弹性 MapReduce

EMR 开发指南

产品文档

腾讯云

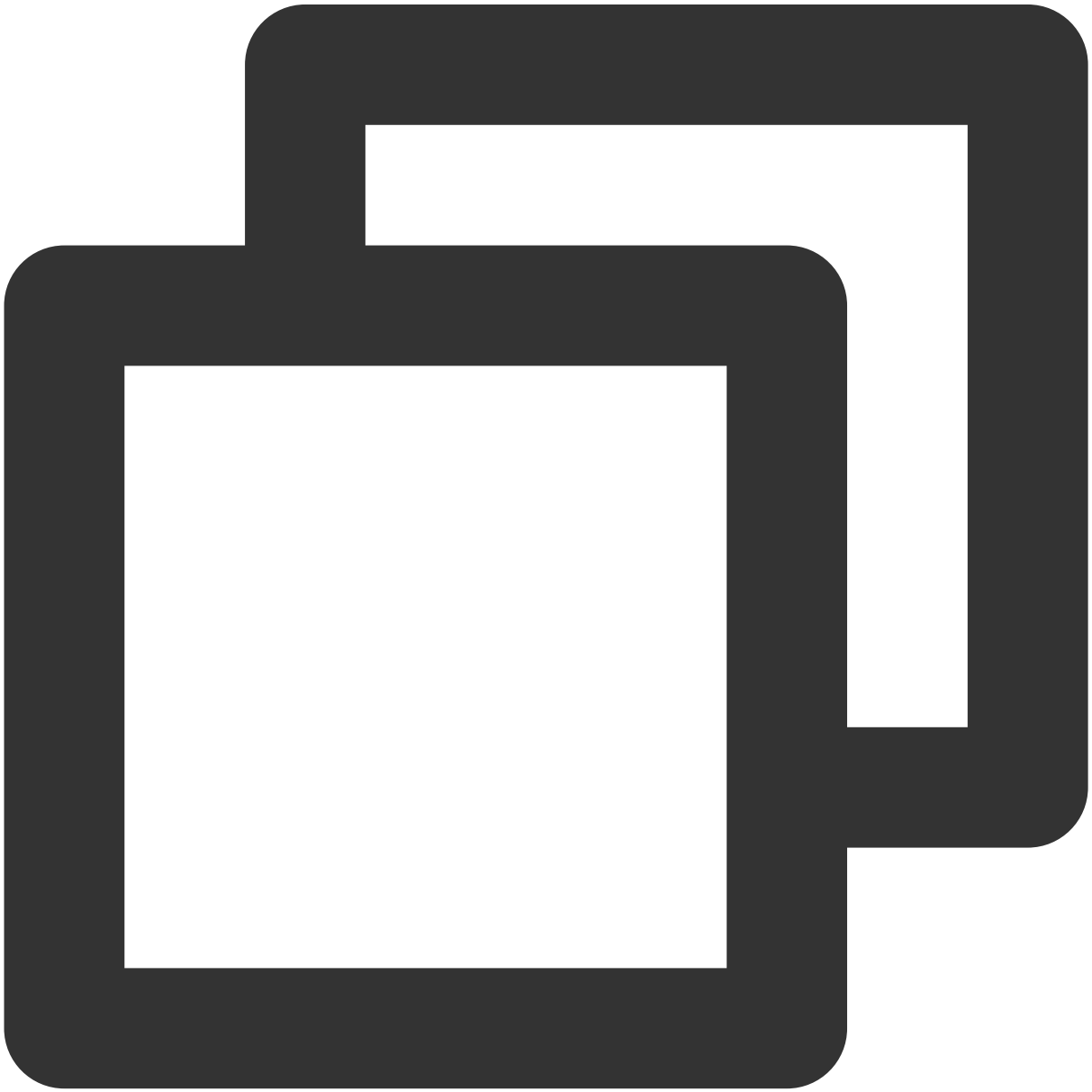
文档目录

RSS 开发指南

EMR 开发指南
Hadoop开发指南
HDFS 常见操作

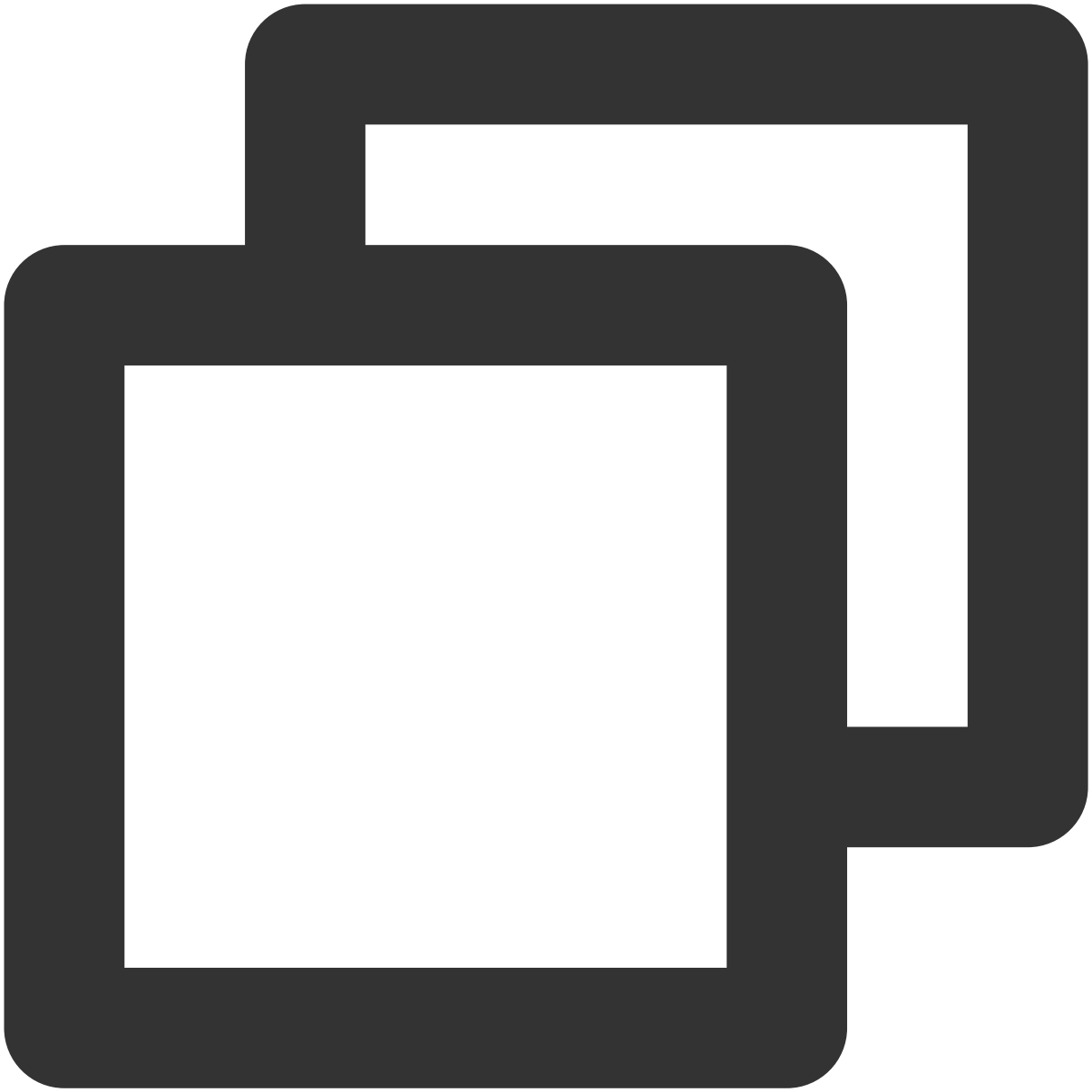
最近更新时间：2021-07-13 11:42:03

腾讯云 EMR 的 Hadoop 集成了腾讯云对象存储，如果您在购买的时候勾选了支持 COS，那么您也可以通过常见的 hadoop 命令操作 COS 上的数据。可通过如下命令操作集群中的数据。



```
#cat 数据
hadoop fs -cat /usr/hive/warehouse/hivewithhdfs.db/record/data.txt
#修改目录或者文件权限
hadoop fs -chmod -R 777 /usr
#改变文件或者目录 owner
hadoop fs -chown -R root:root /usr
#创建文件夹
hadoop fs -mkdir <paths>
#本地文件发送到 HDFS 上
hadoop fs -put <localsrc> ... <dst>
#拷贝本地文件到 HDFS 上
hadoop fs -copyFromLocal <localsrc> URI
#查看文件或者目录的存储使用量
hadoop fs -du URI [URI …]
#删除文件
hadoop fs -rm URI [URI …]
#设置目录或者文件的拷贝数
hadoop fs-setrep [-R] [-w] REP PATH [PATH …]
#检查集群文件坏块
hadoop fsck <path> [-move | -delete | -openforwrite] [-files [-blocks [-locations |
```

更多 HDFS 命令请参考 社区文档，此外如果您的集群是 HA 集群（双 namenode），您可以通过如下命名查看哪个 namenode 是 active 的。



```
#nn1 是 namenode 的 ID，一般为 nn1 和 nn2
hdfs haadmin -getServiceState nn1
#查看当前集群报告
hdfs dfsadmin -report
#namenode 离开安全模式
hdfs dfsadmin -safemode leave
```

HDFS 联邦管理开发指南

最近更新时间：2023-07-12 10:39:36

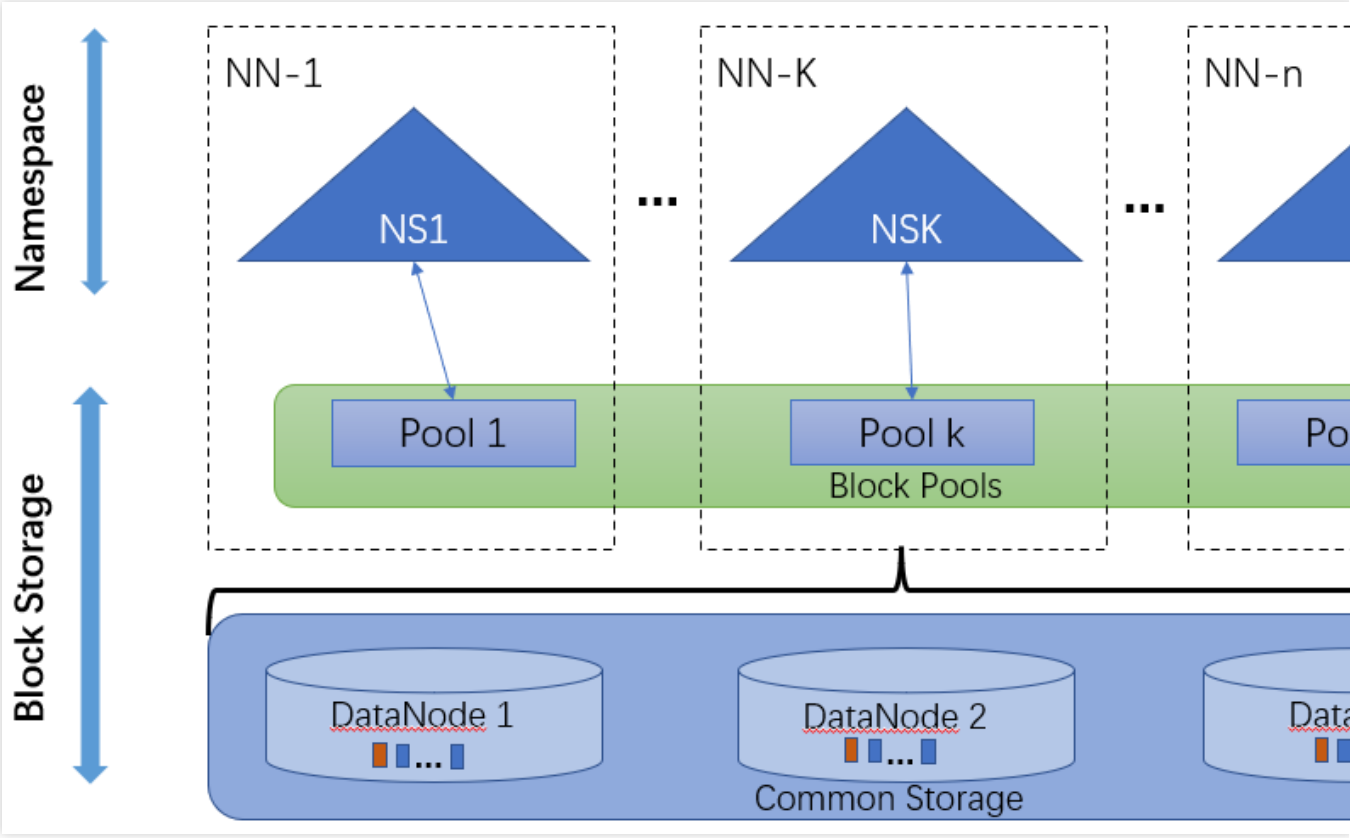
HDFS 联邦管理类型选择

注意

当前 HDFS 联邦管理为白名单开放，如需要可联系工单开通。

HDFS 联邦管理架构

HDFS Federation（联邦）使用多个独立的 NameNode/Namespace 来使 HDFS 的命名服务能够水平扩展，联邦中的 DataNode 被所有的 NameNode 用作公共存储块的地方；每一个 DataNode 都会向所在集群中所有的 NameNode 注册，并且会周期性地发送心跳和块信息报告，同时处理来自所有 NameNode 的指令。



有关更多 HDFS Federation介绍，请参阅 Hadoop 文档。

ViewFs Federation 原理

为了方便管理多个命名空间，HDFS ViewFs Federation 采用了经典的 Client Side Mount Table（社区 ViewFs 功能）。通过客户端侧将应用的不同路径映射到具体的 NameService 上，从而达成存储分离或者性能分离的场景。

HDFS ViewFs Federation 采用 Client-Side Mount Table 分摊文件和负载，该方法更多的需要人工介入（明确的规划）以达到理想的负载均衡。

Router-based Federation 原理

Router-based Federation 提供了一个软件层来管理多个 NameSpace 空间。相较于 ViewFs 通过在客户端维护挂载表信息，Router-based Federation 是真正做到了对客户端的完全透明。因为这部分映射信息将会被额外的保存下来，还会持久化下来。

HDFS 联邦管理配置

主要是要考虑两个方面：所需的 NameService 数量，业务数据目录挂载方式。

NameService 数量的规划原则

1. 单 NameService 的安全存储容量上限为1亿文件，如果超过该数量，该 NameService 的访问速度、读写吞吐量都会严重降低。所以，如果文件存储量（预计）超过1亿文件，则需要新增 NameService。
2. 对于 HDFS 的重度使用（即大量读写 HDFS 中文件）应用，有必要单独划分一个 NameService 来处理其请求，以保证该应用的数据可以独占该 NameService 的所有处理能力，也避免了该应用对其他应用的影响。
3.对于可靠性要求严苛的应用，可以划分一个 NameService，以免其他应用的过高的读写频率导致该应用无法访问 HDFS，造成该应用业务的不稳定。

业务数据目录挂载方式的规划原则

1. 业务数据相关联的服务的数据目录，尽量挂载在同一 NameService 下面。否则，跨 NameService 的文件读写速率较慢，会降低应用的文件存储性能。
2. 数据量大，且对其他服务的业务无关联的，可以直接使用一个 NameService。
3. 对业务量较小的业务，建议直接将其目录挂载在默认的 NameService（HDFS${clusterid}，HDFS 拼接数字形式 clusterid）中，这样就不用做数据迁移，可以降低配置为 Federation 的复杂性。
4. 建议只对全局一级目录进行 NameService 映射以减少配置复杂度。

方案比较

ViewFs Federation 类型

1. 直接使用 ViewFs
优点：统一视图，不同的应用有相同的使用方式。
缺点：ViewFs 挂载表的变更需要所有使用集群的应用同步读取最新的挂载点。
2. 指定 NameService
优先：不需要同步更改所有应用的配置。

