可以选择使用 gzip 或 bzip2 压缩。
- 不支持 `AllowQuotedRecordDelimiters` 属性。如果指定该属性，则查询将失败。
- 使用客户提供的加密密钥进行的 Amazon S3 服务器端加密（SSE-C）与客户端加密都不受支持。
- S3 Select Pushdown 不能替代使用列式或压缩文件格式（如 ORC 或 Parquet）。

启用 S3 Select Pushdown with PrestoDB 或 Trino

要在 Amazon EMR 上启用 PrestoDB 的 S3 Select Pushdown，请使用 `presto-connector-hive` 配置分类以将 `hive.s3select-pushdown.enabled` 设置为 `true`，如以下示例所示。有关更多信

息，请参阅配置应用程序。还必须设置 hive.s3select-pushdown.max-connections 值。对于大部分应用程序，*500* 的默认设置应该能满足需求。有关更多信息，请参阅下面的 了解和调整 hive.s3select-pushdown.max-connections。

对于 EMR 版本 6.1.0 - 6.3.0 或更高版本上的 PrestoSQL，只需将以下示例中的 *presto-connector-hive* 替换为 `prestosql-connector-hive`。

Amazon EMR 6.4.0 及更高版本使用新名称 Trino 而不是 PrestoSQL。如果您使用 Trino，请在以下示例中将 *presto-connector-hive* 替换为 `trino-connector-hive`。

```
[
    {
        "classification": "presto-connector-hive",
        "properties": {
            "hive.s3select-pushdown.enabled": "true",
            "hive.s3select-pushdown.max-connections": "500"
        }
    }
]
```

了解和调整 hive.s3select-pushdown.max-connections

默认情况下，Presto 使用 EMRFS 作为其文件系统。`emrfs-site` 配置分类中的设置 `fs.s3.maxConnections` 指定允许通过 Presto 的 EMRFS 连接到 Amazon S3 的最大客户端连接数。默认情况下，此连接数为 500。S3 Select Pushdown 在访问 Amazon S3 执行谓词操作时绕过 EMRFS。在此示例中，`hive.s3select-pushdown.max-connections` 的值确定从 Worker 节点执行这些操作所允许的最大客户端连接数。但是，Presto 启动的任何未向下推送的发往 Amazon S3 的请求（例如 GET 操作）将继续由 `fs.s3.maxConnections` 的值控制。

如果您的应用程序遇到错误“Timeout waiting for connection from pool”，请增加 `hive.s3select-pushdown.max-connections` 和 `fs.s3.maxConnections` 的值。

添加数据库连接器

在创建集群时，可以使用配置分类来配置 JDBC 连接器属性。配置分类以 `presto-connector` 开头，例如 `presto-connector-postgresql`。可用的配置分类取决于 Amazon EMR 发行版。有关适用于最新版本的配置分类，请参阅 Amazon EMR 5.36.1 的 the section called “配置分类”。如果您使用的是 Amazon EMR 的其他版本，请参阅 Amazon EMR 5.x 发行版 查看配置分类。有关可通过每个连接器配置的属性的详细信息，请参阅 https://prestodb.io/docs/current/connector.html。

Example – 使用 PostgreSQL JDBC 连接器配置集群

要启动已安装和配置的 PostgreSQL 连接器的集群，请首先创建一个指定包含以下内容的配置分类的 JSON 文件（例如 `myConfig.json`）然后将其本地保存。

按照 Presto 文档中 PostgreSQL 连接器主题中所示，根据您的设置替换相应的连接属性。

```
[
  {
    "Classification": "presto-connector-postgresql",
    "Properties": {
      "connection-url": "jdbc:postgresql://example.net:5432/database",
      "connection-user": "MYUSER",
      "connection-password": "MYPASS"
    },
    "Configurations": []
  }
]
```

创建集群时，请按照以下示例中所示，使用 `--configurations` 选项引用 JSON 文件的路径，其中 `myConfig.json` 位于您运行以下命令的同一目录中：

```
aws emr create-cluster --name PrestoConnector --release-label emr-5.36.1 --instance-
type m5.xlarge \
--instance-count 2 --applications Name=Hadoop Name=Hive Name=Pig Name=Presto \
--use-default-roles --ec2-attributes KeyName=myKey \
--log-uri s3://my-bucket/logs --enable-debugging \
--configurations file://myConfig.json
```

通过 Presto on Amazon EMR 使用 SSL/TLS 和配置 LDAPS

使用 Amazon EMR 发行版 5.6.0 及更高版本，您可以启用 SSL/TLS 以帮助保护 Presto 节点之间的内部通信安全。您可以为传输中加密设置安全配置以执行该操作。有关更多信息，请参阅《Amazon EMR 管理指南》中的加密选项和使用安全配置设置集群安全性。

在将安全配置与传输中加密一起使用时，Amazon EMR 会为 Presto 执行以下操作：

- 分发您在整个 Presto 集群中用于传输中加密的加密构件或证书。有关更多信息，请参阅为传输中的数据加密提供证书。
- 使用 `presto-config` 配置分类设置以下属性，该分类对应于 Presto 的 `config.properties` 文件：

  - 在所有节点上将 `http-server.http.enabled` 设置为 `false`，这将禁用 HTTP 以便支持 HTTPS。这要求您在为传输中加密设置安全配置时提供适用于公有和私有 DNS 的证书。执行此操作的一种方法是使用支持多个域的 SAN（使用者备用名称）证书。
  - 设置 `http-server.https.*` 值。有关配置详细信息，请参阅 Presto 文档中的 LDAP 身份验证。
- 对于 EMR 6.1.0 版及更高版本上的 PrestoSQL（Trino），Amazon EMR 会自动配置共享密钥，以实现集群节点之间的安全内部通信。您无需执行任何额外的配置即可启用此安全功能，并且您可以使用自己的私有密钥覆盖配置。有关 Trino 内部身份验证的信息，请参阅 Trino 353 文档：安全的内部通信。

此外，对于 Amazon EMR 发行版 5.10.0 及更高版本，您可以为使用 HTTPS 建立的到 Presto 协调器的客户端连接设置 LDAP 身份验证。该设置使用安全 LDAP（LDAPS）。必须在 LDAP 服务器上启用 TLS，并且 Presto 集群必须使用启用了传输中数据加密的安全配置。需要使用额外的配置。配置选项因使用的 Amazon EMR 发行版而有所不同。有关更多信息，请参阅为 Presto on Amazon EMR 使用 LDAP 身份验证。

默认情况下，Presto on Amazon EMR 使用端口 8446 进行内部 HTTPS 通信。用于内部通信的端口必须与用于通过客户端 HTTPS 访问 Presto 协调器的端口相同。`http-server.https.port` 配置分类中的 `presto-config` 属性指定该端口。

为 Presto on Amazon EMR 使用 LDAP 身份验证

可以按照本节中的步骤配置 LDAP。请参阅每个步骤以了解示例以及指向更多信息的链接。

配置 LDAP 身份验证的步骤

- 步骤 1：收集有关 LDAP 服务器的信息并将服务器证书复制到 Amazon S3
- 步骤 2：设置安全配置
- 步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON
- 步骤 4：创建脚本以复制 LDAP 服务器证书并将其上传到 Amazon S3
- 步骤 5：创建集群

步骤 1：收集有关 LDAP 服务器的信息并将服务器证书复制到 Amazon S3

您需要在下一节中使用这些信息和内容，以便从 LDAP 服务器中配置 LDAP 身份验证。

LDAP 服务器的 IP 地址或主机名

Amazon EMR 主节点上的 Presto 协调器必须能够访问具有指定的 IP 地址或主机名的 LDAP 服务器。默认情况下，Presto 使用 LDAPS 通过端口 636 与 LDAP 服务器通信。如果您的 LDAP 实施需要使用自定义端口，您可以使用 `ldap.url` 属性（Amazon EMR 5.16.0 或更高版本）或者 `authentication.ldap.url`（早期版本）指定该端口。将自定义端口替换为 636，如 步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON 中的 `presto-config` 配置分类示例所示。确保任何防火墙和安全组允许端口 636（或自定义端口）以及端口 8446（或自定义端口）上的入站和出站流量，端口 8446 用于内部集群通信。

LDAP 服务器证书

您必须将证书文件上传到 Amazon S3 中的安全位置。有关更多信息，请参阅《Amazon Simple Storage Service 用户指南》中的如何将文件和文件夹上传到 S3 存储桶。您可以创建一个引导操作，以便在集群启动时将该证书从 Amazon S3 复制到集群中的每个节点。在步骤 4：创建脚本以复制 LDAP 服务器证书并将其上传到 Amazon S3中。示例证书为 *`s3://MyBucket/ldap_server.crt`*。

LDAP 服务器的匿名绑定设置

如果 PrestoDB 禁用了匿名绑定，您需要使用有权限绑定到 LDAP 服务器的账户的用户 ID（UID）和密码，以便 Presto 服务器建立连接。您可以使用 `internal-communication.authentication.ldap.user` 配置分类中的 `internal-communication.authentication.ldap.password` 和 `presto-config` 属性指定 UID 和密码。Amazon EMR 5.10.0 不支持这些设置，因此，在使用该发行版时，LDAP 服务器上必须支持匿名绑定。

请注意，Trino 不需要匿名绑定配置。

获取 LDAP 服务器上的匿名绑定状态

- 从 Linux 客户端中使用 ldapwhoami 命令，如以下示例所示：

  ```
  ldapwhoami -x -H ldaps://LDAPServerHostNameOrIPAddress
  ```

  如果不允许匿名绑定，该命令将返回以下内容：

  ```
  ldap_bind: Inappropriate authentication (48)
  additional info: anonymous bind disallowed
  ```

验证账户是否具有使用简单身份验证的 LDAP 服务器的权限

- 从 Linux 客户端中使用 ldapwhoami 命令，如以下示例所示。该示例使用虚构用户 *presto*，该用户名存储在具有虚构主机名 *ip-xxx-xxx-xxx-xxx.ec2.internal* 的 EC2 实例上运行的 Open LDAP 服务器中。该用户与组织单位 (OU) *admins* 和密码 *123456* 相关联：

  ```
  ldapwhoami -x -w "123456" -D uid=presto,ou=admins,dc=ec2,dc=internal -H ldaps://ip-xxx-xxx-xxx-xxx.ec2.internal
  ```

  如果该账户有效并具有相应的权限，该命令将返回：

  ```
  dn:uid=presto,ou=admins,dc=ec2,dc=internal
  ```

为了清楚起见，步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON 中的示例配置包含该账户，但 5.10.0 示例除外，该发行版不支持该账户。如果 LDAP 服务器使用匿名绑定，请删除 `internal-communication.authentication.ldap.user` 和 `internal-communication.authentication.ldap.password` 名称/值对。

Presto 用户的 LDAP 可分辨名称 (DN)

为 Presto 指定 LDAP 配置时，您可以指定包含 `${USER}` 以及组织单位（OU）和额外域组件（DC）的绑定模式。在密码身份验证期间，Presto 将 `${USER}` 替换为每个用户的实际用户 ID（UID），以便与该绑定模式指定的可分辨名称（DN）相匹配。您需要使用合格用户所属的 OU 及其 DC。例如，要允许 `admins` 域上的 `corp.example.com` OU 中的用户在 Presto 中进行身份验证，您可以将 `${USER},ou=admins,dc=corp,dc=example,dc=com` 指定为用户绑定模式。

> **Note**
>
> 当您使用Amazon CloudFormation时，您需要使用 Fn:: Sub 函数才能将${USER}替换为实际的用户 ID (UID)。有关更多信息，请参阅《Amazon CloudFormation 用户指南》 中的 Fn::Sub 主题。

在使用 Amazon EMR 5.10.0 时，您只能指定一种此类模式。在使用 Amazon EMR 5.11.0 或更高版本时，您可以指定多种模式并以冒号（:）分隔。尝试在 Presto 中进行身份验证的用户先与第一种模式进行比较，然后与第二种模式进行比较，依此类推。有关示例，请参阅 步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON。

步骤 2：设置安全配置

创建一个安全配置并启用传输中加密。有关更多信息，请参阅《Amazon EMR 管理指南》中的创建安全配置。在设置传输中加密时提供的加密构件用于加密 Presto 节点之间的内部通信。有关更多信息，请参阅为传输中的数据加密提供证书。LDAP 服务器证书用于对到 Presto 服务器的客户端连接进行身份验证。

步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON

您可以使用 `presto-config` 配置分类为 LDAP 设置 Presto 属性。根据 Amazon EMR 发行版和安装的 Presto（PrestoDB 或 Trino）的不同，`presto-config` 的格式和内容稍有不同。在本节后面提供了配置差异示例。有关更多信息，请参阅配置应用程序。

以下步骤假定您将 JSON 数据保存到 *`MyPrestoConfig.json`* 文件中。如果使用控制台，请将该文件上传到 Amazon S3 中的安全位置，以便在创建集群时引用该文件。如果使用 Amazon CLI，您可以在本地引用该文件。

Example 采用 PrestoSQL (Trino) 的 Amazon EMR 6.1.0 及更高版本

以下示例使用 步骤 1：收集有关 LDAP 服务器的信息并将服务器证书复制到 Amazon S3 中的 LDAP 主机名，以便在 LDAP 服务器中验证身份以进行绑定。指定了两种用户绑定模式，它指示 LDAP 服务器上的 `admins` OU 和 `datascientists` OU 中的用户可以在 Trino 服务器上作为用户进行身份验证。绑定模式由冒号 (`:`) 分隔。

Amazon EMR 6.4.0 及更高版本使用新名称 Trino 而不是 PrestoSQL。如果您使用 Trino，请在以下配置分类中将 *`prestosql-config`* 替换为 `trino-config`、*`prestosql-password-authenticator`* 和 `trino-password-authenticator`。

```
[
   {
      "Classification":"prestosql-config",
      "Properties":{
         "http-server.authentication.type":"PASSWORD"
      }
   },
   {
      "Classification":"prestosql-password-authenticator",
      "Properties":{
         "password-authenticator.name":"ldap",
         "ldap.url":"ldaps://ip-xxx-xxx-xxx-xxx.ec2.internal:636",
```

```
         "ldap.user-bind-pattern": "uid=${USER},ou=admins,dc=ec2,dc=internal:uid=
${USER},ou=datascientists,dc=ec2,dc=internal"
      }
   }
]
```

Example Amazon EMR 5.16.0 及更高版本

以下示例使用步骤 1：收集有关 LDAP 服务器的信息并将服务器证书复制到 Amazon S3中的 LDAP 用户 ID 和密码以及 LDAP 主机名，以便在 LDAP 服务器中验证身份以进行绑定。指定了两种用户绑定模式，它指示 LDAP 服务器上的 admins OU 和 datascientists OU 中的用户可以在 Presto 服务器上作为用户进行身份验证。绑定模式由冒号 (:) 分隔。

```
[{
        "Classification": "presto-config",
                "Properties": {
                        "http-server.authentication.type": "PASSWORD"
                }
        },
        {
                "Classification": "presto-password-authenticator",
                "Properties": {
                        "password-authenticator.name": "ldap",
                        "ldap.url": "ldaps://ip-xxx-xxx-xxx-xxx.ec2.internal:636",
                        "ldap.user-bind-pattern": "uid=
${USER},ou=admins,dc=ec2,dc=internal:uid=${USER},ou=datascientists,dc=ec2,dc=internal",
                        "internal-communication.authentication.ldap.user": "presto",
                        "internal-communication.authentication.ldap.password": "123456"
                }
        }]
```

Example Amazon EMR 5.11.0 至 5.15.0

这些发行版的 presto-config 配置分类的格式略有不同。以下示例指定与上一示例相同的参数。

```
[{
        "Classification": "presto-config",
                "Properties": {
                        "http-server.authentication.type": "LDAP",
                        "authentication.ldap.url": "ldaps://ip-xxx-xxx-xxx-
xxx.ec2.internal:636",
                        "authentication.ldap.user-bind-pattern": "uid=
${USER},ou=admins,dc=ec2,dc=internal:uid=${USER},ou=datascientists,dc=ec2,dc=internal",
```

```
                        "internal-communication.authentication.ldap.user": "presto",
                        "internal-communication.authentication.ldap.password": "123456"
                }
        }]
```

Example Amazon EMR 5.10.0

Amazon EMR 5.10.0 仅支持匿名绑定，因此，将省略这些条目。此外，只能指定一种绑定模式。

```
[{
        "Classification": "presto-config",
                "Properties": {
                        "http-server.authentication.type": "LDAP",
                        "authentication.ldap.url": "ldaps://ip-xxx-xxx-xxx-
xxx.ec2.internal:636",
                        "ldap.user-bind-pattern": "uid=
${USER},ou=prestousers,dc=ec2,dc=internal"
                }
        }]
```

步骤 4：创建脚本以复制 LDAP 服务器证书并将其上传到 Amazon S3

创建一个脚本以将证书文件复制到集群中的每个节点，然后将其添加到密钥存储中。请使用文本编辑器创建脚本，保存该脚本，然后将其上传到 Amazon S3 中。在步骤 5：创建集群中，脚本文件是作为 *s3://MyBucket/LoadLDAPCert.sh* 引用的。

以下示例脚本使用默认密钥存储密码 *changeit*。我们建议您在创建集群后连接到主节点，并使用 keytool 命令更改密钥存储密码。

```
#!/bin/bash
aws s3 cp s3://MyBucket/ldap_server.crt .
sudo keytool -import -keystore /usr/lib/jvm/jre-1.8.0-openjdk.x86_64/lib/security/
cacerts -trustcacerts -alias ldap_server -file ./ldap_server.crt -storepass changeit -
noprompt
```

步骤 5：创建集群

在创建集群时，您可以指定 Presto 以及希望 Amazon EMR 安装的其他应用程序。以下示例还引用 JSON 中的配置分类属性，但也可以指定内联的配置分类。

使用 Amazon EMR 控制台创建具有 LDAP 身份验证的 Presto 集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 选择 Presto 以及 Amazon EMR 安装的其他应用程序，然后在 Software Configuration (软件配置) 下面选择要使用的 Amazon EMR 发行版。仅 Amazon EMR 5.10.0 和更高版本支持 LDAP 身份验证。
4. 在 Edit software settings (编辑软件设置) 下面，选择 Load JSON from S3 (从 S3 加载 JSON)，输入在步骤 3：使用 LDAP 的 Presto 属性创建配置 JSON中创建的 JSON 配置文件的 Amazon S3 位置，然后选择 Next (下一步)。
5. 配置集群硬件和网络，然后选择下一步。
6. 选择 Bootstrap Actions (引导操作)。对于添加引导操作，请选择自定义操作，然后选择配置并添加。
7. 输入引导操作的名称，输入在中创建的脚本位置步骤 4：创建脚本以复制 LDAP 服务器证书并将其上传到 Amazon S3（如 s3://MyBucket/LoadLDAPCert.sh），然后选择添加。
8. 在常规选项、标签和其他选项下面，选择适合您的应用程序的设置，然后选择下一步。
9. 选择身份验证和加密，然后选择您在中创建的安全配置步骤 2：设置安全配置。
10. 选择适合您的应用程序的其他安全选项，然后选择创建集群。

使用 Amazon CLI 创建具有 LDAP 身份验证的 Presto 集群

- 使用 `aws emr create-cluster` 命令。至少，指定 Presto 应用程序，以及在以前步骤中创建的 Presto 配置分类、引导脚本和安全配置。以下示例将配置文件作为在运行该命令的同一目录中保存的 JSON 文件引用。另一方面，引导脚本必须保存在 Amazon S3 中。下面的示例使用了 `s3://MyBucket/LoadLDAPCert.sh`。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\)。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

```
aws emr create-cluster  --applications Name=presto --release-label emr-5.16.0 \
```

```
--use-default-roles --ec2-attributes KeyName=MyKeyPair,SubnetId=subnet-1234ab5
 \ --instance-count 3 --instance-type m5.xlarge --region us-west-2 --name
 "MyPrestoWithLDAPAuth" \
--bootstrap-actions Name="Distribute LDAP server cert",Path="s3://MyBucket/
LoadLDAPCert.sh" \
--security-configuration MyPrestoLDAPSecCfg --configurations file://
MyPrestoConfig.json
```

激活 Presto 严格模式

在某些情况下，长时间运行的查询会产生极高的成本，并且可能导致 Amazon EMR 使用更多的集群资源。这会与集群上的其他工作负载竞争资源。在 Amazon EMR 版本 6.8 及更高版本中，您可以使用严格模式功能，拒绝以下类型的长时间运行的查询或向您发出警告：

- 分区列上没有谓词的查询，这会导致对大量数据进行表扫描
- 在两个大表之间使用交叉连接的查询
- 对大量行进行无限制排序的查询

在 Presto 完全优化查询计划后，严格模式将开始运行。要根据您的查询需求使用和自定义严格模式，您可以通过以下方式配置 Presto。

Presto 严格模式配置

设置	描述	默认值
`strict-mode-enabled`	开启和关闭严格模式。值 `true` 表示严格模式已开启。	`false`
`strict-mode-fail-query`	如果严格模式检测到可能长时间运行的查询，则会拒绝相关查询。如果值为 `false`，则 Amazon EMR 只会发出警告。	`false`
`strict-mode-restrictions`	指定严格模式开启时将应用的限制。严格模式支持以下限制：MANDATORY_PARTITION_PREDICA	MANDATORY_PARTITION_PREDICATE、DISALLOW_CROSS_JOIN、LIMITED_SORT

设置	描述	默认值
	TE、DISALLOW_CROSS_JOIN 和 LIMITED_SORT。	

要试用严格模式，您可以覆盖这些配置，或者在使用 Presto 客户端时将其设置为会话属性。

在通过 Amazon Web Services Management Console创建集群时设置配置

1. 选择 Create cluster（创建集群）并选择 Amazon EMR 版本 6.8.0，然后选择 Presto 或 Trino。有关更多信息，请参阅 安装 PrestoDB 和 Trino。
2. 直接指定严格模式的配置属性，或者将 JSON 文件上传到 Amazon S3。您也可为您的元数据存储选择 。指定您的 VPC、子网、引导操作、密钥对和安全组。选择 Create cluster（创建集群）以创建集群。
3. 登录到集群的主节点，然后运行 `presto-cli` 或 `trino-cli`。
4. 提交查询。严格模式会验证每个查询并确定它是否在长时间运行。根据您的 `strict-mode-fail-query` 设置，Amazon EMR 会拒绝查询或发出警告。
5. 完成查询后，终止集群并删除您的资源。

通过 Amazon CLI 在正在运行的集群上设置配置

1. 使用 Amazon CLI 登录到集群的主节点，然后运行 `presto-cli` 或 `trino-cli`。
2. 使用所需的值运行以下命令。

```
set session strict_mode_enabled = true;
set session strict_mode_fail_query = false;
set session strict_mode_restrictions = 'DISALLOW_CROSS_JOIN,LIMITED_SORT';
```

注意事项

在使用严格模式时，请注意以下几点：

- 在某些情况下，严格模式可能会拒绝并未消耗大量资源的短时查询。例如，对小表的查询未应用动态筛选，也未用交叉连接替内部连接。这可能导致查询使用强制分区谓词或禁止交叉连接。发生这种情况时，严格模式会拒绝该查询。

- 严格模式检查仅适用于 SELECT、INSERT、CREATE TABLE AS SELECT 和 EXPLAIN ANALYZE 查询类型。
- 您只能通过 Hive 连接器使用对强制分区谓词的限制。

在 Presto 中处理竞价型实例丢失

在 Amazon EMR 中使用竞价型实例时，您可以通过较低的成本利用 Amazon EC2 容量运行大数据工作负载。为换取更低的成本，Amazon EC2 可能会在发出两分钟通知后中断竞价型实例。当您终止某个节点时，Presto 最长可能需要 10 分钟才会返回错误。这会导致错误报告出现不必要的延迟，并可能导致重试。快速终止功能可让您控制 Presto 处理已终止节点的方式。

Presto 协调器的作用是通过定期轮询其状态来跟踪所有 Worker 节点。不使用快速终止功能时，协调器不会向 YARN NodeManager 查询各个节点的状态。这可能导致在查询失败之前出现长时间的重试循环。使用快速终止功能后，一旦轮询未能连接到主机，Presto 协调器将会向 NodeManager 查询节点状态。如果 NodeManager 显示节点处于非活动状态，则 Presto 会放弃继续重试，查询失败并返回 `NODE_DECOMMISSIONED` 错误。

下面的一组配置参数允许您控制和自定义 Presto 在节点终止时的行为。

Presto 的节点故障处理配置

设置	描述	默认值
`query.remote-task.max-backoff-duration`	协调器会继续尝试从 Worker 节点获取远程任务状态的持续时间。	10 分钟
`query.remote-task.quick-terminate-node-failure`	如果协调器无法连接到该节点或在该节点上运行的工作线程，则将激活快速节点故障。`query.remote-task.terminate-on-connect-exception` 的值将决定协调器是必须连接到该节点还是必须连接到工作线程。 节点查询失败，并且 Amazon EMR 会将该节点从可用工作线程列表中移除。发生这种情况	`true`

设置	描述	默认值
	时，您将无法使用该节点来计划新的查询。 如果您将此值设置为 `false`，Presto 会恢复其先前的行为，即 Presto 协调器在将该节点标记为不可用之前再次尝试连接到该节点（对于 `query.remote-task.max-backoff-duration`），并且将该节点上正在进行的查询设为失败。	
`query.remote-task.terminate-on-connect-exception`	如果可以连接到主机但协调器无法连接到主机上的工作进程，则指定 Amazon EMR 是否应作为节点。将此值设置为 `true` 时，则将在无法访问主机时激活快速查询失败。	`false`

Trino 中的容错执行

容错执行是 Trino 中的一种机制，集群可以使用该机制来减少查询失败。为此，它会在查询失败时重试查询或其组件任务。激活容错执行后，中间交换数据会假脱机，并且如果在查询执行期间发生 Worker 中断或其他故障，可被其他 Worker 重用。

有关 Trino 中容错执行的更多信息，请参阅 Trino 博客上的 Project Tardigrade delivers ETL at Trino speeds to early users（Project Tardigrade 以 Trino 般高速向早期用户提供 ETL）。

配置

默认情况下停用容错执行。若要激活该功能，请根据所需的重试策略将 `trino-config` 分类中的 `retry-policy` 配置属性设置为 `QUERY` 或 `TASK`，如下所示。

```
{"classification":
    "trino-config",
```

```
"properties":
    {
    "retry-policy":
        "QUERY"
    }
}
```

QUERY 重试策略指示 Trino 在 Worker 节点上发生错误时自动重试查询。当 Trino 集群的大部分工作负载包含许多小查询时，建议您使用 QUERY 重试策略。

TASK 重试策略指示 Trino 在失败时重试单个查询任务。建议在 Trino 执行大批量查询时使用此策略。集群可以更有效地重试查询中的较小任务，而不是重试整个查询。

交换管理器

交换管理器存储和管理假脱机数据，以实现容错执行。它利用外部存储来存储超出内存缓冲区大小的溢出数据。您可以配置基于文件系统的交换管理器，将假脱机数据存储在指定位置，例如 Amazon S3、Amazon S3 兼容系统或 HDFS。

Amazon EMR 发行版 6.9.0 及更高版本包括用于配置交换管理器的 `trino-exchange-manager` 分类。这些版本还支持 HDFS 进行假脱机。

设置交换管理器

使用 `trino-exchange-manager` 配置分类来配置交换管理器。该分类会在协调器和所有 Worker 节点上创建 `etc/exchange-manager.properties` 配置文件。分类还将 `exchange-manager.name` 配置属性设置为 `filesystem`。

默认情况下，Amazon EMR 发行版 6.9.0 及更高版本使用 HDFS 作为交换管理器。HDFS 在 Amazon EMR EC2 集群中提供，默认情况下，假脱机在 `trino-exchange/` 目录中进行。要使用默认设置，请设置以下配置：

```
{"Classification":
    "trino-exchange-manager"
}
```

如果要提供自定义位置，请在 `trino-exchange-manager` 分类中设置以下属性：

- 将 `exchange.use-local-hdfs` 设置为 `true`。
- 在 HDFS 中将 `exchange.base-directories` 设置为自定义目录位置，例如 `exchange.base-directories=/exchange`。如果自定义目录尚不在 HDFS 中，Amazon EMR 将创建该目录。

HDFS 交换管理器配置

根据内部测试结果，与其他基于云的文件系统相比，我们建议您假脱机到本地 HDFS 以实现更好的查询性能。您可以使用 HDFS，为交换管理器设置以下配置。

配置	描述	默认设置
`exchange.hdfs.block-size`	HDFS 存储的块大小	4MB
`hdfs.config.resources`	配置 HDFS 的文件路径列表	如果 `exchange.use-local-hdfs` 为 `true`，则使用 `core-site.xml` 、`hdfs-site.xml` 文件的路径；否则为 `null`

有关其他容错执行配置属性及如何设置 Amazon S3 或其他 Amazon S3 兼容系统以进行假脱机的信息，请参阅 Trino 文档的容错执行页面。

注意事项和限制

- 如果启用容错执行，则会在设置 `retry-policy` 时禁用不支持 `write` 的连接器的 `write` 操作。在 Amazon EMR 发行版 6.9.0 中，Delta Lake、Hive 和 Iceberg 连接器支持使用 `retry-policy` 的 `write` 操作。
- 如果您使用交换管理器并执行高成本的 I/O 操作，当交换管理器将中间数据假脱机到外部存储时，查询性能可能会下降。

使用采用 Graceful Decommission 的 Presto 自动扩展配置

Amazon EMR 发行版 5.30.0 及更高版本包含一项可用于为某些扩展操作设置宽限期的功能。宽限期允许 Presto 任务在节点因横向缩减大小调整操作或自动扩展策略请求而终止之前继续运行。有关扩展规则的更多信息，请参阅《Amazon EMR 管理指南》中的了解自动扩展规则。采用 Graceful Decommission 的 Presto 弹性伸缩配置可防止在正在停用的节点上计划新任务，同时允许在达到关机超时之前完成已在运行的任务。正在运行的查询将在节点停用之前完成执行。实例集不支持弹性伸缩。

您可以控制在收到自动扩展关闭请求后必须完成 Presto 任务的时间长度。默认情况下，Amazon EMR 的关闭超时为 0 分钟，这意味着如果缩减请求需要，则 Amazon EMR 会立即终止节点及其上运行的任

何 Presto 任务。要为 Amazon EMR 上进行的 Presto 任务设置更长的超时，以允许在缩减集群之前完成正在运行的查询，请使用 `presto-config` 配置分类将 `graceful-shutdown-timeout` 参数设置为大于零的值（单位为秒或分钟）。有关更多信息，请参阅配置应用程序。

例如，将 `graceful-shutdown-timeout` 值增大至 `"30m"` 以指定 30 分钟的超时时段。在关闭超时期结束后，如果标记为停用的节点正在等待查询任务完成，则系统将强制终止该节点，查询失败。如果查询任务在 5 分钟内完成，则系统将在到达 5 分钟时终止标记为停用的节点，前提是其他 YARN 应用程序已完成执行。

Example 采用 Graceful Decommission 的 Presto 自动扩展配置示例

将 `graceful-shutdown-timeout` 值替换为适合您的设置的分钟数。没有最大值。下面的示例将超时值设置为 `1800` 秒（30 分钟）。

```
[
    {
        "classification": "presto-config",
        "properties": {
            "graceful-shutdown-timeout": "1800s"
        }
    }
]
```

限制

PrestoDB Graceful Decommission 不适用于禁用 HTTP 连接的 EMR 集群，例如 `http-server.http.enabled` 设置为 `false` 时。Trino完全不支持 Graceful Decommission，不论 `http-server.http.enabled` 设置如何。

Presto on Amazon EMR 注意事项

运行 Presto on Amazon EMR 时应注意以下限制。

Presto 命令行可执行文件

在 Amazon EMR 中，PrestoDB 和 Trino 均使用相同的命令行可执行文件 `presto-cli`，如以下示例所示。

```
presto-cli --catalog hive
```

不可配置的 Presto 部署属性

您使用的 Amazon EMR 版本决定了可用的 Presto 部署配置。有关这些配置属性的更多信息，请参阅 Presto 文档中的 部署 Presto。下表显示了 Presto `properties` 文件的不同配置选项。

文件	可配置
`log.properties`	PrestoDB：在 Amazon EMR 版本 4.0.0 及更高版本中可配置。使用 `presto-log` 配置分类。 Trino（PrestoSQL）：在 Amazon EMR 版本 6.1.0 及更高版本中可配置。使用 `prestosql-log` 或 `trino-log` 配置分类。
`config.properties`	PrestoDB：在 Amazon EMR 版本 4.0.0 及更高版本中可配置。使用 `presto-config` 配置分类。 Trino（PrestoSQL）：在 Amazon EMR 版本 6.1.0 及更高版本中可配置。使用 `prestosql-config` 或 `trino-config` 配置分类。
`hive.properties`	PrestoDB：在 Amazon EMR 版本 4.1.0 及更高版本中可配置。使用 `presto-connector-hive` 配置分类。 Trino（PrestoSQL）：在 Amazon EMR 版本 6.1.0 及更高版本中可配置。使用 `prestosql-connector-hive` 或 `trino-connector-hive` 配置分类。
`node.properties`	PrestoDB：在 Amazon EMR 版本 5.6.0 及更高版本中可配置。使用 `presto-node` 配置分类。 Trino（PrestoSQL）：在 Amazon EMR 版本 6.1.0 及更高版本中可配置。使用 `prestosql-node` 或 `trino-node` 配置分类。