缺点：不同组件和应用需要明确各自的使用目录，组件路径之间耦合的场景将复杂化。

Router-based Federation 类型配置方式

1. 直接使用 Router-based Federation

优点：统一视图，不同的应用有相同的使用方式。与 ViewFs 相比，Router-based Federation 挂载表的变更直接生效，不需要使用集群的应用同步更新挂载表。

缺点：客户端访问时，增加 DFSRouter 转发层，对性能略有影响。

2. 指定 NameService

优先：不需要同步更改所有应用的配置。

缺点：不同组件和应用需要明确各自的使用目录，组件路径之间耦合的场景将复杂化。

HDFS 联邦管理

最近更新时间：2023-03-15 10:12:39

功能介绍

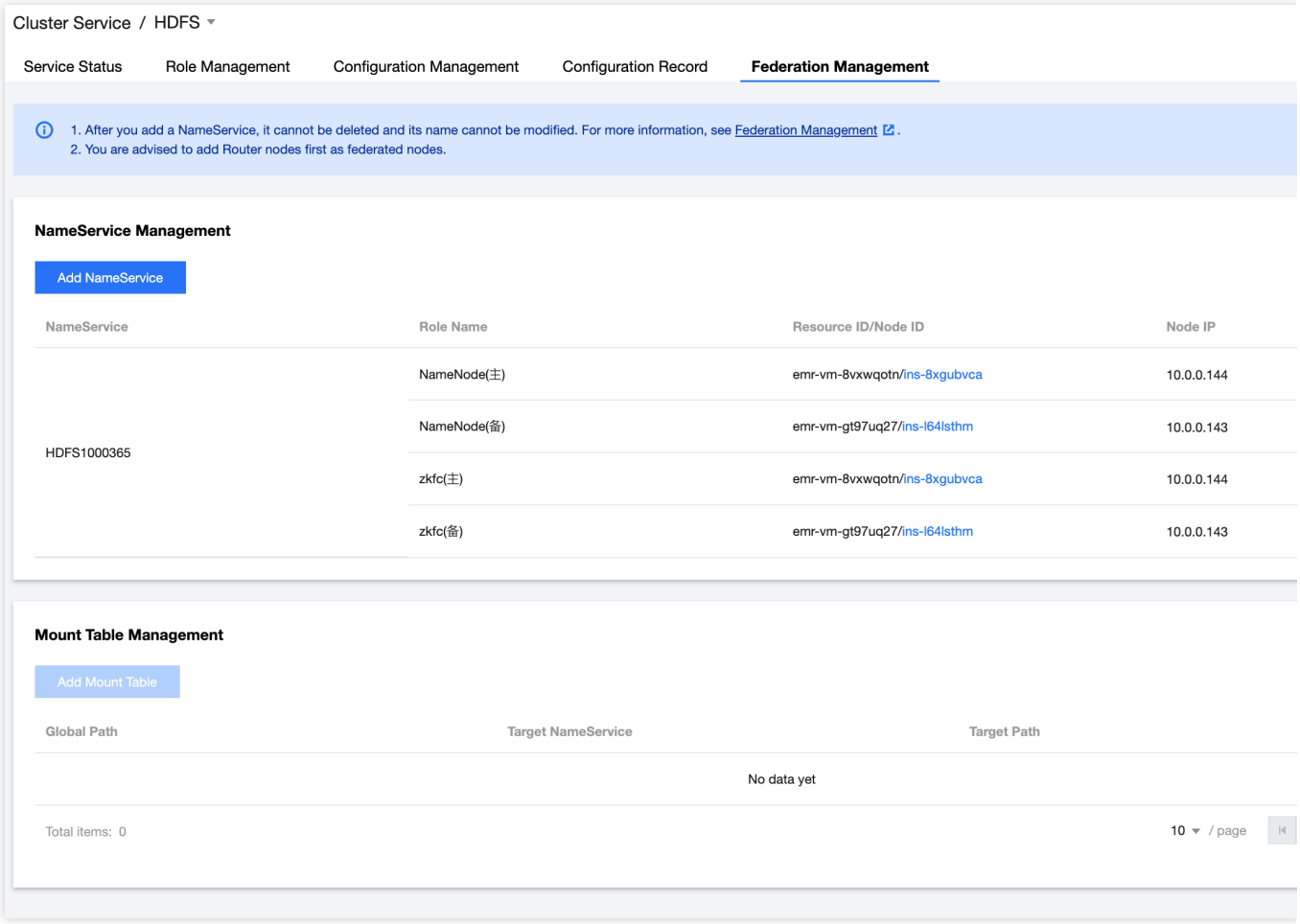
HDFS 联邦管理是基于 HDFS Federation 特性提供的 HDFS 联邦集群部署管理能力，包含 NameService 管理以及挂载表管理。在 Hadoop 集群类型 HA 模式下支持联邦管理，支持 ViewFS Federation 和 Router-based Federation 两种联邦类型选择，联邦类型选择后不可更改。Router 节点会用于新扩展的 NameNode 部署，用作 NameNode 部署后的 Router 节点不支持销毁和节点维度所有角色启停。

注意

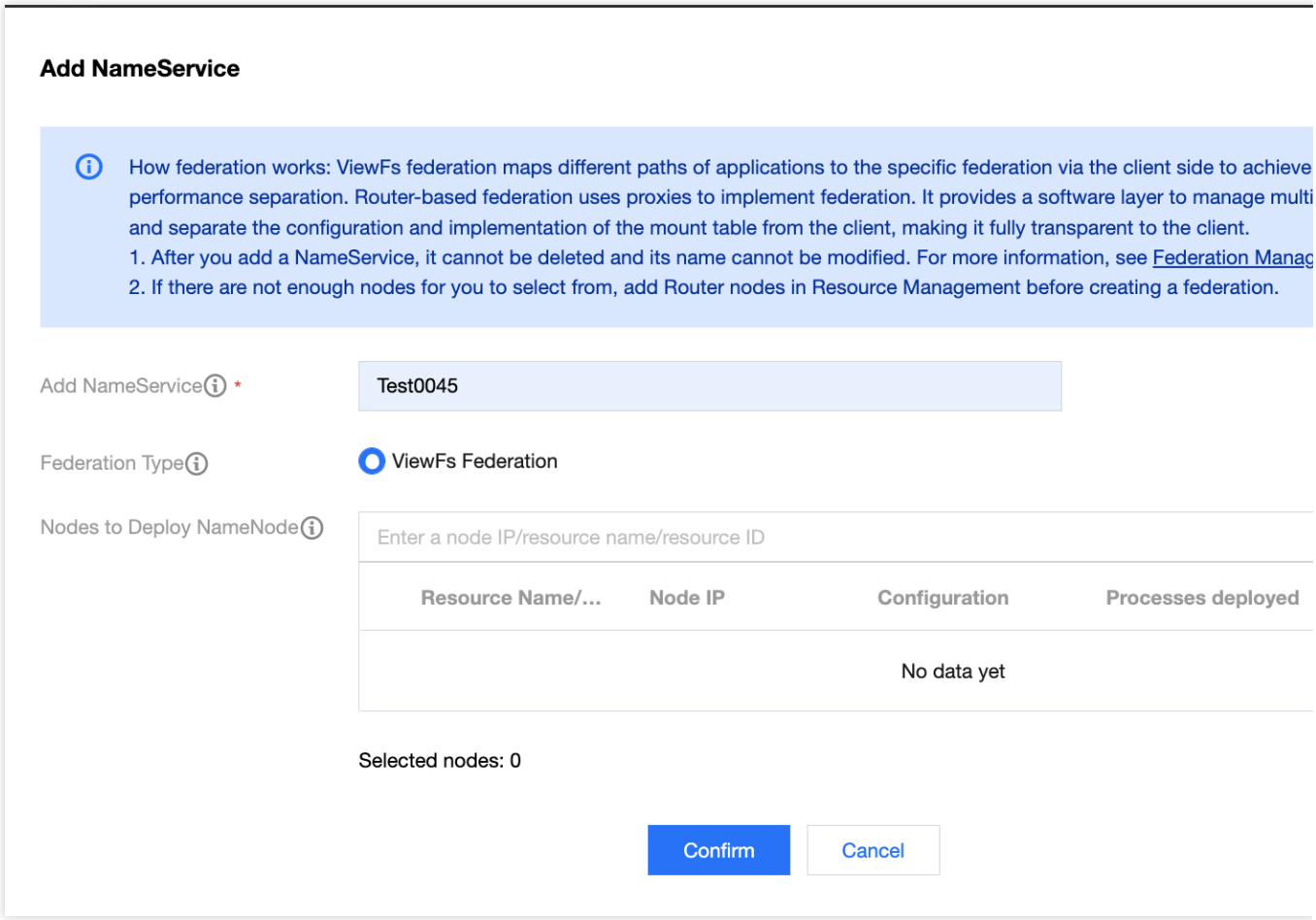
1. 当前 HDFS 联邦管理为白名单开放，如需要可 工单 联系我们开通。
2. EMR 所有产品版本均支持 ViewFs Federation 联邦类型；因 HDFS-2.9.0及以上支持 Router-based Federation 联邦类型，所以 EMR 仅 EMR-V3.x.x 及以上产品版本支持 Router-based Federation 联邦类型, EMR-V2.x.x 系列不支持。
3. 在角色管理页中，将联邦节点 NameNode 角色进程暂停时，会影响集群的扩容操作，需恢复 NameNode 角色进程后再执行扩容。

操作步骤

1. 登录 EMR 控制台，在集群列表中单击对应的**集群 ID/名称**进入**集群详情页**。
2. 在集群详情页中单击**集群服务**，然后选择 **HDFS 组件右上角操作>联邦管理**，即可进入联邦管理页面。



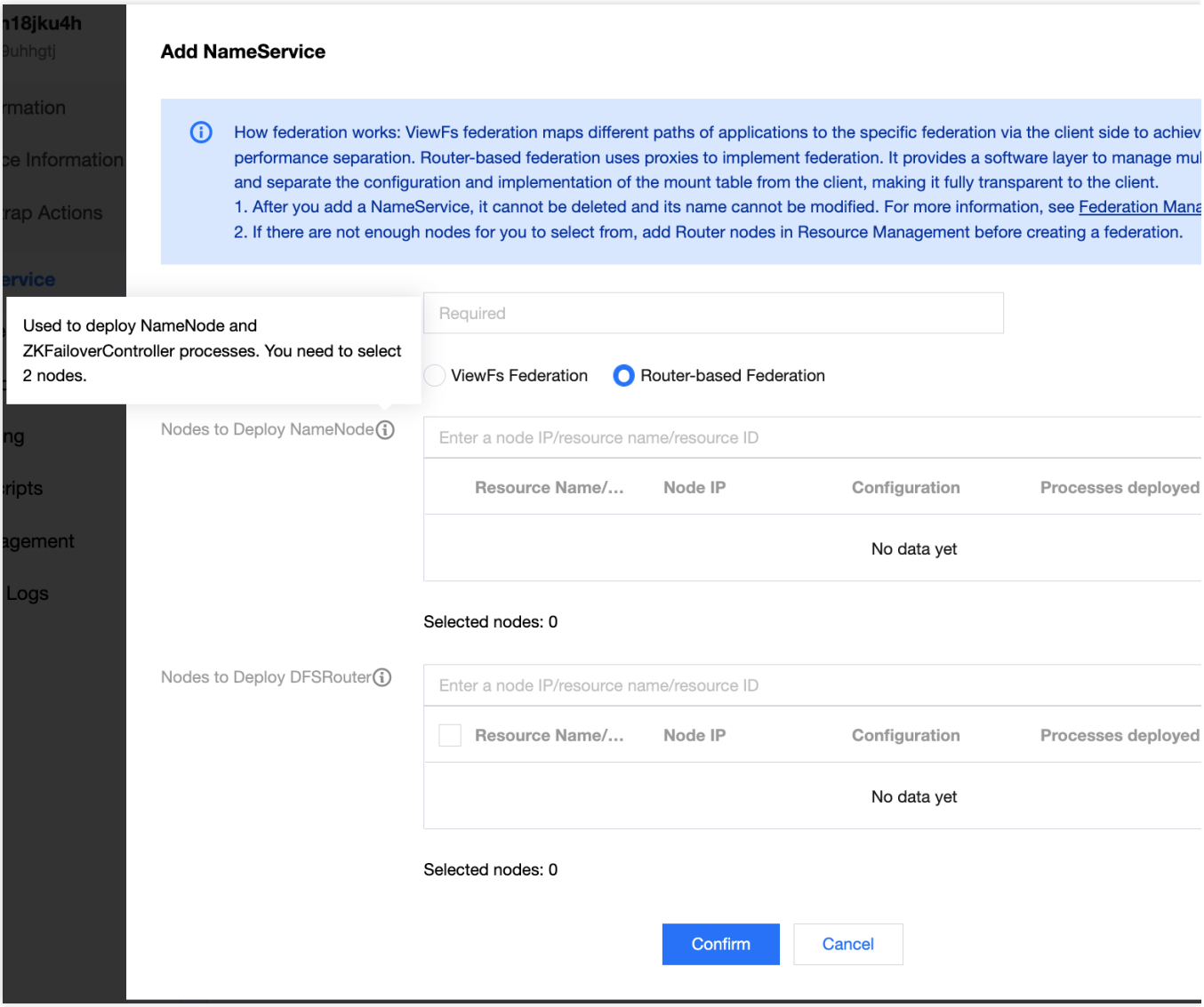
3. 单击**添加 NameService** 即可进行 HDFS 联邦新建，需要输入 NameService 名称，选择联邦类型、NameNode 节点、DFSRouter 节点（Router-based Federation 选择）等。



4. 选择**添加联邦节点**

联邦节点采用集群中的 Router 节点，需先在资源管理页中扩容增加 Router 节点后，再将其设置为联邦节点；NameNode 进程需要选择2个节点，每个节点将会部署 NameNode 进程和 ZKFC 进程。

第一次新建联邦类型 Router-based Federation 时，部署 DFSRouter 进程需要选择至少2个节点；当再次新建联邦时 DFSRouter 节点可复用，节点数量可大于等于0。



注意

HDFS-3.3.0以下的版本，在**非首次**新增 NameService 成功后且联邦类型为 Router-based Federation 时，需在角色管理页重启历史的 DFSRouter 进程（为保证业务正常建议在业务低峰期执行）；HDFS-3.3.0版本及以上支持了热加载配置，无需此操作。

开启了 Kerberos 的集群，添加联邦 NameService 后，提交到 yarn 的任务如要用到新 NameService 上的文件，需要先重启 yarn 的 ResourceManager（为保证业务正常建议在业务低峰期执行）。

设置 NameService 名称后，不可修改也不可以删除且 NameService 名称不能为"nsfed"、“haclusterX”、”ClusterX“等系统关键词。

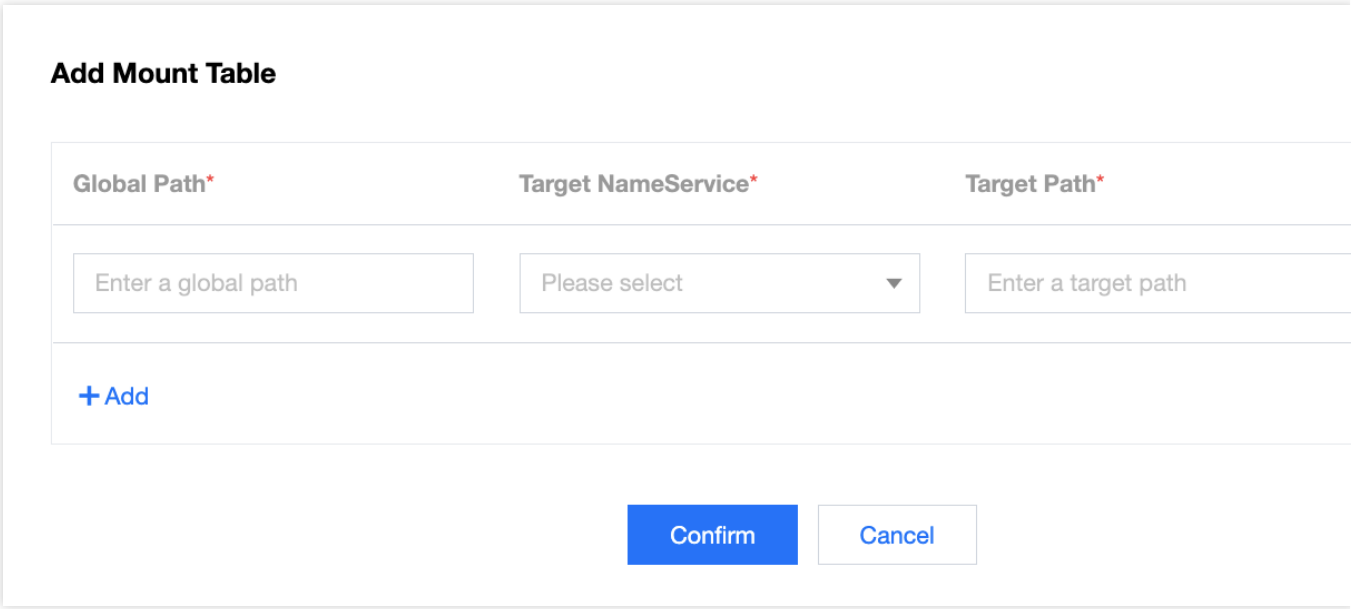
5. 添加挂载表

当新增 NameService 成功后，才能添加挂载表。为了避免配置复杂度过高，对当前集群目录进行映射，建议仅全局一级目录进行NameService 映射，例如挂载“/tmp”，“/user”，“/srv”等；可批量添加挂载路径。

路径：在 ViewFs 统一命名空间的路径名称，在 Router-based Federation 统一命名空间的路径名称，也称挂载点。

目标NameService：挂载点映射到真实路径对应的 NameService。

目标路径：在对应 NameService 上的真实路径，与全局路径的名字无需保持一致。



注意

路径指向：

1.1 登录 NameNode 节点，执行 hdfs dfs -ls /指向的是这个 NameNode 所管理的 namespace 下的路径，对于 ViewFs 联邦需要用 hdfs dfs -ls ViewFs://ClusterX/ 才会指向全局路径，对于 Router-based 联邦需要用 hdfs dfs -ls hdfs://nsfed/ 才会指向全局路径。

1.2 登录其它节点，例如充当客户端的Router节点，hdfs dfs -ls / 指向的是全局路径。

各业务组件数据放在一级目录之下，不支持直接放在根目录下访问，根目录不支持挂载。

默认的 NameService 上是有/emr 目录，需要挂载。

提交 MapReduce 任务

最近更新时间：2022-03-16 17:51:19

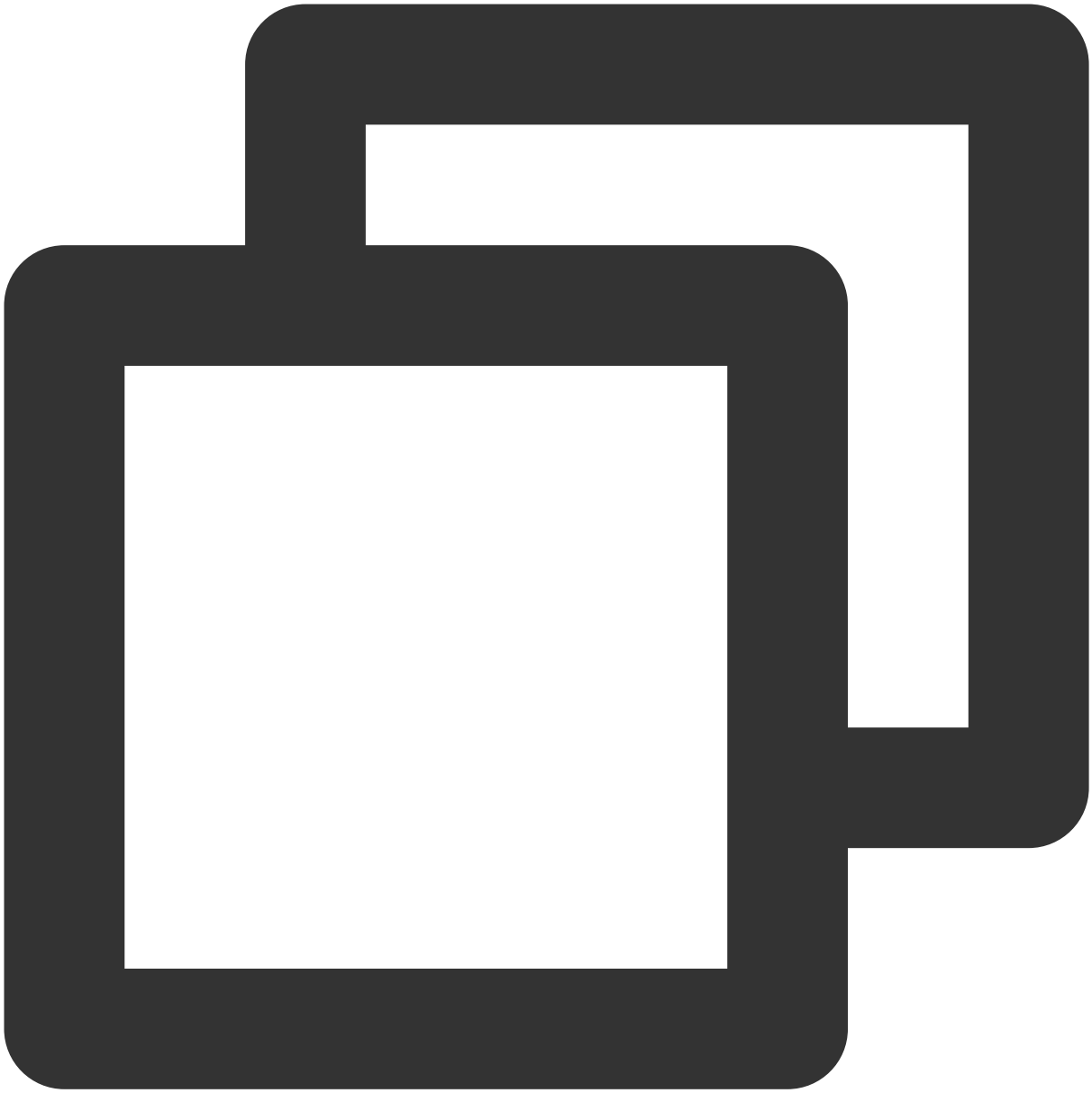
本操作手册只描述了命令行模式下基本的 MapReduce 任务操作以及 MapReduce 计算任务如何访问腾讯云对象存储 COS 上面的数据，详细资料可以参考 社区资料。
本次提交的任务为 wordcount 任务即统计单词个数，提前需要在集群中上传需要统计的文件。
Hadoop 等相关软件路径在 `/usr/local/service/` 下。
相关日志路径在 `/data/emr` 下。

1. 开发准备

由于任务中需要访问腾讯云对象存储 COS，所以需要在 COS 中先 创建一个存储桶（Bucket）。
确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群时在基础配置页面勾选了**开启 COS**，并在下方填写自己的 SecretId 和 SecretKey。SecretId 和 SecretKey 可以在 API 密钥管理界面 查看。如果还没有密钥，请单击**新建密钥**建立一个新的密钥。

2. 登录 EMR 服务器

在做相关操作前需要登录到 EMR 集群中的任意一个机器，建议登录到 Master 节点。EMR 是建立在 Linux 操作系统的腾讯云服务器 CVM 上的，所以在命令行模式下使用 EMR 需要登录 CVM 服务器。
创建 EMR 集群后，在控制台中选择弹性 MapReduce。在**集群资源>资源管理**中单击**Master 节点**，选择 Master 节点的资源 ID，即可进入云服务器控制台并且找到 EMR 对应的云服务器。
登录 CVM 的方法可参见 登录 Linux 实例 。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的**登录**，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。
输入正确后，即可进入 EMR 集群的命令行界面。所有的 Hadoop 操作都在 Hadoop 用户下，登录 EMR 节点后默认在 root 用户，需要切换到 Hadoop 用户。使用如下命令切换用户，并且进入 Hadoop 文件夹下：

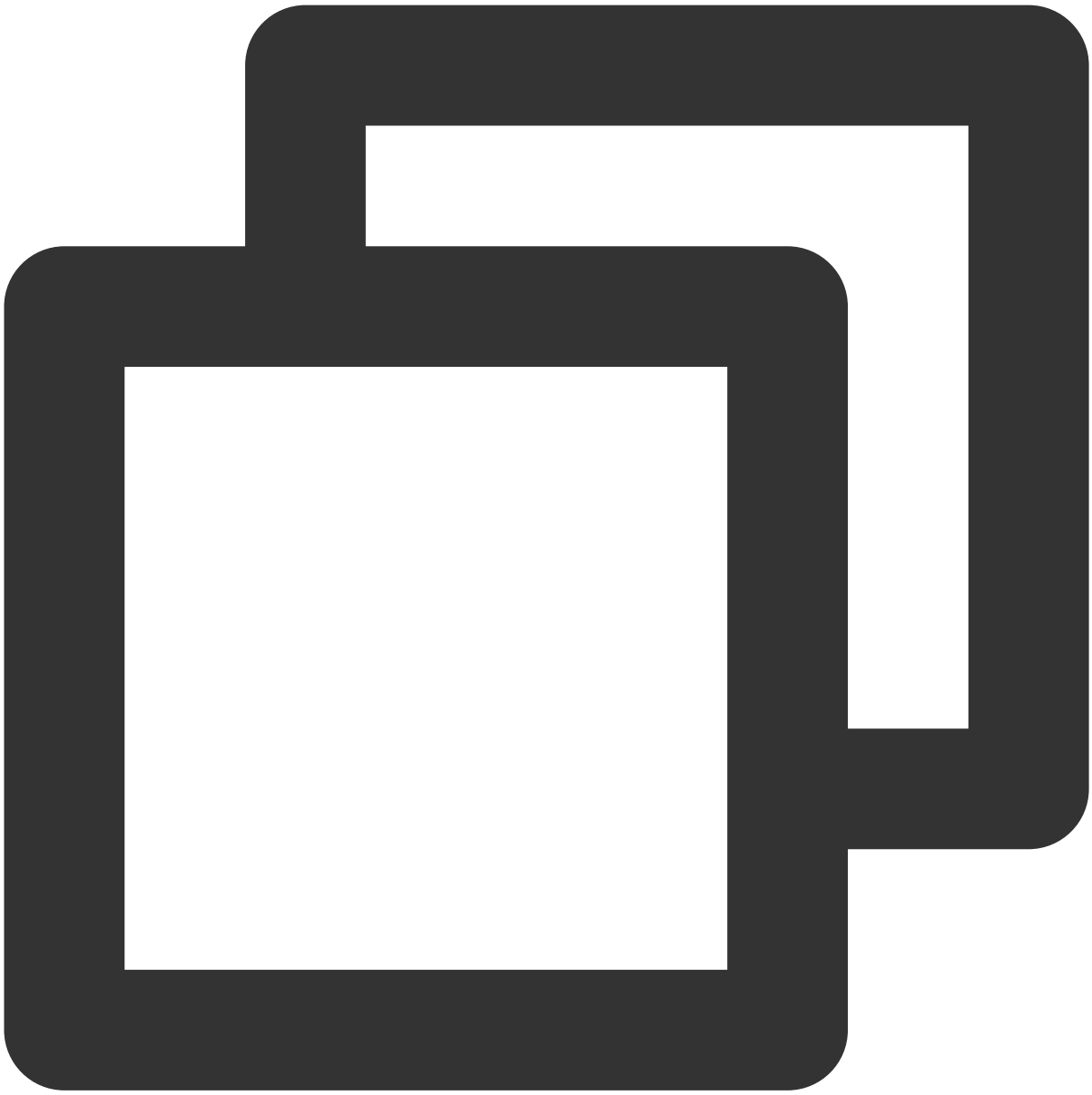


```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/hadoop
[hadoop@172 hadoop]$
```

3. 数据准备

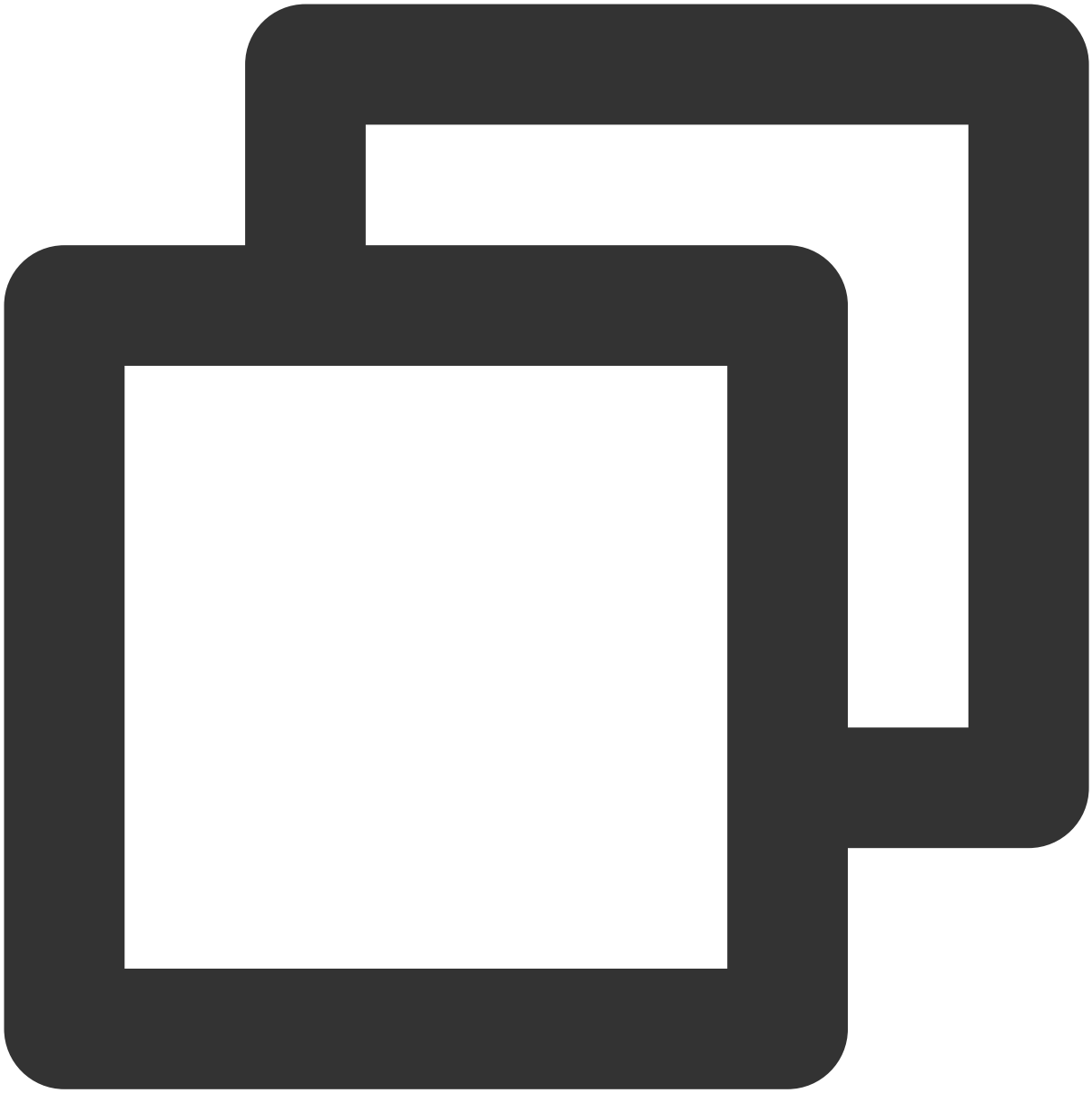
您需要准备统计的文本文件。分为两种方式：**数据存储在 HDFS 集群**和**数据存储在 COS**。