文件	可配置
jvm.config	不可配置。

PrestoDB 和 Trino 安装

继续使用应用程序名称 Presto 在集群上安装 PrestoDB。要在集群上安装 Trino，请使用应用程序名称 Trino（或在 Amazon EMR 早期版本中使用 PrestoSQL）。

您可以安装 PrestoDB 或 Trino，但不能在同一个集群上同时安装两者。如果在尝试创建集群时同时指定了 PrestoDB 和 Trino，则会出现验证错误，并且集群创建请求将会失败。

EMRFS 和 PrestoS3FileSystem 配置

使用 Amazon EMR 版本 5.12.0 及更高版本时，PrestoDB 可以使用 EMRFS。这是默认配置。EMRFS 也是 Amazon EMR 版本 6.1.0 及更高版本中的默认 Trino（PrestoSQL）文件系统。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 EMR 文件系统（EMRFS）。在早期版本的 Amazon EMR 中，唯一的配置选项是 PrestoS3FileSystem。

您可以使用安全配置为 Amazon S3 中的 EMRFS 数据设置加密。您还可以使用 IAM 角色处理 EMRFS 对 Amazon S3 的请求。有关更多信息，请参阅《Amazon EMR 管理指南》中的了解加密选项和为处理 EMRFS 对 Amazon S3 的请求配置 IAM 角色。

> **Note**
>
> 如果您使用 Amazon EMR 版本 5.12.0 查询 Amazon S3 中的底层数据，则可能会出现 Presto 错误。这是因为 Presto 无法从 emrfs-site.xml 提取配置分类值。解决方法是在 usr/lib/presto/plugin/hive-hadoop2/ 下创建一个 emrfs 子目录，并在 usr/lib/presto/plugin/hive-hadoop2/emrfs 中创建一个指向现有 /usr/share/aws/emr/emrfs/conf/emrfs-site.xml 文件的符号链接。然后重新启动 presto-server 进程（首先执行 sudo presto-server stop，然后执行 sudo presto-server start）。

您可以覆盖 EMRFS 默认值并改用 PrestoS3FileSystem。为此，请使用 presto-connector-hive 配置分类将 hive.s3-file-system-type 设置为 PRESTO，如以下示例所示。有关更多信息，请参阅配置应用程序。

```
[
```

```
   {
      "Classification": "presto-connector-hive",
      "Properties": {
         "hive.s3-file-system-type": "PRESTO"
      }
   }
]
```

如果您使用的是 PrestoS3FileSystem，请使用 Trino 的 `presto-connector-hive` 配置分类或 `trino-connector-hive` 来配置 PrestoS3FileSystem 属性。有关可用属性的更多信息，请参阅 Presto 文档中“Hive 连接器”部分的 Amazon S3 配置。这些设置不适用于 EMRFS。

终端用户模拟的默认设置

默认情况下，Amazon EMR 版本 5.12.0 及更高版本支持通过终端用户模拟来访问 HDFS。有关更多信息，请参阅 Presto 文档中的终端用户模拟。要使用 `presto-config` 配置分类更改此设置，请将 `hive.hdfs.impersonation.enabled` 属性设置为 `false`。

Presto Web 界面的默认端口

默认情况下，Amazon EMR 将 Presto 协调器上的 Presto Web 界面配置为使用端口 8889（针对 PrestoDB 和 Trino）。要更改端口，请使用 `presto-config` 配置分类设置 `http-server.http.port` 属性。有关更多信息，请参阅 Presto 文档的部署 Presto 部分中的配置属性。

某些版本中的 Hive 存储桶执行问题

Presto 发行版 152.3 存在一个与 Hive 存储桶执行有关的问题，此问题在某些情况下可能会显著降低 Presto 的查询性能。Amazon EMR 版本 5.0.3、5.1.0 和 5.2.0 包含此版本的 Presto。要解决此问题，请使用 `presto-connector-hive` 配置分类将 `hive.bucket-execution` 属性设置为 `false`，如以下示例所示。

```
[
   {
      "Classification": "presto-connector-hive",
      "Properties": {
         "hive.bucket-execution": "false"
      }
   }
]
```

Presto 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Presto 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Presto 版本信息

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.14.0	0.281	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.13.0	0.281	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.12.0	0.281	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode,

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
		hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.11.1	0.279	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.11.0	0.279	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.10.1	0.278	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.10.0	0.278	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.9.1	0.276	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.9.0	0.276	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.8.1	0.273	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.8.0	0.273	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.7.0	0.272	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.36.1	0.267	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.36.0	0.267	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.6.0	0.267	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.35.0	0.266	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.5.0	0.261	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.4.0	0.254.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.3.1	0.245.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.3.0	0.245.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.2.1	0.238.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.2.0	0.238.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.1.1	0.232	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.1.0	0.232	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-6.0.1	0.230	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-6.0.0	0.230	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.34.0	0.261	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.33.1	0.245.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.33.0	0.245.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.32.1	0.240.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.32.0	0.240.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.31.1	0.238.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.31.0	0.238.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.30.2	0.232	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.30.1	0.232	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker
emr-5.30.0	0.232	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mariadb-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.29.0	0.227	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.28.1	0.227	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.28.0	0.227	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-presto, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.27.1	0.224	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.27.0	0.224	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.26.0	0.220	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.25.0	0.220	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.24.1	0.219	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.24.0	0.219	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.23.1	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.23.0	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.22.0	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.21.2	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.21.1	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.21.0	0.215	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.20.1	0.214	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.20.0	0.214	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.19.1	0.212	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.19.0	0.212	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.18.1	0.210	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.18.0	0.210	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.17.2	0.206	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.17.1	0.206	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.17.0	0.206	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.16.1	0.203	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.16.0	0.203	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.15.1	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.15.0	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.14.2	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.14.1	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.14.0	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.13.1	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.13.0	0.194	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.12.3	0.188	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.12.2	0.188	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.12.1	0.188	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.12.0	0.188	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.11.4	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.11.3	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.11.2	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.11.1	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.11.0	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.10.1	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.10.0	0.187	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.9.1	0.184	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.9.0	0.184	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.8.3	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.8.2	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.8.1	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.8.0	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.7.1	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.7.0	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.6.1	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.6.0	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.5.4	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.5.3	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.5.2	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.5.1	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.5.0	0.170	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.4.1	0.166	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.4.0	0.166	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.3.2	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.3.1	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.3.0	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.2.3	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.2.2	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.2.1	0.157.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.2.0	0.152.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.1.1	0.152.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

Amazon EMR 发行版标签	Presto 版本	随 Presto 安装的组件
emr-5.1.0	0.152.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.0.3	0.152.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker
emr-5.0.0	0.150	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hive-client, hcatalog-server, mysql-server, presto-coordinator, presto-worker

下表列出了 Amazon EMR 每个发行版中包含的 Trino (Presto SQL) 版本，以及随应用程序一起安装的组件。PrestoSQL 从版本 351 开始将其名称变更为 Trino。

Trino (PrestoSQL) 版本信息

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.14.0	422	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.13.0	414	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.12.0	414	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
		cemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.11.1	410	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.11.0	410	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.10.1	403	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.10.0	403	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.9.1	398	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.9.0	398	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.8.1	388	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.8.0	388	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.7.0	378	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.6.0	367	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.5.0	360	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker
emr-6.4.0	359	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-trino, hcatalog-server, mariadb-server, trino-coordinator, trino-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.3.1	350	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker
emr-6.3.0	350	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.2.1	343	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker
emr-6.2.0	343	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker

Amazon EMR 发行版标签	Trino (PrestoSQL) 版本	随 Trino (PrestoSQL) 安装的组件
emr-6.1.1	338	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker
emr-6.1.0	338	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hive-client, hudi, hudi-prestosql, hcatalog-server, mariadb-server, prestosql-coordinator, prestosql-worker

Trino（PrestoSQL）发布说明（按版本分类）

- Amazon EMR 6.9.0 – Trino（PrestoSQL）发布说明

Amazon EMR 6.9.0 – Trino（PrestoSQL）发布说明

Amazon EMR 6.9.0 – Trino（PrestoSQL）新功能

- 为了支持长时间运行的查询，Trino 现在包括容错执行机制。容错执行通过重试失败的查询或其组件任务来减少查询失败。有关更多信息，请参阅Trino 中的容错执行。

Amazon EMR 6.9.0 – Trino（PrestoSQL）更改

Amazon EMR 6.9.0 – PrestoDB 更改

类型	描述
升级	将 PrestoDB 升级到 0.276
升级	对于 Hadoop 3.3.3 的支持
升级	将 Hudi 升级到 0.12.1
特征	Amazon EMR 和 Presto 使用 GCSC API 与 Amazon Lake Formation 集成，以实现交互式工作负载。
特征	Security Configs for PrestoDB 中添加了与 Kerberos 相关的配置，以启用 Keberos。
错误修复	已恢复 OSS 拉取请求 #18115，添加该请求旨在减少 `hdfsConfiguration` 副本数。这导致在使用 EMRFS 或 Hudi 表时 HDFS 配置副本错误。

Amazon EMR 6.9.0 – Trino 更改

类型	描述
升级	将 Trino 升级到 398
升级	对于 Hadoop 3.3.3 的支持

类型	描述
特征	Tardigrade 支持：添加了对 HDFS 和 Amazon S3 上交换假脱机的支持。有关更多信息，请参阅Trino 中的容错执行。
错误修复	当使用 Trino Iceberg 并启用 Glue 目录时，避免在 `iceberg.properties.` 中添加元存储 URI

Amazon EMR 6.9.0 – Trino（PrestoSQL）已知问题

- 对于 Amazon EMR 发行版 6.9.0，Trino 不适用于为 Apache Ranger 启用的集群。如果您需要将 Trino 与 Ranger 结合使用，请联系 Amazon Web Services Support。

Apache Spark

Apache Spark 是一个分布式处理框架和编程模型，可帮助您使用 Amazon EMR 集群进行机器学习、流处理或图形分析。Spark 与 Apache Hadoop 类似，也是一款常用于大数据工作负载的开源、分布式处理系统。但 Spark 与 Hadoop MapReduce 有一些明显的不同。Spark 拥有经过优化的有向无环图 (DAG) 执行引擎并会积极地在内存中缓存数据，这可提高性能，尤其是对于某些算法和交互式查询。

Spark 内在支持使用 Scala、Python 和 Java 编写的应用程序。它还包含几个紧密集成的库，适用于 SQL（Spark SQL）、机器学习（MLlib）、流处理（Spark Streaming）和图形处理（GraphX）。这些工具可让您更轻松地在各种使用案例中充分发挥 Spark 框架的优势。

您可以在 Amazon EMR 集群上与其他 Hadoop 应用程序一同安装 Spark，它还能借助 EMR 文件系统（EMRFS）直接访问 Amazon S3 中的数据。Hive 也与 Spark 集成，以便您使用 HiveContext 对象运行使用 Spark 的 Hive 脚本。Hive 上下文作为 `sqlContext` 包含在 Spark Shell 中。

有关使用 Spark 设置 EMR 集群和分析示例数据集的示例教程，请参阅 Amazon 新闻博客上的 教程：Amazon EMR 入门。

> **Important**
>
> Apache Spark 版本 2.3.1（从 Amazon EMR 发行版 5.16.0 开始提供）解决了 CVE-2018-8024 和 CVE-2018-1334 问题。建议您将 Spark 的早期版本迁移到 Spark 2.3.1 版本或更高版本。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Spark 的版本，以及 Amazon EMR 随 Spark 一起安装的组件。

有关此发行版中随 Spark 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Spark 版本信息

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.14.0	Spark 3.4.1	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-librar

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
		y, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

> **Note**
>
> Amazon EMR 发行版 6.8.0 随附 Apache Spar 3.3.0。此 Spark 发行版使用 Apache Log4j 2 和 `log4j2.properties` 文件，在 Spark 进程中配置 Log4j。如果您在集群中使用 Spark 或使用自定义配置参数创建 EMR 集群，并且希望升级到 Amazon EMR 发行版 6.8.0，则必须迁移到新的 `spark-log4j2` 配置分类和 Apache Log4j 2 的密钥格式。有关更多信息，请参阅从 Apache Log4j 1.x 迁移到 Log4j 2.x。

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Spark 的版本，以及 Amazon EMR 随 Spark 一起安装的组件。

有关此发行版中随 Spark 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Spark 版本信息

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.36.1	Spark 2.4.8	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
		server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

主题

使用 Spark 创建集群

以下程序在 EMR 控制台中使用 Quick Options (快速选项) 创建一个附带 Spark 的集群。

作为替代，您可以使用 Advanced Options (高级选项) 进一步自定义您的集群设置，或是提交步骤以编程方式安装应用程序，然后执行自定义应用程序。利用这些集群创建选项之一，您可以选择使用 Amazon Glue 作为您的 Spark SQL 元存储。参阅 使用 Amazon Glue 数据目录作为 Spark SQL 的元存储 了解更多信息。

启动安装了 Spark 的集群

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 选择 Create cluster (创建集群) 以使用 Quick Options (快速选项)。
3. 输入 Cluster name (集群名称)。
4. 在 Software Configuration (软件配置) 中，选择 Release (发行版) 选项。
5. 在 Applications (应用程序) 中，选择 Spark 应用程序捆绑包。
6. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

> **Note**
>
> 要在创建集群时配置 Spark，请参阅配置 Spark。

使用 Amazon CLI 启动安装了 Spark 的集群

- 使用下面的命令创建集群。

```
aws emr create-cluster --name "Spark cluster" --release-label emr-5.36.1 --
applications Name=Spark \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge --instance-count 3 --use-
default-roles
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

使用 SDK for Java 启动安装了 Spark 的集群

通过 `SupportedProductConfig` 中使用的 `RunJobFlowRequest` 指定 Spark 作为应用程序。

- 下面的实例显示如何通过 Java 使用 Spark 创建集群：

```
import com.amazonaws.AmazonClientException;
import com.amazonaws.auth.AWSCredentials;
import com.amazonaws.auth.AWSStaticCredentialsProvider;
import com.amazonaws.auth.profile.ProfileCredentialsProvider;
import com.amazonaws.services.elasticmapreduce.AmazonElasticMapReduce;
import com.amazonaws.services.elasticmapreduce.AmazonElasticMapReduceClientBuilder;
import com.amazonaws.services.elasticmapreduce.model.*;
import com.amazonaws.services.elasticmapreduce.util.StepFactory;

public class Main {

 public static void main(String[] args) {
  AWSCredentials credentials_profile = null;
  try {
   credentials_profile = new
 ProfileCredentialsProvider("default").getCredentials();
       } catch (Exception e) {
           throw new AmazonClientException(
                   "Cannot load credentials from .aws/credentials file. " +
                   "Make sure that the credentials file exists and the profile
 name is specified within it.",
                   e);
       }

       AmazonElasticMapReduce emr = AmazonElasticMapReduceClientBuilder.standard()
   .withCredentials(new AWSStaticCredentialsProvider(credentials_profile))
   .withRegion(Regions.US_WEST_1)
   .build();

       // create a step to enable debugging in the AWS Management Console
  StepFactory stepFactory = new StepFactory();
  StepConfig enabledebugging = new StepConfig()
      .withName("Enable debugging")
      .withActionOnFailure("TERMINATE_JOB_FLOW")
      .withHadoopJarStep(stepFactory.newEnableDebuggingStep());

       Application spark = new Application().withName("Spark");

       RunJobFlowRequest request = new RunJobFlowRequest()
           .withName("Spark Cluster")
```

```
            .withReleaseLabel("emr-5.20.0")
            .withSteps(enabledebugging)
            .withApplications(spark)
            .withLogUri("s3://path/to/my/logs/")
          .withServiceRole("EMR_DefaultRole")
          .withJobFlowRole("EMR_EC2_DefaultRole")
            .withInstances(new JobFlowInstancesConfig()
                .withEc2SubnetId("subnet-12ab3c45")
                .withEc2KeyName("myEc2Key")
                .withInstanceCount(3)
                .withKeepJobFlowAliveWhenNoSteps(true)
                .withMasterInstanceType("m4.large")
                .withSlaveInstanceType("m4.large")
            );
        RunJobFlowResult result = emr.runJobFlow(request);
    System.out.println("The cluster ID is " + result.toString());
 }
}
```

使用 Amazon EMR 6.x 通过 Docker 运行 Spark 应用程序

借助 Amazon EMR 6.0.0，Spark 应用程序可以使用 Docker 容器来定义其库依赖项，而不是在集群中的单个 Amazon EC2 实例上安装依赖项。要使用 Docker 运行 Spark，您必须首先配置 Docker 注册表，并在提交 Spark 应用程序时定义其他参数。有关更多信息，请参阅配置 Docker 集成。

提交应用程序时，YARN 调用 Docker 来拉取指定的 Docker 映像并在 Docker 容器内运行 Spark 应用程序。这让您可以轻松地定义和隔离依赖项。它可以通过执行任务所需的库来缩短在 Amazon EMR 集群中引导启动或准备实例的时间。

利用 Docker 运行 Spark 时的注意事项

使用 Docker 运行 Spark 时，请确保满足以下先决条件：

- `docker` 程序包和 CLI 仅安装在核心节点和任务节点上。
- 在 Amazon EMR 6.1.0 及更高版本中，您也可以使用以下命令在主节点安装 Docker。
  - ```
    sudo yum install -y docker
    sudo systemctl start docker
    ```
- `spark-submit` 命令应始终从 Amazon EMR 集群上的主实例运行。
    ```

- 用于解析 Docker 映像的 Docker 注册表必须使用带有 `container-executor` 分类键的分类 API 来定义，以便在启动集群时定义其他参数：
  - `docker.trusted.registries`
  - `docker.privileged-containers.registries`
- 要在 Docker 容器中执行 Spark 应用程序，需要以下配置选项：
  - `YARN_CONTAINER_RUNTIME_TYPE=docker`
  - `YARN_CONTAINER_RUNTIME_DOCKER_IMAGE={DOCKER_IMAGE_NAME}`
- 使用 Amazon ECR 检索 Docker 镜像时，必须将集群配置为对自身进行身份验证。为此，您必须使用以下配置选项：
  - YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG={DOCKER_CLIENT_CONFIG_PATH_ON
- 在 EMR 6.1.0 及更高版本中，当自用 ECR 自动身份验证时，您不需要使用列出的命令 YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG={DOCKER_CLIENT_CONFIG_PATH_ON_HDFS
- 与 Spark 一起使用的任何 Docker 映像都必须在 Docker 映像中安装了 Java。

有关先决条件的更多信息，请参阅配置 Docker 集成。

创建 Docker 镜像

Docker 映像是使用 Dockerfile 创建的，该文件定义了要包含在映像中的程序包和配置。以下两个示例 Dockerfile 文件使用 PySpark 和 SparkR。

PySpark Dockerfile

从该 Dockerfile 创建的 Docker 映像包括 Python 3 和 NumPy Python 程序包。此 Dockerfile 使用 Amazon Linux 2 和 Amazon Corretto JDK 8。

```
FROM amazoncorretto:8

RUN yum -y update
RUN yum -y install yum-utils
RUN yum -y groupinstall development

RUN yum list python3*
RUN yum -y install python3 python3-dev python3-pip python3-virtualenv

RUN python -V
RUN python3 -V
```

```
ENV PYSPARK_DRIVER_PYTHON python3
ENV PYSPARK_PYTHON python3

RUN pip3 install --upgrade pip
RUN pip3 install numpy pandas

RUN python3 -c "import numpy as np"
```

SparkR Dockerfile

从该 Dockerfile 创建的 Docker 映像包括 R 和 randomForest CRAN 程序包。此 Dockerfile 包含 Amazon Linux 2 和 Amazon Corretto JDK 8。

```
FROM amazoncorretto:8

RUN java -version

RUN yum -y update
RUN amazon-linux-extras install R4

RUN yum -y install curl hostname

#setup R configs
RUN echo "r <- getOption('repos'); r['CRAN'] <- 'http://cran.us.r-project.org';
 options(repos = r);" > ~/.Rprofile

RUN Rscript -e "install.packages('randomForest')"
```

有关 Dockerfile 语法的更多信息，请参阅 Dockerfile 参考文档。

使用来自 Amazon ECR 的 Docker 镜像

Amazon Elastic Container Registry（Amazon ECR）是一个完全托管式 Docker 容器注册表，可让开发人员轻松地存储、管理和部署 Docker 容器镜像。使用 Amazon ECR 时，必须将集群配置为信任您的 ECR 实例，并且必须配置身份验证，以便集群使用来自 Amazon ECR 的 Docker 镜像。有关更多信息，请参阅配置 YARN 以访问 Amazon ECR。

要确保 EMR 主机可以访问存储在 Amazon ECR 中的镜像，集群必须具有通过 `AmazonEC2ContainerRegistryReadOnly` 策略授予的与实例配置文件关联的权限。有关更多信息，请参阅AmazonEC2ContainerRegistryReadOnly策略。

在此示例中，必须使用以下附加配置创建集群，以确保 Amazon ECR 注册表受信任。将 *123456789123.dkr.ecr.us-east-1.amazonaws.com* 端点替换为您的 Amazon ECR 端点。

```
[
  {
    "Classification": "container-executor",
    "Configurations": [
      {
        "Classification": "docker",
        "Properties": {
          "docker.privileged-containers.registries":
 "local,centos,123456789123.dkr.ecr.us-east-1.amazonaws.com",
          "docker.trusted.registries": "local,centos,123456789123.dkr.ecr.us-
east-1.amazonaws.com"
        }
      }
    ],
    "Properties": {}
  }
]
```

将 PySpark 与 Amazon ECR 结合使用

以下示例使用 PySpark Dockerfile，该文件将被标记并上传到 Amazon ECR。上载 Dockerfile 之后，您可以运行 PySpark 任务并从 Amazon ECR 中引用 Docker 镜像。

启动集群后，使用 SSH 连接到核心节点，并运行以下命令从 PySpark Dockerfile 示例构建本地 Docker 映像。

首先，创建一个目录和一个 Dockerfile。

```
mkdir pyspark
vi pyspark/Dockerfile
```

粘贴 PySpark Dockerfile 的内容，然后运行以下命令来构建 Docker 映像。

```
sudo docker build -t local/pyspark-example pyspark/
```

为示例创建 `emr-docker-examples` ECR 存储库。

```
aws ecr create-repository --repository-name emr-docker-examples
```

标记本地生成的映像并将其上传到 ECR，同时将 *123456789123.dkr.ecr.us-east-1.amazonaws.com* 替换为您的 ECR 终端节点。

```
sudo docker tag local/pyspark-example 123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-examples:pyspark-example
sudo docker push 123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-examples:pyspark-example
```

使用 SSH 连接到主节点并准备文件名为 `main.py` 的 Python 脚本。将以下内容粘贴到 `main.py` 文件中并保存它。

```
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName("docker-numpy").getOrCreate()
sc = spark.sparkContext

import numpy as np
a = np.arange(15).reshape(3, 5)
print(a)
```

要在 EMR 6.0.0 上提交任务，请引用 Docker 镜像的名称。定义其他配置参数，以确保作业执行使用 Docker 作为运行时。使用 Amazon ECR 时，`YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG` 必须引用 `config.json` 文件，其中包含用于对 Amazon ECR 进行身份验证的凭证。

```
DOCKER_IMAGE_NAME=123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-examples:pyspark-example
DOCKER_CLIENT_CONFIG=hdfs:///user/hadoop/config.json
spark-submit --master yarn \
--deploy-mode cluster \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG=$DOCKER_CLIENT_CONFIG \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG=$DOCKER_CLIENT_CONFIG \
--num-executors 2 \
main.py -v
```

在 EMR 6.1.0 及更高版本中，要提交作业，请引用 Docker 镜像的名称。启用 ECR 自动身份验证后，运行以下命令。

```
DOCKER_IMAGE_NAME=123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-
examples:pyspark-example
spark-submit --master yarn \
--deploy-mode cluster \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--num-executors 2 \
main.py -v
```

作业完成后，记下 YARN 应用程序 ID，并使用以下命令获取 PySpark 作业的输出。

```
yarn logs --applicationId application_id | grep -C2 '\[\['
LogLength:55
LogContents:
[[ 0  1  2  3  4]
 [ 5  6  7  8  9]
 [10 11 12 13 14]]
```

将 SparkR 与 Amazon ECR 结合使用

以下示例使用 SparkR Dockerfile，该文件将被标记并上传到 ECR。上载此 Dockerfile 之后，您可以运行 SparkR 任务并从 Amazon ECR 中引用 Docker 镜像。

启动集群后，使用 SSH 连接到核心节点，并运行以下命令从 SparkR Dockerfile 示例构建本地 Docker 映像。

首先，创建一个目录和该 Dockerfile。

```
mkdir sparkr
vi sparkr/Dockerfile
```

粘贴 SparkR Dockerfile 的内容并运行以下命令来构建 Docker 映像。

```
sudo docker build -t local/sparkr-example sparkr/
```

标记本地生成的镜像并将其上载到 Amazon ECR，同时将 *123456789123.dkr.ecr.us-east-1.amazonaws.com* 替换为您的 Amazon ECR 端点。

```
sudo docker tag local/sparkr-example 123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-
docker-examples:sparkr-example
sudo docker push 123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-
examples:sparkr-example
```

使用 SSH 连接到主节点并准备名为 `sparkR.R` 的 R 脚本。将以下内容粘贴到 `sparkR.R` 文件中。

```
library(SparkR)
sparkR.session(appName = "R with Spark example", sparkConfig =
 list(spark.some.config.option = "some-value"))

sqlContext <- sparkRSQL.init(spark.sparkContext)
library(randomForest)
# check release notes of randomForest
rfNews()

sparkR.session.stop()
```

要在 EMR 6.0.0 上提交任务，请引用 Docker 镜像的名称。定义其他配置参数，以确保作业执行使用 Docker 作为运行时。使用 Amazon ECR 时，`YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG` 必须引用 `config.json` 文件，其中包含用于对 ECR 进行身份验证的凭证。

```
DOCKER_IMAGE_NAME=123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-
examples:sparkr-example
DOCKER_CLIENT_CONFIG=hdfs:///user/hadoop/config.json
spark-submit --master yarn \
--deploy-mode cluster \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG=
$DOCKER_CLIENT_CONFIG \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_CLIENT_CONFIG=
$DOCKER_CLIENT_CONFIG \
sparkR.R
```

在 EMR 6.1.0 及更高版本中，要提交作业，请引用 Docker 镜像的名称。启用 ECR 自动身份验证后，运行以下命令。

```
DOCKER_IMAGE_NAME=123456789123.dkr.ecr.us-east-1.amazonaws.com/emr-docker-
examples:sparkr-example
spark-submit --master yarn \
--deploy-mode cluster \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.executorEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_TYPE=docker \
--conf spark.yarn.appMasterEnv.YARN_CONTAINER_RUNTIME_DOCKER_IMAGE=$DOCKER_IMAGE_NAME \
sparkR.R
```

作业完成后，记下 YARN 应用程序 ID，并使用以下命令获取 SparkR 作业的输出。此示例包括测试，以确保 randomForest 库、已安装的版本和发布说明可用。

```
yarn logs --applicationId application_id | grep -B4 -A10 "Type rfNews"
randomForest 4.6-14
Type rfNews() to see new features/changes/bug fixes.
Wishlist (formerly TODO):

* Implement the new scheme of handling classwt in classification.

* Use more compact storage of proximity matrix.

* Allow case weights by using the weights in sampling?

=============================================================================
Changes in 4.6-14:
```

使用 Amazon Glue 数据目录作为 Spark SQL 的元存储

使用 Amazon EMR 发行版 5.8.0 或更高版本，您可以将 Spark SQL 配置为使用 Amazon Glue 数据目录作为元存储。当您需要持久的元数据仓或由不同集群、服务、应用程序和 Amazon 账户共享的元数据仓时，我们建议使用此配置。

Amazon Glue 是一项完全托管式提取、转换和加载（ETL）服务，使您能够轻松且经济高效地对数据进行分类、清理和扩充，并在各种数据存储之间可靠地移动数据。Amazon Glue 数据目录跨各种数据源和数据格式提供统一的元数据存储库，从而不仅与 Amazon EMR 集成，还与 Amazon RDS、Amazon Redshift、Redshift Spectrum、Athena 以及任何与 Apache Hive 元存储兼容的应用程序集成。AmazonGlue 爬网程序能够自动从 Amazon S3 源数据推断架构，从而将关联的元数据存储

在数据目录中。有关数据目录的更多信息，请参阅《Amazon Glue 开发人员指南》中的填充 Amazon Glue 数据目录。

使用 Amazon Glue 需单独付费。在数据目录中存储和访问数据需按月付费；为 Amazon Glue ETL 作业和爬网程序运行时按小时费率付费（按分计费）；为每个预置的开发端点支付每小时费率（按分计费）。数据目录让您最多可免费存储一百万个对象。如果您存储一百万个以上的对象，将需要为超过一百万的每 100,000 个对象支付 1 美元。数据目录中的对象为表、分区或数据库。有关更多信息，请参阅 Glue 定价。

> **Important**
>
> 如果您在 2017 年 8 月 14 日之前使用 Amazon Athena 或 Amazon Redshift Spectrum 创建了表，则数据库和表将存储在 Athena 托管式目录中，该目录与 Amazon Glue 数据目录相互独立。要将 Amazon EMR 与这些表集成，您必须升级到 Amazon Glue 数据目录。有关更多信息，请参阅《Amazon Athena 用户指南》中的升级到 Amazon Glue 数据目录。

指定 Amazon Glue 数据目录作为元存储

您可以使用Amazon Web Services Management Console、Amazon CLI 或 Amazon EMR API 指定 Amazon Glue 数据目录作为元存储。在使用 CLI 或 API 时，您可以使用 Spark 的配置分类指定数据目录。此外，使用 Amazon EMR 5.16.0 及更高版本时，您可以使用配置分类指定其他 Amazon Web Services 账户中的数据目录。在使用控制台时，您可以使用 Advanced Options (高级选项) 或 Quick Options (快速选项) 指定数据目录。

> **Note**
>
> 使用 Amazon Glue 数据目录的选项也适用于 Zeppelin，因为 Zeppelin 安装有 Spark SQL 组件。

New console

使用新控制台指定 Amazon Glue 数据目录作为 Spark 元存储

1. 登录 Amazon Web Services Management Console 并打开 Amazon EMR 控制台，网址为 https://console.aws.amazon.com/emr。
2. 在左侧导航窗格中的 EMR on EC2 下，选择 Clusters（集群），然后选择 Create cluster（创建集群）。

3. 在 Application bundle（应用程序包）下，选择 Spark 或 Custom（自定义）。如果您自定义集群，请确保选择 Zeppelin 或 Spark 作为应用程序之一。
4. 在 Amazon Glue Data Catalog s设置下，选择用于 Spark 表元数据复选框。
5. 选择适用于集群的任何其他选项。
6. 要启动集群，选择 Create cluster（创建集群）。

Old console

使用旧控制台指定 Amazon Glue 数据目录作为 Spark 元存储

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 对于 Release (版本)，选择 emr-5.8.0 或更高版本。
4. 在 Release (版本) 下，选择 Spark 或 Zeppelin。
5. 在 Amazon Glue Data Catalog settings (Amazon Glue 数据目录设置) 下面，选择 Use for Spark table metadata (用于 Spark 表元数据)。
6. 根据需要为您的集群选择其他选项，选择 Next (下一步)，然后根据需要为您的应用程序配置其他集群选项。

Amazon CLI

使用 Amazon CLI 指定 Amazon Glue 数据目录作为 Spark 元存储

有关使用 Amazon CLI 和 EMR API 指定配置分类的更多信息，请参阅配置应用程序。

- 使用 `spark-hive-site` 分类指定 `hive.metastore.client.factory.class` 的值，如下例所示：

```
[
  {
    "Classification": "spark-hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
 "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  }
```

```
]
```

要在其他Amazon账户中指定数据目录，请添加 `hive.metastore.glue.catalogid` 属性，如以下示例所示。将 *`acct-id`* 替换为数据目录的Amazon账户。

```
[
  {
    "Classification": "spark-hive-site",
    "Properties": {
      "hive.metastore.client.factory.class":
 "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory",
      "hive.metastore.glue.catalogid": "acct-id"
    }
  }
]
```

IAM 权限

集群的 EC2 实例配置文件必须具有适用于 Amazon Glue 操作的 IAM 权限。此外，如果您为 Amazon Glue 数据目录对象启用加密，还必须允许该角色加密、解密和生成用于加密的Amazon KMS key。