首先要把本地的数据上传到云服务器。可以使用 scp 或者 sftp 服务来把本地文件上传到 EMR 集群的云服务器中。在本地命令行使用：



```
scp $localfile root@公网IP地址:$remotefolder
```

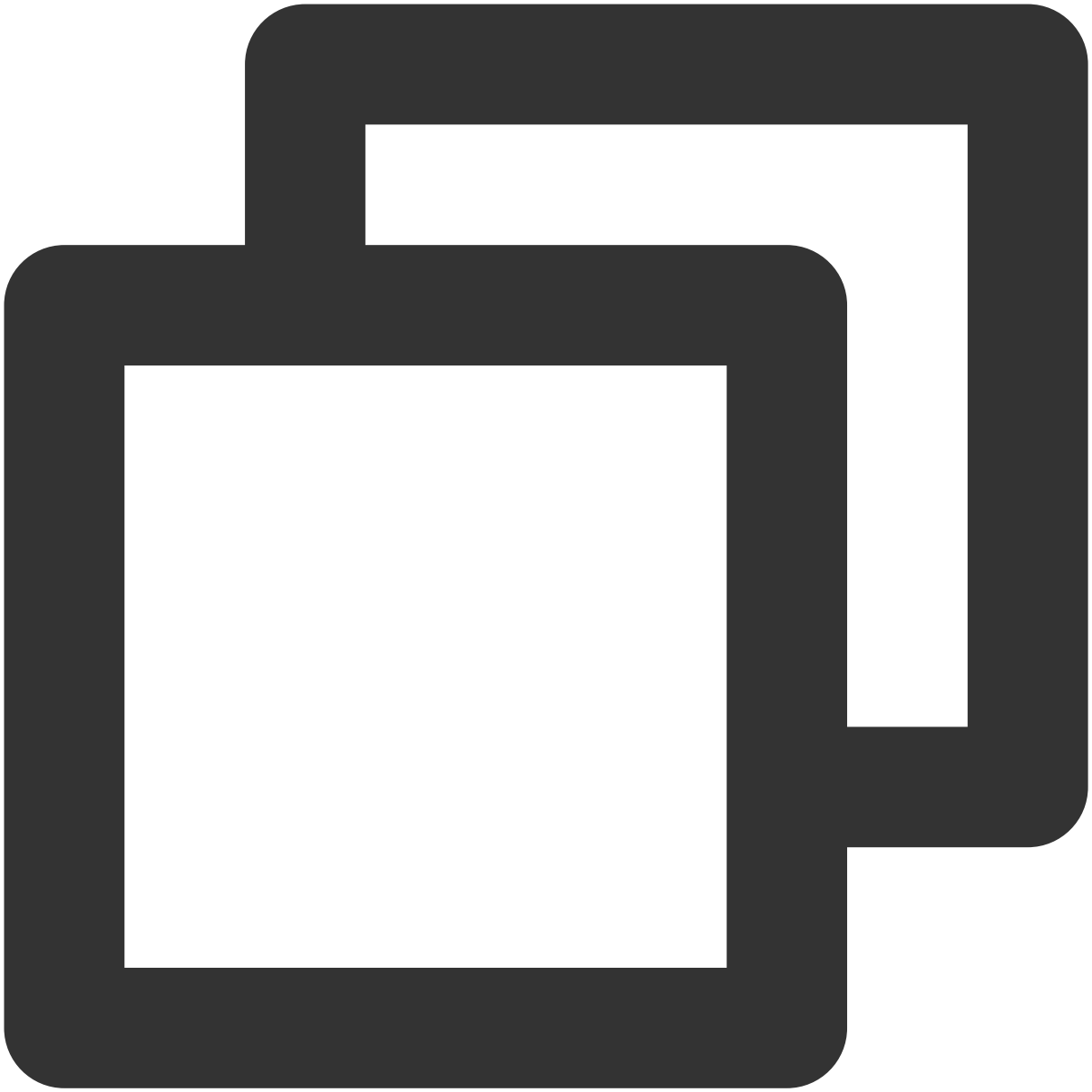
其中，$localfile 是您的本地文件的路径加名称；root 为 CVM 服务器用户名；公网 IP 地址可以在 EMR 控制台的节点信息中或者在云服务器控制台查看；$remotefolder 是您要存放文件的 CVM 服务器路径。
上传成功后，在 EMR 集群命令行中即可查看对应文件夹下是否有相应文件。



```
[hadoop@172 hadoop]$ ls -l
```

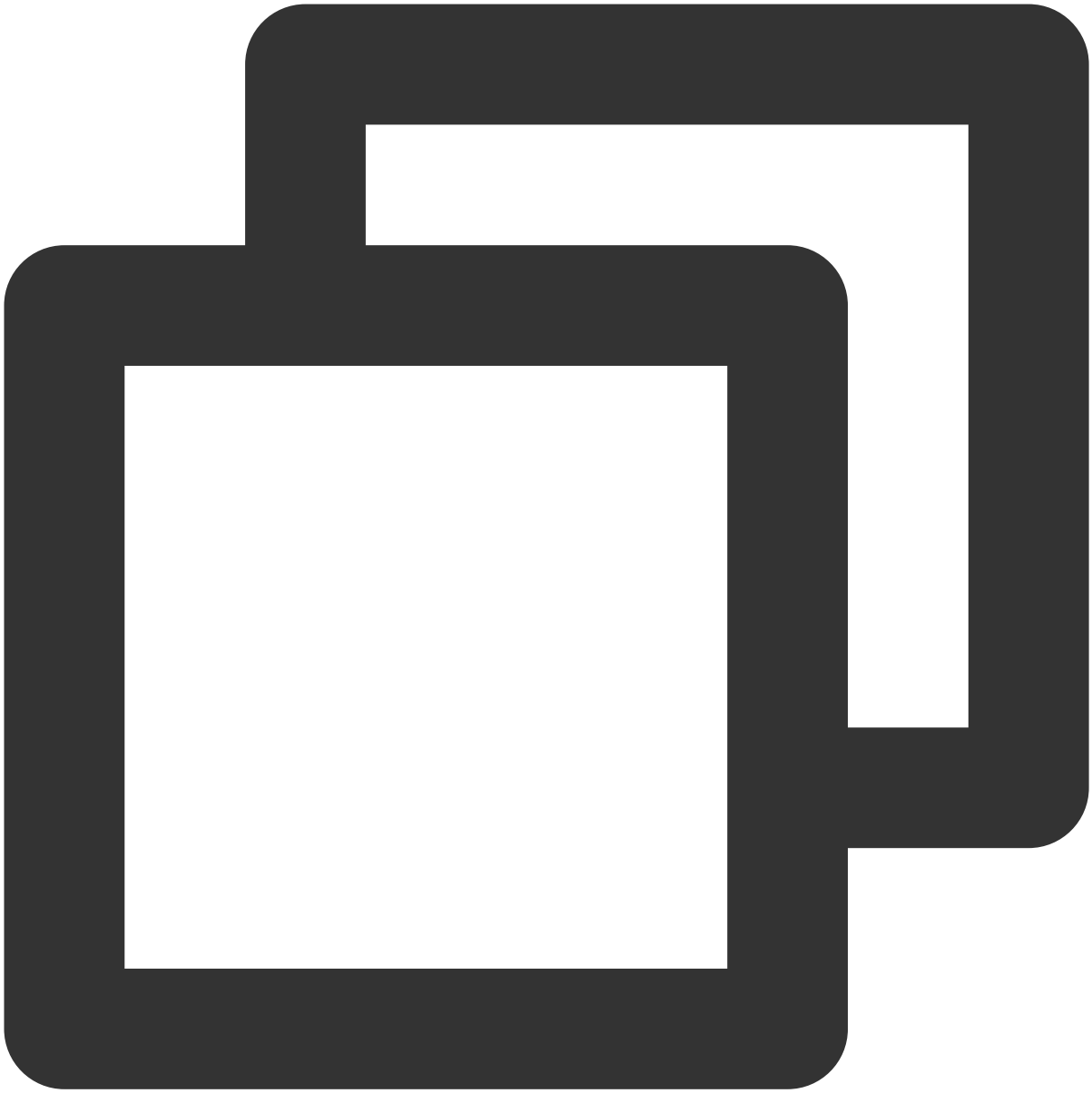
数据存放在 HDFS

将数据上传到腾讯云服务器之后，可以把数据拷贝到 HDFS 集群。这里使用 `/usr/local/service/hadoop` 目录下的 README.txt 文本文件作为说明。通过如下指令把文件拷贝到 Hadoop 集群：



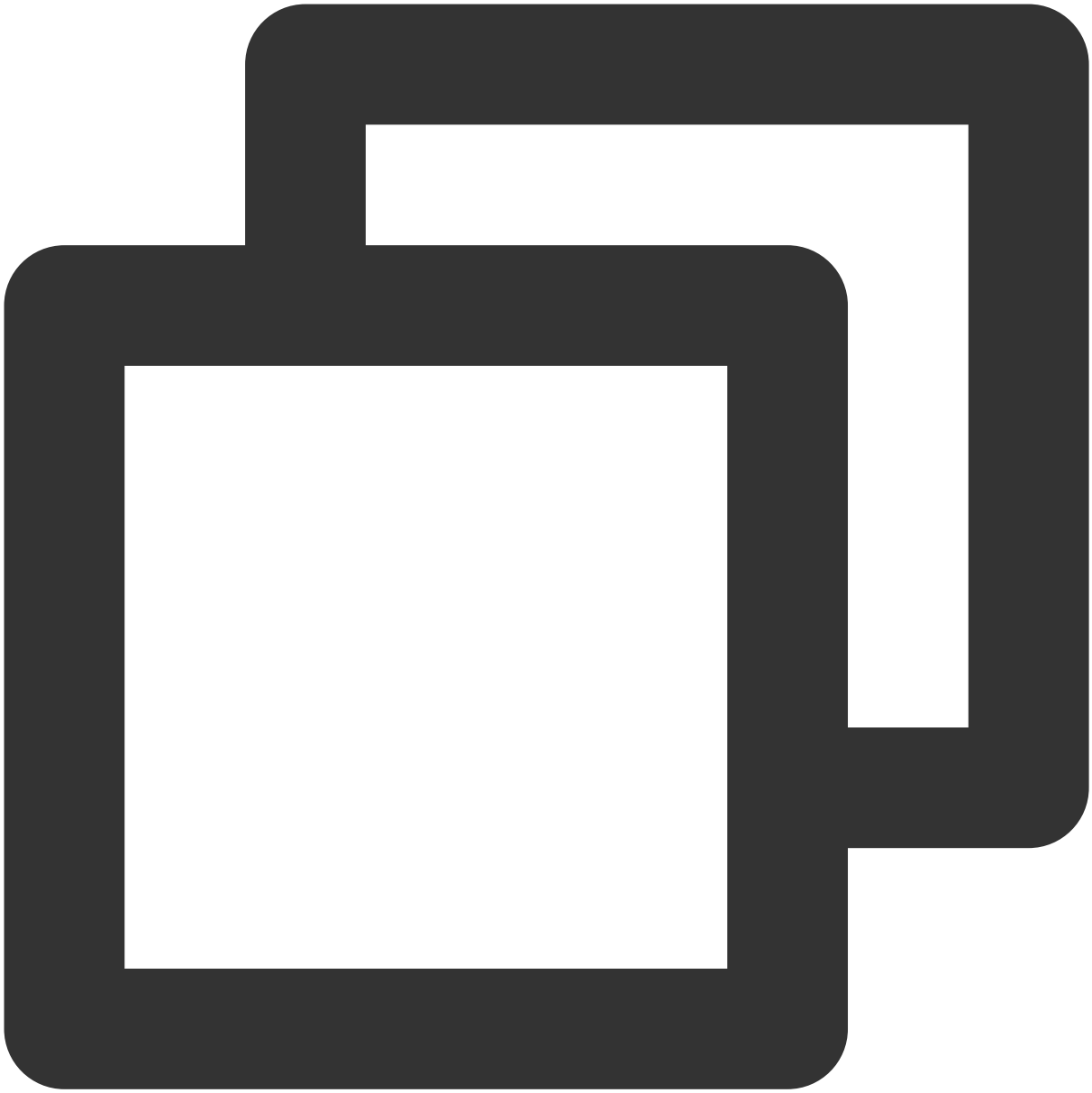
```
[hadoop@172 hadoop]$ hadoop fs -put README.txt /user/hadoop/
```

拷贝完成后使用如下指令查看拷贝好的文件：



```
[hadoop@172 hadoop]$ hadoop fs -ls /user/hadoop
输出：
-rw-r--r-- 3 hadoop supergroup 1366 2018-06-28 11:39 /user/hadoop/README.txt
```

如果 Hadoop下面没有 `/user/hadoop` 文件夹，用户可以自己创建，指令如下：

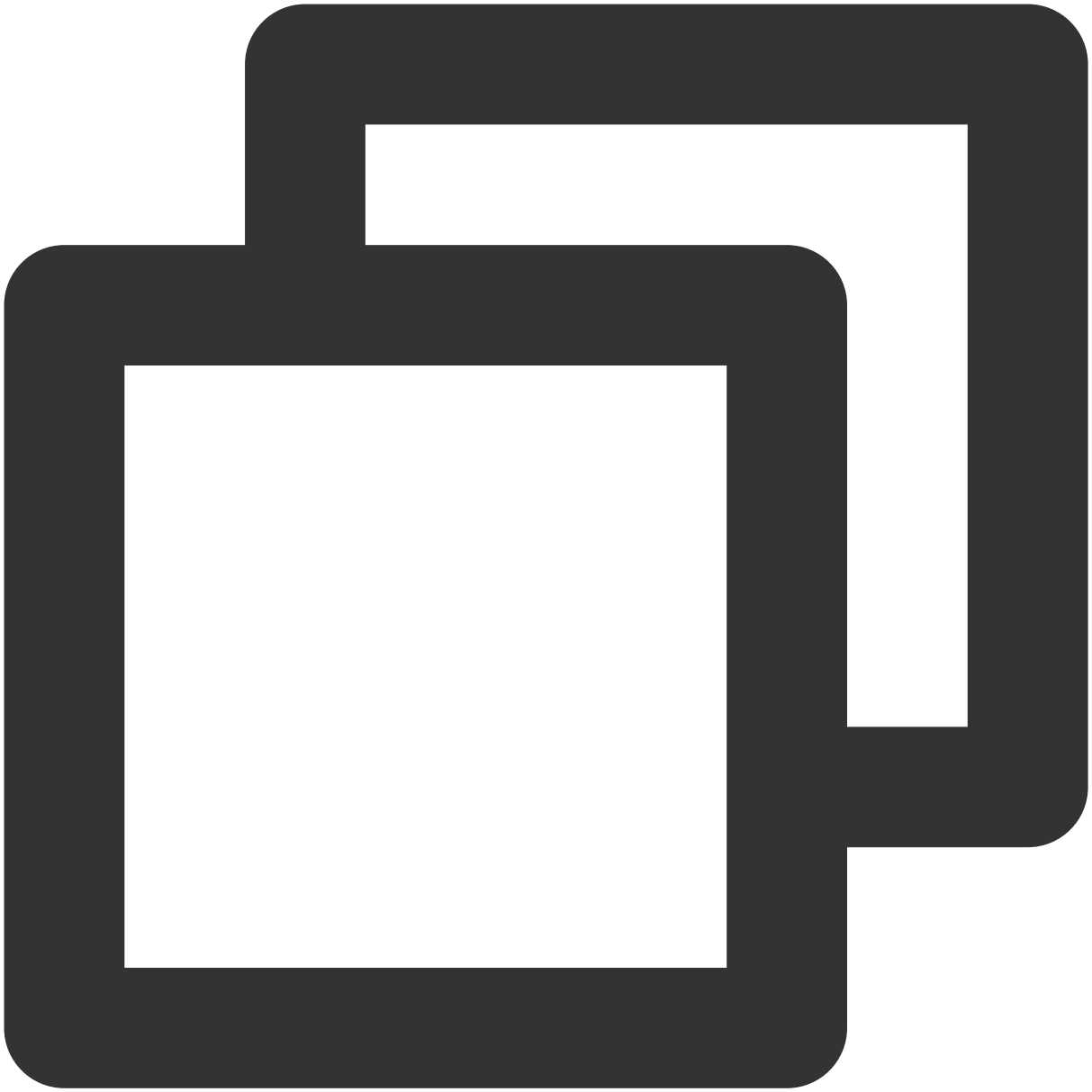


```
[hadoop@172 hadoop]$ hadoop fs -mkdir /user/hadoop
```

更多 Hadoop 指令见 HDFS 常见操作。

数据存放在 COS

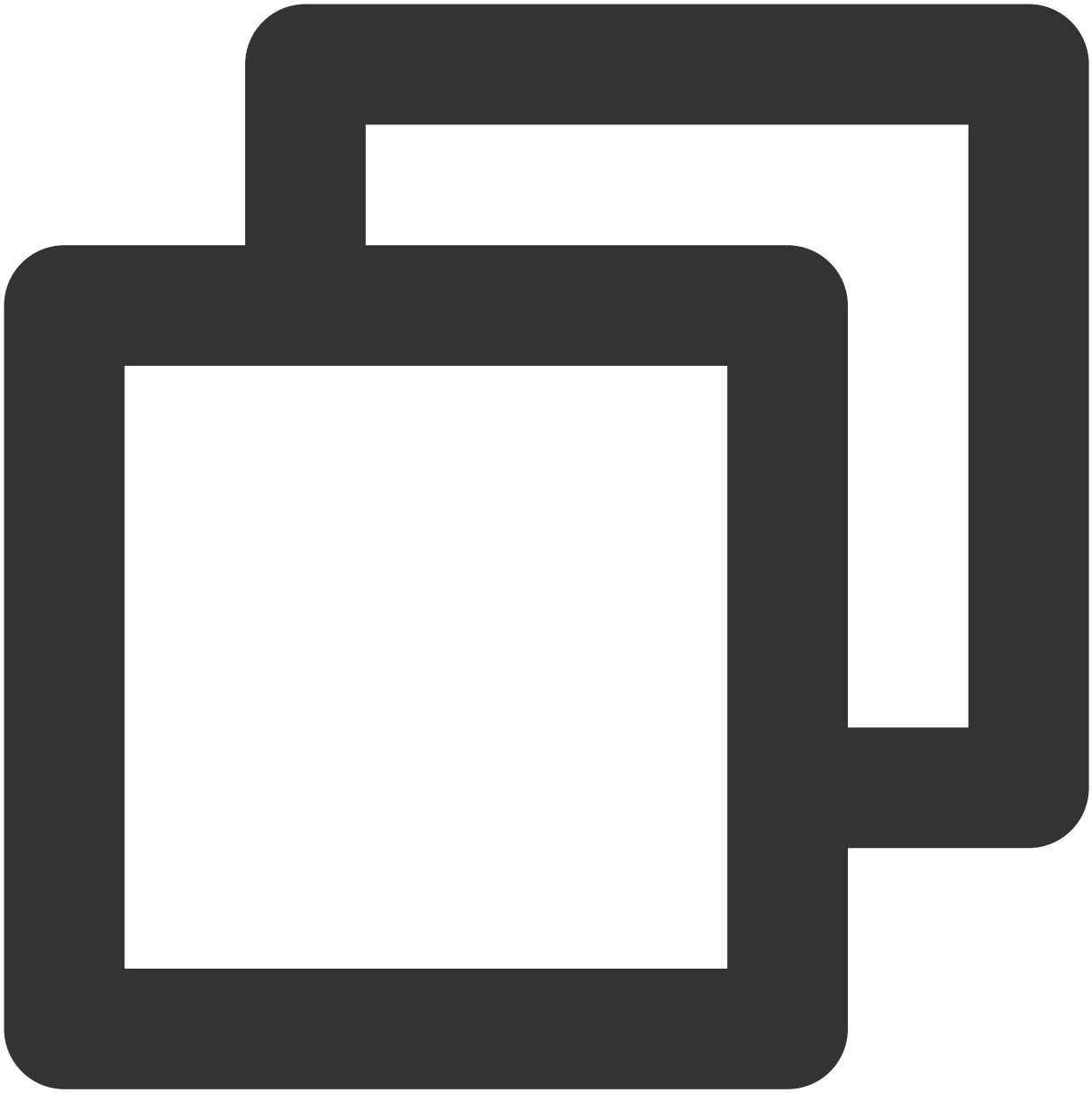
数据存放在 COS 中有两种方式：**在本地通过 COS 的控制台上传**和**在 EMR 集群通过 Hadoop 命令上传**。
在本地通过 COS 控制台直接上传，如果数据文件已经在 COS 可以通过如下命令查看：



```
[hadoop@10 hadoop]$ hadoop fs -ls cosn://$bucketname/README.txt
-rw-rw-rw- 1 hadoop hadoop 1366 2017-03-15 19:09 cosn://$bucketname /README.txt
```

其中 $bucketname 替换成您的储存桶的名字和路径。

在 EMR 集群通过 Hadoop 命令上传，指令如下：



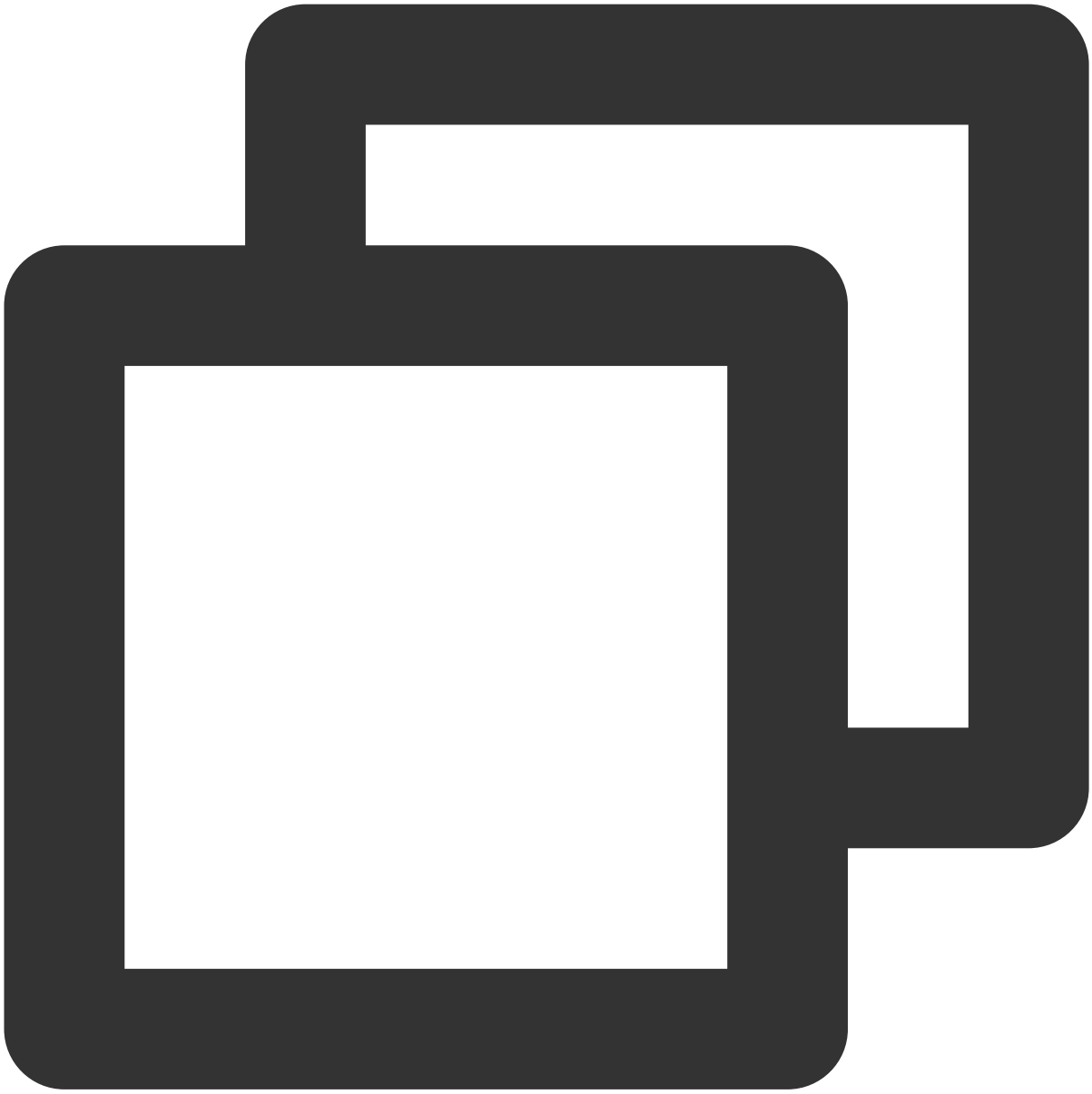
```
[hadoop@10 hadoop]$ hadoop fs -put README.txt cosn:// $bucketname /
[hadoop@10 hadoop]$ bin/hadoop fs -ls cosn:// $bucketname /README.txt
-rw-rw-rw- 1 hadoop hadoop 1366 2017-03-15 19:09 cosn://$bucketname /README.txt
```

4. 通过 MapReduce 提交任务

本次提交的任务是 Hadoop 集群自带的例程 wordcount。wordcount 已被压缩为 jar 包上传到了创建好的 Hadoop 中，用户可以直接调来使用。

统计 HDFS 中的文本文件

进入 `/usr/local/service/hadoop` 目录，和数据准备中一样。通过如下命令来提交任务：

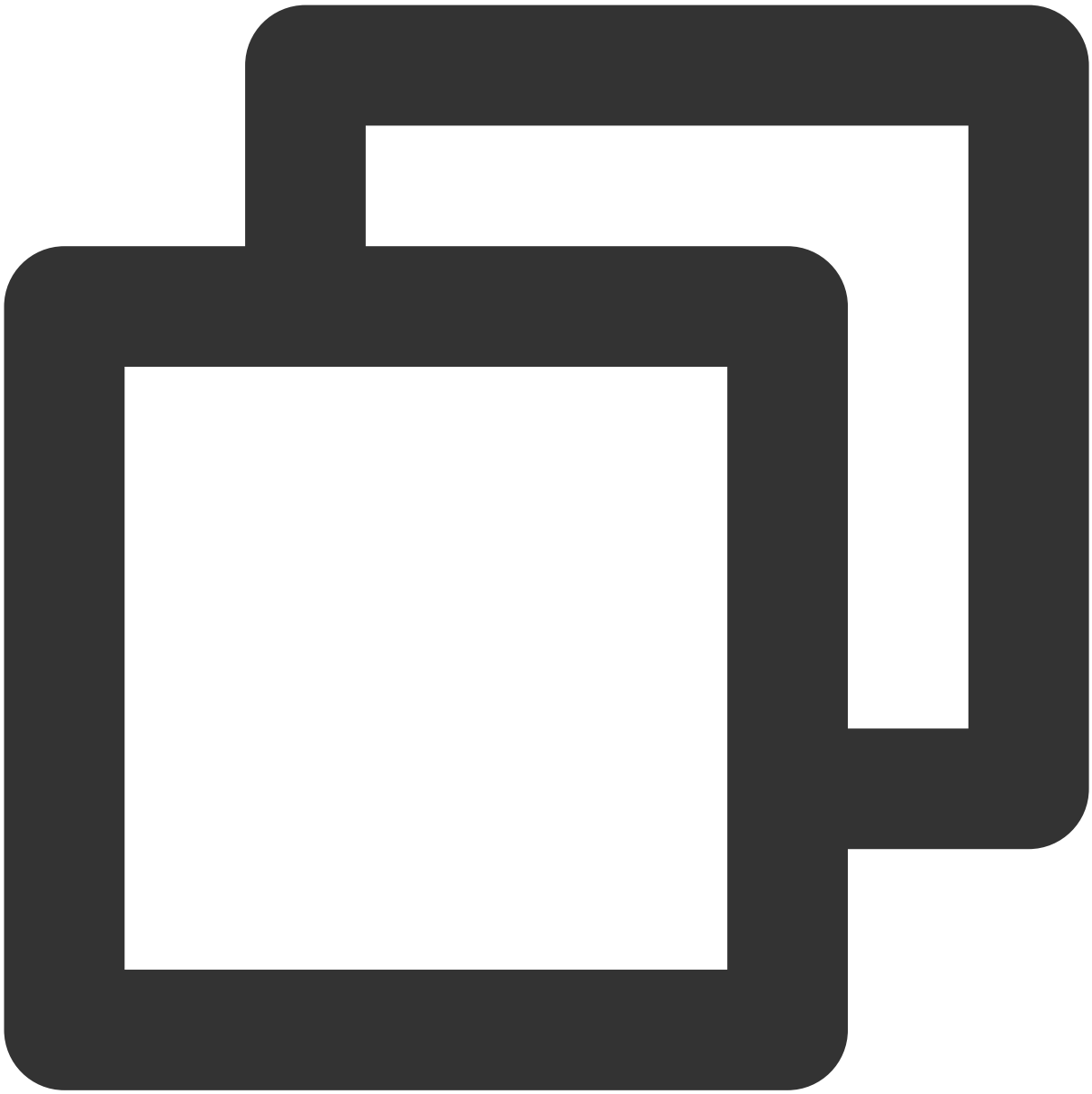


```
[hadoop@10 hadoop]$ bin/yarn jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples
/user/hadoop/README.txt /user/hadoop/output
```

注意

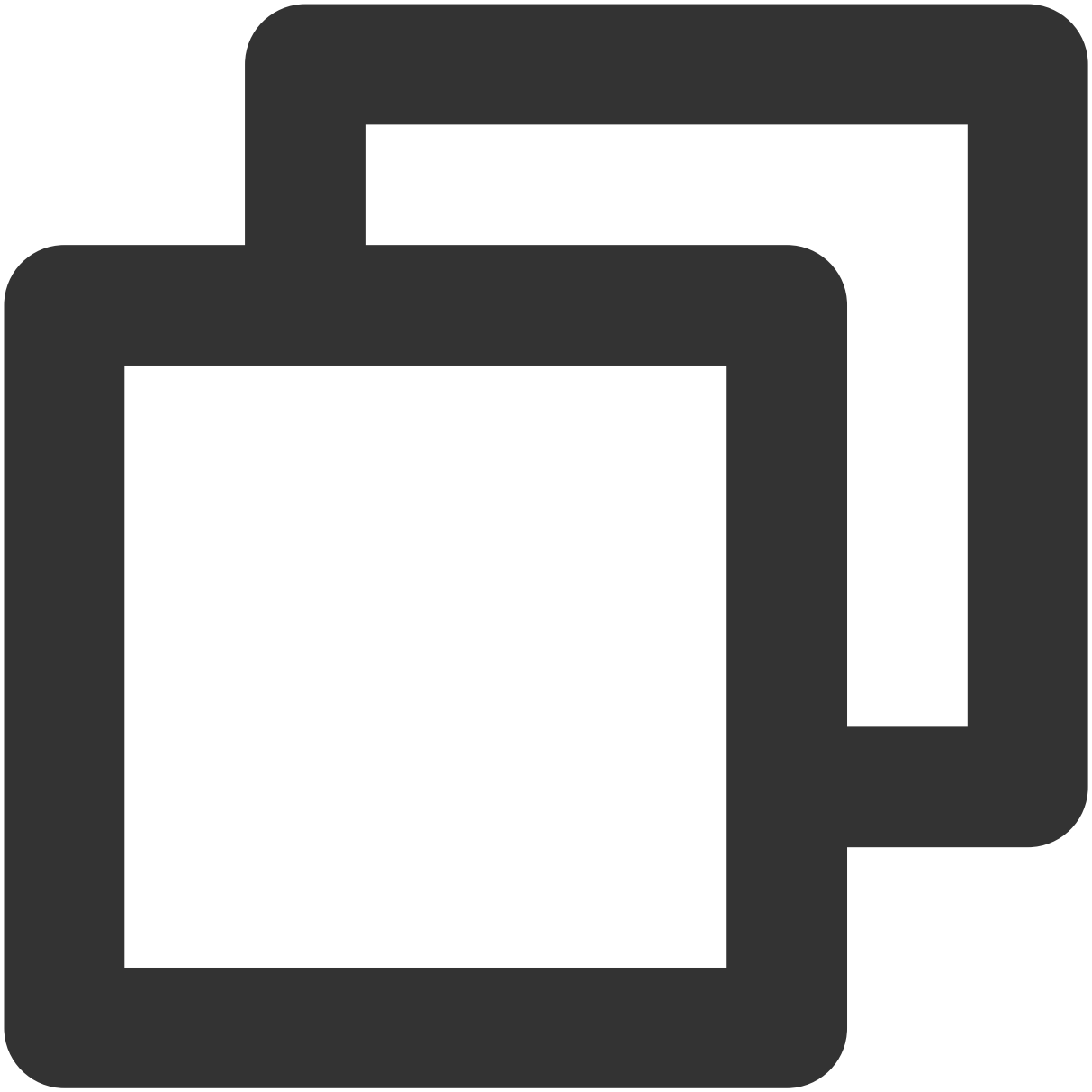
以上整个命令为一条完整的指令，`/user/hadoop/README.txt` 为输入的待处理文件，`/user/hadoop/output` 为输出文件夹，在提交命令之前要保证 output 文件夹尚未创建，否则提交会出错。