适用于 Amazon Glue 操作的权限

如果使用适用于 Amazon EMR 默认的 EC2 实例配置文件，则无需执行任何操作。附加到 `EMR_EC2_DefaultRole` 的 `AmazonElasticMapReduceforEC2Role` 托管式策略允许所有必要 Amazon Glue 操作。但是，如果您指定自定义 EC2 实例配置文件和权限，则必须配置合适的 Amazon Glue 操作。使用 `AmazonElasticMapReduceforEC2Role` 托管策略作为起点。如需了解更多信息，请参阅《Amazon EMR 管理指南》中的集群 EC2 实例的服务角色（EC2 实例配置文件）。

用于加密和解密 Amazon Glue 数据目录的权限

您的实例配置文件需要使用密钥加密和解密数据的权限。如果以下语句适用，您不必配置这些权限：

- 您使用 Amazon Glue 的托管式密钥启用 Amazon Glue Data Catalog 对象的加密。
- 您使用的是同一Amazon Web Services 账户的集群，其作为 AmazonGlue Data Catalog。

否则，您必须将以下语句添加到附加到 EC2 实例配置文件的权限策略。

```
[
    {
        "Version": "2012-10-17",
        "Statement": [
            {
                "Effect": "Allow",
                "Action": [
                    "kms:Decrypt",
                    "kms:Encrypt",
                    "kms:GenerateDataKey"
                ],
                "Resource": "arn:aws:kms:region:acct-
id:key/12345678-1234-1234-1234-123456789012"
            }
        ]
    }
]
```

有关 Amazon Glue 数据目录加密的更多信息，请参阅《Amazon Glue 开发人员指南》中的加密您的数据目录。

基于资源的权限

如果您将 Amazon Glue 与 Amazon EMR 中的 Hive、Spark 或 Presto 结合使用，Amazon Glue 支持使用基于资源的策略来控制对数据目录资源的访问权限。这些资源包括数据库、表、连接和用户定义的函数。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的 Amazon Glue 资源策略。

当使用基于资源的策略来限制从 Amazon EMR 中访问 Amazon Glue 时，在权限策略中指定的委托人必须是与创建集群时指定的 EC2 实例配置文件相关联的角色 ARN。例如，对于附加到目录的基于资源的策略，您可以使用以下示例中显示的格式为集群 EC2 实例的默认服务角色指定角色 ARN，将 *EMR_EC2_DefaultRole* 指定为 `Principal`：

```
arn:aws:iam::acct-id:role/EMR_EC2_DefaultRole
```

acct-id 可以与 Amazon Glue 账户 ID 不同。这允许从不同账户中的 EMR 集群进行访问。您可以指定多个委托人，且每个委托人都可以来自不同的账户。

使用 Amazon Glue 数据目录时的注意事项

在使用 Amazon Glue 数据目录作为 Spark 的元存储时，请考虑以下项目：

- 当您创建表时，具有没有位置 URI 的默认数据库会导致失败。作为解决方法，请在您使用 LOCATION 时使用 s3://*EXAMPLE-DOC-BUCKET* 子句指定一个存储桶位置，如 CREATE TABLE。或者，在除默认数据库之外的数据库内创建表。
- 不支持在 Amazon Glue 中重命名表。
- 当您创建 Hive 表而不指定 LOCATION 时，表数据存储在通过 hive.metastore.warehouse.dir 属性指定的位置。默认情况下，这是 HDFS 中的一个位置。如果另一个集群需要访问该表，则它将失败，除非它有足够的权限访问创建该表的集群。此外，由于 HDFS 存储是暂时性的，因此如果集群终止，表数据将丢失，并且必须重新创建该表。建议您在使用 Amazon Glue 创建 Hive 表时，指定 Amazon S3 中的一个 LOCATION。此外，也可以使用 hive-site 配置分类来为 hive.metastore.warehouse.dir 指定 Amazon S3 中的位置，它适用于所有 Hive 表。如果表在 HDFS 位置创建，并且创建该表的集群仍在运行，您可以在 Amazon Glue 中更新 Amazon S3 中表的位置。有关更多信息，请参阅《Amazon Glue 开发人员指南》中的使用 Amazon Glue 控制台上的表。
- 不支持包含引号和撇号的分区值，例如 PARTITION (owner="Doe's").
- emr-5.31.0 及更高版本支持列统计数据。
- 不支持使用 Hive 授权。作为替代方案，考虑使用基于 Amazon Glue 资源的策略。有关更多信息，请参阅将用于 Amazon EMR 访问的基于资源的策略用于 Amazon Glue 数据目录。

配置 Spark

您可以使用配置分类配置 Amazon EMR 上的 Spark。有关使用配置分类的更多信息，请参阅 配置应用程序。

Amazon EMR 上的 Spark 的配置分类包括如下：

- **spark** – 将 maximizeResourceAllocation 属性设置为 true 或 false。在设置为 true 时，Amazon EMR 将基于集群硬件配置自动配置 spark-defaults 属性。有关更多信息，请参阅使用 maximizeResourceAllocation。
- **spark-defaults** – 在 spark-defaults.conf 文件中设置值。有关更多信息，请参阅 Spark 文档中的 Spark 配置。
- **spark-env** – 在 spark-env.sh 文件中设置值。有关更多信息，请参阅 Spark 文档中的环境变量。
- **spark-hive-site** – 在 hive-site.xml 中为 Spark 设置值。
- **spark-log4j** – （Amazon EMR 版本 6.7.x 及更低版本）在 log4j.properties 文件中设置值。有关更多信息，请参阅 Github 上的 log4j.properties.template 文件。

- **spark-log4j2** – （Amazon EMR 版本 6.8.0 及更高版本）在 log4j2.properties 文件中设置值。有关更多信息，请参阅 Github 上的 log4j2.properties.template 文件。
- **spark-metrics** – 在 metrics.properties 文件中设置值。有关设置和更多信息，请参阅 Github 上的 metrics.properties.template 文件和 Spark 文档中的指标。

> **Note**
>
> 如果您要将 Spark 工作负载从另一个平台迁移到 Amazon EMR，我们建议您在添加自定义配置之前通过Amazon EMR 设置的 Spark 默认值检测工作负载。大多数客户都通过我们的默认设置见证了性能有所改善

主题

Amazon EMR 设置的 Spark 默认值

下表说明 Amazon EMR 如何在 spark-defaults 中设置影响应用程序的默认值。

Amazon EMR 设置的 Spark 默认值

设置	描述	默认值
spark.executor.memory	每个执行程序进程要使用的内存量。例如，1g、2g。	此设置由集群中的核心实例和任务实例类型决定。
spark.executor.cores	要对每个执行程序使用的内核的数量。	此设置由集群中的核心实例和任务实例类型决定。

设置	描述	默认值
spark.dynamicAllocation.enabled	如果为 true，则使用动态资源分配，以基于工作负载增大和减小注册到应用程序的执行程序的数目。	true（使用 Amazon EMR 4.4.0 及更高版本） **Note** Spark Shuffle Service 由 Amazon EMR 自动配置。
spark.sql.hive.advancedPartitionPredicatePushdown.enabled	如果为 true，则启用高级分区谓词下推到 Hive 元数据仓。	true
spark.sql.hive.stringLikePartitionPredicatePushdown.enabled	将 startsWith 、contains 和 endsWith 筛选条件向下推送到 Hive 元数据仓中。 **Note** Glue 不支持 startsWith 、contains 或 endsWith 谓词下推。如果您使用的是 Glue 元数据仓，并且由于这些函数的谓词向下推而遇到错误，请将此配置设置为 false。	true

在 Amazon EMR 6.1.0 上配置 Spark 垃圾回收

使用 spark.driver.extraJavaOptions 和 spark.executor.extraJavaOptions 设置自定义垃圾回收配置会导致 Amazon EMR 6.1 中的驱动程序或执行程序启动失败，因为垃圾回收

配置与 Amazon EMR 6.1.0 的配置存在冲突。对于 Amazon EMR 6.1.0，默认垃圾回收配置通过 `spark.driver.defaultJavaOptions` 和 `spark.executor.defaultJavaOptions` 设置。此配置仅适用于 Amazon EMR 6.1.0。与垃圾回收无关的 JVM 选项（例如用于配置日志记录（`-verbose:class`）的选项）仍然可以通过 `extraJavaOptions` 设置。有关更多信息，请参阅 Spark 应用程序属性。

使用 maximizeResourceAllocation

您可以使用 `spark` 配置分类将 `maximizeResourceAllocation` 设置为 `true`，以将执行程序配置为使用集群中的每个节点上尽可能多的资源。`maximizeResourceAllocation` 特定于 Amazon EMR。当您启用 `maximizeResourceAllocation`，EMR 会计算核心实例组中实例上的执行程序可用的最大计算和内存资源。然后，它将根据计算出的最大值设置相应的 `spark-defaults` 设置。

> **Note**
>
> 您不应在集群上将 `maximizeResourceAllocation` 选项与其他分布式应用程序（如 HBase）一起使用。Amazon EMR 对分布式应用程序使用自定义 YARN 配置，这可能与 `maximizeResourceAllocation` 冲突并导致 Spark 应用程序失败。

以下是一个 `maximizeResourceAllocation` 设置为 `true` 的 Spark 分类配置。

```
[
  {
    "Classification": "spark",
    "Properties": {
      "maximizeResourceAllocation": "true"
    }
  }
]
```

启用 **spark-defaults**时在 **maximizeResourceAllocation** 中配置的设置

设置	描述	Value
spark.default.parallelism	在用户未设置的情况下由转换(如联接、reduceByKey 和并行化) 返回的 RDD 中的分区数。	对 YARN 容器可用的 CPU 内核数的 2 倍。

设置	描述	Value
spark.driver.memory	要用于驱动程序进程（即初始化 SparkContext）的内存量。（例如，1g、2g）。	基于集群中的实例类型配置设置。但是，由于 Spark 驱动程序可在主实例或某个核心实例（例如，分别在 YARN 客户端和集群模式中）上运行，因此将根据这两个实例组中的实例类型的较小者进行设置。
spark.executor.memory	每个执行者进程要使用的内存量。（例如，1g、2g）	基于集群中的核心和任务实例类型配置设置。
spark.executor.cores	要对每个执行程序使用的内核的数量。	基于集群中的核心和任务实例类型配置设置。
spark.executor.instances	执行程序数。	基于集群中的核心和任务实例类型配置设置。除非同时将 `spark.dynamicAllocation.enabled` 显式设置为 true，否则将设置。

配置节点停用行为

在使用 Amazon EMR 发行版 5.9.0 或更高版本时，Amazon EMR 上的 Spark 包含一组功能，有助于确保 Spark 正常处理因手动调整大小或自动扩展策略请求引起的节点终止。Amazon EMR 会在 Spark 中实施拒绝名单机制，该机制的基础是 YARN 停用机制。此机制有助于确保不会在即将停用的节点上计划新任务，同时允许正在运行的任务完成。此外，有些功能可以在节点终止导致随机数据块丢失时帮助更快地恢复 Spark 任务。可以更快触发并优化重新计算进程，从而加快重新计算和减少阶段重试，并防止因丢失随机数据块引发的提取失败所导致的任务失败。

> **Important**
>
> Amazon EMR 发行版 5.11.0 中添加了 `spark.decommissioning.timeout.threshold` 设置，用于提升使用竞价型实例时的 Spark 恢复能力。在早期发行版中，当节点使用竞价型实例且该实例因出价而终止时，Spark 可能无法正常地处理终止。任务可能失败，而且随机重新计算可能花费大量时间。为此，如果您使用竞价型实例，建议使用发行版 5.11.0 或更高版本。

Spark 节点停用设置

设置	描述	默认值
spark.blacklist.decommissioning.enabled	当设置为 true，YARN 中的 Spark 拒绝名单节点处于 decommissioning 状态。Spark 不在于该节点上运行的执行程序上安排新任务。允许已经在运行的任务完成。	true
spark.blacklist.decommissioning.timeout	处于 decommissioning 状态的节点被加入拒绝名单的时间量。默认情况下，此值设置成一小时，这也是 yarn.resourcemanager.decommissioning.timeout 的默认值。要确保节点在其整个退役周期内都处于拒绝名单中，请将此值设置为等于或大于 yarn.resourcemanager.decommissioning.timeout 。在停用超时过期后，节点转为 decommissioned 状态，并且 Amazon EMR 可能终止节点的 EC2 实例。如果超时过期后有任何任务仍在运行，这些任务会丢失或被终止并在于其他节点上运行的其他执行程序上重新安排。	1h
spark.decommissioning.timeout.threshold	在 Amazon EMR 发行版 5.11.0 或更高版本中可用。以秒为单位指定。当某个节点转为停用状态时，如果主机将在等于或小于此值的时段后停用，Amazon EMR 不仅会将此	20s

设置	描述	默认值
	节点加入拒绝名单，还会清除主机状态（通过 spark.resourceManager.cleanupExpiredHost 指定），而不会等待此节点转换为停用状态。这使 Spark 能够更好地处理竞价型实例终止，因为无论 yarn.resourcemager.decommissioning.timeout 的值如何，Spot 实例都会在 20 秒的超时时间后停用，因此可能没有足够的时间提供其他节点来随机读取文件。	
spark.resourceManager.cleanupExpiredHost	当设置为 true 时，Spark 会注销处于 decommissioned 状态的节点上的执行程序存储的所有已缓存数据和随机数据块。这会加速恢复过程。	true
spark.stage.attempt.ignoreOnDecommissionFetchFailure	当设置为 true 时，有助于防止 Spark 进入失败阶段并因为从退役节点获取失败次数过多而导致任务失败。从处于 decommissioned 状态的节点获取随机数据块失败的次数不会计入连续获取失败的最大数量。	true

Spark ThriftServer 环境变量

Spark 将 Hive Thrift 服务器端口环境变量 HIVE_SERVER2_THRIFT_PORT 设置为 10001。

更改 Spark 默认设置

您可以使用 `spark-defaults` 配置分类或 `spark` 配置分类中的 `maximizeResourceAllocation` 设置更改 `spark-defaults.conf` 中的默认值。

以下过程说明如何使用 CLI 或控制台修改设置。

使用 CLI 创建一个 spark.executor.memory 设为 2g 的集群

- 使用以下命令创建一个安装了 Spark 且 `spark.executor.memory` 设为 2g 的集群，该集群引用存储在 Amazon S3 中的 `myConfig.json` 文件。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Spark \
--instance-type m5.xlarge --instance-count 2 --service-role EMR_DefaultRole_V2
 --ec2-attributes InstanceProfile=EMR_EC2_DefaultRole --configurations https://
s3.amazonaws.com/mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

`myConfig.json`:

```
[
    {
      "Classification": "spark-defaults",
      "Properties": {
        "spark.executor.memory": "2G"
      }
    }
  ]
```

使用控制台创建一个 spark.executor.memory 设为 2g 的集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。

3. 选择 Spark。
4. 在 Edit software settings (编辑软件设置) 下，将 Enter configuration (输入配置) 保留选中状态并输入以下配置：

```
classification=spark-defaults,properties=[spark.executor.memory=2G]
```

5. 选择其他选项，选择 ，然后选择 Create cluster (创建集群)。

设置 maximizeResourceAllocation

- 使用 Amazon CLI 创建一个安装了 Spark 且 `maximizeResourceAllocation` 设为 true 的集群，该集群引用存储在 Amazon S3 中的 `myConfig.json` 文件 。

```
aws emr create-cluster --release-label emr-5.36.1 --applications Name=Spark \
--instance-type m5.xlarge --instance-count 2 --service-role EMR_DefaultRole_V2
 --ec2-attributes InstanceProfile=EMR_EC2_DefaultRole --configurations https://
s3.amazonaws.com/mybucket/myfolder/myConfig.json
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符 (\ )。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号 (^)。

`myConfig.json`:

```
[
  {
    "Classification": "spark",
    "Properties": {
      "maximizeResourceAllocation": "true"
    }
  }
]
```

> **Note**
>
> 对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

从 Apache Log4j 1.x 迁移到 Log4j 2.x

Apache Spark 版本 3.2.x 及更早版本使用旧版 Apache Log4j 1.x 和 `log4j.properties` 文件在 Spark 进程中配置 Log4j。Apache Spark 版本 3.3.0 及更高版本使用 Apache Log4j 2.x 和 `log4j2.properties` 文件在 Spark 进程中配置 Log4j。

如果您使用低于 6.8.0 的 Amazon EMR 版本配置了 Apache Spark Log4j，则必须删除旧版 `spark-log4j` 配置分类并迁移到 `spark-log4j2` 配置分类和密钥格式，然后才能升级到 Amazon EMR 6.8.0 或更高版本。在 Amazon EMR 版本 6.8.0 及更高版本中，旧版 `spark-log4j` 分类会导致集群创建失败并出现 `ValidationException` 错误。不会因为与 Log4j 不兼容相关的故障而向您收费，但您必须删除已失效的 `spark-log4j` 配置分类才能继续。

有关从 Apache Log4j 1.x 迁移到 Log4j 2.x 的更多信息，请参阅 Github 上的《Apache Log4j 迁移指南》和 Spark Log4j 2 模板。

> **Note**
>
> 对于 Amazon EMR，Apache Spark 使用 `log4j2.properties` 文件，而不是《Apache Log4j 迁移指南》中所述的 .xml 文件。此外，我们不建议使用 Log4j 1.x 桥接方法转换为 Log4j 2.x。

优化 Spark 性能

Amazon EMR 为 Spark 提供多项性能优化功能。本主题详细介绍了各个优化功能。

有关如何设置 Spark 配置的更多信息，请参阅 配置 Spark。

自适应查询执行

自适应查询执行是一个根据运行时统计信息重新优化查询计划的框架。自 Amazon EMR 5.30.0 起，以下来自 Apache Spark 3 的自适应查询执行优化可用于 Spark 2 的 Apache EMR 运行时。

- 自适应连接转换
- 随机分区的自适应合并

自适应连接转换

自适应连接转换根据查询阶段的运行时大小，将 sort-merge-join 操作转换为 broadcast-hash-joins 操作，以此来提高查询性能。当连接的一端小到足以在所有执行程序之间高效地广播其输出时，Broadcast-hash-joins 往往表现更好，从而避免了对联结的两端进行随机交换和排序的需要。自适应连接转换扩大了 Spark 自动执行 broadcast-hash-joins 时的适用情况范围。

该功能已默认启用。可以通过将 `spark.sql.adaptive.enabled` 设置为 `false` 来禁用它，同时会禁用自适应查询执行框架。Spark 决定在连接的其中一端的运行时大小统计数据不超过 `spark.sql.autoBroadcastJoinThreshold`（其默认值为 10485760 字节，即 10MiB）时，将 sort-merge-join 转换为 broadcast-hash-join。

随机分区的自适应合并

随机分区的自适应合并通过合并小的连续随机分区来避免产生太多小任务的开销，从而提高查询性能。这样，您就可以预先配置更多的初始随机分区，然后在运行时将其减少到目标大小，从而提高拥有更均匀分配的随机分区的可能性。

此功能默认情况下已启用，除非 `spark.sql.shuffle.partitions` 采用显式设置。可以通过将 `spark.sql.adaptive.coalescePartitions.enabled` 设置为 `true` 来启用它。初始数量的随机分区和目标分区大小都可以分别使用 `spark.sql.adaptive.coalescePartitions.minPartitionNum` 和 `spark.sql.adaptive.advisoryPartitionSizeInBytes` 属性进行优化。有关此功能的相关 Spark 属性的详细信息，请参阅下表。

Spark 自适应合并分区属性

属性	默认值	描述
`spark.sql.adaptive.coalescePartitions.enabled`	true，除非 `spark.sql.shuffle.partitions` 为显式设置	如果为 true 且 spark.sql.adaptive.enabled 为 true，则 Spark 将根据目标大小

属性	默认值	描述
		合并连续的随机分区（通过 spark.sql.adaptive .advisoryPartition SizeInBytes 指定），以避免过多的小任务。
spark.sql.adaptive .advisoryPartition SizeInBytes	64MB	合并时，随机分区q的指导大小（按字节计算）。此配置仅在 spark.sql.adaptive .enabled 和 spark.sql .adaptive.coalesce Partitions.enabled 两者都为 true 时才有效。
spark.sql.adaptive .coalescePartition s.minPartitionNum	25	合并后的最小随机分区数。此配置仅在 spark.sql .adaptive.enabled 和 spark.sql.adaptive .coalescePartition s.enabled 两者都为 true 时才有效。
spark.sql.adaptive .coalescePartition s.initialPartition Num	1000	合并前的随机分区的初始数量。此配置仅在 spark.sql .adaptive.enabled 和 spark.sql.adaptive .coalescePartition s.enabled 两者都为 true 时才有效。

动态分区修剪

动态分区修剪通过针对特定的查询更准确地选择表中需要读取和处理的特定分区来提高作业性能。通过减少读取和处理的数据量，可节省大量的作业执行时间。对于 Amazon EMR 5.26.0，此功能已默认启

用。对于 Amazon EMR 5.24.0 和 5.25.0，您可以在 Spark 中或在创建集群时，通过设置 Spark 属性 `spark.sql.dynamicPartitionPruning.enabled` 来启用此功能。

Spark 动态分区修剪分区属性

属性	默认值	描述
`spark.sql.dynamicPartitionPruning.enabled`	`true`	如果为 true，则启用动态分区修剪。
`spark.sql.optimizer.dynamicPartitionPruning.enforceBroadcastReuse`	`true`	为 `true` 时，Spark 会在查询执行之前执行防御性检查，以确保动态修剪筛选条件中广播交换的重复使用不会被以后的准备规则（如用户定义的列式规则）中断。当重用被中断且此配置是 `true` 时，Spark 会删除受影响的动态修剪筛选条件，以防止发生性能和正确性问题。当动态修剪筛选条件的广播交换从相应连接操作的广播交换产生不同且不一致的结果时，可能会出现正确性问题。将此配置设置为 `false` 应谨慎执行；它允许解决如下类似场景：当用户定义的列式规则中断重用时。启用“自适应查询执行”后，将始终强制执行广播重用。

这种优化功能在 Spark 2.4.2 的现有功能基础之上进行改进，只支持向下推送可以在计划时解析的静态谓词。

以下是 Spark 2.4.2 中静态谓词向下推送的示例。

```
partition_col = 5
```

```
partition_col IN (1,3,5)

partition_col between 1 and 3

partition_col = 1 + 3
```

动态分区修剪允许 Spark 引擎在运行时动态地推断哪些分区需要读取，哪些分区可以安全地消除。例如，以下查询涉及两个表：store_sales 表，其中包含所有店铺的全部总销售额（按区域分区）；以及 store_regions 表，其中包含每个国家/地区的区域映射。这些表包含有关分布于全球的存储的数据，但我们只查询北美的数据。

```
select ss.quarter, ss.region, ss.store, ss.total_sales
from store_sales ss, store_regions sr
where ss.region = sr.region and sr.country = 'North America'
```

如果没有动态分区修剪，此查询将读取所有区域，然后过滤出与子查询的结果匹配的区域子集。使用动态分区修剪，此查询将只读取和处理子查询中返回的区域的分区。这样，通过减少数据存储和处理较少的记录，节省了时间和资源。

展平标量子查询

这种优化功能通过对同一个表执行标量子查询来提高查询的性能。对于 Amazon EMR 5.26.0，此功能已默认启用。借助 Amazon EMR 5.24.0 和 5.25.0，您可以在 Spark 中或在创建集群时，通过设置 Spark 属性 spark.sql.optimizer.flattenScalarSubqueriesWithAggregates.enabled 来启用此功能。当此属性设置为 true 时，查询优化程序会展平使用相同关系的聚合标量子查询（如果可能）。标量子查询通过以下方法展平：将子查询中存在的任何谓词推送到聚合函数，然后执行一个聚合（针对所有聚合函数，按每个关系）。

以下示例是一个将受益于此优化的查询示例。

```
select (select avg(age) from students                    /* Subquery 1 */
                where age between 5 and 10) as group1,
       (select avg(age) from students                    /* Subquery 2 */
                where age between 10 and 15) as group2,
       (select avg(age) from students                    /* Subquery 3 */
                where age between 15 and 20) as group3
```

此优化将之前的查询重写为：

```
select c1 as group1, c2 as group2, c3 as group3
```

```
from (select avg (if(age between 5 and 10, age, null)) as c1,
             avg (if(age between 10 and 15, age, null)) as c2,
             avg (if(age between 15 and 20, age, null)) as c3 from students);
```

请注意，重写的查询将只读取一次 student 表，而三个子查询的谓词将推送到 avg 函数中。

DISTINCT Before INTERSECT

这种优化可优化使用 INTERSECT 时的联接。对于 Amazon EMR 5.26.0，此功能已默认启用。借助 Amazon EMR 5.24.0 和 5.25.0，您可以在 Spark 中或在创建集群时，通过设置 Spark 属性 spark.sql.optimizer.distinctBeforeIntersect.enabled 来启用此功能。使用 INTERSECT 的查询会自动转换为使用左半联接。当此属性设置为 true 时，如果查询优化程序检测到 DISTINCT 运算符可以进行左半联接 BroadcastHashJoin（而非 SortMergeJoin），它会将 DISTINCT 运算符推送到 INTERSECT 的子级。

以下示例是一个将受益于此优化的查询示例。

```
(select item.brand brand from store_sales, item
     where store_sales.item_id = item.item_id)
intersect
(select item.brand cs_brand from catalog_sales, item
     where catalog_sales.item_id = item.item_id)
```

如果没有启用此属性 spark.sql.optimizer.distinctBeforeIntersect.enabled，则查询将被重写，如下所示。

```
select distinct brand from
  (select item.brand brand from store_sales, item
     where store_sales.item_id = item.item_id)
left semi join
   (select item.brand cs_brand from catalog_sales, item
     where catalog_sales.item_id = item.item_id)
 on brand <=> cs_brand
```

当您启用此属性 spark.sql.optimizer.distinctBeforeIntersect.enabled 时，查询将被重写，如下所示。

```
select brand from
  (select distinct item.brand brand from store_sales, item
```

```
     where store_sales.item_id = item.item_id)
left semi join
   (select distinct item.brand cs_brand from catalog_sales, item
     where catalog_sales.item_id = item.item_id)
 on brand <=> cs_brand
```

Bloom 筛选条件连接

这种优化可以通过使用从联接另一端的值生成的 Bloom 筛选条件对联接的一端进行预筛选，来提高部分联接的性能。对于 Amazon EMR 5.26.0，此功能已默认启用。借助 Amazon EMR 5.25.0，您可以在 Spark 中或在创建集群时，通过将 Spark 属性 `spark.sql.bloomFilterJoin.enabled` 设置为 `true` 来启用此功能。

下面是一个可以受益于 Bloom 筛选条件的示例查询。

```
select count(*)
from sales, item
where sales.item_id = item.id
and item.category in (1, 10, 16)
```

启用此功能后，Bloom 筛选条件将根据所有类别位于要查询的类别集中的项目ID构建。扫描销售表时，Bloom 筛选条件用于确定哪些销售属于肯定不在 Bloom 筛选条件定义的集中的项目。借此，可以尽早筛选出这些被标识的销售。

优化的连接重新排序

这项优化通过将涉及带筛选条件的表的联接进行重新排序来提高查询性能。对于 Amazon EMR 5.26.0，此功能已默认启用。对于 Amazon EMR 5.25.0，您可以通过将 Spark 配置参数 `spark.sql.optimizer.sizeBasedJoinReorder.enabled` 设置为 true 来启用此功能。Spark 的默认行为是从左到右联接表，如查询中所列。此策略可能会错过首先使用筛选条件执行较小联接的机会，以便之后利用更昂贵的联接。

下面的示例查询报告了一个国家/地区所有商店的所有退回商品。如果不经过优化的联接重新排序，Spark 首先会联接两个大型表 `store_sales` 和 `store_returns`，然后将其与 `store` 联接，最终再联接 `item`。

```
select ss.item_value, sr.return_date, s.name, i.desc,
from store_sales ss, store_returns sr, store s, item i
where ss.id = sr.id and ss.store_id = s.id and ss.item_id = i.id
```

```
and s.country = 'USA'
```

经过优化的联接重新排序，Spark 首先会联接 store_sales 与 store，因为 store 有一个筛选条件并且小于 store_returns 和 broadcastable。然后，Spark 会联接 store_returns，最后联接 item。如果 item 有一个筛选条件并且可广播，则其也符合重新排序的条件，这会使 store_sales 与 store 联接，之后联接 item，并在最后联接 store_returns。

Spark 结果片段缓存

Amazon EMR 6.6.0 及更高版本包含可选的 Spark 结果片段缓存功能，该功能可自动缓存结果片段。这些结果片段是查询子树的结果的一部分，其存储在您选择的 Amazon S3 存储桶中。存储的查询结果片段将在后续查询执行时重复使用，从而加快查询速度。

结果片段缓存的工作原理是分析 Spark SQL 查询并将符合条件的结果片段缓存在指定的 S3 位置。在后续查询运行中，系统会自动检测并从 S3 中获取可用的查询结果片段。结果片段缓存不同于结果集缓存，其中，后续查询必须与原始查询完全匹配才能从缓存返回结果。当用于重复以静态数据子集为目标的查询时，结果片段缓存可显著提高性能。

请考虑以下查询，它计算 2022 年之前的订单：

```
select
    l_returnflag,
    l_linestatus,
    count(*) as count_order
from
    lineitem
where
    l_shipdate <= current_date
    and year(l_shipdate) == '2022'
group by
    l_returnflag,
    l_linestatus
```

随着时间推移，此查询需要每天运行以报告当年的总销售额。如果没有结果片段缓存，则需要每天重新计算一年中所有日期的结果。随着时间推移，查询速度会变慢，并且在年底最慢，届时将需要重新计算所有 365 天的结果。

当您激活结果片段缓存时，将使用缓存中一年所有以前日期的结果。每天，该功能只能重新计算一天的结果。在该功能计算结果片段后，该功能将缓存片段。因此，启用缓存的查询时间很快，并且每次后续查询都保持不变。

启用 Spark 结果片段缓存

要启用 Spark 结果片段缓存，请执行以下步骤：

1. 在 Amazon S3 中创建缓存存储桶并授权 EMRFS 的读/写访问。有关更多信息，请参阅授予对 Amazon S3 中的 EMRFS 数据的访问权。
2. 设置 EMR Spark 配置以启用该功能。

```
spark.subResultCache.enabled = true
spark.subResultCache.fs.root.path = s3://DOC-EXAMPLE-BUCKET/cache_dir/
```

3. 为存储桶启用 S3 生命周期管理以自动清理缓存文件。
4. 或者，配置 reductionRationThreshold 和 maxBufferSize 属性以进一步调整该功能。

```
spark.sql.subResultCache.reductionRatioThreshold
spark.sql.subResultCache.maxBufferSize
```

使用结果片段缓存时的注意事项

当您使用已缓存在 Amazon S3 中的结果而不是重新计算它们时，所节省的成本会随着使用相同缓存结果的次数而增加。对于具有大表扫描后跟筛选条件或散列聚合，并且将结果大小减少至少 8 倍（即输入大小:结果的比率至少为 8:1）的查询将从此功能受益最多。输入和结果之间的缩减率越大，成本效益就越大。只要生成结果的成本高于从 Amazon S3 获取结果的成本，缩减率较小、但在表扫描和筛选条件或聚合之间包含昂贵计算步骤的查询也将受益。默认情况下，结果片段缓存仅在检测到缩减率至少为 8:1 时才生效。

当查询重复使用缓存的结果时，此功能的好处最大。滚动和增量窗口查询就是很好的例子。例如，一个 30 天滚动窗口查询已经运行了 29 天，它只需要从其原始输入源提取 1/30 的目标数据，并将使用前 29 天的缓存结果片段。增量窗口查询将受益更多，因为窗口的开始保持固定：在每次调用查询时，需要从输入源读取的处理比例较小。

以下是使用结果片段缓存时的其他注意事项：

- 如果查询的目标不是具有相同查询片段的相同数据，则缓存命中率较低，因此不会从此功能受益。
- 如果查询的缩减率较低且不包含昂贵的计算步骤，则将导致缓存结果的读取开销与初始处理的开销大致相同。
- 由于写入缓存的成本，第一个查询将始终显示较小的回归。

- 结果片段缓存功能仅适用于 Parquet 文件。不支持其他文件格式。
- 结果片段缓存功能缓冲区将仅尝试缓存文件拆分大小为 128 MB 或更大的扫描。在默认 Spark 配置下，如果扫描大小（正扫描的所有文件的总大小）除以执行程序内核数小于 128 MB，则结果片段缓存将被禁用。如果设置了下面所列的任何 Spark 配置，则文件拆分大小将为：

```
min(maxPartitionBytes, max(openCostInBytes, scan size / minPartitionNum))
```

  - spark.sql.leafNodeDefaultParallelism（默认值为 spark.default.parallelism)
  - spark.sql.files.minPartitionNum（默认值为 spark.sql.leafNodeDefaultParallelism）
  - spark.sql.files.openCostInBytes
  - spark.sql.files.maxPartitionBytes
- 结果片段缓存功能以 RDD 分区粒度缓存。前面描述的默认 8:1 缩减率是按每个 RDD 分区评估的。与每 RDD 缩减率始终低于 8:1 的工作负载相比，每 RDD 缩减率大于和低于 8:1 的工作负载的性能优势可能更小。
- 默认情况下，结果片段缓存功能对缓存的每个 RDD 分区使用 16MB 写入缓冲区。如果每个 RDD 分区的缓存超过 16MB，则确定无法进行写入的成本可能会导致性能下降。
- 默认情况下，结果片段缓存不会尝试缓存缩减率小于 8:1 的 RDD 分区结果，并将写入缓冲区限制为 16MB，但这两个值都可以通过以下配置进行调整：