执行完成后，通过如下命令查看执行输出文件：



```
[hadoop@10 hadoop]$ bin/hadoop fs -ls /user/hadoop/output
Found 2 items
-rw-r--r-- 3 hadoop supergroup 0 2017-03-15 19:52 /user/hadoop/output/_SUCCESS
-rw-r--r-- 3 hadoop supergroup 1306 2017-03-15 19:52 /user/hadoop/output/part-r-000
```

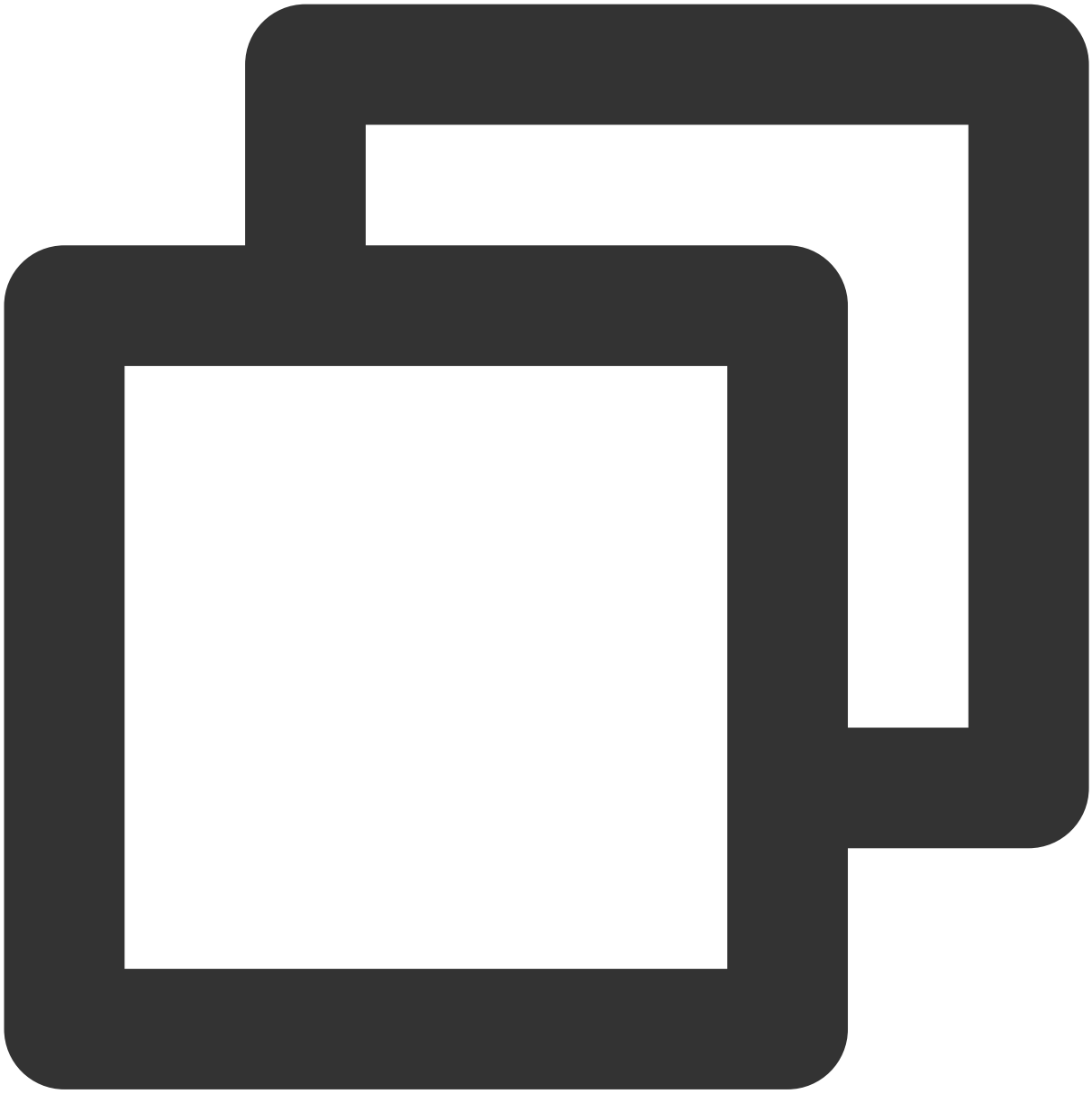
通过如下指令查看 part-r-00000 中的统计结果：



```
[hadoop@10 hadoop]$ bin/hadoop fs -cat /user/hadoop/output/part-r-00000
(BIS),  1
(ECCN)  1
(TSU)   1
(see    1
5D002.C.1,      1
740.13) 1
<http://www.wassenaar.org/>     1
……
```

统计 COS 中的文本文件

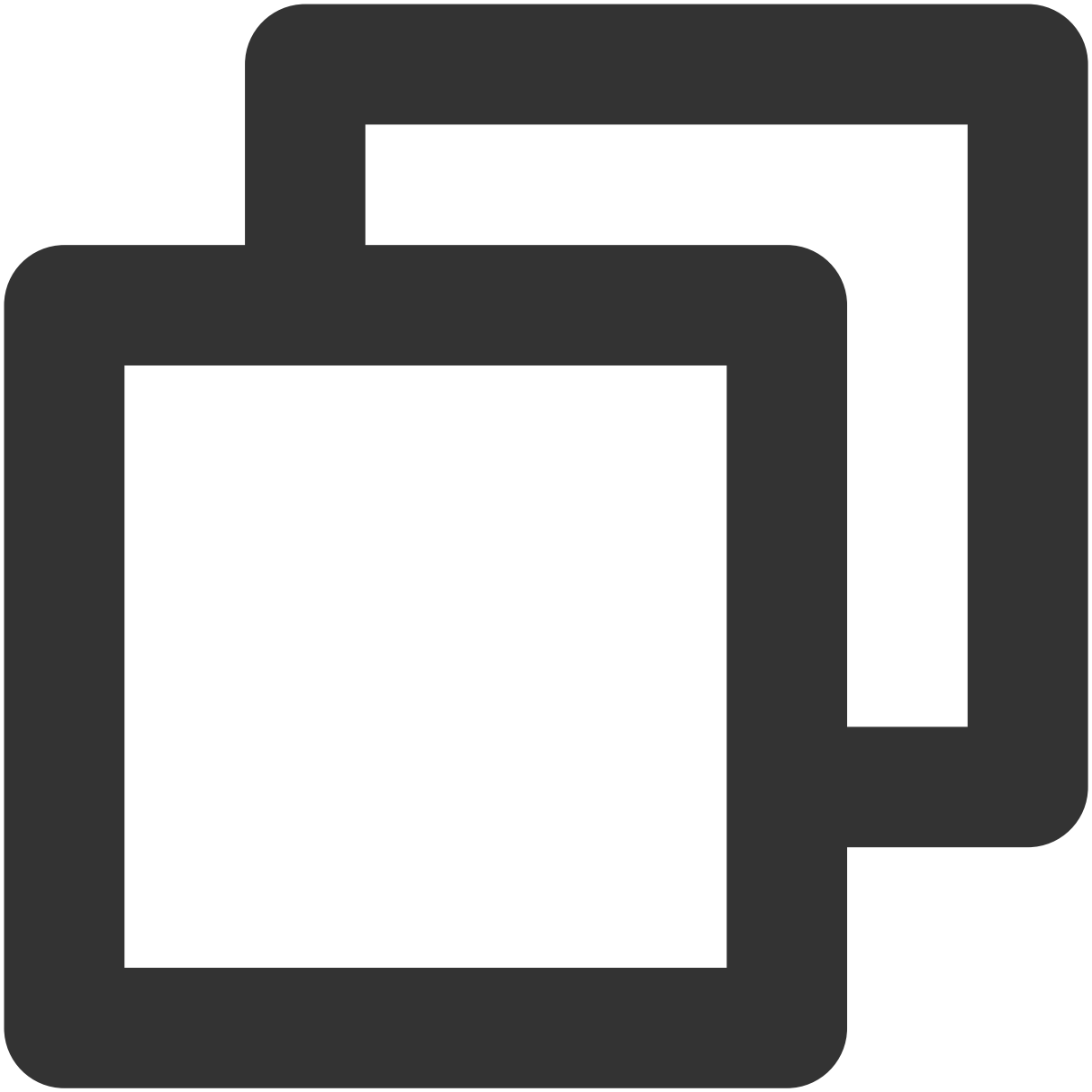
进入 `/usr/local/service/hadoop` 目录，通过如下命令来提交任务：



```
[hadoop@10 hadoop]$ bin/yarn jar ./share/hadoop/mapreduce/hadoop-mapreduce-examples
cosn://$bucketname/README.txt /user/hadoop/output
```

命令的输入文件改为了 `cosn:// $bucketname /README.txt` ，即处理 COS 中的文件，其中 $bucketname 为您的存储桶的名字加路径。依然输出到 HDFS 集群中，也可以选择输出到 COS 中。查看输出的方法和上文一样。

查看任务日志



```
#查看任务状态
bin/mapred job -status jobid
#查看任务日志
yarn logs -applicationId id
```

自动扩容 Task 节点不分配 ApplicationMaster

最近更新时间：2023-07-12 10:37:09

功能介绍

在自动扩缩容场景下，缩容规则触发时，即将被销毁的 Task 节点若正在运行着 ApplicationMaster（AM），那么节点在销毁的同时，运行的 AM 也会随之销毁，会导致当前 Job 整体失败。
在自动扩缩容场景下，通过扩容添加的 Task 节点默认不分配 ApplicationMaster，保证缩容Task节点时，运行的ApplicationMaster不被销毁，Job 可以正常进行。

功能特性

通过对 YARN 的源码改造以及对自动扩容流程增加配置项，实现自动扩容添加的 Task 节点将不会被分配 AM，AM 将会全部分配到非自动扩容的 Task 节点，自动扩容添加的 Task 节点仅承担计算任务；当对 Task 节点进行缩容时，只有正在运行的计算任务的节点会被销毁，但对应 AM 仍然保持存活状态，AM 可以重新在其他节点重试相关的计算任务，当前整个 Job 可以继续进行下去。