```
spark.sql.subResultCache.reductionRatioThreshold (default: 8.0)
spark.sql.subResultCache.maxBufferSize (default: 16MB, max: 64MB)
```

- 使用相同 EMR 发行版的多个集群可以共享同一个缓存位置。为了确保结果的正确性，结果片段缓存将不使用不同发行版的 Amazon EMR 写入的缓存结果。
- 对于 Spark Streaming 用例或当使用 RecordServer、Apache Ranger 或 Amazon Lake Formation 时，将自动禁用结果片段缓存。
- 结果片段缓存读/写使用 EMRFS 和 Amazon S3 存储桶。支持 CSE/SSE S3/SSE KMS 加密。

使用 Nvidia Spark-RAPIDS Accelerator for Spark

对于 Amazon EMR 发行版 6.2.0 及更高版本，您可以使用 Nvidia 的 RAPIDS Accelerator for Apache Spark 插件来通过 EC2 图形处理器（GPU）实例类型加速 Spark。Rapids Accelerator 将通过 GPU 加速您的 Apache Spark 3.0 数据科学管道，无需更改代码，并将加快数据处理和模型训练，同时大幅降低基础设施成本。

以下部分将引导您完成配置 EMR 集群来使用 Spark-RAPIDS Plugin for Spark。

选择实例类型

要将 Nvidia Spark-RAPIDS 插件用于 Spark，核心实例组和任务实例组必须使用符合 Spark-RAPIDS 的硬件要求的 EC2 GPU 实例类型。要查看 EMR 支持的 GPU 实例类型的完整列表，请参阅《Amazon EMR 管理指南》中的支持的实例类型。主实例组的实例类型可以是 GPU 或非 GPU 类型，但不支持 ARM 实例类型。

为集群设置应用程序配置

1。使 Amazon EMR 能在您的新集群上安装插件

要安装插件，请在创建集群时提供以下配置：

```
{
 "Classification":"spark",
 "Properties":{
  "enableSparkRapids":"true"
 }
}
```

2. 将 YARN 配置为使用 GPU

有关在 YARN 上使用 GPU 的详细信息，请参阅 Apache Hadoop 文档中的 Using GPU On YARN。下面是示例配置：

```
{
 "Classification":"yarn-site",
 "Properties":{
  "yarn.nodemanager.resource-plugins":"yarn.io/gpu",
  "yarn.resource-types":"yarn.io/gpu",
  "yarn.nodemanager.resource-plugins.gpu.allowed-gpu-devices":"auto",
  "yarn.nodemanager.resource-plugins.gpu.path-to-discovery-executables":"/usr/bin",
  "yarn.nodemanager.linux-container-executor.cgroups.mount":"true",
  "yarn.nodemanager.linux-container-executor.cgroups.mount-path":"/sys/fs/cgroup",
  "yarn.nodemanager.linux-container-executor.cgroups.hierarchy":"yarn",
  "yarn.nodemanager.container-
executor.class":"org.apache.hadoop.yarn.server.nodemanager.LinuxContainerExecutor"
 }
},
{
 "Classification":"container-executor",
 "Properties":{
```

```
 },
 "Configurations":[
  {
   "Classification":"gpu",
   "Properties":{
    "module.enabled":"true"
   }
  },
  {
   "Classification":"cgroups",
   "Properties":{
    "root":"/sys/fs/cgroup",
    "yarn-hierarchy":"yarn"
   }
  }
 ]
}
```

3. 将 Spark 配置为使用 RAPIDS

以下是使 Spark 能够使用 RAPIDS 插件所需的配置：

```
{
 "Classification":"spark-defaults",
 "Properties":{
  "spark.plugins":"com.nvidia.spark.SQLPlugin",
  "spark.executor.resource.gpu.discoveryScript":"/usr/lib/spark/scripts/gpu/
getGpusResources.sh",
  "spark.executor.extraLibraryPath":"/usr/local/cuda/targets/x86_64-linux/lib:/usr/
local/cuda/extras/CUPTI/lib64:/usr/local/cuda/compat/lib:/usr/local/cuda/lib:/usr/
local/cuda/lib64:/usr/lib/hadoop/lib/native:/usr/lib/hadoop-lzo/lib/native:/docker/usr/
lib/hadoop/lib/native:/docker/usr/lib/hadoop-lzo/lib/native"
 }
}
```

在您的集群上启用 Spark RAPIDS 插件后，XGBoost 文档中提供的 XGBoost4J-Spark 库也可以使用。您可以使用以下配置将 XGBoost 与您的 Spark 任务集成：

```
{
 "Classification":"spark-defaults",
 "Properties":{
  "spark.submit.pyFiles":"/usr/lib/spark/jars/xgboost4j-spark_3.0-1.4.2-0.3.0.jar"
```

```
 }
}
```

有关可用于优化 GPU 加速的 EMR 集群的其他 Spark 配置，请参阅 Nvidia.github.io 文档中的 RAPIDS Accelerator for Apache Spark Tuning Guide。

4. 配置 YARN 容量调度器

必须配置 `DominantResourceCalculator` 来启用 GPU 调度和隔离。有关详细信息，请参阅 Apache Hadoop 文档中的 Using GPU On YARN。

```
{
 "Classification":"capacity-scheduler",
 "Properties":{
  "yarn.scheduler.capacity.resource-
calculator":"org.apache.hadoop.yarn.util.resource.DominantResourceCalculator"
 }
}
```

5. 创建一个 JSON 文件以包含您的所有配置

您可以创建一个 JSON 文件，在其中包含您的配置，以便为 Spark 集群使用 RAPIDS 插件。您稍后在启动集群时需提供该文件。

文件可以存储在本地或 S3 上。有关如何为集群提供应用程序配置的详细信息，请参阅配置应用程序。

下面是一个文件示例，名称为 `my-configurations.json`。您可以将其用作模板，来开始构建自己的配置。

```
[
 {
  "Classification":"spark",
  "Properties":{
   "enableSparkRapids":"true"
  }
 },
 {
  "Classification":"yarn-site",
  "Properties":{
   "yarn.nodemanager.resource-plugins":"yarn.io/gpu",
   "yarn.resource-types":"yarn.io/gpu",
   "yarn.nodemanager.resource-plugins.gpu.allowed-gpu-devices":"auto",
   "yarn.nodemanager.resource-plugins.gpu.path-to-discovery-executables":"/usr/bin",
```

```
    "yarn.nodemanager.linux-container-executor.cgroups.mount":"true",
    "yarn.nodemanager.linux-container-executor.cgroups.mount-path":"/sys/fs/cgroup",
    "yarn.nodemanager.linux-container-executor.cgroups.hierarchy":"yarn",
    "yarn.nodemanager.container-
executor.class":"org.apache.hadoop.yarn.server.nodemanager.LinuxContainerExecutor"
   }
  },
  {
   "Classification":"container-executor",
   "Properties":{

   },
   "Configurations":[
    {
     "Classification":"gpu",
     "Properties":{
      "module.enabled":"true"
     }
    },
    {
     "Classification":"cgroups",
     "Properties":{
      "root":"/sys/fs/cgroup",
      "yarn-hierarchy":"yarn"
     }
    }
   ]
  },
  {
   "Classification":"spark-defaults",
   "Properties":{
    "spark.plugins":"com.nvidia.spark.SQLPlugin",
    "spark.executor.resource.gpu.discoveryScript":"/usr/lib/spark/scripts/gpu/
getGpusResources.sh",
    "spark.executor.extraLibraryPath":"/usr/local/cuda/targets/x86_64-linux/lib:/usr/
local/cuda/extras/CUPTI/lib64:/usr/local/cuda/compat/lib:/usr/local/cuda/lib:/usr/
local/cuda/lib64:/usr/lib/hadoop/lib/native:/usr/lib/hadoop-lzo/lib/native:/docker/usr/
lib/hadoop/lib/native:/docker/usr/lib/hadoop-lzo/lib/native",
    "spark.submit.pyFiles":"/usr/lib/spark/jars/xgboost4j-spark_3.0-1.4.2-0.3.0.jar",
    "spark.rapids.sql.concurrentGpuTasks":"1",
    "spark.executor.resource.gpu.amount":"1",
    "spark.executor.cores":"2",
    "spark.task.cpus":"1",
    "spark.task.resource.gpu.amount":"0.5",
```

```
   "spark.rapids.memory.pinnedPool.size":"0",
   "spark.executor.memoryOverhead":"2G",
   "spark.locality.wait":"0s",
   "spark.sql.shuffle.partitions":"200",
   "spark.sql.files.maxPartitionBytes":"512m"
  }
 },
 {
  "Classification":"capacity-scheduler",
  "Properties":{
   "yarn.scheduler.capacity.resource-
calculator":"org.apache.hadoop.yarn.util.resource.DominantResourceCalculator"
  }
 }
]
```

为您的集群添加引导操作

为了在 GPU 上使用 YARN，您需要在集群上打开对 YARN 的 cgroup 权限，这可以使用 EMR 引导操作脚本来完成。

有关如何在创建集群时提供引导操作脚本的更多信息，请参阅《Amazon EMR 管理指南》中的引导操作基础。

下面是一个名为 `my-bootstap-action.sh` 的示例脚本：

```
#!/bin/bash

set -ex

sudo chmod a+rwx -R /sys/fs/cgroup/cpu,cpuacct
sudo chmod a+rwx -R /sys/fs/cgroup/devices
```

启动您的集群。

最后一步是使用上述集群配置启动您的集群。以下是一个通过 EMR CLI 启动集群的命令示例：

```
 aws emr create-cluster \
--release-label emr-6.2.0 \
--applications Name=Hadoop Name=Spark \
--service-role EMR_DefaultRole_V2 \
```

```
--ec2-attributes KeyName=my-key-pair,InstanceProfile=EMR_EC2_DefaultRole \
--instance-groups InstanceGroupType=MASTER,InstanceCount=1,InstanceType=m4.4xlarge \

                  InstanceGroupType=CORE,InstanceCount=1,InstanceType=g4dn.2xlarge \

                  InstanceGroupType=TASK,InstanceCount=1,InstanceType=g4dn.2xlarge \
--configurations file:///my-configurations.json \
--bootstrap-actions Name='My Spark Rapids Bootstrap action',Path=s3://my-bucket/my-
bootstrap-action.sh
```

访问 Spark Shell

Spark Shell 基于 Scala REPL (读取-求值-输出-循环)。它让您能够以交互方式创建 Spark 程序并将工作提交到框架。您可以通过 SSH 连接主节点并调用 `spark-shell`，从而访问 Spark Shell。有关如何连接到主节点的更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。以下示例使用存储在 Amazon S3 中的 Apache HTTP Server 访问日志。

> **Note**
>
> 这些示例中使用的存储桶对能够访问美国东部（弗吉尼亚北部）的客户端可用。

默认情况下，Spark Shell 创建其自己的 SparkContext 对象（称作 `sc`）。如果 REPL 中需要，您可以使用此上下文。sqlContext 在 Shell 中也可用，它是一个 HiveContext。

Example 使用 Spark Shell 统计存储在 Amazon S3 中的某个文件中的某个字符串的出现次数

本示例使用 `sc` 读取 Amazon S3 中的 textFile。

```
scala> sc
res0: org.apache.spark.SparkContext = org.apache.spark.SparkContext@404721db

scala> val textFile = sc.textFile("s3://elasticmapreduce/samples/hive-ads/tables/
impressions/dt=2009-04-13-08-05/ec2-0-51-75-39.amazon.com-2009-04-13-08-05.log")
```

Spark 创建 textFile 及关联的数据结构。然后，示例会统计此日志文件中包含字符串“cartoonnetwork.com”的行数：

```
scala> val linesWithCartoonNetwork = textFile.filter(line =>
 line.contains("cartoonnetwork.com")).count()
```

```
linesWithCartoonNetwork: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at
 filter at <console>:23
<snip>
<Spark program runs>
scala> linesWithCartoonNetwork
res2: Long = 9
```

Example 使用基于 Python 的 Spark Shell 统计存储在 Amazon S3 中的某个文件中的某个字符串的出现次数

Spark 还包含一个基于 Python 的 Shell `pyspark`，您可以用它来设计以 Python 编写的 Spark 程序的原型。与使用 `spark-shell` 的方法一样，在主节点上调用 `pyspark` 即可；它包含同样的 SparkContext 对象。

```
>>> sc
<pyspark.context.SparkContext object at 0x7fe7e659fa50>
>>> textfile = sc.textFile("s3://elasticmapreduce/samples/hive-ads/tables/impressions/
dt=2009-04-13-08-05/ec2-0-51-75-39.amazon.com-2009-04-13-08-05.log")
```

Spark 创建 textFile 及关联的数据结构。然后，示例会统计此日志文件中包含字符串“cartoonnetwork.com”的行数。

```
>>> linesWithCartoonNetwork = textfile.filter(lambda line: "cartoonnetwork.com" in
 line).count()
15/06/04 17:12:22 INFO lzo.GPLNativeCodeLoader: Loaded native gpl library from the
 embedded binaries
15/06/04 17:12:22 INFO lzo.LzoCodec: Successfully loaded & initialized native-lzo
 library [hadoop-lzo rev EXAMPLE]
15/06/04 17:12:23 INFO fs.EmrFileSystem: Consistency disabled, using
 com.amazon.ws.emr.hadoop.fs.s3n.S3NativeFileSystem as filesystem implementation
<snip>
<Spark program continues>
>>> linesWithCartoonNetwork
9
```

将 Amazon SageMaker Spark 用于机器学习

当使用 Amazon EMR 发行版 5.11.0 及更高版本时，`aws-sagemaker-spark-sdk` 组件将随 Spark 一起安装。此组件安装 Amazon SageMaker Spark 和用于 Spark 与 Amazon SageMaker 集成的关联依赖项。您可以使用 Amazon SageMaker Spark 构造使用 Amazon SageMaker 阶段的 Spark 机器学

习 (ML) 管道。有关更多信息，请参阅 GitHub 上的 Amazon SageMaker Spark README 和《Amazon SageMaker 开发人员指南》中的将 Apache Spark 与 Amazon SageMaker 结合使用。

编写 Spark 应用程序

可使用 Scala、Java 或 Python 来编写 Spark 应用程序。Apache Spark 文档的 Spark 示例主题包含多个 Spark 应用程序示例。下面所示为三个内在支持的应用程序中的 Estimating Pi 示例。您还可以在 `$SPARK_HOME/examples` 和 GitHub 查看完整的示例。有关如何为 Spark 构建 JAR 的更多信息，请参阅 Apache Spark 文档中的快速入门主题。

Scala

为避免发生 Scala 兼容性问题，建议您在为 Amazon EMR 集群编译 Spark 应用程序时使用正确的 Scala 版本的 Spark 依赖项。您应该使用的 Scala 版本取决于您的集群上安装的 Spark 版本。例如，Amazon EMR 发行版 5.30.1 使用 Spark 2.4.5，该版本是使用 Scala 2.11 构建的。如果您的集群使用 Amazon EMR 发行版 5.30.1，请使用 Scala 2.11 的 Spark 依赖项。有关 Spark 使用的 Scala 版本的更多信息，请参阅 Apache Spark 文档

```
package org.apache.spark.examples
import scala.math.random
import org.apache.spark._

/** Computes an approximation to pi */
object SparkPi {
  def main(args: Array[String]) {
    val conf = new SparkConf().setAppName("Spark Pi")
    val spark = new SparkContext(conf)
    val slices = if (args.length > 0) args(0).toInt else 2
    val n = math.min(100000L * slices, Int.MaxValue).toInt // avoid overflow
    val count = spark.parallelize(1 until n, slices).map { i =>
      val x = random * 2 - 1
      val y = random * 2 - 1
      if (x*x + y*y < 1) 1 else 0
    }.reduce(_ + _)
    println("Pi is roughly " + 4.0 * count / n)
    spark.stop()
  }
}
```

Java

```
package org.apache.spark.examples;

import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.api.java.function.Function;
import org.apache.spark.api.java.function.Function2;

import java.util.ArrayList;
import java.util.List;

/**
 * Computes an approximation to pi
 * Usage: JavaSparkPi [slices]
 */
public final class JavaSparkPi {

  public static void main(String[] args) throws Exception {
    SparkConf sparkConf = new SparkConf().setAppName("JavaSparkPi");
    JavaSparkContext jsc = new JavaSparkContext(sparkConf);

    int slices = (args.length == 1) ? Integer.parseInt(args[0]) : 2;
    int n = 100000 * slices;
    List<Integer> l = new ArrayList<Integer>(n);
    for (int i = 0; i < n; i++) {
      l.add(i);
    }

    JavaRDD<Integer> dataSet = jsc.parallelize(l, slices);

    int count = dataSet.map(new Function<Integer, Integer>() {
      @Override
      public Integer call(Integer integer) {
        double x = Math.random() * 2 - 1;
        double y = Math.random() * 2 - 1;
        return (x * x + y * y < 1) ? 1 : 0;
      }
    }).reduce(new Function2<Integer, Integer, Integer>() {
      @Override
      public Integer call(Integer integer, Integer integer2) {
        return integer + integer2;
```

```
      }
    });

    System.out.println("Pi is roughly " + 4.0 * count / n);

    jsc.stop();
  }
}
```

Python

```
import argparse
import logging
from operator import add
from random import random

from pyspark.sql import SparkSession

logger = logging.getLogger(__name__)
logging.basicConfig(level=logging.INFO, format="%(levelname)s: %(message)s")


def calculate_pi(partitions, output_uri):
    """
    Calculates pi by testing a large number of random numbers against a unit circle
    inscribed inside a square. The trials are partitioned so they can be run in
    parallel on cluster instances.

    :param partitions: The number of partitions to use for the calculation.
    :param output_uri: The URI where the output is written, typically an Amazon S3
                       bucket, such as 's3://example-bucket/pi-calc'.
    """

    def calculate_hit(_):
        x = random() * 2 - 1
        y = random() * 2 - 1
        return 1 if x**2 + y**2 < 1 else 0

    tries = 100000 * partitions
    logger.info(
        "Calculating pi with a total of %s tries in %s partitions.", tries, partitions
    )
```

```
    with SparkSession.builder.appName("My PyPi").getOrCreate() as spark:
        hits = (
            spark.sparkContext.parallelize(range(tries), partitions)
            .map(calculate_hit)
            .reduce(add)
        )
        pi = 4.0 * hits / tries
        logger.info("%s tries and %s hits gives pi estimate of %s.", tries, hits, pi)
        if output_uri is not None:
            df = spark.createDataFrame([(tries, hits, pi)], ["tries", "hits", "pi"])
            df.write.mode("overwrite").json(output_uri)


if __name__ == "__main__":
    parser = argparse.ArgumentParser()
    parser.add_argument(
        "--partitions",
        default=2,
        type=int,
        help="The number of parallel partitions to use when calculating pi.",
    )
    parser.add_argument(
        "--output_uri", help="The URI where output is saved, typically an S3 bucket."
    )
    args = parser.parse_args()

    calculate_pi(args.partitions, args.output_uri)
```

使用 Amazon S3 提高 Spark 性能

Amazon EMR 提供一些功能，有助于优化使用 Spark 查询、读取和写入保存在 Amazon S3 中的数据的性能。

S3 Select 可通过将处理“向下推送”到 Amazon S3 来提高某些应用程序中 CSV 和 JSON 文件的查询性能。

经 EMRFS S3 优化的提交程序是 OutputCommitter 类的替代项，该程序使用 EMRFS 的分段上传功能，提高使用 Spark SQL、DataFrame 和 Datasetss 向 Amazon S3 写入 Parquet 文件的性能。

主题

将 S3 Select 与 Spark 结合使用以提高查询性能

在 Amazon EMR 发行版 5.17.0 及更高版本中，您可以将 S3 Select 与 Amazon EMR 上的 Spark 结合使用。S3 Select 可让应用程序仅从对象检索数据子集。对于 Amazon EMR，筛选要处理的大型数据集的计算工作是从集群“向下推送”到 Amazon S3，这可以在某些应用程序中提高性能和减少 Amazon EMR 与 Amazon S3 之间传输的数据量。

S3 Select 支持使用 `s3selectCSV` 和 `s3selectJSON` 值来指定数据格式的 CSV 和 JSON 文件。有关更多信息以及示例，请参阅 在代码中指定 S3 Select。

S3 Select 是否适合我的应用程序？

建议您分别在使用和不使用 S3 Select 的情况下测试您的应用程序，以查看 S3 Select 是否适用于您的应用程序。

使用以下准则来确定您的应用程序是否为使用 S3 Select 的候选项：

- 您的查询将筛选掉原始数据集的一半以上的数据。
- 您在 Amazon S3 和 Amazon EMR 集群之间的网络连接具有良好的传输速度和可用带宽。Amazon S3 不压缩 HTTP 响应，因此响应大小可能会根据压缩的输入文件而增大。

注意事项和限制

- 使用客户提供的加密密钥进行的 Amazon S3 服务器端加密（SSE-C）与客户端加密都不受支持。
- 不支持 `AllowQuotedRecordDelimiters` 属性。如果指定该属性，则查询将失败。
- 仅支持采用 UTF-8 格式的 CSV 和 JSON 文件。不支持多行 CSV。
- 仅支持未压缩文件或 gzip 文件。
- 不支持 Spark CSV 和 JSON 选项（如 `nanValue`、`positiveInf`、`negativeInf`）以及与损坏记录相关的选项（例如，failfast 和 dropmalformed 模式）。
- 不支持在十进制数中使用逗号 (,)。例如，不支持 `10,000`，支持 `10000`。
- 不支持最后一行中的注释字符。

- 文件末尾的空行不会被处理。
- 以下筛选条件不会向下推送到 Amazon S3：
  - 聚合函数（如 `COUNT()` 和 `SUM()`）。
  - 对属性进行 `CAST()` 的筛选条件。例如，`CAST(stringColumn as INT) = 1`。
  - 具有作为对象或很复杂的属性的筛选条件。例如，`intArray[1] = 1`，`objectColumn.objectNumber = 1`。
  - 值不是文本值的筛选条件。例如，`intColumn1 = intColumn2`
  - 仅支持 S3 Select 支持的数据类型，但存在记录的限制。

在代码中指定 S3 Select

以下示例演示了如何使用 Scala、SQL、R 和 PySpark 指定适用于 CSV 的 S3 Select。您可以通过同样的方法使用适用于 JSON 的 S3 Select。有关选项、默认值和限制的列表，请参阅选项。

PySpark

```
spark
  .read
  .format("s3selectCSV") // "s3selectJson" for Json
  .schema(...) // optional, but recommended
  .options(...) // optional
  .load("s3://path/to/my/datafiles")
```

R

```
read.df("s3://path/to/my/datafiles", "s3selectCSV", schema, header = "true",
 delimiter = "\t")
```

Scala

```
spark
  .read
  .format("s3selectCSV") // "s3selectJson" for Json
  .schema(...) // optional, but recommended
  .options(...) // optional. Examples:
  // .options(Map("quote" -> "\'", "header" -> "true")) or
  // .option("quote", "\'").option("header", "true")
  .load("s3://path/to/my/datafiles")
```

SQL

```
CREATE TEMPORARY VIEW MyView (number INT, name STRING) USING s3selectCSV OPTIONS
 (path "s3://path/to/my/datafiles", header "true", delimiter "\t")
```

选项

使用 s3selectCSV 和 s3selectJSON 时，有以下选项可用。如果未指定，将使用默认值。

使用 S3selectCSV 时的选项

选项	默认值	用量
compression	"none"	指示是否使用了压缩。"gzip" 是除 "none" 之外唯一受支持的设置。
delimiter	","	指定字段分隔符。
quote	'\"'	指定引号字符。不支持指定空字符串，这样做会导致 XML 格式不正确错误。
escape	'\\'	指定转义字符。
header	"false"	"false" 指定不存在标头。"true" 指定第一行中存在标头。仅支持第一行中的标头，不支持标头前面的空行。
comment	"#"	指定注释字符。无法禁用注释标记。换句话说，不支持值 \u0000。
nullValue	""	

使用 S3selectJSON 时的选项

选项	默认值	用量
`compression`	`"none"`	指示是否使用了压缩。`"gzip"` 是除 `"none"` 之外唯一受支持的设置。
`multiline`	“false”	`"false"` 指定 JSON 采用 S3 Select `LINES` 格式，这意味着输入数据中的每个行都包含单个 JSON 对象。`"true"` 指定 JSON 采用 S3 Select `DOCUMENT` 格式，这意味着 JSON 对象可跨输入数据中的多个行。

使用经 EMRFS S3 优化的提交程序

经 EMRFS S3 优化的提交程序是 OutputCommitter 的一种实施替代项，该程序已针对在使用 EMRFS 时向 Amazon S3 写入文件进行了优化。通过避免在任务处理和任务提交阶段列出并重命名在 Amazon S3 中完成的操作来提高应用程序性能。提交程序适用于 Amazon EMR 发行版 5.19.0 及更高版本，在 Amazon EMR 5.20.0 及更高版本中将默认启用。提交程序用于 Spark 任务，该任务使用 Spark SQL、DataFrame 或 Dataset。从 Amazon EMR 6.4.0 开始，此提交程序可用于所有常见格式，包括 parquet、ORC 和基于文本的格式（包括 CSV 和 JSON）。对于 Amazon EMR 6.4.0 之前的发行版，仅支持 Parquet 格式。在某些情况下，不使用提交程序。有关更多信息，请参阅经 EMRFS S3 优化的提交程序的要求。

主题

- 经 EMRFS S3 优化的提交程序的要求
- 经 EMRFS S3 优化的提交程序和分段上传
- 任务优化注意事项
- 为 Amazon EMR 5.19.0 启用经 EMRFS S3 优化的提交程序

经 EMRFS S3 优化的提交程序的要求

满足以下条件时，将使用经 EMRFS S3 优化的提交程序：

- 您可以使用 Spark SQL、DataFrame 或 Dataset 运行 Spark 任务，从而将文件写入 Amazon S3。从 Amazon EMR 6.4.0 开始，此提交程序可用于所有常见格式，包括 parquet、ORC 和基于文本的格式（包括 CSV 和 JSON）。对于 Amazon EMR 6.4.0 之前的发行版，仅支持 Parquet 格式。
- 分段上传在 Amazon EMR 中已启用。这是默认模式。有关更多信息，请参阅经 EMRFS S3 优化的提交程序和分段上传。
- 使用 Spark 的内置文件格式支持。内置文件格式支持用于以下情况：
  - 对于 Hive 元存储表，当 `spark.sql.hive.convertMetastoreParquet` 设置为 `true` 时，可用于 Parquet 表，或 `spark.sql.hive.convertMetastoreOrc` 设置为 `true` 时，可用于 EMR 6.4.0 或更高版本的 Orc 表。这些是默认设置。
  - 当任务写入文件格式数据来源或表时，例如，使用 `USING parquet` 子句创建目标表。
  - 当作业写入未分区的 Hive 元存储 Parquet 表时。Spark 的内置 Parquet 支持不支持分区的 Hive 表，这是一个已知限制。有关更多信息，请参阅 Apache Spark SQL、DataFrame 和 Dataset 指南中的 Hive 元存储 Parquet 表转换。
- 写入默认分区位置的 Spark 任务操作，例如 `${table_location}/k1=v1/k2=v2/`，使用提交程序。如果任务操作写入自定义分区位置，则不使用提交程序，例如，如果使用 `ALTER TABLE SQL` 命令设置自定义分区位置。
- 必须使用 Spark 的以下值：
  - `spark.sql.parquet.fs.optimized.committer.optimization-enabled` 属性必须设置为 `true`。这是 Amazon EMR 5.20.0 及更高版本的默认设置。对于 Amazon EMR 5.19.0，默认值是 `false`。有关配置此值的信息，请参阅为 Amazon EMR 5.19.0 启用经 EMRFS S3 优化的提交程序。
  - 如果写入非分区的 Hive 元存储表，则仅支持 Parquet 和 Orc 文件格式。如果写入非分区的 Paquet Hive 元存储表，则须将 `spark.sql.hive.convertMetastoreParquet` 设置为 `true`。如果写入非分区的 Orc Hive 元存储表，则 `spark.sql.hive.convertMetastoreOrc` 须设置为 `true`。这些是默认设置。
  - `spark.sql.parquet.output.committer.class` 必须设置为 `com.amazon.emr.committer.EmrOptimizedSparkSqlParquetOutputCommitter`。这是默认设置。
  - 必须将 `spark.sql.sources.commitProtocolClass` 设置为 `org.apache.spark.sql.execution.datasources.SQLEmrOptimizedCommitProtocol` 或

`org.apache.spark.sql.execution.datasources.SQLHadoopMapReduceCommitProtocol`。是 EMR 5.x 系列 5.30.0 及更高版本以及 EMR 6.x 系列 6.2.0 及更高版本的默认设置。`org.apache.spark.sql.execution.datasources.SQLHadoopMapReduceCommitProtoco` 是以前 EMR 版本的默认设置。

- 如果 Spark 作业用动态分区列覆盖分区的 Parquet 数据集，则 `partitionOverwriteMode` 写入选项和 `spark.sql.sources.partitionOverwriteMode` 必须设置为 `static`。这是默认设置。

> **Note**
>
> Spark 2.4.0 中引入了 `partitionOverwriteMode` 写入选项。对于随附了 Amazon EMR 版本 5.19.0 的 Spark 版本 2.3.2，请设置 `spark.sql.sources.partitionOverwriteMode` 属性。

当不使用经 EMRFS S3 优化的提交程序时

通常，经 EMRFS S3 优化的提交程序不会在以下情况下使用。

情况	为什么不使用提交程序
当您向 HDFS 写入时	提交程序只支持使用 EMRFS 写入 Amazon S3。
当您使用 S3A 文件系统时	提交程序只支持 EMRFS。
当使用 MapReduce 或 Spark 的 RDD API 时	提交程序只支持使用 SparkSQL、DataFrame 或数据集 API。

以下 Scala 示例演示了防止经 EMRFS S3 优化的提交程序被整个使用（第一个示例）和部分使用（第二个示例）的一些其他情况。

Example – 动态分区覆盖模式

以下 Scala 示例指示 Spark 使用不同的提交算法，这完全阻止了经 EMRFS S3 优化的提交程序的使用。该代码将 `partitionOverwriteMode` 属性设置为 `dynamic`，以仅覆盖您的数据所写入到的分区。然后，通过 `partitionBy` 指定动态分区列，并将写入模式设置为 `overwrite`。

```
val dataset = spark.range(0, 10)
  .withColumn("dt", expr("date_sub(current_date(), id)"))

dataset.write.mode("overwrite")
  .option("partitionOverwriteMode", "dynamic")
  .partitionBy("dt")
  .parquet("s3://EXAMPLE-DOC-BUCKET/output")
```

您必须配置所有三项设置，以避免使用经 EMRFS S3 优化的提交程序。当您执行此操作时，Spark 会执行在 Spark 的提交协议中指定的不同提交算法。对于早于 5.30.0 的 Amazon EMR 5.x 发行版和早于 6.2.0 的 Amazon EMR 6.x 发行版，提交协议使用 Spark 的暂存目录，该目录是在以 `.spark-staging` 开头的输出位置下创建的临时目录。该算法按顺序对分区目录进行重命名，这可能会对性能产生负面影响。有关 Amazon EMR 发行版 5.30.0 及更高版本和 6.2.0 及更高版本的更多信息，请参阅使用经 EMRFS S3 优化的提交协议。

Spark 2.4.0 中的算法遵循以下步骤：

1. 任务尝试将其输出写入 Spark 的暂存目录下的分区目录，例如 `${outputLocation}/spark-staging-${jobID}/k1=v1/k2=v2/`。
2. 对于写入的每个分区，任务尝试都跟踪相对分区路径，例如 `k1=v1/k2=v2`。
3. 任务成功完成后，它会为驱动程序提供它跟踪的所有相对分区路径。
4. 完成所有任务后，作业提交阶段将收集成功任务尝试在 Spark 的暂存目录下写入的所有分区目录。Spark 使用目录树重命名操作按顺序将这些目录的每一个都重命名为其最终输出位置。
5. 暂存目录会在作业提交阶段完成之前删除。

Example – 自定义分区位置

在此示例中，Scala 代码插入到两个分区中。一个分区具有自定义分区位置。另一个分区使用默认分区位置。经 EMRFS S3 优化的提交程序仅用于将任务输出写入到使用默认分区位置的分区。

```
val table = "dataset"
val location = "s3://bucket/table"

spark.sql(s"""
  CREATE TABLE $table (id bigint, dt date)
  USING PARQUET PARTITIONED BY (dt)
  LOCATION '$location'
```

```
""")