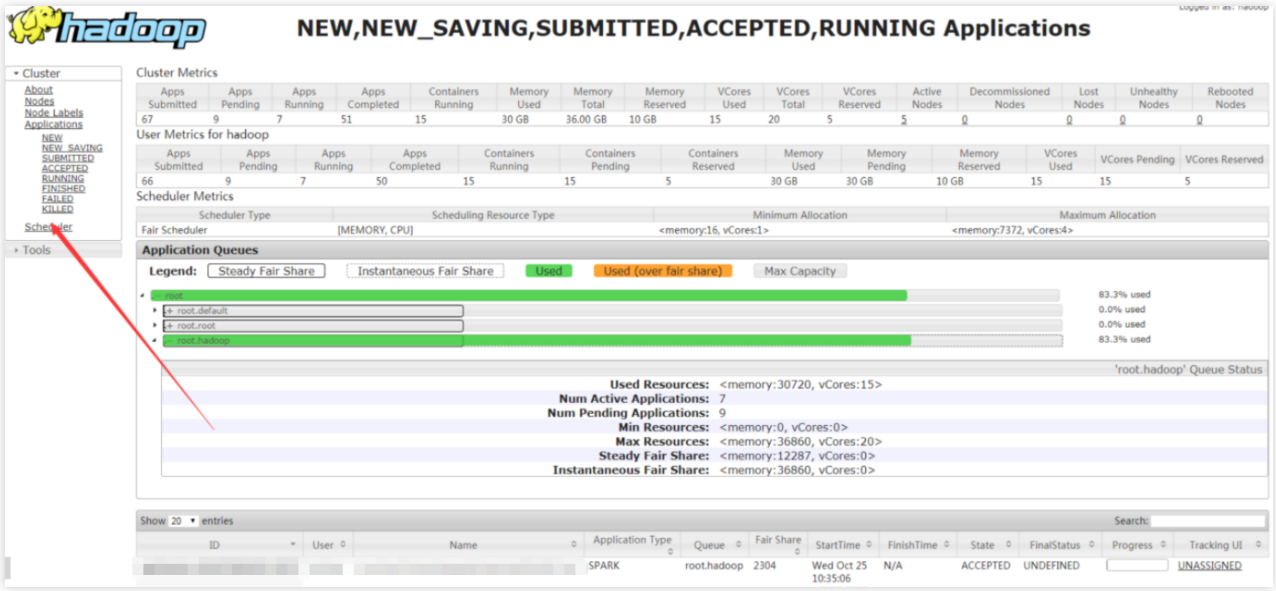
适用范围

仅对自动扩容的 Task 节点生效，手动扩容不生效。

YARN 任务队列管理

最近更新时间：2021-10-08 15:17:07

您可以登录 YARN 自带的 YARNwebUI，登录方式为通过 EMR 提供的快捷入口，查看快捷入口请参考 软件 WebUI 入口，登录后您将看到如下图：



从图中可以看到整个集群的一些监控信息：

应用信息：9个等待，7个执行，51个完成，总结67个；其中有15个 container 正在执行。

内存信息：总共36G，使用了30G，保留10G。

虚拟核信息：总共20个，使用了15个，保留5个。

节点信息：可用节点5个，退服0个，丢失节点0个，不健康节点0个，重启节点0个。

调度器信息：使用 Fair Scheduler（公平调度器），最大最小的内存和 CPU 分配信息。

其中 Application Queues 栏包含了集群的任务队列信息（以 root.hadoop 队列为例）：

使用的资源：内存30720M（30G），虚拟核15个。

正在执行的应用7个。

等待的应用9个。

占用最小资源：0M，0核。

占用最大资源：36860M，20核。

队列资源的稳定的公共共享。

队列的瞬时公平共享资源。

该队列占用资源的比例为83.3%。

YARN 标签调度实践

最近更新时间：2023-03-15 10:12:39

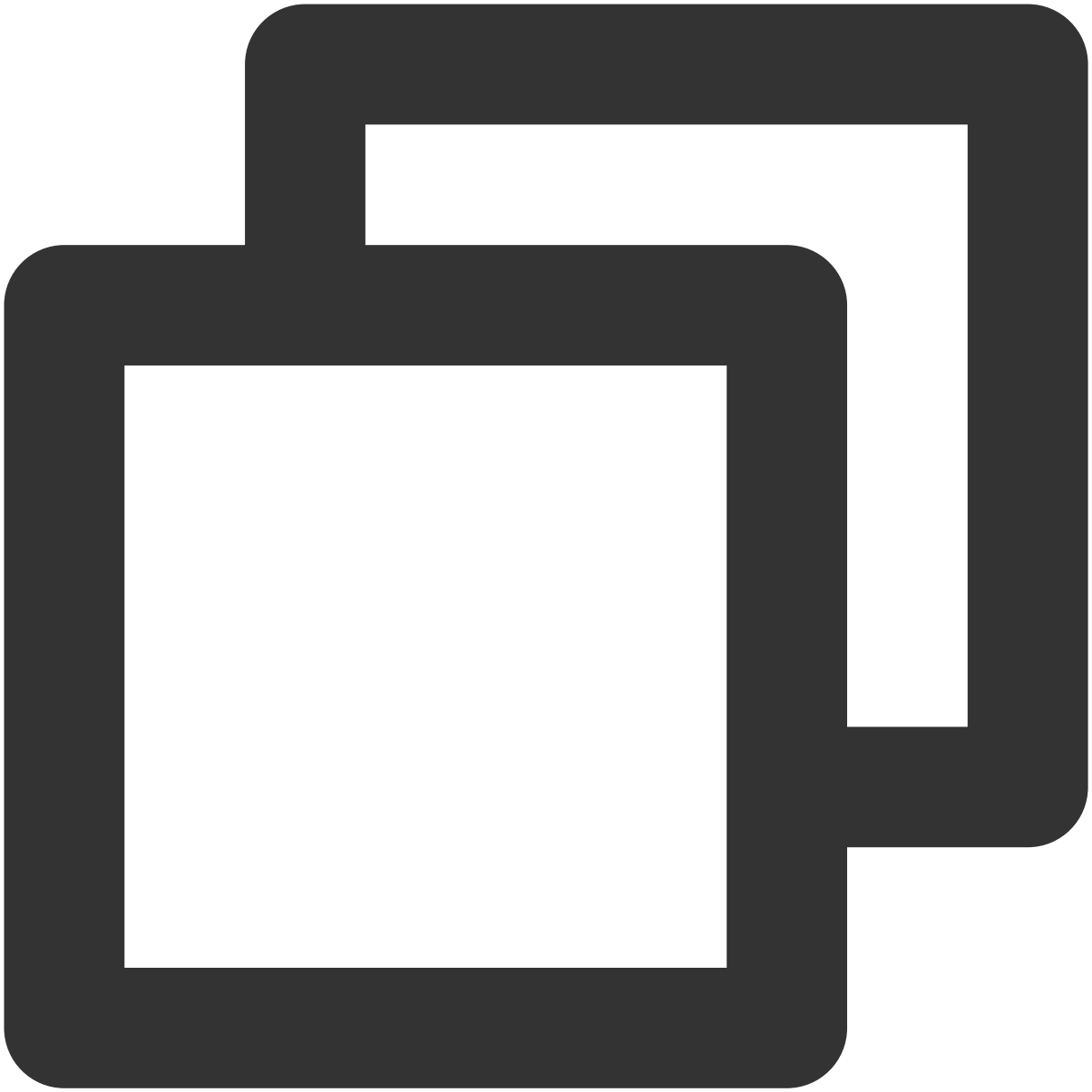
功能介绍

Spark on Yarn 使用 Yarn 作为资源调度器，在 Hadoop2.7.2 之后 Yarn 在容器调度器 Capacity Scheduler 基础上增加了标签调度。
Capacity Scheduler 是一个多租户资源调度器，其核心思想是 Hadoop 集群中的可用资源由多个组织共享，这些组织根据自己的计算需求共同为集群提供计算能力。Capacity Scheduler 具有分层队列、容量保证、安全性、弹性资源、多租户、基于资源的调度、映射等特性。集群的所有资源全部分配给多个队列，提交到队列的所有应用程序都可以使用分配给队列的资源，其他组织未使用的空闲资源可以非抢占式的分配给运行低于容量的队列上的应用程序，确保应用程序获得资源容量下限的同时能弹性的满足不同应用程序的资源需求。
Capacity Scheduler 将集群资源粗略的分配给不同的队列，不能指定队列中应用程序的运行位置。标签调度在 Capacity Scheduler 之上通过给集群各节点打上不同的 Node Label 进行更细粒度的资源划分，应用程序可以指定运行的位置。节点标签具有以下特点：
1. 一个节点只有一个节点标签（也即属于一个节点分区），集群按节点分区被划分为多个不相交的子集群。
2. 根据匹配策略，节点分区具有两种类型：独占和非独占。独占的节点分区将容器分配给完全匹配节点分区的节点；非独占的节点分区将空闲资源共享给请求 default 分区的容器。
3. 每个队列通过设置可访问的节点标签来指定应用程序运行的节点分区。
4. 可以设置每个节点分区内不同队列的资源占比。
5. 在资源请求中指定所需的节点标签，仅在节点具有相同标签时才分配。如果未指定节点标签要求（可通过配置项修改队列的默认标签名），则仅在属于 default 分区的节点上分配此类资源请求。

配置说明

1. 设置 ResourceManager 启用 Capacity Scheduler

标签调度无法单独使用，只能配合 Capacity Scheduler 使用。Yarn 使用 Capacity Scheduler 作为默认调度器，如果您现在正使用其他调度器，请先启动 Capacity Scheduler。在 `${HADOOP_HOME}/etc/hadoop/yarn-site.xml` 中设置：



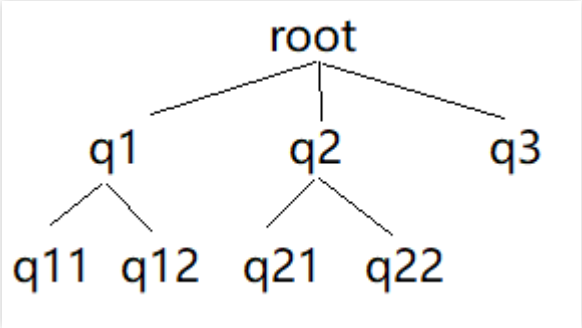
```
<property>
    <name>yarn.resourcemanager.scheduler.class</name>
    <value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.capacity.Capacit
</property>
```

2. 配置 Capacity Scheduler 参数

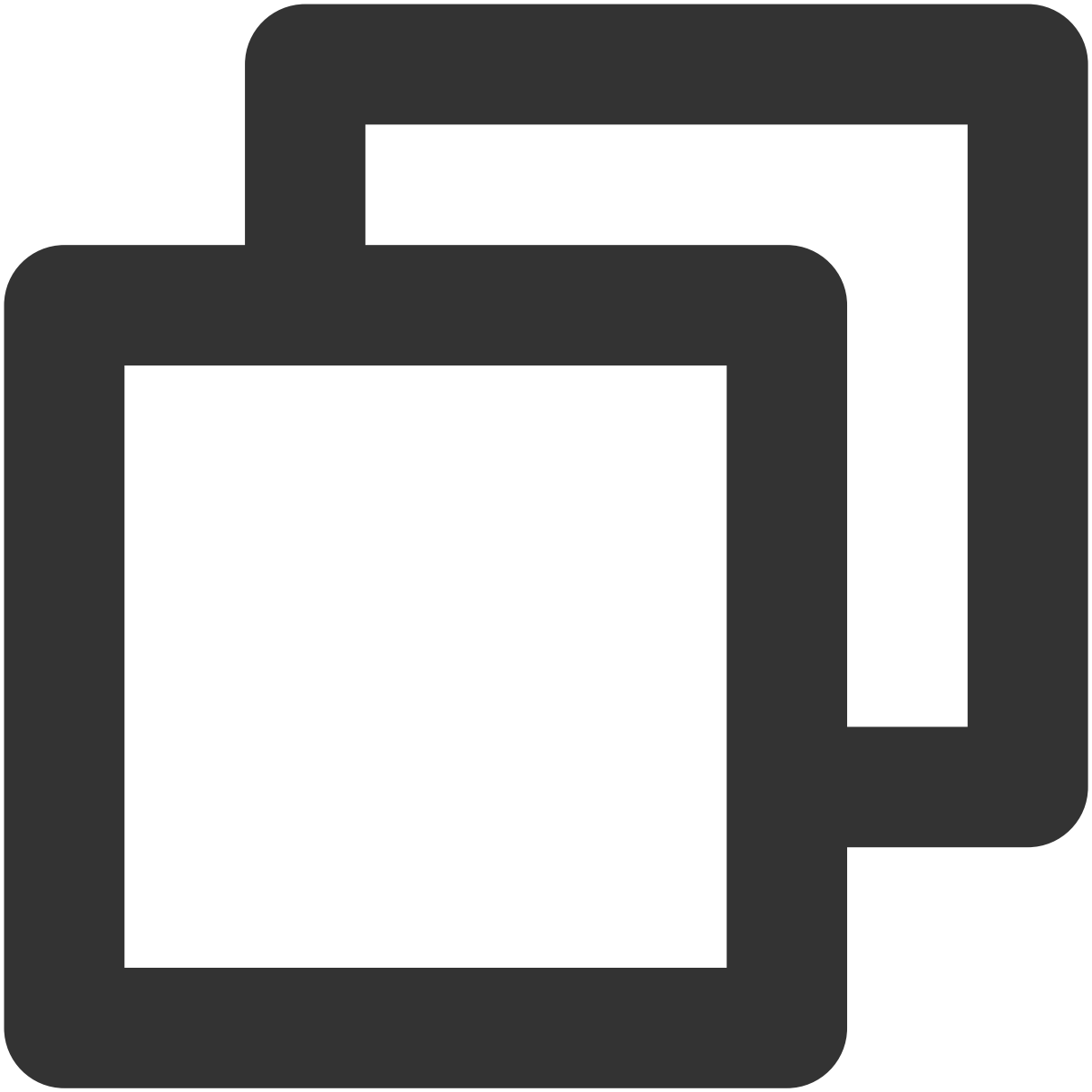
在 `${HADOOP_HOME}/etc/hadoop/capacity-scheduler.xml` 中设置 `yarn.scheduler.capacity.root` 是 Capacity Scheduler 预定义的根队列，其他队列均为根队列的子队列。

所有队列以树的形式组织。`yarn.scheduler.capacity.<queue-path>.queues` 用于设置 queue-path 路径下的子队列，使用逗号分隔。

示例：



上图结构配置如下：



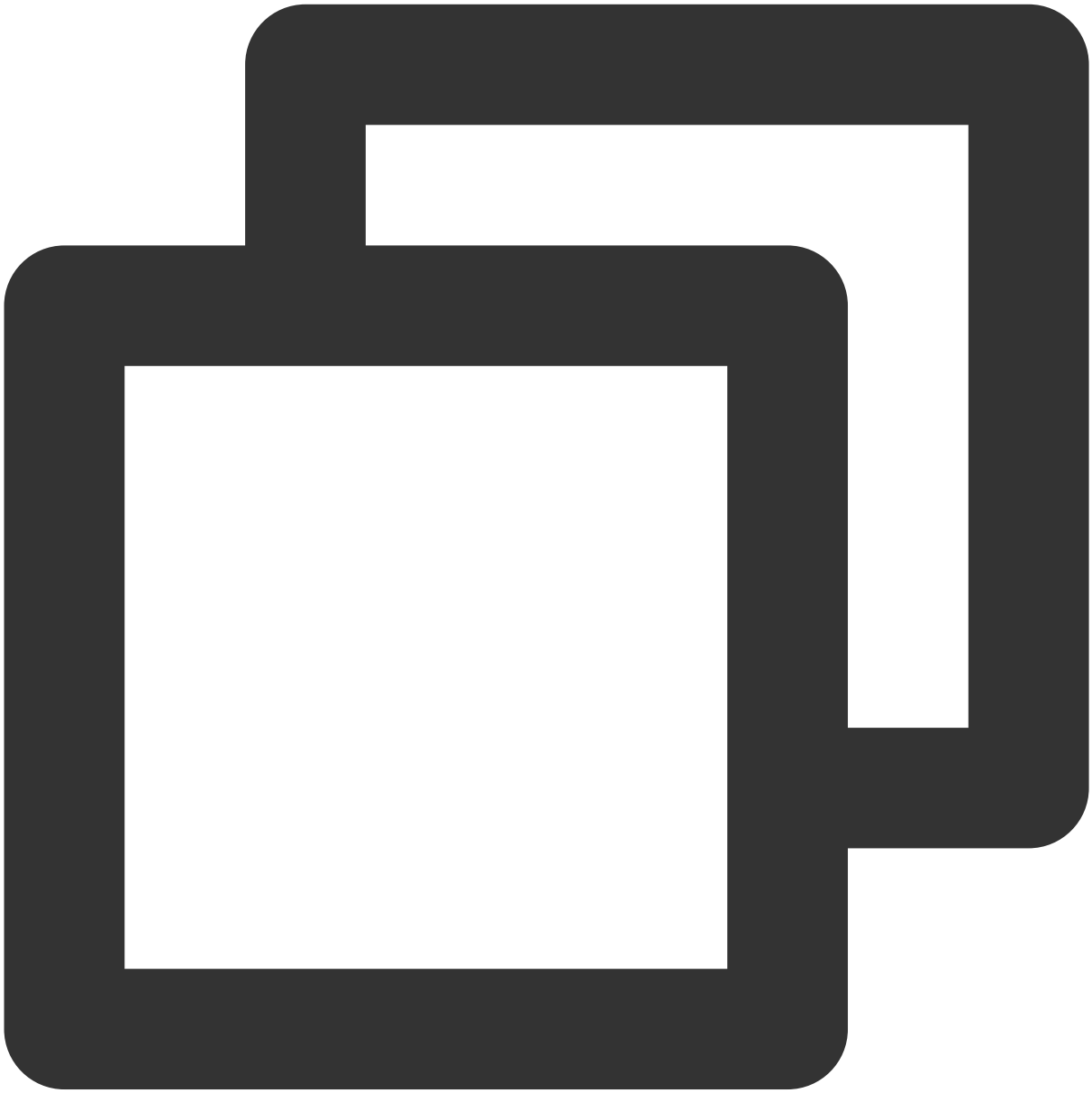
```
<property>
    <name>yarn.scheduler.capacity.root.queues</name>
    <value>q1,q2,q3</value>
</property>
<property>
    <name>yarn.scheduler.capacity.root.q1.queues</name>
    <value>q11,q12</value>
</property>
<property>
    <name>yarn.scheduler.capacity.root.q2.queues</name>
    <value>q21,q22</value>
```

```
</property>
```

其他 Capacity Scheduler 配置请查询文档。

3. 设置 ResourceManager 启用 Node Label

在 `conf/yarn-site.xml` 中设置。



```
<property>
    <name>yarn.node-labels.fs-store.root-dir</name>
    <value>hdfs://namenode:port/path-to-store/node-labels/</value>
</property>
```

```
<property>
    <name>yarn.node-labels.enabled</name>
    <value>true</value>
</property>
<property>
    <name>yarn.node-labels.configuration-type</name>
    <value>centralized or delegated-centralized or distributed</value>
</property>
```

注意

1. 确保已创建 `yarn.node-labels.fs-store.root-dir` ，且 ResourceManager 有权访问。
2. 若将节点标签保存在 RM 的本地文件系统，可使用 `file://home/yarn/node-label` 等路径。但是为了保证集群的高可用，避免 RM 宕机而丢失标签信息，建议将标签信息保存在 HDFS 上。
3. 在 hadoop2.8.2 下需要配置 `yarn.node-labels.configuration-type` 配置项。