// Add a partition using a custom location
val customPartitionLocation = "s3://bucket/custom"
spark.sql(s"""
  ALTER TABLE $table ADD PARTITION (dt='2019-01-28')
  LOCATION '$customPartitionLocation'
""")

// Add another partition using default location
spark.sql(s"ALTER TABLE $table ADD PARTITION (dt='2019-01-29')")

def asDate(text: String) = lit(text).cast("date")

spark.range(0, 10)
  .withColumn("dt",
    when($"id" > 4, asDate("2019-01-28")).otherwise(asDate("2019-01-29")))
  .write.insertInto(table)
```

Scala 代码创建以下 Amazon S3 对象：

```
custom/part-00001-035a2a9c-4a09-4917-8819-e77134342402.c000.snappy.parquet
custom_$folder$
table/_SUCCESS
table/dt=2019-01-29/part-00000-035a2a9c-4a09-4917-8819-e77134342402.c000.snappy.parquet
table/dt=2019-01-29_$folder$
table_$folder$
```

当写入到自定义位置的分区时，Spark 会使用类似于上一个示例的提交算法，如下所述。与前面的示例一样，该算法会导致顺序重命名，这可能会影响性能。

1. 在将输出写入自定义位置的分区时，任务会写入到 Spark 的暂存目录下的文件中，该目录是在最终输出位置下创建的。该文件的名称包含一个随机 UUID，以防止文件冲突。任务尝试跟踪每个文件以及最终所需的输出路径。
2. 在任务成功完成后，它会为驱动程序提供这些文件及其最终所需的输出路径。
3. 所有任务都完成后，作业提交阶段会按顺序将为自定义位置的分区写入的所有文件重命名为其最终输出路径。
4. 暂存目录会在作业提交阶段完成之前删除。

经 EMRFS S3 优化的提交程序和分段上传

要使用经 EMRFS S3 优化的提交程序，则必须在 Amazon EMR 中启用分段上传。默认启用分段上传。需要时，您可以重新启用它。有关更多信息，请参阅《Amazon EMR 管理指南》中的为 Amazon S3 配置分段上传。

经 EMRFS S3 优化的提交程序利用分段上传类似于事务的特征，来确保由任务写入的文件在任务提交后尝试仅显示在作业的输出位置。通过以这种方式使用分段上传，提交程序可通过默认的 FileOutputCommitter 算法版本 2 提高任务提交性能。使用经 EMRFS S3 优化的提交程序时，有一些与传统分段上传行为的关键区别需要考虑：

- 无论文件大小如何，分段上传都会执行。这不同于 EMRFS 的默认行为，其中 `fs.s3n.multipart.uploads.split.size` 属性在触发分段上传时，控制文件大小。
- 在任务提交或中止之前，分段上传在较长时间内都保持在未完成状态。这不同于 EMRFS 的默认行为，其中分段上传在任务完成写入给定文件时完成。

由于这些区别，如果 Spark Executor JVM 在任务正在运行或将数据写入到 Amazon S3 时发生崩溃或被终止，未完成的分段上传更可能被留下来。因此，当您使用经 EMRFS S3 优化的提交程序时，请务必遵循管理失败的分段上传的最佳实践。有关更多信息，请参阅《Amazon EMR 管理指南》中使用 Amazon S3 存储桶的最佳实践。

任务优化注意事项

在任务被提交或中止之前，经 EMRFS S3 优化的提交程序会占用少量内存来存放任务尝试写入的每个文件。在大多数作业中，占用的内存量可以忽略不计。对于包含写入大量文件的长期任务的作业，提交程序占用的内存可能很大，需要调整为 Spark 执行程序分配的内存。您可以使用 `spark.executor.memory` 属性调整执行程序内存。作为指导，编写 100000 个文件的单个任务通常需要额外的 100 MB 内存。有关更多信息，请参阅 Apache Spark 配置文档中的应用程序属性。

为 Amazon EMR 5.19.0 启用经 EMRFS S3 优化的提交程序

如果您使用的是 Amazon EMR 5.19.0，则可以在从 Spark 中创建集群时手动将 `spark.sql.parquet.fs.optimized.committer.optimization-enabled` 属性设置为 `true`（如果您使用的是 Amazon EMR）。

在创建集群时启用经 EMRFS S3 优化的提交程序

使用 spark-defaults 配置分类将 spark.sql.parquet.fs.optimized.committer.optimization-enabled 属性设置为 true。有关更多信息，请参阅配置应用程序。

从 Spark 启用经 EMRFS S3 优化的提交程序

您可以将 spark.sql.parquet.fs.optimized.committer.optimization-enabled 设置为 true，方法是在 SparkConf 中对其进行硬编码，并在 Spark Shell 或 spark-submit 和 spark-sql 工具或 conf/spark-defaults.conf 中将其作为 --conf 参数进行传递。有关更多信息，请参阅 Apache Spark 文档中的 Spark 配置。

以下示例显示了如何在运行 spark-sql 命令的同时启用提交程序。

```
spark-sql \
  --conf spark.sql.parquet.fs.optimized.committer.optimization-enabled=true \
  -e "INSERT OVERWRITE TABLE target_table SELECT * FROM source_table;"
```

使用经 EMRFS S3 优化的提交协议

经 EMRFS S3 优化的提交协议是一种 FileCommitProtocol 实施替代项，该程序已经过优化，可在使用 EMRFS 时通过 Spark 动态分区覆盖将文件写入 Amazon S3。该协议可在 Spark 动态分区覆盖任务提交阶段避免 Amazon S3 中的重命名操作，从而提高应用程序性能。

请注意，使用经 EMRFS S3 优化的提交程序 也可以通过避免重命名操作来提高性能。但是，它不适用于动态分区覆盖情况，同时提交协议的改进仅针对动态分区覆盖情况。

提交协议适用于 Amazon EMR 发行版 5.30.0 及更高版本，在 Amazon EMR 6.2.0 及更高版本中将默认启用。从发行版 5.31.0 开始，Amazon EMR 增加了并行性改进。协议用于使用 Spark SQL、DataFrame 或 Dataset 的 Spark 任务。在某些情况下，不使用提交协议。有关更多信息，请参阅经 EMRFS S3 优化的提交协议的要求。

主题

- 经 EMRFS S3 优化的提交协议的要求
- 经 EMRFS S3 优化的提交协议和分段上传
- 任务优化注意事项

经 EMRFS S3 优化的提交协议的要求

满足以下条件时，将使用经 EMRFS S3 优化的提交协议：

- 您可以运行使用 Spark SQL、DataFrame 或 Dataset 的 Spark 任务来覆盖分区表。
- 您可以运行分区覆盖模式为 `dynamic` 的 Spark 任务。
- 分段上传在 Amazon EMR 中已启用。这是默认模式。有关更多信息，请参阅经 EMRFS S3 优化的提交协议和分段上传。
- EMRFS 的文件系统缓存已启用。这是默认模式。检查设置 `fs.s3.impl.disable.cache` 是否设置为 `false`。
- 使用 Spark 的内置数据来源支持。内置数据来源支持用于以下情况：
  - 当任务写入内置数据来源或表时。
  - 当任务写入 Hive 元存储 Parquet 表时。当 `spark.sql.hive.convertInsertingPartitionedTable` 和 `spark.sql.hive.convertMetastoreParquet` 都设置为 true 时，就会发生这种情况。这些是默认设置。
  - 当任务写入 Hive 元存储 ORC 表时。当 `spark.sql.hive.convertInsertingPartitionedTable` 和 `spark.sql.hive.convertMetastoreOrc` 都设置为 `true` 时，就会发生这种情况。这些是默认设置。
- 写入默认分区位置的 Spark 任务操作，例如 `${table_location}/k1=v1/k2=v2/`，使用提交协议。如果任务操作写入自定义分区位置，则不使用协议，例如，如果使用 `ALTER TABLE SQL` 命令设置自定义分区位置。
- 必须使用 Spark 的以下值：
  - `spark.sql.sources.commitProtocolClass` 必须设置为 `org.apache.spark.sql.execution.datasources.SQLEmrOptimizedCommitProtocol`。这是 Amazon EMR 发行版 5.30.0 及更高版本和 Amazon EMR 发行版 6.2.0 及更高版本的默认设置。
  - `partitionOverwriteMode` 写入选项或 `spark.sql.sources.partitionOverwriteMode` 必须设置为 `dynamic`。默认设置为 `static`。

> **Note**
>
> Spark 2.4.0 中引入了 `partitionOverwriteMode` 写入选项。对于随附了 Amazon EMR 版本 5.19.0 的 Spark 版本 2.3.2，请设置 `spark.sql.sources.partitionOverwriteMode` 属性。

- 如果 Spark 任务覆盖了 Hive 元存储 Parquet 表，则 `spark.sql.hive.convertMetastoreParquet`、`spark.sql.hive.convertInsertingPartit` 和 `spark.sql.hive.convertMetastore.partitionOverwriteMode` 必须设置为 `true`。这些是默认设置。
- 如果 Spark 任务覆盖了 Hive 元存储 ORC 表，则 `spark.sql.hive.convertMetastoreOrc`、`spark.sql.hive.convertInsertingPartitione` 和 `spark.sql.hive.convertMetastore.partitionOverwriteMode` 必须设置为 `true`。这些是默认设置。

Example – 动态分区覆盖模式

在此 Scala 示例中已触发优化。首先，将 `partitionOverwriteMode` 属性设置为 `dynamic`。这只会覆盖您正在写入数据的那些分区。然后，通过 `partitionBy` 指定动态分区列，并将写入模式设置为 `overwrite`。

```
val dataset = spark.range(0, 10)
  .withColumn("dt", expr("date_sub(current_date(), id)"))

dataset.write.mode("overwrite")                 // "overwrite" instead of "insert"
  .option("partitionOverwriteMode", "dynamic")  // "dynamic" instead of "static"
  .partitionBy("dt")                            // partitioned data instead of
 unpartitioned data
  .parquet("s3://EXAMPLE-DOC-BUCKET/output")    // "s3://" to use EMR file system,
 instead of "s3a://" or "hdfs://"
```

当不使用经 EMRFS S3 优化的提交协议时

通常，经 EMRFS S3 优化的提交协议的工作原理与开源默认 Spark SQL 提交协议 `org.apache.spark.sql.execution.datasources.SQLHadoopMapReduceCommitProtocol` 的工作原理相同。以下情况下不会进行优化。

情况	为什么不使用提交协议
当您向 HDFS 写入时	提交协议只支持使用 EMRFS 写入 Amazon S3。
当您使用 S3A 文件系统时	提交协议只支持 EMRFS。
当使用 MapReduce 或 Spark 的 RDD API 时	提交协议只支持使用 SparkSQL、DataFrame 或 Dataset API。
当没有触发动态分区覆盖时	提交协议仅优化动态分区覆盖情况。有关其他情况，请参阅 使用经 EMRFS S3 优化的提交程序。

以下 Scala 示例演示了经 EMRFS S3 优化的提交协议委托给 `SQLHadoopMapReduceCommitProtocol` 的一些其他情况。

Example – 具有自定义分区位置的动态分区覆盖模式

在此示例中，Scala 程序以动态分区覆盖模式覆盖两个分区。一个分区具有自定义分区位置。另一个分区使用默认分区位置。经 EMRFS S3 优化的提交协议仅改进了使用默认分区位置的分区。

```
val table = "dataset"
val inputView = "tempView"
val location = "s3://bucket/table"

spark.sql(s"""
  CREATE TABLE $table (id bigint, dt date)
  USING PARQUET PARTITIONED BY (dt)
  LOCATION '$location'
""")

// Add a partition using a custom location
val customPartitionLocation = "s3://bucket/custom"
spark.sql(s"""
  ALTER TABLE $table ADD PARTITION (dt='2019-01-28')
  LOCATION '$customPartitionLocation'
""")

// Add another partition using default location
spark.sql(s"ALTER TABLE $table ADD PARTITION (dt='2019-01-29')")
```

```
def asDate(text: String) = lit(text).cast("date")

spark.range(0, 10)
  .withColumn("dt",
    when($"id" > 4, asDate("2019-01-28")).otherwise(asDate("2019-01-29")))
  .createTempView(inputView)

// Set partition overwrite mode to 'dynamic'
spark.sql(s"SET spark.sql.sources.partitionOverwriteMode=dynamic")

spark.sql(s"INSERT OVERWRITE TABLE $table SELECT * FROM $inputView")
```

Scala 代码创建以下 Amazon S3 对象：

```
custom/part-00001-035a2a9c-4a09-4917-8819-e77134342402.c000.snappy.parquet
custom_$folder$
table/_SUCCESS
table/dt=2019-01-29/part-00000-035a2a9c-4a09-4917-8819-e77134342402.c000.snappy.parquet
table/dt=2019-01-29_$folder$
table_$folder$
```

> **Note**
>
> 在早期 Spark 版本中写入自定义分区位置可能会导致数据丢失。在此示例中，分区 `dt='2019-01-28'` 将丢失。有关详细信息，请参阅 SPARK-35106。此问题已在 Amazon EMR 发行版 5.33.0 及更高版本（6.0.x 和 6.1.x 除外）中得到修复。

当写入到自定义位置的分区时，Spark 会使用类似于上一个示例的提交算法，如下所述。与前面的示例一样，该算法会导致顺序重命名，这可能会影响性能。

Spark 2.4.0 中的算法遵循以下步骤：

1. 在将输出写入自定义位置的分区时，任务会写入到 Spark 的暂存目录下的文件中，该目录是在最终输出位置下创建的。该文件的名称包含一个随机 UUID，以防止文件冲突。任务尝试跟踪每个文件以及最终所需的输出路径。
2. 在任务成功完成后，它会为驱动程序提供这些文件及其最终所需的输出路径。
3. 所有任务都完成后，作业提交阶段会按顺序将为自定义位置的分区写入的所有文件重命名为其最终输出路径。

4. 暂存目录会在作业提交阶段完成之前删除。

经 EMRFS S3 优化的提交协议和分段上传

要在经 EMRFS S3 优化的提交协议中使用动态分区覆盖优化，必须在 Amazon EMR 中启用分段上传。默认启用分段上传。需要时，您可以重新启用它。有关更多信息，请参阅《Amazon EMR 管理指南》中的为 Amazon S3 配置分段上传。

在动态分区覆盖期间，经 EMRFS S3 优化的提交协议利用分段上传类似于事务的特征，来确保由任务写入的文件在任务提交后尝试仅显示在任务的输出位置。通过以这种方式使用分段上传，提交协议提高了默认 `SQLHadoopMapReduceCommitProtocol` 的任务提交性能。使用经 EMRFS S3 优化的提交协议时，需要考虑一些与传统分段上传行为的关键区别：

- 无论文件大小如何，分段上传都会执行。这不同于 EMRFS 的默认行为，其中 `fs.s3n.multipart.uploads.split.size` 属性在触发分段上传时，控制文件大小。
- 在任务提交或中止之前，分段上传在较长时间内都保持在未完成状态。这不同于 EMRFS 的默认行为，其中分段上传在任务完成写入给定文件时完成。

由于这些区别，如果 Spark Executor JVM 在任务正在运行或将数据写入 Amazon S3 时发生崩溃或被终止，或者 Spark Driver JVM 在任务正在运行时发生崩溃或被终止，未完成的分段上传更可能会被留下来。因此，当您使用经 EMRFS S3 优化的提交协议时，请务必遵循管理失败的分段上传的最佳实践。有关更多信息，请参阅《Amazon EMR 管理指南》中使用 Amazon S3 存储桶的最佳实践。

任务优化注意事项

在 Spark 执行程序上，经 EMRFS S3 优化的提交协议会占用少量内存来存储任务尝试写入的每个文件，直到任务提交或中止。在大多数作业中，占用的内存量可以忽略不计。

在 Spark 驱动程序上，经 EMRFS S3 优化的提交协议需要内存来存储每个已提交文件的元数据信息，直到任务提交或中止。在大多数任务中，默认的 Spark 驱动程序内存设置可以忽略不计。

对于包含写入大量文件的长期任务的作业，提交协议占用的内存可能很大，需要调整分配给 Spark 的内存，尤其是 Spark 执行程序。您可以使用 Spark 驱动程序的 `spark.driver.memory` 属性和 `spark.executor.memory` 属性来优化内存。作为指导，编写 100000 个文件的单个任务通常需要额外的 100 MB 内存。有关更多信息，请参阅 Apache Spark 配置文档中的应用程序属性。

使用 EMRFS 重试 Amazon S3 请求

本主题提供有关使用 EMRFS 向 Amazon S3 发出请求时可以使用的重试策略的信息。当您的请求速率提高时，S3 会尝试扩展以支持新的速率。在此过程中，S3 可以限制请求并返回 `503 Slow Down` 错误。为了提高 S3 请求的成功率，您可以通过在 `emrfs-site` 配置中配置属性以调整重试策略。

您可以通过以下方法调整重试策略。

- 提高默认指数退避重试策略的最大重试限制。
- 启用和配置加性增长/加速递减 (AIMD) 重试策略。Amazon EMR 发行版 6.4.0 及更高版本支持 AIMD。

使用默认的指数退避策略

默认情况下，EMRFS 使用指数退避策略来重试 Amazon S3 请求。默认 EMRFS 重试限制为 15。为避免 S3 `503 Slow Down` 错误，您可以在创建新集群时、在正在运行的群集上或应用程序运行时提高重试限制。

要提高重试限制，您必须在您的 `emrfs-site` 配置中更改 `fs.s3.maxRetries` 的值。以下示例配置将 `fs.s3.maxRetries` 设置为自定义值 30。

```
[
    {
      "Classification": "emrfs-site",
      "Properties": {
        "fs.s3.maxRetries": "30"
      }
    }
]
```

有关使用配置对象的更多信息，请参阅 配置应用程序。

使用 AIMD 重试策略

在 Amazon EMR 发行版 6.4.0 及更高版本中，EMRFS 支持基于加性增长/加速递减（AIMD）模型的替代重试策略。当您使用大型 Amazon EMR 集群时，AIMD 重试策略尤其有用。

AIMD 使用有关最近成功请求的数据计算自定义请求率。此策略减少了受限请求的数量和每个请求所需的总尝试次数。

要启用 AIMD 重试策略，必须在您的 emrfs-site 配置中将 fs.s3.aimd.enabled 属性设置为 true，如以下示例所示。

```
[
    {
      "Classification": "emrfs-site",
      "Properties": {
        "fs.s3.aimd.enabled": "true"
      }
    }
]
```

有关使用配置对象的更多信息，请参阅 配置应用程序。

高级 AIMD 重试设置

您可以配置下表中列出的属性，以便在使用 AIMD 重试策略时优化重试行为。对于大多数使用案例，我们建议您使用默认值。

高级 AIMD 重试策略属性

属性	默认值	描述
fs.s3.aimd.increaseIncrement	0.1	控制连续请求成功时请求速率的增长速度。
fs.s3.aimd.reductionFactor	2	控制 Amazon S3 返回 503 响应时请求速率降低的速度。默认因子 2 将请求率降低一半。
fs.s3.aimd.minRate	0.1	设置请求经历 S3 持续限制时的请求速率的下限。
fs.s3.aimd.initialRate	5500	设置初始请求速率，然后该速率将根据您为 fs.s3.aimd.increaseIncrement 和 fs.s3.aimd.reductionFactor 指定的值变化。

属性	默认值	描述
		初始速率也用于 GET 请求，并针对 PUT 请求按比例 (3500/5500) 扩展。
`fs.s3.aimd.adjustWindow`	2	控制调整请求速率的频率，以响应数量衡量。
`fs.s3.aimd.maxAttempts`	100	设置尝试请求的最大尝试次数。

添加 Spark 步骤

您可以使用 Amazon EMR 步骤向安装在 EMR 集群上的 Spark 框架提交工作。有关更多信息，请参阅《Amazon EMR 管理指南》中的步骤。在控制台和 CLI 中，您使用 Spark 应用程序步骤 (代表您将 `spark-submit` 脚本作为步骤运行) 来完成此操作。借助 API，您通过 `spark-submit` 使用步骤调用 `command-runner.jar`。

有关向 Spark 提交应用程序的更多信息，请参阅 Apache Spark 文档中的提交应用程序主题。

使用控制台提交 Spark 步骤

1. 通过以下链接打开 Amazon EMR 控制台：https://console.aws.amazon.com/emr。
2. 在 Cluster List (集群列表) 中，选择您的集群的名称。
3. 滚动到 Steps (步骤) 部分并展开它，然后选择 Add step (添加步骤)。
4. 在 Add Step (添加步骤) 对话框中：

   - 对于 Step type (步骤类型)，选择 Spark application (Spark 应用程序)。
   - 对于 Name（名称），接受原定设置名称（Spark application）或键入一个新名称。
   - 对于 Deploy mode (部署模式)，选择 Client (客户端) 或 Cluster (集群) 模式。客户端模式在集群的主实例上启动驱动程序，而集群模式在集群上启动驱动程序。对于客户端模式，驱动程序的日志输出将显示在步骤日志中，而对于集群模式，驱动程序的日志输出将显示在第一个 YARN 容器的日志中。有关更多信息，请参阅 Apache Spark 文档中的集群模式概览。
   - 指定所需的 Spark-submit options。有关 `spark-submit` 选项的更多信息，请参阅使用 spark-submit 启动应用程序。
   - 对于 Application location (应用程序位置)，指定应用程序的本地或 S3 URI 路径。

- 对于 Arguments (参数)，将该字段保留为空白。
- 对于 Action on failure (出现故障时的操作)，接受默认选项 Continue (继续)。

5. 选择 Add（添加）。步骤会出现在控制台中，其状态为“Pending”。
6. 步骤的状态会随着步骤的运行从“Pending”变为“Running”，再变为“Completed”。要更新状态，请选择 Actions (操作) 列上方的 Refresh (刷新) 图标。
7. 如果您配置了日志记录，则这一步的结果将放在 Amazon EMR 控制台的 Cluster Details (集群详细信息) 页面上，位于您的步骤旁边的 Log Files (日志文件) 下方。在启动集群时，您可以选择在配置的日志存储桶中查找步骤信息。

使用 Amazon CLI 向 Spark 提交工作

在创建集群时提交步骤，或使用 `aws emr add-steps` 子命令在现有集群中提交步骤。

1. 使用 `create-cluster`，如以下示例所示。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Add Spark Step Cluster" --release-label emr-5.36.1
 --applications Name=Spark \
--ec2-attributes KeyName=myKey --instance-type m5.xlarge --instance-count 3 \
--steps Type=Spark,Name="Spark Program",ActionOnFailure=CONTINUE,Args=[--
class,org.apache.spark.examples.SparkPi,/usr/lib/spark/examples/jars/spark-
examples.jar,10] --use-default-roles
```

作为替代方法，您可使用 `command-runner.jar`，如以下示例所示。

```
aws emr create-cluster --name "Add Spark Step Cluster" --release-label emr-5.36.1 \
--applications Name=Spark --ec2-attributes KeyName=myKey --instance-type m5.xlarge
 --instance-count 3 \
--steps Type=CUSTOM_JAR,Name="Spark Program",Jar="command-
runner.jar",ActionOnFailure=CONTINUE,Args=[spark-example,SparkPi,10] --use-default-
roles
```

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

2. 或者，向正在运行的集群添加步骤。使用 `add-steps`。

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF --steps
 Type=Spark,Name="Spark Program",ActionOnFailure=CONTINUE,Args=[--
class,org.apache.spark.examples.SparkPi,/usr/lib/spark/examples/jars/spark-
examples.jar,10]
```

作为替代方法，您可使用 `command-runner.jar`，如以下示例所示。

```
aws emr add-steps --cluster-id j-2AXXXXXXGAPLF --steps Type=CUSTOM_JAR,Name="Spark
 Program",Jar="command-runner.jar",ActionOnFailure=CONTINUE,Args=[spark-
example,SparkPi,10]
```

使用 SDK for Java 向 Spark 提交工作

1. 下面的示例显示如何通过 Java 添加步骤至带有 Spark 的集群：

```
AWSCredentials credentials = new BasicAWSCredentials(accessKey, secretKey);
AmazonElasticMapReduce emr = new AmazonElasticMapReduceClient(credentials);

StepFactory stepFactory = new StepFactory();
AmazonElasticMapReduceClient emr = new AmazonElasticMapReduceClient(credentials);
AddJobFlowStepsRequest req = new AddJobFlowStepsRequest();
req.withJobFlowId("j-1K48XXXXXXHCB");

List<StepConfig> stepConfigs = new ArrayList<StepConfig>();

HadoopJarStepConfig sparkStepConf = new HadoopJarStepConfig()
   .withJar("command-runner.jar")
   .withArgs("spark-submit","--executor-memory","1g","--
class","org.apache.spark.examples.SparkPi","/usr/lib/spark/examples/jars/spark-
examples.jar","10");

StepConfig sparkStep = new StepConfig()
```

```
    .withName("Spark Step")
    .withActionOnFailure("CONTINUE")
    .withHadoopJarStep(sparkStepConf);

stepConfigs.add(sparkStep);
req.withSteps(stepConfigs);
AddJobFlowStepsResult result = emr.addJobFlowSteps(req);
```

2. 您可以通过查看该步骤的日志来了解该步骤的结果。如果您已启用日志记录，则可通过以下方式在Amazon Web Services Management Console中执行此操作：选择 Steps (步骤)，选择您的步骤，然后为 Log files (日志文件) 选择 `stdout` 或 `stderr`。要查看可用日志，请选择 View Logs (查看日志)。

覆盖 Spark 默认配置设置

建议您为不同的应用程序覆盖 Spark 默认配置值。您可以在提交应用程序时使用步骤完成此操作 (实质上是向 `spark-submit` 传递选项)。例如，您可能需要通过更改 `spark.executor.memory` 来更改为执行者进程分配的内存。您可以为 `--executor-memory` 开关提供类似下面的参数：

```
spark-submit --executor-memory 1g --class org.apache.spark.examples.SparkPi /usr/lib/
spark/examples/jars/spark-examples.jar 10
```

同样地，您也可以调节 `--executor-cores` 和 `--driver-memory`。在步骤中，您可以向步骤提供以下参数：

```
--executor-memory 1g --class org.apache.spark.examples.SparkPi /usr/lib/spark/examples/
jars/spark-examples.jar 10
```

您还可以使用 `--conf` 选项调节没有内置开关的设置。有关可调节的其他设置的更多信息，请参阅 Apache Spark 文档中的动态加载 Spark 属性主题。

查看 Spark 应用程序历史记录

您可以在控制台中，使用集群详细信息页面上的 Application user interfaces (应用程序用户界面) 选项卡查看 Spark、YARN 应用程序和 Tez UI 详细信息。通过使用 Amazon EMR 应用程序用户界面（UI），您可以更轻松地对活动的作业和任务历史记录进行故障排查和分析。

有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。

访问 Spark Web UI

您可以查看 Spark Web UI，具体做法是：按照操作步骤在 Amazon EMR 管理指南中一个名为 Connect to the cluster (连接到集群) 的部分中创建 SSH 隧道或创建代理，然后导航到集群的 YARN ResourceManager。在 Tracking UI (跟踪 UI) 下选择适合您的应用程序的链接。如果您的应用程序正在运行，您将看到 ApplicationMaster。这会将您带到主应用程序的 Web UI，在端口 20888 上，无论驱动程序位于何处都是如此。如果您在 YARN 客户端模式下运行，则驱动程序可能会位于集群的主节点上。如果您在 YARN 集群模式下运行应用程序，则驱动程序会位于集群上该应用程序的 ApplicationMaster 中。如果您的应用程序已完成，您将看到 History（历史记录），它会将您带到 Spark HistoryServer UI 中 EMR 集群的主节点，端口号为 18080。这适用于已经完成的应用程序。您还可以直接导航到 Spark HistoryServer UI，网址为 http://*master-public-dns-name*:18080/。

在 Amazon EMR 发行版 5.25.0 或更高版本中，可以从控制台访问 Spark 历史记录服务器 UI，而无需通过 SSH 连接设置 Web 代理。有关更多信息，请参阅查看持久性应用程序用户界面。

将适用于 Apache Spark 的 Amazon Redshift 集成与 Amazon EMR 结合使用

在 Amazon EMR 发行版 6.4.0 及更高版本中，每个版本的映像都包含 Apache Spark 和 Amazon Redshift 之间的连接器。通过该连接器，您可以在 Amazon EMR 上使用 Spark 处理存储在 Amazon Redshift 中的数据。对于 Amazon EMR 发行版 6.4.0 至 6.8.0，集成基于 `spark-redshift` 开源连接器。对于 Amazon EMR 发行版 6.9.0 及更高版本，适用于 Apache Spark 的 Amazon Redshift 集成已从社区版本迁移到本地集成。

主题

- 使用适用于 Apache Spark 的 Amazon Redshift 集成启动 Spark 应用程序
- 使用适用于 Apache Spark 的 Amazon Redshift 集成进行身份验证
- 在 Amazon Redshift 中进行读取和写入
- 使用 Spark 连接器时的注意事项和限制

使用适用于 Apache Spark 的 Amazon Redshift 集成启动 Spark 应用程序

对于 Amazon EMR 版本 6.4 至 6.9，您必须使用 `--jars` 或 `--packages` 选项来指定要使用以下哪个 JAR 文件。`--jars` 选项指定存储在本地、HDFS 中或使用 HTTP/S 的依赖项。要查看 `--jars` 选项支持的其他文件位置，请参阅 Spark 文档中的高级依赖管理。`--packages` 选项指定存储在公共 Maven 存储库中的依赖项。

- spark-redshift.jar
- spark-avro.jar
- RedshiftJDBC.jar
- minimal-json.jar

Amazon EMR 6.10.0 及更高版本不需要 minimal-json.jar 依赖关系，并且默认情况下会自动将其他依赖项安装到每个集群。以下示例显示了如何使用适用于 Apache Spark 的 Amazon Redshift 集成启动 Spark 应用程序。

Amazon EMR 6.10.0 +

以下示例显示了如何在 Amazon EMR 版本 6.10 和更高版本上使用 spark-redshift 连接器启动 Spark 应用程序。

```
spark-submit my_script.py
```

Amazon EMR 6.4.0 - 6.9.x

要在 Amazon EMR 版本 6.4 至 6.9 上通过 spark-redshift 连接器启动 Spark 应用程序，必须使用 --jars 或 --packages 选项，如以下例所示。请注意，--jars 选项列出的路径是 JAR 文件的默认路径。

```
spark-submit \
  --jars /usr/share/aws/redshift/jdbc/RedshiftJDBC.jar,/usr/share/aws/redshift/
spark-redshift/lib/spark-redshift.jar,/usr/share/aws/redshift/spark-redshift/lib/
spark-avro.jar,/usr/share/aws/redshift/spark-redshift/lib/minimal-json.jar \
  my_script.py
```

使用适用于 Apache Spark 的 Amazon Redshift 集成进行身份验证

使用 Amazon Secrets Manager 检索凭证并连接到 Amazon Redshift

以下代码示例显示了如何使用 Amazon Secrets Manager 检索凭证，以通过 Python 中 Apache Spark 的 PySpark 接口连接到 Amazon Redshift 集群。

```
from pyspark.sql import SQLContext
import boto3
```

```
sc = # existing SparkContext
sql_context = SQLContext(sc)

secretsmanager_client = boto3.client('secretsmanager')
secret_manager_response = secretsmanager_client.get_secret_value(
    SecretId='string',
    VersionId='string',
    VersionStage='string'
)
username = # get username from secret_manager_response
password = # get password from secret_manager_response
url = "jdbc:redshift://redshifthost:5439/database?user=" + username + "&password=" +
 password

# Read data from a table
df = sql_context.read \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("dbtable", "my_table") \
    .option("tempdir", "s3://path/for/temp/data") \
    .load()
```

使用 IAM 检索凭证并连接到 Amazon Redshift

您可以使用 Amazon Redshift 提供的 JDBC 版本 2 驱动程序，通过 Spark 连接器连接到 Amazon Redshift。要使用 Amazon Identity and Access Management（IAM），请将您的 JDBC URL 配置为使用 IAM 身份验证。要从 Amazon EMR 连接到 Redshift 集群，您必须授予 IAM 角色权限以检索临时 IAM 凭证。将以下权限分配到 IAM 角色，使其能够检索凭证并运行 Amazon S3 操作。

- Redshift:GetClusterCredentials（适用于预置的 Redshift 集群）
- Redshift:DescribeClusters（适用于预置的 Redshift 集群）
- Redshift:GetWorkgroup（适用于 Amazon Redshift Serverless 工作组）
- Redshift:GetCredentials（适用于 Amazon Redshift Serverless 工作组）
- s3:GetBucket
- s3:GetBucketLocation
- s3:GetObject
- s3:PutObject
- s3:GetBucketLifecycleConfiguration

有关 GetClusterCredentials 的更多信息，请参阅 GetClusterCredentials 的资源策略。

您还必须确保 Amazon Redshift 可以在 COPY 和 UNLOAD 操作期间担任 IAM 角色。

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
                "Service": "redshift.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
    ]
}
```

以下示例在 Spark 和 Amazon Redshift 之间使用 IAM 身份验证：

```
from pyspark.sql import SQLContext
import boto3

sc = # existing SparkContext
sql_context = SQLContext(sc)

url = "jdbc:redshift:iam//redshift-host:redshift-port/db-name"
iam_role_arn = "arn:aws:iam::account-id:role/role-name"

# Read data from a table
df = sql_context.read \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("aws_iam_role", iam_role_arn) \
    .option("dbtable", "my_table") \
    .option("tempdir", "s3a://path/for/temp/data") \
    .mode("error") \
    .load()
```

在 Amazon Redshift 中进行读取和写入

以下代码示例使用 PySpark，通过数据来源 API 和 SparkSQL 从 Amazon Redshift 数据库读取和写入示例数据。

Data source API

使用 PySpark，通过数据来源 API 从 Amazon Redshift 数据库读取和写入示例数据。

```
import boto3
from pyspark.sql import SQLContext

sc = # existing SparkContext
sql_context = SQLContext(sc)

url = "jdbc:redshift:iam://redshifthost:5439/database"
aws_iam_role_arn = "arn:aws:iam::accountID:role/roleName"

df = sql_context.read \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("dbtable", "tableName") \
    .option("tempdir", "s3://path/for/temp/data") \
    .option("aws_iam_role", "aws_iam_role_arn") \
    .load()

df.write \
    .format("io.github.spark_redshift_community.spark.redshift") \
    .option("url", url) \
    .option("dbtable", "tableName_copy") \
    .option("tempdir", "s3://path/for/temp/data") \
    .option("aws_iam_role", "aws_iam_role_arn") \
    .mode("error") \
    .save()
```

SparkSQL

使用 PySpark，通过 SparkSQL 在 Amazon Redshift 数据库中读取和写入示例数据。

```
import boto3
import json
import sys
import os
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .enableHiveSupport() \
    .getOrCreate()
```

```
url = "jdbc:redshift:iam://redshifthost:5439/database"
aws_iam_role_arn = "arn:aws:iam::accountID:role/roleName"

bucket = "s3://path/for/temp/data"
tableName = "tableName" # Redshift table name

s = f"""CREATE TABLE IF NOT EXISTS {tableName} (country string, data string)
    USING io.github.spark_redshift_community.spark.redshift
    OPTIONS (dbtable '{tableName}', tempdir '{bucket}', url '{url}', aws_iam_role
 '{aws_iam_role_arn}' ); """

spark.sql(s)

columns = ["country" ,"data"]
data = [("test-country","test-data")]
df = spark.sparkContext.parallelize(data).toDF(columns)

# Insert data into table
df.write.insertInto(tableName, overwrite=False)
df = spark.sql(f"SELECT * FROM {tableName}")
df.show()
```

使用 Spark 连接器时的注意事项和限制

- 建议您为从 Spark on Amazon EMR 到 Amazon Redshift 的 JDBC 连接启用 SSL。
- 作为最佳实践，建议您在 Amazon Secrets Manager 中管理 Amazon Redshift 集群的凭证。有关示例，请参阅使用 Amazon Secrets Manager 检索用于连接到 Amazon Redshift 的凭证。
- 建议使用参数 `aws_iam_role` 为 Amazon Redshift 身份验证参数传递 IAM 角色。
- 参数 `tempformat` 目前不支持 Parquet 格式。
- `tempdir` URI 指向 Amazon S3 位置。此临时目录不会自动清理，因此可能会增加额外成本。
- 请考虑以下针对 Amazon Redshift 的建议：
  - 建议阻止对 Amazon Redshift 集群的公有访问。
  - 建议启用 Amazon Redshift 审计日志记录。
  - 建议启用 Amazon Redshift 静态加密。
- 请考虑以下针对 Amazon S3 的建议：
  - 建议阻止对 Amazon S3 存储桶的公有访问。

- 建议使用 Amazon S3 服务器端加密以加密使用的 Amazon S3 存储桶。
- 建议使用 Amazon S3 生命周期策略定义 Amazon S3 存储桶的保留规则。
- Amazon EMR 始终验证从开源导入到映像中的代码。出于安全原因，我们不支持从 Spark 到 Amazon S3 的以下身份验证方法：
  - 在 `hadoop-env` 配置分类中设置 Amazon 访问密钥
  - 在 `tempdir` URI 中编码 Amazon 访问密钥

有关使用连接器及其支持参数的更多信息，请参阅以下资源：

- Amazon Redshift Management Guide（《Amazon Redshift 管理指南》）中的 Amazon Redshift integration for Apache Spark（适用于 Apache Spark 的 Amazon Redshift 集成）
- Github 上的 `spark-redshift` 社区存储库

Spark 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Spark 的版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

> **Important**
>
> Apache Spark 版本 2.3.1（从 Amazon EMR 发行版 5.16.0 开始提供）解决了 CVE-2018-8024 和 CVE-2018-1334 问题。建议您将 Spark 的早期版本迁移到 Spark 2.3.1 版本或更高版本。

Spark 版本信息

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.14.0	3.4.1	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode,

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
		hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.13.0	3.4.1	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.12.0	3.4.0	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.11.1	3.3.2	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.11.0	3.3.2	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.10.1	3.3.1	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.10.0	3.3.1	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.9.1	3.3.0	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.9.0	3.3.0	aws-sagemaker-spark-sdk, delta, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.8.1	3.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.8.0	3.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.7.0	3.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.36.1	2.4.8	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.36.0	2.4.8	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.6.0	3.2.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.35.0	2.4.8	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.5.0	3.1.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, iceberg, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.4.0	3.1.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.3.1	3.1.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.3.0	3.1.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.2.1	3.0.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.2.0	3.0.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.1.1	3.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.1.0	3.0.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-6.0.1	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-6.0.0	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.34.0	2.4.8	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.33.1	2.4.7	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.33.0	2.4.7	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.32.1	2.4.7	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.32.0	2.4.7	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.31.1	2.4.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.31.0	2.4.6	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.30.2	2.4.5	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.30.1	2.4.5	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.30.0	2.4.5	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-notebook-env, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.29.0	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.28.1	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.28.0	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.27.1	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.27.0	2.4.4	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.26.0	2.4.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.25.0	2.4.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.24.1	2.4.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.24.0	2.4.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.23.1	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.23.0	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.22.0	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.21.2	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.21.1	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.21.0	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.20.1	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.20.0	2.4.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.19.1	2.3.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.19.0	2.3.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.18.1	2.3.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.18.0	2.3.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, nginx, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.17.2	2.3.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.17.1	2.3.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.17.0	2.3.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, emr-s3-select, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.16.1	2.3.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.16.0	2.3.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.15.1	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.15.0	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.14.2	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.14.1	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.14.0	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.13.1	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.13.0	2.3.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.12.3	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.12.2	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.12.1	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.12.0	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.11.4	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.11.3	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.11.2	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.11.1	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.11.0	2.2.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.10.1	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.10.0	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.9.1	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.9.0	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.8.3	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.8.2	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.8.1	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.8.0	2.2.0	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.7.1	2.1.1	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.7.0	2.1.1	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.6.1	2.1.1	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.6.0	2.1.1	emrfs, emr-goodies, emr-ddb, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.5.4	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.5.3	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.5.2	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.5.1	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.5.0	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.4.1	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.4.0	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.3.2	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.3.1	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.3.0	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.2.3	2.0.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.2.2	2.0.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.2.1	2.0.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.2.0	2.0.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.1.1	2.0.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-5.1.0	2.0.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.0.3	2.0.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-5.0.0	2.0.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.9.6	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.9.5	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.9.4	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.9.3	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.9.2	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.9.1	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.8.5	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.8.4	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.8.3	1.6.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.8.2	1.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.8.0	1.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.7.4	1.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.7.2	1.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.7.1	1.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.7.0	1.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.6.0	1.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.5.0	1.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.4.0	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.3.0	1.6.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.2.0	1.5.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave
emr-4.1.0	1.5.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Amazon EMR 发行版标签	Spark 版本	随 Spark 安装的组件
emr-4.0.0	1.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave

Apache Sqoop

Apache Sqoop 是用于在 Amazon S3、Hadoop、HDFS 和 RDBMS 数据库之间传输数据的工具。有关更多信息，请参阅 Apache Sqoop 网站。Amazon EMR 发行版 5.0.0 及更高版本包含 Sqoop。早期发行版包含包含 Sqoop，将其用作沙盒应用程序。有关更多信息，请参阅Amazon EMR 4.x 发行版。

下表列出了 Amazon EMR 6.x 系列的最新发行版本附带的 Sqoop 的版本，以及 Amazon EMR 随 Sqoop 一起安装的组件。

有关此发行版中随 Sqoop 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Sqoop 版本信息

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.14.0	Sqoop 1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

下表列出了 Amazon EMR 5.x 系列的最新发行版本附带的 Sqoop 的版本，以及 Amazon EMR 随 Sqoop 一起安装的组件。

有关此发行版中随 Sqoop 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Sqoop 版本信息

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.36.1	Sqoop 1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode,

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
		hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

主题

Amazon EMR 上的 Sqoop 注意事项

在 Amazon EMR 上运行 Sqoop 时，请考虑以下项目。

使用 Sqoop 与 HCatalog 集成

Amazon EMR 上的 Sqoop 支持 Sqoop-HCatalog 集成。在使用 Sqoop 将输出写入到 Amazon S3 中 HCatalog 表中时，可以通过设置 `mapred.output.direct.NativeS3FileSystem` 和将 `mapred.output.direct.EmrFileSystem` 属性设置为 `false` 来禁用 Amazon EMR 直接写入。有关更多信息，请参阅使用 HCatalog。您可使用 Hadoop `-D mapred.output.direct.NativeS3FileSystem=false` 和 `-D mapred.output.direct.EmrFileSystem=false` 命令。如果您未禁用直接写入，则不会发生错误，但在 Amazon S3 中创建表时不会写入任何数据。

Sqoop JDBC 和数据库支持

默认情况下，Sqoop 已安装 MariaDB 和 PostgreSQL 驱动程序。为 Sqoop 安装的 PostgreSQL 驱动程序仅适用于 PostgreSQL 8.4。要安装 Sqoop 的备用 JDBC 连接器集，请连接到集群的主节点并将这些连接器安装在 `/usr/lib/sqoop/lib` 中。下面是各个 JDBC 连接器的链接：

- MariaDB：关于 MariaDB Connector/J。
- PostgreSQL：PostgreSQL JDBC 驱动程序。

- SQLServer：下载适用于 SQL Server 的 Microsoft JDBC 驱动程序。
- MySQL：下载 Connector/J
- Oracle：从 Oracle Maven 存储库获取 Oracle JDBC 驱动程序和 UCP

Sqoop 支持的数据库列在以下 URL 中：http://sqoop.apache.org/docs/*version*/SqoopUserGuide.html#_supported_databases，其中的 *version* 是您正在使用的 Sqoop 的版本，例如 1.4.6。如果 JDBC 连接字符串与此列表中的不匹配，您必须指定驱动程序。

例如，您可使用以下命令导出到 Amazon Redshift 数据库表（适用于 JDBC 4.1）：

```
sqoop export --connect jdbc:redshift://$MYREDSHIFTHOST:5439/mydb --table mysqoopexport
 --export-dir s3://mybucket/myinputfiles/ --driver com.amazon.redshift.jdbc41.Driver --
username master --password Mymasterpass1
```

您可同时使用 MariaDB 和 MySQL 连接字符串，但如果您指定 MariaDB 连接字符串，则需要指定驱动程序：

```
sqoop export --connect jdbc:mariadb://$HOSTNAME:3306/mydb --table mysqoopexport
 --export-dir s3://mybucket/myinputfiles/ --driver org.mariadb.jdbc.Driver --
username master --password Mymasterpass1
```

如果您使用安全套接字层加密来访问您的数据库，则需要使用与以下 Sqoop 导出示例中类似的 JDBC URI：

```
sqoop export --connect jdbc:mariadb://$HOSTNAME:3306/mydb?
verifyServerCertificate=false&useSSL=true&requireSSL=true --table mysqoopexport
 --export-dir s3://mybucket/myinputfiles/ --driver org.mariadb.jdbc.Driver --
username master --password Mymasterpass1
```

如需详细了解有关在 RDS 中使用 SSL 加密功能，请参阅《Amazon RDS 用户指南》中的使用 SSL 来加密与数据库实例的连接。

有关更多信息，请参阅 Apache Sqoop 文档。

Sqoop 发行历史记录

下表列出了 Amazon EMR 的每个发行版本中所包含的 Sqoop 的版本，以及在安装应用程序时一同安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Sqoop 版本信息

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.14.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.13.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.12.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
		timeline-server, mariadb-server, sqoop-client
emr-6.11.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.11.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.10.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.10.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.9.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.9.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.8.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.8.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.7.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.36.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.36.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.6.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.35.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.5.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.4.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.3.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.3.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.2.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.2.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-6.1.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-6.1.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.34.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.33.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.33.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.32.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.32.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.31.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.31.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.30.2	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.30.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.30.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mariadb-server, sqoop-client
emr-5.29.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.28.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.28.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.27.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.27.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.26.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.25.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.24.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.24.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.23.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.23.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.22.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.21.2	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.21.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.21.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.20.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.20.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.19.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.19.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.18.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.18.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.17.2	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.17.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.17.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.16.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.16.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.15.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.15.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.14.2	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.14.1	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.14.0	1.4.7	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.13.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.13.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.12.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.12.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.12.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.12.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.11.4	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.11.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.11.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.11.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.11.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.10.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.10.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.9.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.9.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.8.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.8.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.8.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.8.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.7.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.7.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.6.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.6.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, mysql-server, sqoop-client
emr-5.5.4	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.5.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.5.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.5.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.5.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.4.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.4.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.3.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.3.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.3.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.2.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.2.2	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.2.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.2.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.1.1	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.1.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

Amazon EMR 发行版标签	Sqoop 版本	随 Sqoop 安装的组件
emr-5.0.3	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client
emr-5.0.0	1.4.6	emrfs, emr-ddb, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, mysql-server, sqoop-client

TensorFlow

TensorFlow 是一种开源符号数学库，用于机器智能和深度学习应用程序。有关更多信息，请参阅 TensorFlow 网站。TensorFlow 在 Amazon EMR 发行版 5.17.0 及更高版本中提供。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 TensorFlow 版本，以及 Amazon EMR 随 TensorFlow 一起安装的组件。

有关此发行版中随 TensorFlow 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 TensorFlow 版本信息

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.14.0	TensorFlow 2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 TensorFlow 版本，以及 Amazon EMR 随 TensorFlow 一起安装的组件。

有关此发行版中随 TensorFlow 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 TensorFlow 版本信息

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.36.1	TensorFlow 2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodema

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
		nager, hadoop-yarn-resour cemanager, hadoop-yarn-timeline-server, tensorflow

使用的 TensorFlow 版本因 Amazon EC2 实例类型而异

Amazon EMR 使用的 TensorFlow 库版本因您为集群选择的实例类型而异。下表按实例类型列出版本。

EC2 实例类型	TensorFlow 版本
M5 和 C5	利用英特尔 MKL 进行优化的 TensorFlow 1.9.0
P2	采用 CUDA 9.2、cuDNN 7.1 的 Tensorflow 1.9.0
P3	采用 CUDA 9.2、cuDNN 7.1、NCCL 2.2.13 的 Tensorflow 1.9.0 NVIDIA NCCL 仅适用于 P3 实例。最终用户许可协议（EULA）：在 Amazon EMR 上使用 Nvidia 组件，即表示您同意产品 EULA 中列出的条款和条件。
所有其他地址	Tensorflow 1.9.0

安全性

除了遵循安全使用 TensorFlow 指南外，还建议您在私有子网中启动集群，以帮助您限制对受信任源的访问。有关更多信息，请参阅《Amazon EMR 管理指南》中的 Amazon VPC 选项。

使用 TensorBoard

TensorBoard 是一套用于 TensorFlow 程序的可视化工具。有关更多信息，请参阅 Tensorflow 网站上的 Tensorflow：可视化学习。

要将 TensorBoard 与 Amazon EMR 结合使用，您必须在集群主节点上启动 TensorBoard。

在 Amazon EMR 上将 Tensorboard 与 Tensorflow 结合使用

1. 使用 SSH 连接到集群的主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 键入以下命令，在主节点上启动 TensorFboard。将 */my/log/directory* 替换为您使用摘要写入器生成和存储摘要数据的主节点上的目录。

   Amazon EMR 5.19.0 and later

   ```
   python3 -m tensorboard.main --logdir=/home/hadoop/tensor --bind_all
   ```

   Amazon EMR 5.18.1 and earlier

   ```
   python3 -m tensorboard.main --logdir=/my/log/dir
   ```

   默认情况下，主节点使用端口 6006 和主公共 DNS 名称托管 TensorBoard。启动 TensorBoard 后，命令行输出将显示可用于连接到 TensorBoard 的 URL，如以下示例所示：

   ```
   TensorBoard 1.9.0 at http://master-public-dns-name:6006 (Press CTRL+C to quit)
   ```
3. 设置来自受信任客户端对主节点上 Web 界面的访问权限。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看 Amazon EMR 集群上托管的 Web 界面。
4. 打开 TensorBoard（`http://master-public-dns-name:6006`）。

TensorFlow 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 TensorFlow 的版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

TensorFlow 版本信息

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.14.0	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode,

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
		hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.13.0	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.12.0	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.11.1	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.11.0	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.10.1	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.10.0	2.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.9.1	2.10.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.9.0	2.10.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.8.1	2.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.8.0	2.9.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.7.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.36.1	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.36.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.6.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.35.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.5.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.4.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.3.1	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.3.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.2.1	2.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.2.0	2.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.1.1	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.1.0	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-6.0.1	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-6.0.0	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.34.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.33.1	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.33.0	2.4.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.32.1	2.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.32.0	2.3.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.31.1	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.31.0	2.1.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.30.2	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.30.1	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.30.0	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.29.0	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.28.1	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.28.0	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.27.1	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.27.0	1.14.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.26.0	1.13.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.25.0	1.13.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.24.1	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.24.0	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.23.1	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.23.0	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.22.0	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.21.2	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.21.1	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.21.0	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.20.1	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.20.0	1.12.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.19.1	1.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.19.0	1.11.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.18.1	1.9.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.18.0	1.9.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Amazon EMR 发行版标签	TensorFlow 版本	随 TensorFlow 安装的组件
emr-5.17.2	1.9.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.17.1	1.9.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow
emr-5.17.0	1.9.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tensorflow

Apache Tez

Apache Tez 是一个框架，用于创建用于处理数据的复杂有向无环图（DAG）。在某些用例中，您可以将其用作 Hadoop MapReduce 的替代方案。例如，您可以将 Pig 和 Hive 工作流与 Hadoop MapReduce 一起运行，也可以使用 Tez 作为执行引擎。有关更多信息，请参阅 https://tez.apache.org/。Amazon EMR 4.7.0 及更高版本包括 Tez。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Tez 的版本，以及 Amazon EMR 随 Tez 一起安装的组件。

有关此发行版中随 Tez 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Tez 版本信息

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.14.0	Tez 0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Tez 的版本，以及 Amazon EMR 随 Tez 一起安装的组件。

有关此发行版中随 Tez 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Tez 版本信息

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.36.1	Tez 0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
		hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

主题

使用 Tez 创建集群

要安装 Tez，请在创建集群时选择 Apache Tez 作为应用程序。

使用控制台创建安装了 Tez 的集群

1. 导航到 Amazon EMR 新控制台，然后从侧面导航栏中选择切换到旧控制台。有关切换到旧控制台后预期情况的更多信息，请参阅 Using the old console。
2. 依次选择 Create cluster (创建集群)、Go to advanced options (转到高级选项)。
3. 在软件配置下，选择 emr-4.7.0 或更高的版本。
4. 选择 Tez 以及您希望 Amazon EMR 安装的其他应用程序。
5. 根据需要选择其它选项，然后选择 Create cluster (创建集群)。

使用 Amazon CLI 创建包含 Tez 的集群

- 使用 `create-cluster` 命令以及 `-- applications` 选项指定 Tez。以下示例将创建一个安装了 Tez 的集群。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --name "Cluster with Tez" --release-label emr-5.36.1 \
--applications Name=Tez --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3 --use-default-roles
```

配置 Tez

您可以使用 `tez-site` 配置分类设置值来自定义 Tez，该配置分类将配置 `tez-site.xml` 配置文件中的设置。有关更多信息，请参阅 Apache Tez 文档中的 TezConfiguration。要将 Hive 或 Pig 更改为使用 Tez 执行引擎，请根据需要使用 `hive-site` 和 `pig-properties` 配置分类。示例如下。

Example 示例：自定义 Tez 根日志记录级别，并将 Tez 设置为 Hive 和 Pig 的执行引擎

下面显示的 `create-cluster` 命令将创建一个安装了 Tez、Hive 和 Pig 的集群。该命令引用存储在 Amazon S3 中的文件 `myConfig.json`，该文件为将 `tez.am.log.level` 设置为 DEBUG 的 `tez-site` 分类指定属性，以及将执行引擎设置为 Tez，以便 Hive 和 Pig 使用 `hive-site` 和 `pig-properties` 配置分类。

> **Note**
>
> 为了便于读取，包含 Linux 行继续符（\）。它们可以通过 Linux 命令删除或使用。对于 Windows，请将它们删除或替换为脱字号（^）。

```
aws emr create-cluster --release-label emr-5.36.1 \
--applications Name=Tez Name=Hive Name=Pig --ec2-attributes KeyName=myKey \
--instance-type m5.xlarge --instance-count 3 \
--configurations https://s3.amazonaws.com/mybucket/myfolder/myConfig.json --use-
default-roles
```

下面显示的是 `myConfig.json` 的内容示例。

```
[
    {
      "Classification": "tez-site",
      "Properties": {
        "tez.am.log.level": "DEBUG"
      }
    },
    {
      "Classification": "hive-site",
      "Properties": {
        "hive.execution.engine": "tez"
      }
    },
    {
      "Classification": "pig-properties",
      "Properties": {
        "exectype": "tez"
      }
    }
  ]
```

> **Note**
>
> 对于 Amazon EMR 5.21.0 及更高版本，您可以覆盖集群配置，并为运行的集群中的每个实例组指定额外的配置分类。要完成此操作，您可以使用 Amazon EMR 控制台、Amazon Command Line Interface（Amazon CLI）或 Amazon SDK。有关更多信息，请参阅为运行的集群中的实例组提供配置。

Tez Web UI

Tez 拥有自己的 Web 用户界面。要查看 Web UI，请参阅以下 URL。

```
http://masterDNS:8080/tez-ui
```

要在 Tez Web UI 上启用 Hive 查询选项卡，请设置以下配置。

```
[
  {
    "Classification": "hive-site",
```

```
    "Properties": {
      "hive.exec.pre.hooks": "org.apache.hadoop.hive.ql.hooks.ATSHook",
      "hive.exec.post.hooks": "org.apache.hadoop.hive.ql.hooks.ATSHook",
      "hive.exec.failure.hooks": "org.apache.hadoop.hive.ql.hooks.ATSHook"
    }
  }
]
```

您还可以在控制台中，使用集群详细信息页面中的 Application user interfaces (应用程序用户界面) 选项卡查看 Tez、Spark 和 YARN 应用程序。Amazon EMR 应用程序用户界面（UI）在集群外托管，并且在集群终止后可用。它们不要求您设置 SSH 连接或 Web 代理，因此您可以更轻松地分析活动的作业和作业历史记录并排除故障。

有关更多信息，请参阅《Amazon EMR 管理指南》中的查看应用程序历史记录。

时间线服务器

YARN 时间线服务器配置为在安装 Tez 时运行。要查看通过时间线服务器使用 Tez 或 MapReduce 执行引擎提交的任务，请使用 URL `http://master-public-DNS:8188` 查看 Web UI。有关更多信息，请参阅《Amazon EMR 管理指南》中的查看 Amazon EMR 集群上托管的 Web 界面。

Tez 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Tez 的版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Tez 版本信息

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.14.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
		timeline-server, tez-on-yarn, tez-on-worker
emr-6.13.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker
emr-6.12.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker
emr-6.11.1	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.11.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker
emr-6.10.1	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker
emr-6.10.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn, tez-on-worker

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.9.1	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.9.0	0.10.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.8.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.8.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.7.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.36.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.36.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.6.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.35.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.5.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.4.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.3.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.3.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.2.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.2.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.1.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.1.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-6.0.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-6.0.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.34.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.33.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.33.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.32.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.32.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.31.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.31.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.30.2	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.30.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.30.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.29.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.28.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.28.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.27.1	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.27.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.26.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.25.0	0.9.2	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.24.1	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.24.0	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.23.1	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.23.0	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.22.0	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.21.2	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.21.1	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.21.0	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.20.1	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.20.0	0.9.1	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.19.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.19.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.18.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.18.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.17.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.17.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.17.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.16.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.16.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.15.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.15.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.14.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.14.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.14.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.13.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.13.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.12.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.12.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.12.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.12.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.11.4	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.11.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.11.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.11.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.11.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.10.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.10.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.9.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.9.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.8.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.8.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.8.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.8.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.7.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.7.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.6.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.6.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.5.4	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.5.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.5.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.5.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.5.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.4.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.4.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.3.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.3.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.3.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.2.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.2.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.2.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.2.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.1.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.1.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-5.0.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-5.0.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.9.6	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-4.9.5	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.9.4	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.9.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-4.9.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.9.1	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.8.5	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-4.8.4	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.8.3	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.8.2	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-4.8.0	0.8.4	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.7.4	0.8.3	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.7.2	0.8.3	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Amazon EMR 发行版标签	Tez 版本	随 Tez 安装的组件
emr-4.7.1	0.8.3	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn
emr-4.7.0	0.8.3	emrfs, emr-goodies, hadoop-client, hadoop-mapred, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, tez-on-yarn

Tez 发布说明（按版本分类）

主题

- Amazon EMR 6.14.0 – Tez 发布说明
- Amazon EMR 6.13.0 – Tez 发布说明
- Amazon EMR 6.12.0 – Tez 发布说明
- Amazon EMR 6.11.0 – Tez 发布说明
- Amazon EMR 6.10.0 – Tez 发布说明
- Amazon EMR 6.9.0 – Tez 发布说明
- Amazon EMR 6.8.0 – Tez 发布说明
- Amazon EMR 6.7.0 – Tez 发布说明
- Amazon EMR 6.6.0 – Tez 发布说明

Amazon EMR 6.14.0 – Tez 发布说明

Amazon EMR 6.14.0 – Tez 更改

类型	描述
改进	将 Tez 中的 TLS 版本升级到 1.2

Amazon EMR 6.13.0 – Tez 发布说明

Amazon EMR 6.13.0 – Tez 更改

类型	描述
错误修复	恢复 TEZ-4295：无法解压缩数据。缓冲区长度太小。
错误修复	恢复 TEZ-4302：`CodecUtils` 中的 `NullPointerException` ，具有 `GzipCodec` 。
错误修复	恢复 TEZ-4234：压缩机可能导致 `Buffer.limit` 中的 `IllegalArgumentException` ，其中限制超过容量。
错误修复	恢复 TEZ-4135：在执行内存中读取时改善内存分配。

Amazon EMR 6.12.0 – Tez 发布说明

Amazon EMR 6.12.0 – Tez 更改

类型	描述
改进	添加了对 JDK 11 和 JDK 17 运行时系统的支持

类型	描述
错误修复	TEZ-4492：更新 Bowerrc 以使用 bower.herokuapp 镜像以避免 Bower Registry CERT_EXPIRE 问题（BOWER-2608）
升级	将 Surefire 升级到 3.0.0-M7

Amazon EMR 6.11.0 – Tez 发布说明

Amazon EMR 6.11.0 – Tez 更改

类型	描述
错误	修复了在清理随机数据的顶点等级时无效的顶点状态转换
错误	修复了随机数据的 DAG 或顶点等级清理不起作用的问题
改进	默认启用 tez.am.dag.cleanup.on.completion 以清除已完成 DAG 的随机数据

Amazon EMR 6.10.0 – Tez 发布说明

Amazon EMR 6.10.0 – Tez 更改

类型	描述
特征	默认情况下启用 `tez.runtime.transfer.data-via-events.enabled`
逆向移植	TEZ-4450：修复通过数据移动事件传输随机数据时随机数据获取失败的问题
逆向移植	TEZ-4460：修复从 Tez Shuffle Handler 获取随机数据时出现的读取超时错误

类型	描述
逆向移植	TEZ-4455：在 ShuffleHandler 管道中添加 LoggingHandler 以获得更好的可调试性
错误	修复启用了抢占任务后 Tez 任务间歇性卡住的问题

Amazon EMR 6.9.0 – Tez 发布说明

Amazon EMR 6.9.0 – Tez 更改

类型	描述
升级	Tez 已升级到 0.10.2。有关更多信息，请参阅 change log for Apache Tez 0.10.2（Apache Tez 0.10.2 的更改日志）。
升级	将 Hadoop 升级到 3.3.3。
错误	由于 TEZ-4450，默认情况下禁用 `tez.runtime.transfer.data-via-events.enabled` 。

Amazon EMR 6.8.0 – Tez 发布说明

Amazon EMR 6.8.0 – Tez 更改

类型	描述
逆向移植	TEZ-3363：删除 Shuffle Handler 的顶点级别中间数据
逆向移植	TEZ-4129：删除 Shuffle Handler 的尝试失败中间尝试数据

类型	描述
逆向移植	TEZ-4430：修复了 tez.task.launch.cmd-opts 属性不起作用的问题

Amazon EMR 6.7.0 – Tez 发布说明

Amazon EMR 6.7.0 – Tez 更改

类型	描述
逆向移植	TEZ-4403：将 SLF4J 版本升级到 1.7.36
逆向移植	TEZ-4405：将 log4j 1.x 替换为 reload4j
逆向移植	TEZ-4411：Tez 构建失败：找不到 FileSaver.js

Amazon EMR 6.6.0 – Tez 发布说明

Amazon EMR 6.6.0 – Tez 更改

类型	描述
逆向移植	TEZ-3918：修复了 tez.task.log.level 属性无法正常工作的问题。
逆向移植	TEZ-4353：将 commons-io 更新到 2.8.0。
逆向移植	TEZ-4114：从 tez 中删除直接的 jetty 依赖项。
逆向移植	TEZ-4323：通过 TEZ-4114 从 dist 软件包中删除 Jetty jar。

Apache Zeppelin

使用 Apache Zeppelin 作为用于交互式数据探索的笔记本。有关 Zeppelin 的更多信息，请参阅 https://zeppelin.apache.org/。Amazon EMR 发行版 5.0.0 及更高版本包含 Zeppelin。早期发行版包含 Zeppelin，将其用作沙盒应用程序。有关更多信息，请参阅Amazon EMR 4.x 发行版。

要访问 Zeppelin Web 界面，请设置连接到主节点的 SSH 隧道和代理连接。有关更多信息，请参阅查看 EMR 集群上托管的 Web 界面。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 Zeppelin 版本，以及 Amazon EMR 随 Zeppelin 一起安装的组件。

有关此发行版中随 Zeppelin 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 Zeppelin 版本信息

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.14.0	Zeppelin 0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 Zeppelin 版本，以及 Amazon EMR 随 Zeppelin 一起安装的组件。

有关此发行版中随 Zeppelin 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 Zeppelin 版本信息

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.36.1	Zeppelin 0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

主题

在 Amazon EMR 上使用 Zeppelin 时的注意事项

- 使用相同的 SSH 隧道方法连接到 Zeppelin 以连接到主节点上的其他 Web 服务器。Zeppelin 服务器可以在端口 8890 上找到。
- Amazon EMR 发行版 5.0.0 及更高版本上的 Zeppelin 支持 Shiro 身份验证。
- Amazon EMR 发行版 5.8.0 及更高版本上的 Zeppelin 支持使用 Amazon Glue 数据目录作为 Spark SQL 的元存储。有关更多信息，请参阅使用 Amazon Glue 数据目录作为 Spark SQL 的元存储。
- Zeppelin 不使用集群的 `spark-defaults.conf` 配置文件中定义的部分设置（即使在您将 `spark.dynamicAllocation.enabled` 设置为 `true` 时，其指示 YARN 动态分配执行者也是如此）。您必须使用 Zeppelin Interpreter (解释器) 选项卡设置执行者设置 (如内存和内核)，然后为要使用的设置重新启动解释器。

- Amazon EMR 版本 6.10.0 及更高版本支持 Apache Zeppelin 与 Apache Flink 集成。参阅 在 Amazon EMR 中通过 Zeppelin 使用 Flink 作业 了解更多信息。
- Amazon EMR 上的 Zeppelin 不支持 SparkR 解释器。

Zeppelin 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 Zeppelin 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

Zeppelin 版本信息

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.14.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.13.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resour

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
		cemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.12.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.11.1	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.11.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.10.1	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.10.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.9.1	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.9.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.8.1	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.8.0	0.10.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.7.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.36.1	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.36.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.6.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, hudi, hudi-spark, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.35.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.5.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.4.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.3.1	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.3.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.2.1	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.2.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.1.1	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.1.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-6.0.1	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-6.0.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.34.0	0.10.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.33.1	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.33.0	0.9.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.32.1	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.32.0	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.31.1	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.31.0	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.30.2	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.30.1	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.30.0	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.29.0	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.28.1	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.28.0	0.8.2	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.27.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.27.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.26.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.25.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.24.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.24.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.23.1	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.23.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.22.0	0.8.1	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, livy-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.21.2	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.21.1	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.21.0	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.20.1	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.20.0	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.19.1	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.19.0	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.18.1	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.18.0	0.8.0	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.17.2	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.17.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.17.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.16.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.16.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.15.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.15.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.14.2	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.14.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.14.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.13.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.13.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, r, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.12.3	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.12.2	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.12.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.12.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.11.4	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.11.3	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.11.2	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.11.1	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.11.0	0.7.3	aws-sagemaker-spark-sdk, emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.10.1	0.7.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.10.0	0.7.3	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.9.1	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.9.0	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.8.3	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.8.2	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.8.1	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.8.0	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.7.1	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.7.0	0.7.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.6.1	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.6.0	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.5.4	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.5.3	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.5.2	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.5.1	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.5.0	0.7.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.4.1	0.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.4.0	0.7.0	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.3.2	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.3.1	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.3.0	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.2.3	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.2.2	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.2.1	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.2.0	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.1.1	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.1.0	0.6.2	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Amazon EMR 发行版标签	Zeppelin 版本	随 Zeppelin 安装的组件
emr-5.0.3	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server
emr-5.0.0	0.6.1	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, spark-client, spark-history-server, spark-on-yarn, spark-yarn-slave, zeppelin-server

Apache ZooKeeper

Apache ZooKeeper 是用于维护配置信息、命名、提供分布式同步以及提供组服务的集中式服务。有关 ZooKeeper 的更多信息，参阅 http://zookeeper.apache.org/。

下表列出了 Amazon EMR 6.x 系列的最新发行版附带的 ZooKeeper 版本，以及 Amazon EMR 随 ZooKeeper 一起安装的组件。

有关此发行版中随 ZooKeeper 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-6.14.0 的 ZooKeeper 版本信息

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.14.0	ZooKeeper 3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

下表列出了 Amazon EMR 5.x 系列的最新发行版附带的 ZooKeeper 版本，以及 Amazon EMR 随 ZooKeeper 一起安装的组件。

有关此发行版中随 ZooKeeper 安装的组件版本，请参阅 Release 6.14.0 Component Versions。

emr-5.36.1 的 ZooKeeper 版本信息

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.36.1	ZooKeeper 3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
		server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

主题

ZooKeeper 发行历史记录

下表列出了 Amazon EMR 每个发行版中包含的 ZooKeeper 版本，以及随应用程序一起安装的组件。有关每个发行版本中的组件版本，请参阅Amazon EMR 5.x 发行版或者Amazon EMR 4.x 发行版中的“组件版本”部分。

ZooKeeper 版本信息

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.14.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.13.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
		server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.12.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.11.1	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.11.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.10.1	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.10.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.9.1	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.9.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.8.1	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.8.0	3.5.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.7.0	3.5.7	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.36.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.36.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.6.0	3.5.7	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.35.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.5.0	3.5.7	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.4.0	3.5.7	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.3.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.3.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.2.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.2.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.1.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.1.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-6.0.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-6.0.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.34.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.33.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.33.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.32.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.32.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.31.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.31.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.30.2	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.30.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.30.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.29.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.28.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.28.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.27.1	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.27.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.26.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.25.0	3.4.14	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.24.1	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.24.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.23.1	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.23.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.22.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.21.2	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.21.1	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.21.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.20.1	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.20.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.19.1	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.19.0	3.4.13	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.18.1	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.18.0	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.17.2	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.17.1	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.17.0	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.16.1	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.16.0	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.15.1	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.15.0	3.4.12	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.14.2	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.14.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.14.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.13.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.13.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.12.3	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.12.2	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.12.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.12.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.11.4	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.11.3	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.11.2	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.11.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.11.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.10.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.10.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.9.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.9.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.8.3	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.8.2	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.8.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.8.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.7.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.7.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.6.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.6.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, hadoop-yarn-timeline-server, zookeeper-client, zookeeper-server
emr-5.5.4	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.5.3	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.5.2	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.5.1	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.5.0	3.4.10	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.4.1	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.4.0	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.3.2	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.3.1	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.3.0	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.2.3	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.2.2	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.2.1	3.4.9	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.2.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.1.1	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.1.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

Amazon EMR 发行版标签	ZooKeeper 版本	随 ZooKeeper 安装的组件
emr-5.0.3	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server
emr-5.0.0	3.4.8	emrfs, emr-goodies, hadoop-client, hadoop-hdfs-datanode, hadoop-hdfs-library, hadoop-hdfs-namenode, hadoop-httpfs-server, hadoop-kms-server, hadoop-yarn-nodemanager, hadoop-yarn-resourcemanager, zookeeper-client, zookeeper-server

连接器和实用工具

Amazon EMR 提供多个连接器和实用工具来访问作为数据源的其它 Amazon 服务。您通常可在一个程序内访问这些服务中的数据。例如，您可在 Hive 查询、Pig 脚本或 MapReduce 应用程序中指定 Kinesis 流，然后在相应数据上操作。

主题

使用 Amazon EMR 导出、导入、查询和连接 DynamoDB 中的表格

> **Note**
>
> Amazon EMR-DynamoDB 连接器在 GitHub 上是开源的。有关更多信息，请参阅 https://github.com/awslabs/emr-dynamodb-connector。

DynamoDB 是一项完全托管的 NoSQL 数据库服务，提供快速而可预测的性能，能够实现无缝扩展。开发人员可以创建数据库表，并可以不受限制地增加请求流量或存储空间。DynamoDB 可自动将表的数据和流量分布到足够多的服务器中，以便处理客户指定的容量请求和数据存储量，同时还能保持性能一致、访问高效。使用 Amazon EMR 和 Hive 可以快速有效地处理大量数据，如 DynamoDB 中存储的数据。有关 DynamoDB 的更多信息，请参阅 Amazon DynamoDB 开发人员指南。

Apache Hive 是一种可用于查询 map reduce 集群的软件层，使用的是一种名为 HiveQL 的类似 SQL 的简化查询语言。它在 Hadoop 架构的顶层运行。有关 Hive 和 HiveQL 的更多信息，请转至 HiveQL 语言手册。有关 Hive 和 Amazon EMR 的更多信息，请参阅Apache Hive。

您可将 Amazon EMR 与自定义版本的 Hive（包含与 DynamoDB 的连接）配合使用来对存储在 DynamoDB 中的数据执行操作：

- 将 DynamoDB 数据加载到 Hadoop Distributed File System（HDFS）中并用作输入 Amazon EMR 集群的数据。
- 使用类似 SQL 语句（HiveQL）查询实时 DynamoDB 数据。

- 连接 DynamoDB 中存储的数据并导出这些数据或查询连接的数据。
- 将存储在 DynamoDB 中的数据导出到 Amazon S3。
- 将存储在 Amazon S3 中的数据导入到 DynamoDB。

> **Note**
>
> Amazon EMR-DynamoDB 连接器不支持配置为使用 Kerberos 身份验证的集群。

为了执行以下每项任务，您将启动 Amazon EMR 集群，指定 DynamoDB 中数据的位置，并发出 Hive 命令来操作 DynamoDB 中的数据。

有多种启动 Amazon EMR 集群的方法：您可以使用 Amazon EMR 控制台、命令行界面（CLI），或者使用 Amazon SDK 或 Amazon EMR API 对您的集群进行编程。您还可以选择交互运行 Hive 集群还是从脚本运行。在本节中，我们将介绍如何从 Amazon EMR 控制台和 CLI 启动交互式 Hive 集群。

通过交互的方式使用 Hive 是测试查询性能和调试应用程序的良好方式。确定将定期运行的 Hive 命令集之后，请考虑创建一个 Hive 脚本，让 Amazon EMR 来运行。

> **Warning**
>
> DynamoDB 表上的 Amazon EMR 读取或写入操作不利于既定的预置吞吐量，有可能增加预置吞吐量例外情况出现的频率。对于大量请求，Amazon EMR 会使用指数回退实施重试，以管理 DynamoDB 表中的请求负载。如果您与其它流量同时运行 Amazon EMR 任务，就可能导致超出分配的预置吞吐量级别。您可以查看 Amazon CloudWatch 中的 ThrottleRequests 指标来监控这一操作。如果请求负载过高，您可以重新启动集群，将读取百分比设置或写入百分比设置设置为较低的值，从而限制 Amazon EMR 操作。有关 DynamoDB 吞吐量设置的详细信息，请参阅预置吞吐量。
>
> 如果表配置为按需模式，则应先将表更改回预配置模式，再运行导出或导入操作。管道需要吞吐量比率才能计算 DynamoDBtable 中要使用的资源。按需模式删除预置的吞吐量。要预置吞吐量容量，您可以使用 Amazon CloudWatch Events 指标来评估表使用的聚合吞吐量。

主题

- 设置 Hive 表来运行 Hive 命令
- 用于在 DynamoDB 中导出、导入和查询数据的 Hive 命令示例
- 在 DynamoDB 中优化 Amazon EMR 操作的性能

设置 Hive 表来运行 Hive 命令

Apache Hive 是一种可用于使用类似 SQL 的语言查询 Amazon EMR 集群中包含的数据的数据仓库应用程序。有关 Hive 的更多信息，请参阅 http://hive.apache.org/。

下面的程序假定您已创建集群并指定了 Amazon EC2 密钥对。要了解如何开始创建集群，请参阅《Amazon EMR 管理指南》中的 Amazon EMR 入门。

配置 Hive 来使用 MapReduce

在您使用 Amazon EMR 上的 Hive 查询 DynamoDB 表时，如果 Hive 使用默认执行引擎 Tez，则可能会出现错误。因此，在按照本节中所述创建具有与 DynamoDB 集成的 Hive 的集群时，我们建议您使用将 Hive 设置为使用 MapReduce 的配置分类。有关更多信息，请参阅配置应用程序。

以下代码段显示用于将 MapReduce 设置为 Hive 执行引擎的配置分类和属性：

```
[
                {
                    "Classification": "hive-site",
                    "Properties": {
                        "hive.execution.engine": "mr"
                    }
                }
            ]
```

以交互方式运行 Hive 命令

1. 连接到主节点。有关更多信息，请参阅《Amazon EMR 管理指南》中的使用 SSH 连接到主节点。
2. 当命令提示输入当前主节点时，键入 `hive`。

   您应看到 Hive 提示符：`hive>`
3. 输入用于将 Hive 应用程序中的表映射到 DynamoDB 中的数据的 Hive 命令。该表充当对 Amazon DynamoDB 中存储的数据的引用；数据未存储在本地的 Hive 中，每次运行命令时，使用此表的任何查询将针对 DynamoDB 中的实时数据运行，从而占用此表的读取或写入容量。如果您需要对同一数据集运行多个 Hive 命令，请考虑先将其导出。

   下面说明将 Hive 表映射到 DynamoDB 表的语法。

```
CREATE EXTERNAL TABLE hive_tablename
 (hive_column1_name column1_datatype, hive_column2_name column2_datatype...)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodb_tablename",
"dynamodb.column.mapping" =
 "hive_column1_name:dynamodb_attribute1_name,hive_column2_name:dynamodb_attribute2_name..."
```

当您在 Hive 中从 DynamoDB 创建表时，必须使用关键字 EXTERNAL 将该表创建为外部表。外部表与内部表之间的区别是：删除内部表时，将删除内部表中的数据。当连接 Amazon DynamoDB 时，这不是所需行为，因此仅支持外部表。

例如，以下 Hive 命令在 Hive 中创建名为 hivetable1 的表，该表引用名为 dynamodbtable1 的 DynamoDB 表。DynamoDB 表 dynamodbtable1 具有 hash-and-range 主键架构。哈希键元素是 name (字符串类型)。范围键元素是 year (数字类型)。每个项目都有 holidays (字符串集类型) 的属性值。

```
CREATE EXTERNAL TABLE hivetable1 (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");
```

第 1 行使用 HiveQL CREATE EXTERNAL TABLE 语句。对于 hivetable1，您需要为 DynamoDB 表中的每个属性名称/值对建立一列，并提供数据类型。这些值不区分大小写，并且您可以为列提供任何名称 (保留字除外)。

第 2 行使用 STORED BY 语句。STORED BY 的值是用于处理 Hive 与 DynamoDB 之间连接的类的名称。该值应设置为 `'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'`。

第 3 行使用 TBLPROPERTIES 语句将“hivetable1”与 DynamoDB 中相应的表和架构相关联。为 TBLPROPERTIES 提供 `dynamodb.table.name` 参数和 `dynamodb.column.mapping` 参数的值。这些值是区分大小写的。

> **Note**
>
> 此表的所有 DynamoDB 属性名称必须在 Hive 表中有对应的列。根据您的 Amazon EMR 版本，如果不存在一对一映射，则会发生以下情况：

- 在 Amazon EMR 5.27.0 及更高版本上，连接器具有验证功能，以确保 DynamoDB 属性名称与 Hive 表中的列之间一一对应。如果不存在一对一映射，则会发生错误。
- 在 Amazon EMR 5.26.0 及更低版本上，Hive 表将不包含来自 DynamoDB 的名称/值对。如果您未映射 DynamoDB 主键属性，则 Hive 将生成错误。如果您未映射非主键属性，则不会生成错误，但您将无法查看 Hive 表中的数据。如果数据类型不匹配，则值为空。

然后，您可以开始对 hivetable1 运行 Hive 操作。根据 hivetable1 运行的查询也根据 DynamoDB 账户的 DynamoDB 表 dynamodbtable1 在内部运行，在每次执行运行时消耗读取或写入单位。

对 DynamoDB 表运行 Hive 查询时，您需要确保已预置足量的读取容量单位。

例如，假设您为 DynamoDB 表预配置了 100 个读取容量单位。这将允许您每秒执行 100 次读取或读取 409600 字节。如果该表包含 20GB 的数据（21474836480 字节）并且您的 Hive 查询执行全表扫描，则可以估算执行查询将花费多长时间：

21474836480/409600 = 52429 秒 = 14.56 小时

减少所需时间的唯一方法是调整源 DynamoDB 表的读取容量单位。添加更多 Amazon EMR 节点将不会有帮助。

在 Hive 输出中，当一个或多个映射器进程已完成时，将更新完成百分比。对于预配置的读取容量设置较低的大型 DynamoDB 表，完成百分比输出可能会很长时间不更新；在上面的示例中，作业将在几个小时内显示为完成 0%。有关作业进度的详细状态，请转到 Amazon EMR 控制台；您将可以查看单个映射器任务状态和数据读取统计数据。您还可以登录主节点的 Hadoop 界面，查看 Hadoop 统计数据。该界面将向您显示单个映射任务状态和一些数据读取统计数据。有关更多信息，请参阅以下主题：

- 托管在主节点 (master node) 上的 Web 界面
- 查看 Hadoop Web 界面

有关用于执行从 DynamoDB 导出或导入数据和联接表等任务的示例 HiveQL 语句的更多信息，请参阅用于在 DynamoDB 中导出、导入和查询数据的 Hive 命令示例。

取消 Hive 请求

执行 Hive 查询时，来自服务器的初始响应包含用于取消请求的命令。要在此过程中随时取消请求，请使用服务器响应中的 Kill 命令。

1. 输入 `Ctrl+C` 可退出命令行客户端。
2. 在 Shell 提示符下，输入服务器对您的请求的初始响应中的 Kill 命令。

   或者，您也可以从主节点的命令行运行以下命令来终止 Hadoop 任务，其中 *job-id* 是 Hadoop 任务的标识符，可从 Hadoop 用户界面检索到。

   ```
   hadoop job -kill job-id
   ```

Hive 和 DynamoDB 的数据类型

下表显示了可用的 Hive 数据类型、它们对应的默认 DynamoDB 类型以及它们也可以映射到的备用 DynamoDB 类型。

Hive 类型	默认 DynamoDB 类型	备用 DynamoDB 类型
字符串	字符串	
bigint 或 double	数字 (N)	
binary	二进制 (B)	
布尔值	boolean (BOOL)	
数组	list (L)	数字集 (NS)、字符串集 (SS) 或二进制集 (BS)
map<string,string>	项目	map (M)
map<string,?>	map (M)	
	null (NULL)	

如果要将 Hive 数据作为对应的备用 DynamoDB 类型写入，或者您的 DynamoDB 数据包含备用 DynamoDB 类型的属性值，则可以使用 dynamodb.type.mapping 参数指定列和 DynamoDB 类型。以下示例显示了用于指定备用类型映射的语法。

```
CREATE EXTERNAL TABLE hive_tablename (hive_column1_name column1_datatype,
 hive_column2_name column2_datatype...)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodb_tablename",
"dynamodb.column.mapping" =
 "hive_column1_name:dynamodb_attribute1_name,hive_column2_name:dynamodb_attribute2_name...",
"dynamodb.type.mapping" = "hive_column1_name:dynamodb_attribute1_datatype");
```

类型映射参数是可选的，仅必须为使用备用类型的列指定它。

例如，以下 Hive 命令在 Hive 中创建名为 hivetable2 的表，该表引用 DynamoDB 表 dynamodbtable2。它与 hivetable1 相似，不同之处在于它将 col3 列映射到字符串集 (SS) 类型。

```
CREATE EXTERNAL TABLE hivetable2 (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable2",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays",
"dynamodb.type.mapping" = "col3:SS");
```

在 Hive 中，hivetable1 和 hivetable2 是相同的。但是，将这些表中的数据写入其对应的 DynamoDB 表时，dynamodbtable1 将包含列表，而 dynamodbtable2 将包含字符串集。

如果要将 Hive null 值作为 DynamoDB null 类型的属性写入，则您可以使用 dynamodb.null.serialization 参数来写入。以下示例显示了用于指定 null 序列化的语法。

```
CREATE EXTERNAL TABLE hive_tablename (hive_column1_name column1_datatype,
 hive_column2_name column2_datatype...)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodb_tablename",
"dynamodb.column.mapping" =
 "hive_column1_name:dynamodb_attribute1_name,hive_column2_name:dynamodb_attribute2_name...",
"dynamodb.null.serialization" = "true");
```

空序列化参数是可选的，如果未指定，则设置为 false。请注意，无论参数设置如何，DynamoDB null 属性都将作为 Hive 中的 null 值进行读取。仅当将空序列化参数指定为 true 时，才能将具有 null 值的 Hive 集合写入 DynamoDB。否则，将出现 Hive 错误。

就精度而言，Hive 中的 bigint 类型与 Java long 类型相同，而 Hive double 类型与 Java double 类型相同。这意味着，如果您有精度高于 Hive 数据类型所提供精度的数值数据存储在 DynamoDB 中，则使用 Hive 导出、导入或引用 DynamoDB 数据会导致精度损失或 Hive 查询失败。

从 DynamoDB 导出到 Amazon Simple Storage Service（Amazon S3）或 HDFS 的二进制类型作为 Base64 编码的字符串进行存储。如果您要将 Amazon S3 或 HDFS 中的数据导入到 DynamoDB 二进制类型，则该数据应编码为 Base64 字符串。

Hive 选项

您可以设置以下 Hive 选项来管理从 Amazon DynamoDB 的数据传出。这些选项只针对当前 Hive 会话保留。如果您在集群上关闭 Hive 命令提示符并稍后重新打开，则这些设置将恢复为默认值。

Hive 选项	描述
`dynamodb.throughput.read.percent`	设置读取操作的速率，在为您的表分配的范围内保持 DynamoDB 预配置的吞吐速率。该值介于 `0.1` 到 `1.5` 之间（包含端点）。 值 0.5 是默认读取速率，这意味着，Hive 将在表的整个资源中尝试占用一半的预配读取量。增加此值使之高于 0.5 将提高读取请求速率。减少此值使之低于 0.5 将降低读取请求速率。此读取速率是近似值。实际读取速率取决于 DynamoDB 中是否存在统一分配的键等因素。 如果您发现 Hive 操作经常超出您预配的吞吐量，或者如果过多限制了实时读取流量，则可以减少此值使之低于 `0.5`。如果您有足够的容量并希望 Hive 操作的速度更快，请将此值设置为高于 `0.5`。如果您认为有可用的输入/输出操作未使用，则还可以通过将此值设置到最高 1.5 来进行超额预订。
`dynamodb.throughput.write.percent`	设置写入操作的速率，在为您的表分配的范围内保持 DynamoDB 预配置的吞吐速率。该值介于 `0.1` 到 `1.5` 之间（包含端点）。 值 0.5 是默认写入速率，这意味着，Hive 将在表的整个资源中尝试占用一半的预配写入量。增加此值使之高于 0.5 将提高写入请求速率。减少此值使之低于 0.5 将降

Hive 选项	描述
	低写入请求速率。此写入速率是近似值。实际写入速率取决于 DynamoDB 中是否存在统一分配的键等因素 如果您发现 Hive 操作经常超出您预配的吞吐量，或者如果过多限制了实时写入流量，则可以减少此值使之低于 0.5。如果您有足够的容量并希望 Hive 操作的速度更快，请将此值设置为高于 0.5。如果您认为有可用的输入/输出操作未使用或者这是到表的初始数据上载，还没有实时流量，则还可以通过将此值设置到最高 1.5 来进行超额预订。
dynamodb.endpoint	为 DynamoDB 服务指定终端节点。有关可用 DynamoDB 终端节点的更多信息，请参阅区域和终端节点。
dynamodb.max.map.tasks	指定在从 DynamoDB 读取数据时，映射任务的最大数量。此值必须等于或大于 1。
dynamodb.retry.duration	指定要用作重试 Hive 命令的超时时间的分钟数。此值必须是大于或等于 0 的整数。默认超时持续时间为 2 分钟。

这些选项是使用 SET 命令设置的，如以下示例所示。

```
SET dynamodb.throughput.read.percent=1.0;

INSERT OVERWRITE TABLE s3_export SELECT *
FROM hiveTableName;
```

用于在 DynamoDB 中导出、导入和查询数据的 Hive 命令示例

以下示例使用 Hive 命令执行将数据导出到 Amazon S3 或 HDFS、将数据导入到 DynamoDB、连接表、查询表等操作。

对 Hive 表执行的操作将引用 DynamoDB 中存储的数据。Hive 命令受到 DynamoDB 表预置的吞吐量设置约束，并且检索的数据包括 DynamoDB 处理 Hive 操作请求时写入到 DynamoDB 表的数据。如果数据检索过程需要很长一段时间，则自 Hive 命令开始执行以来，Hive 命令返回的某些数据可能已在 DynamoDB 中更新。

Hive 命令 `DROP TABLE` 和 `CREATE TABLE` 仅对 Hive 中的本地表进行操作，而不会在 DynamoDB 中创建或删除表。如果 Hive 查询引用 DynamoDB 中的表，则在您运行查询之前，该表必须已存在。有关在 DynamoDB 中创建和删除表的更多信息，请参阅 Amazon DynamoDB 开发人员指南中的在 DynamoDB 中处理表。

> **Note**
>
> 当您将 Hive 表映射到 Amazon S3 中的位置时，请勿将其映射到存储桶的根路径 s3://mybucket，否则这会在 Hive 将数据写入到 Amazon S3 时导致错误。而是应将表映射到存储桶的子路径 s3://mybucket/mypath。

从 DynamoDB 中导出数据

可以使用 Hive 从 DynamoDB 中导出数据。

将 DynamoDB 表导出到 Amazon S3 存储桶

- 创建一个引用 DynamoDB 中存储的数据的 Hive 表。然后，您可以调用 INSERT OVERWRITE 命令将数据写入到外部目录。在以下示例中，*s3://bucketname/path/subpath/* 是 Amazon S3 中的有效路径。调整 CREATE 命令中的列和数据类型来匹配 DynamoDB 中的值。可以使用此命令在 Amazon S3 中创建 DynamoDB 数据的存档。

```
CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");


INSERT OVERWRITE DIRECTORY 's3://bucketname/path/subpath/' SELECT *
FROM hiveTableName;
```

使用格式设置将 DynamoDB 表导出到 Amazon S3 存储桶

- 创建引用 Amazon S3 中的位置的外部表。此表在下面显示为 s3_export。在调用 CREATE 期间，为此表指定行格式设置。然后，当您使用 INSERT OVERWRITE 将数据从 DynamoDB 导出到 s3_export 时，数据将以指定的格式写出。在以下示例中，数据以逗号分隔值 (CSV) 的格式写出。

```
CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");


CREATE EXTERNAL TABLE s3_export(a_col string, b_col bigint, c_col array<string>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION 's3://bucketname/path/subpath/';

INSERT OVERWRITE TABLE s3_export SELECT *
FROM hiveTableName;
```

在不指定列映射的情况下将 DynamoDB 表导出到 Amazon S3 存储桶

- 创建一个引用 DynamoDB 中存储的数据的 Hive 表。此例与前面的示例类似，只是不指定列映射。该表必须正好具有类型为 map<string, string> 的一个列。如果您随后在 Amazon S3 中创建 EXTERNAL 表，可以调用 INSERT OVERWRITE 命令将数据从 DynamoDB 写入到 Amazon S3。可以使用此命令在 Amazon S3 中创建 DynamoDB 数据的存档。由于没有列映射，因此您无法查询以此方式导出的表。在 Hive 0.8.1.5 或更高版本（在 Amazon EMR AMI 2.2.x 及其更高版本上受支持）中导出数据而不指定列映射。

```
CREATE EXTERNAL TABLE hiveTableName (item map<string,string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1");

CREATE EXTERNAL TABLE s3TableName (item map<string, string>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
LOCATION 's3://bucketname/path/subpath/';

INSERT OVERWRITE TABLE s3TableName SELECT *
```

```
FROM hiveTableName;
```

使用数据压缩将 DynamoDB 表导出到 Amazon S3 存储桶

- Hive 提供多个可以在 Hive 会话期间设置的压缩编解码器。这样做会导致导出的数据以指定的格式进行压缩。以下示例使用 Lempel-Ziv-Oberhumer (LZO) 算法压缩导出的文件。

```
SET hive.exec.compress.output=true;
SET io.seqfile.compression.type=BLOCK;
SET mapred.output.compression.codec = com.hadoop.compression.lzo.LzopCodec;


CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");


CREATE EXTERNAL TABLE lzo_compression_table (line STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
LOCATION 's3://bucketname/path/subpath/';

INSERT OVERWRITE TABLE lzo_compression_table SELECT *
FROM hiveTableName;
```

可用的压缩编解码器包括：

- org.apache.hadoop.io.compress.GzipCodec
- org.apache.hadoop.io.compress.DefaultCodec
- com.hadoop.compression.lzo.LzoCodec
- com.hadoop.compression.lzo.LzopCodec
- org.apache.hadoop.io.compress.BZip2Codec
- org.apache.hadoop.io.compress.SnappyCodec

将 DynamoDB 表导出到 HDFS

- 使用以下 Hive 命令，其中 *hdfs:///directoryName* 是有效的 HDFS 路径，而 *hiveTableName* 为 Hive 中引用 DynamoDB 的表。此导出操作比将 DynamoDB 表导出到 Amazon S3 速度快，因为将数据导出到 Amazon S3 时，Hive 0.7.1.1 将 HDFS 用作中间步骤。以下示例还显示了如何将 dynamodb.throughput.read.percent 设置为 1.0 以提高读取请求速率。

```
CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");

SET dynamodb.throughput.read.percent=1.0;

INSERT OVERWRITE DIRECTORY 'hdfs:///directoryName' SELECT * FROM hiveTableName;
```

  您还可以按上面所示的用于导出到 Amazon S3 的方法，使用格式设置和压缩将数据导出到 HDFS。为此，只需将上面示例中的 Amazon S3 目录替换为 HDFS 目录。

在 Hive 中读取不可打印的 UTF-8 字符数据

- 创建表时，您可以使用 STORED AS SEQUENCEFILE 子句在 Hive 中读取和写入不可打印的 UTF-8 字符数据。SequenceFile 是 Hadoop 二进制文件格式；您需要使用 Hadoop 来读取此文件。以下示例显示了如何将数据从 DynamoDB 导出到 Amazon S3 中。可以使用此功能处理不可打印的 UTF-8 编码字符。

```
CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");


CREATE EXTERNAL TABLE s3_export(a_col string, b_col bigint, c_col array<string>)
STORED AS SEQUENCEFILE
LOCATION 's3://bucketname/path/subpath/';
```

```
INSERT OVERWRITE TABLE s3_export SELECT *
FROM hiveTableName;
```

将数据导入到 DynamoDB

使用 Hive 将数据写入到 DynamoDB 中时，应确保写入容量单位数大于集群中的映射器数。例如，在 m1.xlarge EC2 实例上运行的集群在每个实例上生成 8 个映射器。对于具有 10 个实例的集群，这意味着生成 80 个映射器。如果写入容量单位数不大于集群中的映射器数，则 Hive 写入操作可能会占用所有写入吞吐量，或者尝试占用超过预配置值的吞吐量。有关每种 EC2 实例类型生成的映射器数的更多信息，请参阅 配置 Hadoop。

Hadoop 中的映射器数由输入的拆分数控制。如果拆分数过小，写入命令可能无法占用所有可用的写入吞吐量。

如果具有相同键的项目在目标 DynamoDB 表中存在，则将覆盖该项目。如果目标 DynamoDB 表中不存在具有该键的项目，则将插入该项目。

要将数据从 Amazon S3 导入 DynamoDB

- 您可以使用 Amazon EMR（Amazon EMR）和 Hive 将数据从 Amazon S3 写入到 DynamoDB。

```
CREATE EXTERNAL TABLE s3_import(a_col string, b_col bigint, c_col array<string>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION 's3://bucketname/path/subpath/';

CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");

INSERT OVERWRITE TABLE hiveTableName SELECT * FROM s3_import;
```

在不指定列映射时将表从 Amazon S3 存储桶导入到 DynamoDB 中

- 创建一个引用 Amazon S3 中存储数据的 EXTERNAL 表，该数据是以前从 DynamoDB 中导出的。在导入之前，请确保该表存在于 DynamoDB 中，并且该表与以前导出的 DynamoDB 表具有相同的键架构。此外，该表还必须正好具有类型为 `map<string, string>` 的一个列。如果您随后创

建一个链接到 DynamoDB 的 Hive 表，则可以调用 INSERT OVERWRITE 命令将数据从 Amazon S3 写入到 DynamoDB 中。由于没有列映射，因此您无法查询以此方式导入的表。在 Hive 0.8.1.5 或更高版本（在 Amazon EMR AMI 2.2.3 及其更高版本上受支持）中可以在不指定列映射时导入数据。

```
CREATE EXTERNAL TABLE s3TableName (item map<string, string>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY '\t' LINES TERMINATED BY '\n'
LOCATION 's3://bucketname/path/subpath/';

CREATE EXTERNAL TABLE hiveTableName (item map<string,string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1");

INSERT OVERWRITE TABLE hiveTableName SELECT *
FROM s3TableName;
```

将表从 HDFS 导入到 DynamoDB 中

- 可以使用 Amazon EMR 和 Hive 将数据从 HDFS 写入到 DynamoDB 中。

```
CREATE EXTERNAL TABLE hdfs_import(a_col string, b_col bigint, c_col array<string>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION 'hdfs:///directoryName';

CREATE EXTERNAL TABLE hiveTableName (col1 string, col2 bigint, col3 array<string>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "dynamodbtable1",
"dynamodb.column.mapping" = "col1:name,col2:year,col3:holidays");

INSERT OVERWRITE TABLE hiveTableName SELECT * FROM hdfs_import;
```

查询 DynamoDB 中的数据

以下示例显示了您可以使用 Amazon EMR 查询 DynamoDB 中存储数据的各种方式。

查找映射列的最大值 (**max**)

- 使用如下 Hive 命令。在第一个命令中，CREATE 语句创建了一个引用 DynamoDB 中存储数据的 Hive 表。然后，SELECT 语句使用该表查询 DynamoDB 中存储的数据。以下示例查找给定客户提交的最大订单。

```
CREATE EXTERNAL TABLE hive_purchases(customerId bigint, total_cost double,
 items_purchased array<String>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "Purchases",
"dynamodb.column.mapping" =
 "customerId:CustomerId,total_cost:Cost,items_purchased:Items");

SELECT max(total_cost) from hive_purchases where customerId = 717;
```

使用 **GROUP BY** 子句聚合数据

- 可以使用 GROUP BY 子句收集多条记录的数据。此子句通常与聚合函数 (如 sum、count、min 或 max) 一起使用。以下示例返回提交了三个以上订单的客户的最大订单列表。

```
CREATE EXTERNAL TABLE hive_purchases(customerId bigint, total_cost double,
 items_purchased array<String>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "Purchases",
"dynamodb.column.mapping" =
 "customerId:CustomerId,total_cost:Cost,items_purchased:Items");

SELECT customerId, max(total_cost) from hive_purchases GROUP BY customerId HAVING
 count(*) > 3;
```

连接两个 DynamoDB 表

- 以下示例将两个 Hive 表映射到 DynamoDB 中存储的数据。然后，它对这两个表调用联接。连接在集群上计算并返回。连接不在 DynamoDB 中进行。此示例返回提交了两个以上订单的客户及其购买物的列表。

```
CREATE EXTERNAL TABLE hive_purchases(customerId bigint, total_cost double,
 items_purchased array<String>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "Purchases",
"dynamodb.column.mapping" =
 "customerId:CustomerId,total_cost:Cost,items_purchased:Items");

CREATE EXTERNAL TABLE hive_customers(customerId bigint, customerName string,
 customerAddress array<String>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "Customers",
"dynamodb.column.mapping" =
 "customerId:CustomerId,customerName:Name,customerAddress:Address");

Select c.customerId, c.customerName, count(*) as count from hive_customers c
JOIN hive_purchases p ON c.customerId=p.customerId
GROUP BY c.customerId, c.customerName HAVING count > 2;
```

联接来自不同源的两个表

- 在以下示例中，Customer_S3 是加载了 Amazon S3 中存储的 CSV 文件的 Hive 表，而 hive_purchases 是引用了 DynamoDB 中的数据的表。以下示例将 Amazon S3 中以 CSV 文件格式存储的客户数据与 DynamoDB 中存储的订单数据连接在一起，以返回一组数据，这些数据表示名称中包含“Miller”的客户提交的订单。

```
CREATE EXTERNAL TABLE hive_purchases(customerId bigint, total_cost double,
 items_purchased array<String>)
STORED BY 'org.apache.hadoop.hive.dynamodb.DynamoDBStorageHandler'
TBLPROPERTIES ("dynamodb.table.name" = "Purchases",
"dynamodb.column.mapping" =
 "customerId:CustomerId,total_cost:Cost,items_purchased:Items");

CREATE EXTERNAL TABLE Customer_S3(customerId bigint, customerName string,
 customerAddress array<String>)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION 's3://bucketname/path/subpath/';
```

```
Select c.customerId, c.customerName, c.customerAddress from
Customer_S3 c
JOIN hive_purchases p
ON c.customerid=p.customerid
where c.customerName like '%Miller%';
```

> **Note**
>
> 在上述示例中，为了提高清晰性和完整性，在每个示例中均包括了 CREATE TABLE 语句。针对给定 Hive 表运行多个查询或执行导出操作时，只需在 Hive 会话的开始创建表一次即可。

在 DynamoDB 中优化 Amazon EMR 操作的性能

对 DynamoDB 表进行的 Amazon EMR 操作作为读取操作，并受表预置的吞吐量设置约束。Amazon EMR 实施其逻辑来努力平衡 DynamoDB 表上的负载，从而最大程度降低超出您预置的吞吐量的可能性。每个 Hive 查询结束时，Amazon EMR 均返回有关用于处理查询的集群的信息，包括超出预置的吞吐量的次数。您可以使用此信息以及有关 DynamoDB 吞吐量的 CloudWatch 指标，在后续请求中更好地管理 DynamoDB 表上的负载。

以下因素会影响 Hive 在处理 DynamoDB 表时的查询性能。

预置的读取容量单位

对 DynamoDB 表运行 Hive 查询时，您需要确保已预置足量的读取容量单位。

例如，假设您为 DynamoDB 表预配置了 100 个读取容量单位。这将允许您每秒执行 100 次读取或读取 409600 字节。如果该表包含 20GB 的数据（21474836480 字节）并且您的 Hive 查询执行全表扫描，则可以估算执行查询将花费多长时间：

21474836480/409600 = 52429 秒 = 14.56 小时

减少所需时间的唯一方法是调整源 DynamoDB 表的读取容量单位。将更多节点添加到 Amazon EMR 集群不会有所帮助。

在 Hive 输出中，当一个或多个映射器进程已完成时，将更新完成百分比。对于预配置的读取容量设置较低的大型 DynamoDB 表，完成百分比输出可能会很长时间不更新；在上面的示例中，作业将在几个小时内显示为完成 0%。有关作业进度的详细状态，请转到 Amazon EMR 控制台；您将可以查看单个映射器任务状态和数据读取统计数据。

您还可以登录主节点的 Hadoop 界面，查看 Hadoop 统计数据。该界面将向您显示单个映射任务状态和一些数据读取统计数据。有关更多信息，请参阅《Amazon EMR 管理指南》中的托管在主节点上的 Web 页面。

读取百分比设置

默认情况下，Amazon EMR 根据当前的预置吞吐量管理对您的 DynamoDB 表的请求负载。但是，当 Amazon EMR 返回的作业相关信息中包括预置的吞吐量远远超出响应数时，您可以在设置 Hive 表时使用 `dynamodb.throughput.read.percent` 参数调整默认读取速率。有关设置读取百分比参数的更多信息，请参阅 Hive 选项。

写入百分比设置

默认情况下，Amazon EMR 根据当前的预置吞吐量管理对您的 DynamoDB 表的请求负载。但是，当 Amazon EMR 返回的作业相关信息中所含预置的吞吐量远远超出响应数时，您可以在设置 Hive 表时使用 `dynamodb.throughput.write.percent` 参数调整默认写入速率。有关设置写入百分比参数的更多信息，请参阅Hive 选项。

重试持续时间设置

默认情况下，如果在两分钟（默认重试时间间隔）内没有返回结果，Amazon EMR 将重新运行 Hive 查询。在运行 Hive 查询时，您可以通过设置 `dynamodb.retry.duration` 参数来调整此时间间隔。有关设置写入百分比参数的更多信息，请参阅Hive 选项。

映射任务数

Hadoop 为了处理导出和查询 DynamoDB 中所存储数据的请求而启动的映射器守护进程的读取速率控制在每秒最多 1 MiB 之内，以限制所用的读取容量。如果在 DynamoDB 上有更多预置的吞吐量可用，则可以通过增加映射器守护进程数来改善 Hive 导出和查询操作的性能。为此，您可以增加中的 EC2 实例数，或者增加每个 EC2 实例上运行的映射器守护进程数。

您可以通过停止当前集群，然后使用更大的 EC2 实例数重新启动它，来增加该集群中的 EC2 实例数。可在 Configure EC2 Instances (配置 EC2 实例) 对话框中指定 EC2 实例数（如果您从 Amazon EMR 控制台启动集群），也可以使用 `--num-instances` 选项指定 EC2 实例数（如果您从 CLI 启动）。

实例上运行的映射任务数取决于 EC2 实例类型。有关受支持 EC2 实例类型及每种实例类型提供的映射器数的更多信息，请参阅 任务配置。其中，每个受支持的配置都有一个“任务配置”部分。

增加映射器守护程序数的另一个方法是，将 Hadoop 的 `mapreduce.tasktracker.map.tasks.maximum` 配置参数更改为更大的值。此方法的优点是无需增加 EC2 实例的数量或大小即可为您提供更多映射器，从而

为您节省资金。缺点是将此值设置得过大可能导致集群中的 EC2 实例用尽内存。
要设置 mapreduce.tasktracker.map.tasks.maximum，请启动集群并为 mapreduce.tasktracker.map.tasks.maximum 指定一个值，作为 mapred-site 配置分类的属性。如下例所示。有关更多信息，请参阅配置应用程序。

```
{
    "configurations": [
    {
        "classification": "mapred-site",
        "properties": {
            "mapred.tasktracker.map.tasks.maximum": "10"
        }
    }
    ]
}
```

并行数据请求

从多个用户或多个应用程序向单个表发出的多个数据请求可能会耗尽预配置的读取吞吐量并降低性能。

处理持续时间

DynamoDB 中的数据一致性取决于在每个节点上执行读取和写入操作的顺序。当正在进行 Hive 查询时，其它应用程序可能会将新数据加载到 DynamoDB 表，或者修改或删除现有数据。在这种情况下，Hive 查询的结果可能无法反映查询运行时对数据所做的更改。

避免超出吞吐量

针对 DynamoDB 运行 Hive 查询时，请注意不要超出您的预置吞吐量，因为这会用尽应用程序调用 DynamoDB::Get 时所需的容量。为确保不会发生这种情况，您应通过查看日志并在 Amazon CloudWatch 中监控指标，定期监控读取量并限制应用程序对 DynamoDB::Get 的调用。

请求时间

调度 Hive 查询以便在对 DynamoDB 表的需求较低时访问 DynamoDB 表，可以改善性能。举例来说，如果应用程序的大多数用户住在旧金山，您可以选择在太平洋标准时间凌晨 4 点导出每日数据。（此时，大多数用户都已睡着且未在更新 DynamoDB 数据库中的记录）

基于时间的表

如果将数据组织为一系列基于时间的 DynamoDB 表（例如，每天一个表），您可以在该表不再处于活动状态时导出数据。您可以利用此方法将数据持续备份到 Amazon S3 中。

已存档数据

如果您计划针对 DynamoDB 中存储的数据运行多个 Hive 查询，并且您的应用程序可以接纳已存档数据，那么您可能会希望将数据导出到 HDFS 或 Amazon S3，然后针对数据的副本（而非 DynamoDB）运行 Hive 查询。这将节省读取操作和预配置的吞吐量。

Kinesis

Amazon EMR 集群可以使用 Hadoop 生态系统中的熟悉工具（如 Hive、Pig、MapReduce、Hadoop Streaming API 和 Cascading）直接读取和处理 Amazon Kinesis 流。您还可以将 Amazon Kinesis 中的实时数据与正在运行的集群中 Amazon S3、Amazon DynamoDB 和 HDFS 上的现有数据进行连接。您可以直接将 Amazon EMR 中的数据加载到 Amazon S3 或 DynamoDB 来进行后处理。有关 Amazon Kinesis 服务亮点和定价的信息，请参阅 Amazon Kinesis。

可以对 Amazon EMR 和 Amazon Kinesis 集成执行哪些操作？

Amazon EMR 和 Amazon Kinesis 之间的集成使某些方案更简单，例如：

- 流式处理日志分析 – 您可以分析流式处理 Web 日志，以便每隔几分钟按区域、浏览器和访问域生成前 10 个错误类型的列表。
- 客户参与 – 您可以编写查询将 Amazon Kinesis 中的点击流数据与存储在 DynamoDB 表中的广告活动信息进行连接，以确定显示在特定网站上的最有效广告类别。
- 即席交互式查询 – 您可以定期将 Amazon Kinesis 流中的数据加载到 HDFS 中，并以本地 Impala 表的形式提供该数据以进行快速的交互式分析查询。

对 Amazon Kinesis 流进行检查点分析

用户可以定期对 Amazon Kinesis 流进行批量分析，这些分析称为迭代。因为使用序列号检索 Amazon Kinesis 流数据记录，所以，可通过 Amazon EMR 在 DynamoDB 表中存储的开始和结束序列号来定义迭代边界。例如，当 `iteration0` 结束时，它在 DynamoDB 中存储结束序列号，这样在 `iteration1` 作业开始时，它可以检索流的后续数据。迭代在流数据中的这种映射称为检查点操作。有关更多信息，请参阅 Kinesis 连接器。

如果对迭代进行了检查点操作且作业未能处理某个迭代，则 Amazon EMR 会尝试重新处理该迭代中的记录。

通过检查点功能，您可以：

- 从运行于相同的流和逻辑名称之上的前一个查询处理的序列号之后，开始数据处理
- 重新处理 Kinesis 中由之前的查询处理的同一批数据

要启用检查点操作，请在脚本中将 `kinesis.checkpoint.enabled` 参数设置为 `true`。此外，请配置以下参数：

配置设置	描述
kinesis.checkpoint.metastore.table.name	用于存储检查点信息的 DynamoDB 表名称
kinesis.checkpoint.metastore.hash.key.name	DynamoDB 表的哈希键名称
kinesis.checkpoint.metastore.hash.range.name	DynamoDB 表的范围键名称
kinesis.checkpoint.logical.name	当前处理的逻辑名称
kinesis.checkpoint.iteration.no	与逻辑名称关联的处理的迭代编号
kinesis.rerun.iteration.without.wait	用来指示是否可以重新运行失败的迭代而不等待超时的布尔值；默认值为 `false`

Amazon DynamoDB 表的预置 IOPS 建议

Amazon Kinesis 的 Amazon EMR 连接器使用 DynamoDB 数据库作为对元数据进行检查点操作的支持。必须先在 DynamoDB 中创建表，才能以检查点时间间隔使用 Amazon EMR 集群的 Amazon Kinesis 流中的数据。该表必须与 Amazon EMR 集群位于相同区域中。以下是为您应当为 DynamoDB 表预置的 IOPS 数的一般建议；`j` 应当是可同时运行的最大 Hadoop 任务数（具有不同的逻辑名称+迭代编号组合），`s` 是任何作业将处理的最大分片数：

对于 Read Capacity Units：`j*s/5`

对于 Write Capacity Units：`j*s`

性能注意事项

Amazon Kinesis 分片吞吐量与 Amazon EMR 集群中节点的实例大小以及流中的记录大小成正比。建议在主节点和核心节点上使用 m5.xlarge 或更大的实例。

借助 Amazon EMR 安排 Amazon Kinesis 分析

如果要对活动 Amazon Kinesis 流分析数据，由于任何迭代都受超时和最长持续时间限制，您应经常运行分析，以便从流定期收集详细信息，这十分重要。可以通过多种方式定期执行该类脚本和查询；但建议针对此类周期性任务使用 Amazon Data Pipeline。有关更多信息，请参阅《Amazon Data Pipeline 开发人员指南》中的 Amazon Data Pipeline PigActivity 和 Amazon Data Pipeline HiveActivity。

S3DistCp (s3-dist-cp)

Apache DistCp 是一款开源工具，可以用于复制大量数据。S3DistCp 类似于 DistCp，但经过优化后可以处理 Amazon，尤其是 Amazon S3。Amazon EMR 4.0 及更高版本中的 S3DistCp 命令为 `s3-dist-cp`，您可以将其作为集群或命令行中的步骤添加。使用 S3DistCp 能有效地从 Amazon S3 复制大量数据到 HDFS，供 Amazon EMR 集群中的后续步骤进行处理。您还可以使用 S3DistCp 在 Amazon S3 存储桶之间或从 HDFS 向 Amazon S3 复制数据。若要在存储桶之间和 Amazon 账户之间并行复制大量对象，S3DistCp 可扩展性更强，也更高效。

有关演示 S3DistCP 在实际场景中灵活性的特定命令，请参阅 Amazon 大数据博客中的 Seven tips for using S3DistCp。

与 DistCp 一样，S3DistCp 使用 MapReduce 以分布式方式进行复制。它在几个服务器之间共享复制、错误处理、恢复和报告任务。有关 Apache DistCp 开源项目的更多信息，请参阅 Apache Hadoop 文档中的 DistCp 指南。

如果 S3DistCp 无法复制部分或全部指定文件，集群步骤会失败，并返回一个非零错误代码。如果发生此情况，S3DistCp 将不会清理部分复制的文件。

> **Important**
>
> S3DistCp 不支持包含下划线字符的 Amazon S3 存储桶名称。
> S3DistCp 不支持连结 Parquet 文件。请改用 PySpark。有关更多信息，请参阅在 Amazon EMR 中串连 parquet 文件。
> 为避免在使用 S3DistCP 将单个文件（而不是目录）从 S3 复制到 HDFS 时出现复制错误，请使用 Amazon EMR 版本 5.33.0 或更高版本或 Amazon EMR 6.3.0 或更高版本。

S3DistCp 选项

虽然 S3DistCp 与 DistCp 类似，但前者支持不同的选项组来改变复制和压缩数据的方式。

调用 S3DistCp 时，您可以指定下表中所述的选项。选项是用参数列表添加到步骤的。下表是 S3DistCp 参数的示例。

选项	描述	必填
`--src=LOCATION`	待复制数据的位置。可以是 HDFS 或 Amazon S3 位置。 示例：`--src=s3://` *`DOC-EXAMPLE-BUCKET1`* `/logs/j-3GYXXXXXX9IOJ/node` **Important** S3DistCp 不支持包含下划线字符的 Amazon S3 存储桶名称。	是
`--dest=LOCATION`	数据的目标位置。可以是 HDFS 或 Amazon S3 位置。 示例：`--dest=hdfs:///output` **Important** S3DistCp 不支持包含下划线字符的 Amazon S3 存储桶名称。	是
`--srcPattern=PATTERN`	该正则表达式会筛选 `--src` 的部分数据的复制操作。如果指定的既不是 `--srcPattern` 也不是 `--groupBy` ，那么会将 `--src` 的所有数据复制到 `--dest`。 如果正则表达式参数包含特殊字符，如星号（`*`），那么必须将正则表达式或整个 `--args` 字符串括在单引号（`'`）中。 示例：`--srcPattern=.*daemons.*-hadoop-.*`	否

选项	描述	必填
`--groupBy=PATTERN`	该正则表达式让 S3DistCp 连接匹配该表达式的文件。例如，您可以使用此选项把一个小时内写入的所有日志文件组合成为单个文件。已连接文件的文件名是与该分组的正则表达式相匹配的值。 圆括号表明文件应使用的分组方式，与圆括号内的语句匹配的所有项目都组合成单个输出文件。如果正则表达式不包括圆括号内的语句，则集群将在执行 S3DistCp 步骤时失败，并返回错误。 如果正则表达式参数包含特殊字符，如星号 (*)，那么必须将正则表达式或整个 `--args` 字符串括在单引号 (') 中。 如果指定 `--groupBy` ，则仅复制与指定的模式匹配的文件。您不需要同时指定 `--groupBy` 和 `--srcPattern` 。 示例：`--groupBy=.*subnetid.*([0-9]+-[0-9]+-[0-9]+-[0-9]+).*`	否
`--targetSize=SIZE`	要根据 `--groupBy` 选项创建的文件的大小，以兆字节 (MiB) 为单位。此值必须是整数。如果设置了 `--targetSize` ，则 S3DistCp 尝试匹配此大小；所复制的文件的实际大小可能大于或小于此值。基于数据文件大小聚合任务，因此目标文件大小将可能匹配源数据文件大小。 如果由 `--groupBy` 连接的文件大于 `--targetSize` 的值，则将这些文件分解成部分文件，并在末尾附加一个数值按顺序命名。例如，连接组成 `myfile.gz` 的一个文件将被分解成部分文件，如：`myfile0.gz` 、`myfile1.gz` 等。 示例：`--targetSize=2`	否

选项	描述	必填
--appendToLastFile	指定将文件从 Amazon S3 复制到已存在的 HDFS 时 S3DistCp 的行为。它向现有文件附加新文件数据。如果将 --appendToLastFile 与 --groupBy 结合使用，则将新数据附加到与相同的组匹配的文件。与 --targetSize 结合使用时，此选项也具有 --groupBy。 行为	否
--outputCodec=CODEC	指定用于所复制文件的压缩编解码器。取值可以是：gzip、gz、lzo、snappy 或 none。您可以使用此选项，例如，把 Gzip 压缩的输入文件转换为 LZO 压缩的输出文件，或对文件进行解压缩，作为复制操作的一部分。如果选择输出编解码器，则系统会在文件名末尾追加适当的扩展名 (例如，对于 gz 和 gzip，将追加扩展名 .gz)。如果不为 --outputCodec 指定值，则复制的文件不会出现压缩方面的变化。 示例：--outputCodec=lzo	否
--s3ServerSideEncryption	确保目标数据使用 SSL 传输，并在 Amazon S3 中使用 Amazon 服务端密钥自动加密。使用 S3DistCp 检索数据时，不会自动取消加密对象。如果尝试将未加密的对象复制到需要加密的 Amazon S3 存储桶，则操作将失败。有关更多信息，请参阅使用数据加密。 示例：--s3ServerSideEncryption	否
--deleteOnSuccess	如果复制操作成功，此选项会让 S3DistCp 从源位置删除已复制的文件。如果您以计划任务的形式将输出文件 (如日志文件) 从一个位置复制到另一个位置，又不想复制两次相同的文件，那么这个选项会非常有用。 示例：--deleteOnSuccess	否

选项	描述	必填
`--disableMultipartUpload`	禁用分段上载。 示例：`--disableMultipartUpload`	否
`--multipartUploadChunkSize=SIZE`	Amazon S3 分段上传中每个分段的大小，以 MiB 为单位。S3DistCp 在复制的数据大于 `multipartUploadChunkSize` 时使用分段上传。要提高作业性能，可以增加每个分段的大小。默认大小为 128 MiB。 示例：`--multipartUploadChunkSize=1000`	否
`--numberFiles`	在输出文件之前加上序号。计数从 0 开始，除非 `--startingIndex` 指定一个不同的值。 示例：`--numberFiles`	否
`--startingIndex=INDEX`	使用 `--numberFiles` 指定序列中的第一个数字。 示例：`--startingIndex=1`	否
`--outputManifest=FILENAME`	创建使用 Gzip 压缩的文本文件，其中包含由 S3DistCp 复制的所有文件的列表。 示例：`--outputManifest=manifest-1.gz`	否
`--previousManifest=PATH`	读取清单文件，该文件使用 `--outputManifest` 标志在此前调用 S3DistCp 期间创建。设置 `--previousManifest` 标志时，S3DistCp 从复制操作中排除清单所列文件。如果同时指定 `--outputManifest` 和 `--previousManifest`，则之前清单中列出的文件也会出现在新的清单文件中，但不会复制这些文件。 示例：`--previousManifest=/usr/bin/manifest-1.gz`	否

选项	描述	必填
`--requirePreviousManifest`	需要对 S3DistCp 进行的上一次调用期间创建的之前清单。如果将它设置为 false，则不指定之前的清单也不会生成错误。默认值为 true。	否
`--copyFromManifest`	反转 `--previousManifest` 的行为，让 S3DistCp 使用指定的清单文件作为待复制文件的列表，代替复制时要排除的文件的列表。 示例：`--copyFromManifest --previousManifest=/usr/bin/manifest-1.gz`	否
`--s3Endpoint=ENDPOINT`	指定上载文件时要使用的 Amazon S3 终端节点。此选项会同时设置源位置和目标位置的终端节点。如果未设置，则默认终端节点是 `s3.amazonaws.com` 。有关 Amazon S3 终端节点的列表，请参阅区域和终端节点。 示例：`--s3Endpoint=s3.eu-west-1.amazonaws.com`	否
`--storageClass=CLASS`	目标为 Amazon S3 时要使用的存储类。有效值是 STANDARD 和 REDUCED_REDUNDANCY。如果不指定此选项，则 S3DistCp 尝试保留存储类。 示例：`--storageClass=STANDARD`	否

选项	描述	必填
`--srcPrefixesFile=PATH`	Amazon S3（s3://）、HDFS（hdfs:///）或本地文件系统（file:/）中包含 `src` 前缀的文本文件（每行一个前缀）。 如果提供了 `srcPrefixesFile` ，S3DistCp 不会列出 src 路径。相反，它将生成一个源列表，以作为列出在此文件中指定的所有前缀的组合结果。与 src 相比较的相对路径 (而不是这些前缀) 将用于生成目标路径。如果还指定了 `srcPattern` ，则会将它应用于源前缀的组合列表结果以进一步筛选输入。如果使用了 `copyFromManifest` ，则会复制清单中的对象并忽略 `srcPrefixesFile` 。 示例：`--srcPrefixesFile=PATH`	否

除了上述选项外，S3DistCp 还执行 Tool interface，也就是说，支持通用选项。

添加 S3DistCp 作为集群中的步骤

您可以添加 S3DistCp 作为集群中的步骤进行调用。可以使用控制台、CLI 或 API 在启动时向集群或是向正在运行的集群添加步骤。以下示例说明如何向正在运行的集群添加 S3DistCp 步骤。有关向集群添加步骤的更多信息，请参阅《Amazon EMR 管理指南》中的向集群提交工作。

使用 Amazon CLI 向正在运行的集群添加 S3DistCp 步骤

有关在 Amazon CLI 中使用 Amazon EMR 命令的更多信息，请参阅 Amazon CLI 命令参考。

- 要向集群添加调用 S3DistCp 的步骤，请将实参传递给指定 S3DistCp 应如何执行复制操作的形参。

  以下示例将守护进程日志从 Amazon S3 复制到 `hdfs:///output`。在以下命令中：

  - `--cluster-id` 指定集群
  - `Jar` 是 S3DistCp JAR 文件的位置。如需有关如何使用 command-runner.jar 在群集上运行命令的示例，请参阅提交自定义 JAR 步骤以运行脚本或命令。

- `Args` 是传递到 S3DistCp 的选项名称–值对的逗号分隔的列表。有关可用选项的完整列表，请参阅S3DistCp 选项。

要向正在运行的集群添加 S3DistCp 复制步骤，请在某个 JSON 文件（本例中为 *myStep.json*）中输入以下内容并保存到 Amazon S3 或本地文件系统中。将 *j-3GYXXXXXX9IOK* 替换为您的集群 ID，并将 *mybucket* 替换为您的 Amazon S3 存储桶名称。

```
[
    {
        "Name":"S3DistCp step",
        "Args":["s3-dist-cp","--s3Endpoint=s3.amazonaws.com","--src=s3://mybucket/
logs/j-3GYXXXXXX9IOJ/node/","--dest=hdfs:///output","--srcPattern=.*[a-zA-Z,]+"],
        "ActionOnFailure":"CONTINUE",
        "Type":"CUSTOM_JAR",
        "Jar":"command-runner.jar"
    }
]
```

```
aws emr add-steps --cluster-id j-3GYXXXXXX9IOK --steps file://./myStep.json
```

Example 从 Amazon S3 向 HDFS 复制日志文件

此示例还说明如何通过向正在运行的集群添加步骤来将 Amazon S3 存储桶中存储的日志文件复制到 HDFS 中。在此示例中，`--srcPattern` 选项用于限制复制到守护程序日志的数据。

要使用 `--srcPattern` 选项将日志文件从 Amazon S3 复制到 HDFS，请在某个 JSON 文件（本例中为 *myStep.json*）中输入以下内容并保存到 Amazon S3 或本地文件系统中。将 *j-3GYXXXXXX9IOK* 替换为您的集群 ID，并将 *mybucket* 替换为您的 Amazon S3 存储桶名称。

```
[
    {
        "Name":"S3DistCp step",
        "Args":["s3-dist-cp","--s3Endpoint=s3.amazonaws.com","--src=s3://mybucket/logs/
j-3GYXXXXXX9IOJ/node/","--dest=hdfs:///output","--srcPattern=.*daemons.*-hadoop-.*"],
        "ActionOnFailure":"CONTINUE",
        "Type":"CUSTOM_JAR",
        "Jar":"command-runner.jar"
    }
]
```

在 S3DistCP 作业失败后清理

如果 S3DistCp 无法复制部分或全部指定文件，则此命令或集群步骤会失败，并返回一个非零错误代码。如果发生此情况，S3DistCp 将不会清理部分复制的文件。您必须手动删除它们。

部分复制的文件保存到具有 S3DistCp 作业唯一标识符的子目录中的 HDFS `tmp` 目录下。您可以在作业的标准输出中找到此 ID。

例如，对于具有 ID `4b1c37bb-91af-4391-aaf8-46a6067085a6` 的 S3DistCP 作业，您可以连接到集群的主节点，并运行以下命令以查看与作业关联的输出文件。

```
hdfs dfs -ls /tmp/4b1c37bb-91af-4391-aaf8-46a6067085a6/output
```

该命令将返回与以下类似的文件列表：

```
Found 8 items
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:03 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/_SUCCESS
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:02 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00000
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:02 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00001
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:02 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00002
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:03 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00003
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:03 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00004
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:03 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00005
-rw-r--r--   1 hadoop hadoop          0 2018-12-10 06:03 /tmp/4b1c37bb-91af-4391-
aaf8-46a6067085a6/output/part-r-00006
```

然后，您可以运行以下命令来删除目录和所有内容。

```
hdfs dfs rm -rf /tmp/4b1c37bb-91af-4391-aaf8-46a6067085a6
```

在 Amazon EMR 集群上运行命令和脚本

本主题介绍如何作为步骤在集群上运行命令或脚本。作为步骤运行命令或脚本是将工作提交到集群的其中一种方式，在以下情况下非常有用：

- 当您未允许 SSH 访问您的 Amazon EMR 集群时
- 当你想运行 bash 或 shell 命令来排除集群故障时

您可以在创建集群时或者在集群处于 WAITING 状态时运行脚本。要在步骤处理开始前运行脚本，您可使用引导操作。有关引导操作更多信息，请参阅《Amazon EMR 管理指南》中的创建引导操作以安装其它软件。

Amazon EMR 提供以下工具来帮助您运行脚本、命令和其他集群上的程序。您可以使用 Amazon EMR 管理控制台或 Amazon CLI 调用所有工具。

`command-runner.jar`

位于集群的 Amazon EMR AMI 上。您可以使用 `command-runner.jar` 在集群上运行命令。你没有使用其完整路径指定 `command-runner.jar`。

`script-runner.jar`

Amazon S3 托管在 `s3://<region>.elasticmapreduce/libs/script-runner/script-runner.jar`，其中 *<region>* 是 Amazon EMR 集群所在的区域。您可以使用 `script-runner.jar` 以在集群上运行本地或 Amazon S3 上保存的脚本。当您提交步骤时，您必须指定 `script-runner.jar` 的完整 URI。

提交自定义 JAR 步骤以运行脚本或命令

以下Amazon CLI 示例说明了 Amazon EMR 的 `command-runner.jar` 和 `script-runner.jar` 一些常见使用场景。

Example ：使用 **`command-runner.jar`** 在集群上运行命令

当您使用 `command-runner.jar`时，您可以在步骤的参数列表中指定命令、选项和值。

以下 Amazon CLI 示例将步骤提交至调用 `command-runner.jar` 的运行中的集群。Args 列表指定的命令下载名为*`my-script.sh`*的脚本，下载方式是从 Amazon S3 进入 hadoop 用户主目录。然后，该命令修改脚本权限并运行*`my-script.sh`*。

当您使用 Amazon CLI时，Args 列表中的项目应该用逗号分隔开来，列表元素之间没有空格。例如，使用 Args=[example-command,example-option,"example option value"] 而不是 Args=[example-command, example-option, "example option value"]。

```
aws emr add-steps \
--cluster-id j-2AXXXXXXGAPLF \
--steps Type=CUSTOM_JAR,Name="Download a script from S3, change its permissions, and
 run it",ActionOnFailure=CONTINUE,Jar=command-runner.jar,Args=[bash,-c,"aws s3 cp s3://
EXAMPLE-DOC-BUCKET/my-script.sh /home/hadoop; chmod u+x /home/hadoop/my-script.sh; cd /
home/hadoop; ./my-script.sh"]
```

Example ：使用 **script-runner.jar**在集群上运行脚本

当您使用 script-runner.jar时，在步骤的参数列表中指定想要运行的脚本。

以下 Amazon CLI 示例将步骤提交至调用 script-runner.jar 的运行中的集群。在此情况下，称为 *my-script.sh* 的脚本存储在 Amazon S3。您还可以指定存储在集群主节点 (master node) 的本地脚本。

```
aws emr add-steps \
--cluster-id j-2AXXXXXXGAPLF \
--steps Type=CUSTOM_JAR,Name="Run a script from S3 with script-
runner.jar",ActionOnFailure=CONTINUE,Jar=s3://us-west-2.elasticmapreduce/libs/script-
runner/script-runner.jar,Args=[s3://EXAMPLE-DOC-BUCKET/my-script.sh]
```

其他使用 command-runner.jar 的方法

您还可以借助类似于 spark-submit 或 hadoop-streaming 的工具，使用 command-runner.jar 将工作提交至集群。当您使用 command-runner.jar 启动应用程序时，您指定 CUSTOM_JAR 作为步骤类型而不是使用类似于 SPARK、STREAMING 或者 PIG 的值。工具可用性取决于您在集群上安装的应用程序。

以下示例命令借助 spark-submit 使用command-runner.jar 提交步骤。Args 列表指定 spark-submit 作为命令，接着是附加参数和值的 Spark 应用程序 *my-app.py* Amazon S3 URI。

```
aws emr add-steps \
--cluster-id j-2AXXXXXXGAPLF \
--steps Type=CUSTOM_JAR,Name="Run spark-submit using command-
runner.jar",ActionOnFailure=CONTINUE,Jar=command-runner.jar,Args=[spark-submit,S3://
DOC-EXAMPLE-BUCKET/my-app.py,ArgName1,ArgValue1,ArgName2,ArgValue2]
```

下表列出了通过 command-runner.jar 您可以使用的其他工具。

工具名称	描述
hadoop-streaming	提交 Hadoop 流式处理程序。在控制台和一些开发工具包中，这是流步骤。
hive-script	运行 Hive 脚本。在控制台和开发工具包中，这是 Hive 步骤。
pig-script	运行 Pig 脚本。在控制台和开发工具包中，这是 Pig 步骤。
spark-submit	运行 Spark 应用程序。在控制台中，这是 Spark 步骤。
hadoop-lzo	在目录上运行 Hadoop LZO 索引器。
s3-dist-cp	将大量数据从 Amazon S3 分布式复制到 HDFS。有关更多信息，请参阅S3DistCp (s3-dist-cp)。

Amazon 术语表

有关最新的 Amazon 术语，请参阅《Amazon Web Services 词汇表参考》中的 Amazon 词汇表。