4. 配置 Node Label

在 `etc/hadoop/capacity-scheduler.xml` 中设置。

配置项	描述
yarn.scheduler.capacity. `<queue-path>` .capacity	设置队列可以访问属于 DEFAULT 分区的节点的百分比。**每个 parent 队列下直接子队列的 DEFAULT 容量总和必须等于100。**
yarn.scheduler.capacity. `<queue-path>` .accessible-node-labels	设置队列可以访问的特定标签列表，用逗号分隔，如“HBASE, STORM”意味着队列可以访问标签 HBASE 和 STORM。所有队列都可以在没有标签的情况下访问节点，如果不指定此字段，则将从其父字段继承。如果用户想限制队列仅访问没有标签的节点，该字段只需留空即可。
yarn.scheduler.capacity. `<queue-path>` .accessible-node-labels. `<label>` .capacity	设置队列可以访问属于 `<label>` 分区的节点百分比。**每个 parent 队列下直接子队列的 `<label>` 容量总和必须等于100，默认为0。**
yarn.scheduler.capacity. `<queue-path>` .accessible-node-labels. `<label>` .maximum-capacity	与 Capacity Scheduler 配置项 yarn.scheduler.capacity. `<queue-path>` .maximum-capacity 类似，它指定了 `<queue-path>` 在 `<label>` 分区的最大容量，默认为100。
yarn.scheduler.capacity. `<queue-path>` .default-node-label-expression	当资源请求未指定节点标签时，应用将被提交到该值对应的分区。默认情况下，该值为空，即应用程序将被分配没有标签的节点上的容器。

应用示例

准备工作

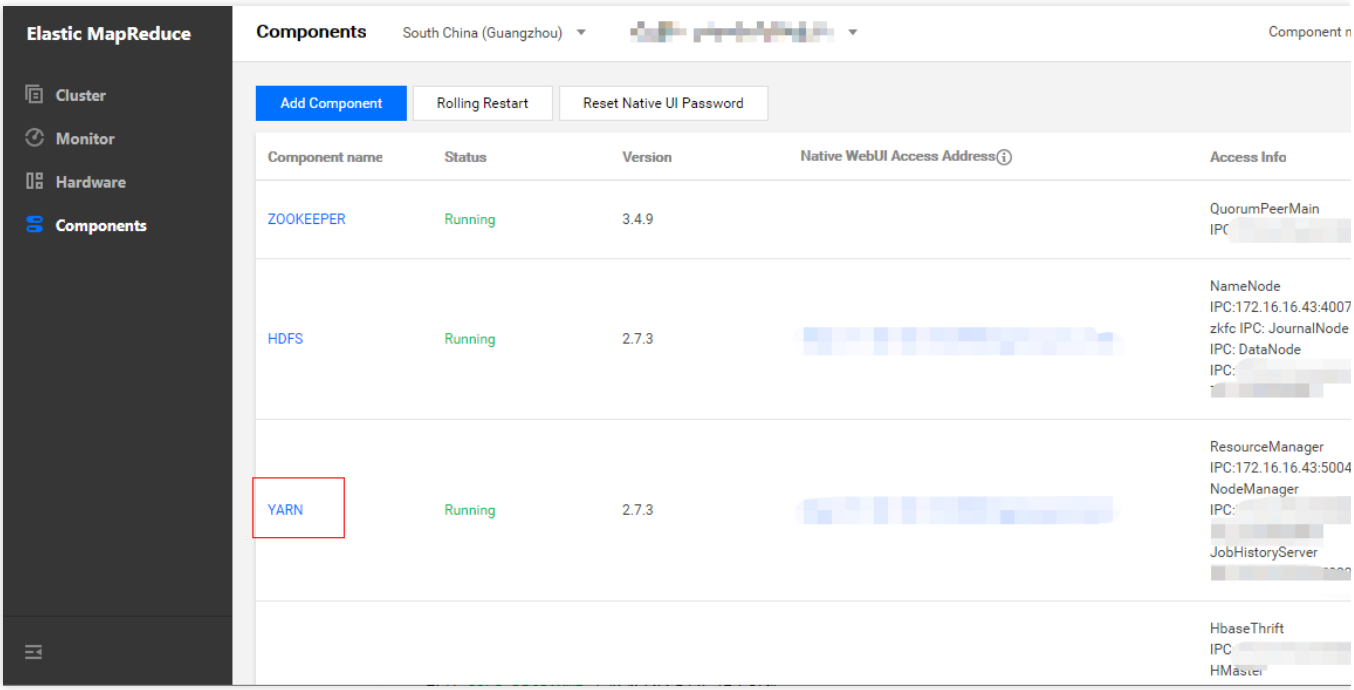
1. 准备集群

确认您已经开通了腾讯云，并且创建了一个 EMR 集群。

2. 检查 YARN 组件配置

在“集群服务”页面中，选择 YARN 组件进入组件管理界面，然后切换至配置管理标签页，修改 `yarn-site.xml` 中相关参数，保存并重启所有 YARN 组件。在角色管理标签页确认 ResourceManager 服务所在节点 IP，之后切换至配置管理标签页修改 `yarn-site.xml` 中相关参数，保存并重启所有 YARN 组件。

在集群列表中单击集群实例 ID，进入集群信息页面，然后单击左侧菜单栏**集群服务**，选择 YARN 组件管理中**操作 > 配置管理**。



确认 RM 的 IP 地址。

在 YARN 组件“配置管理”页面，选择**维度范围**为节点维度，选择节点为 RM 的 IP 地址，单击**修改配置**修改 RM 所在节点 `yarn-site.xml` 的 `yarn.resourcemanager.scheduler.class` 参数。

在 Capacity-Scheduler.xml 中配置 Node Label 与队列的映射关系和占比

1. 创建存储节点标签的 HDFS 目录。

```
[root@172 ~]# cd /usr/local/service/hadoop/
[root@172 hadoop]# su hadoop
[hadoop@172 hadoop]$ hadoop fs -mkdir -p /yarn/node-labels
```

2. 在 `core-site.xml` 中获取 NN 的 IP 和 Port。

```
<property>
        <name>fs.defaultFS</name>
        <value>hdfs://172.16.32.47:4007</value>
</property>
```

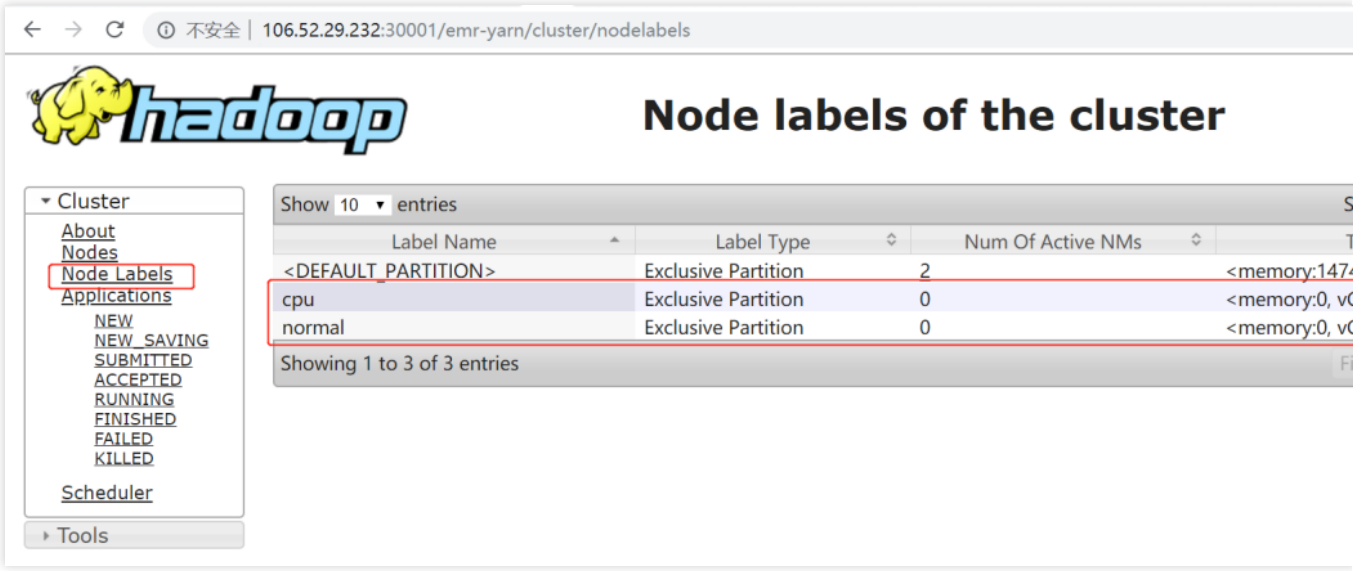
3. 在 master 节点 `yarn-site.xml` 中新建配置项后，重启 ResourceManager。

yarn.node-labels.fs-store.root-dir	hdfs://172.16
yarn.node-labels.enabled	true
yarn.node-labels.configuration-type	centralized

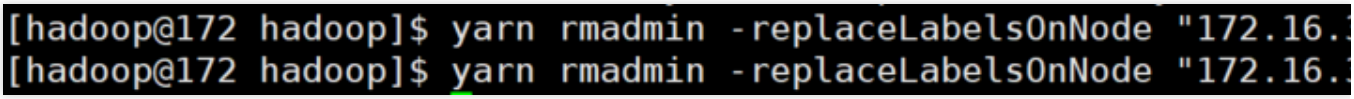
4. 使用 `yarn rmadmin -addToClusterNodeLabels` 命令新增标签。

```
[hadoop@172 hadoop]$ yarn cluster --list-node-labels
Node Labels:
[hadoop@172 hadoop]$ yarn rmadmin -addToClusterNodeLabels "
[hadoop@172 hadoop]$ yarn cluster --list-node-labels
Node Labels: <normal:exclusivity=true>,<cpu:exclusivity=tru
```

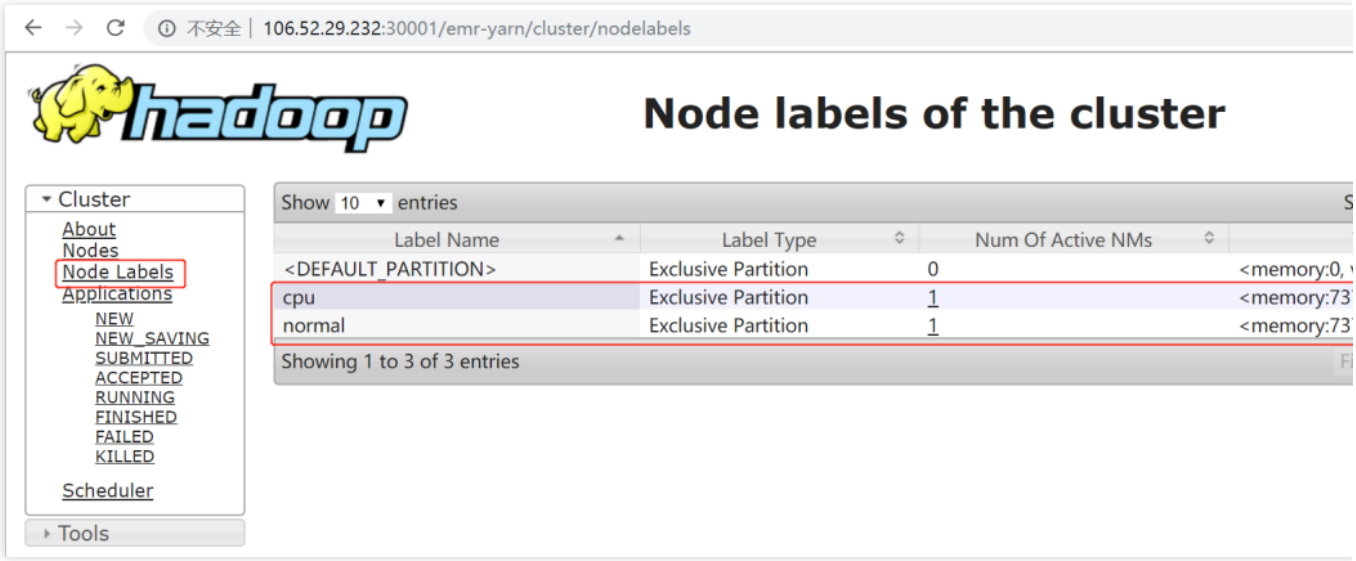
打开 YARN 组件的 WebUI 界面，在 NodeLabels 面板中可以看到集群所有的标签。



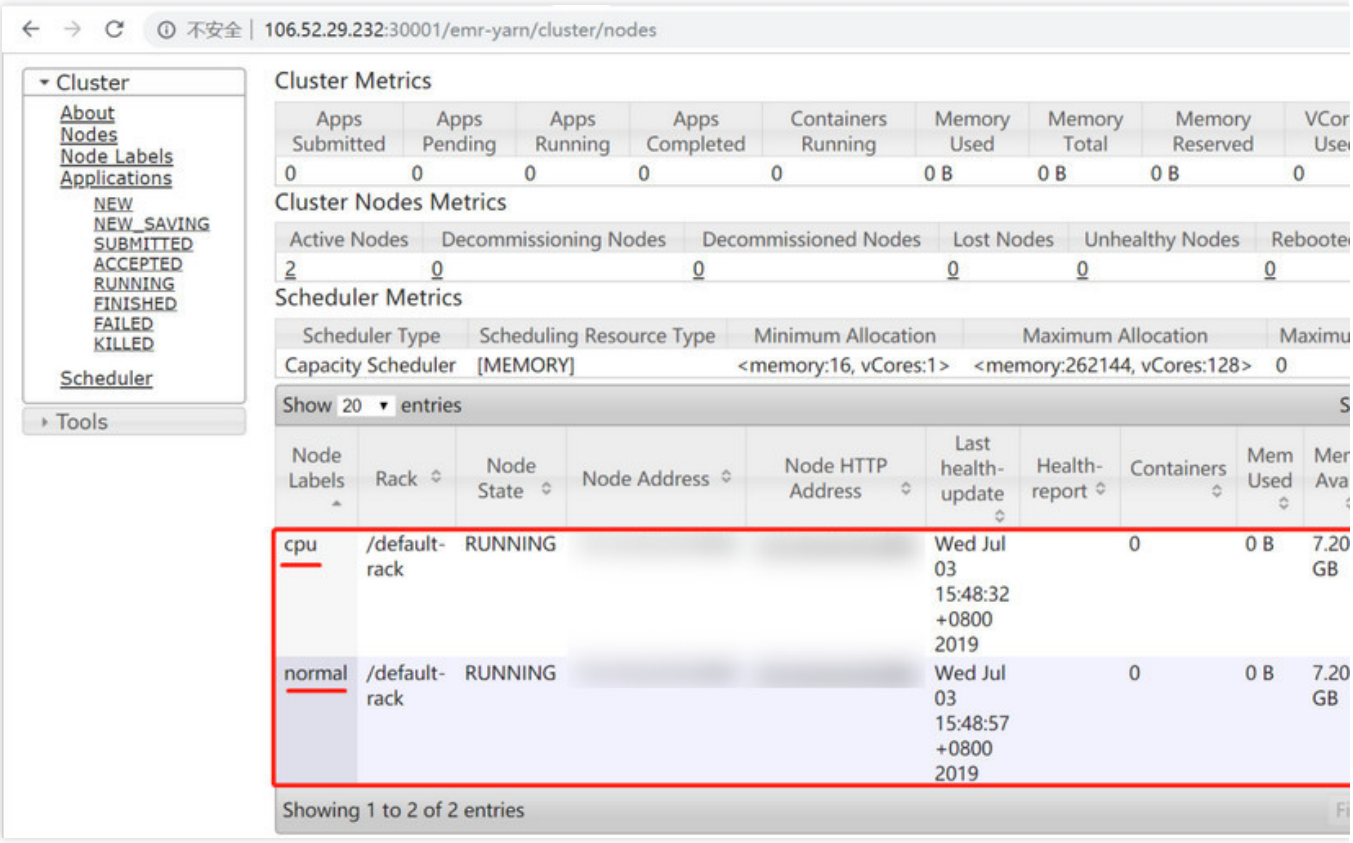
5. 使用 `yarn rmadmin -replaceLabelsOnNode` 命令给节点打标签。



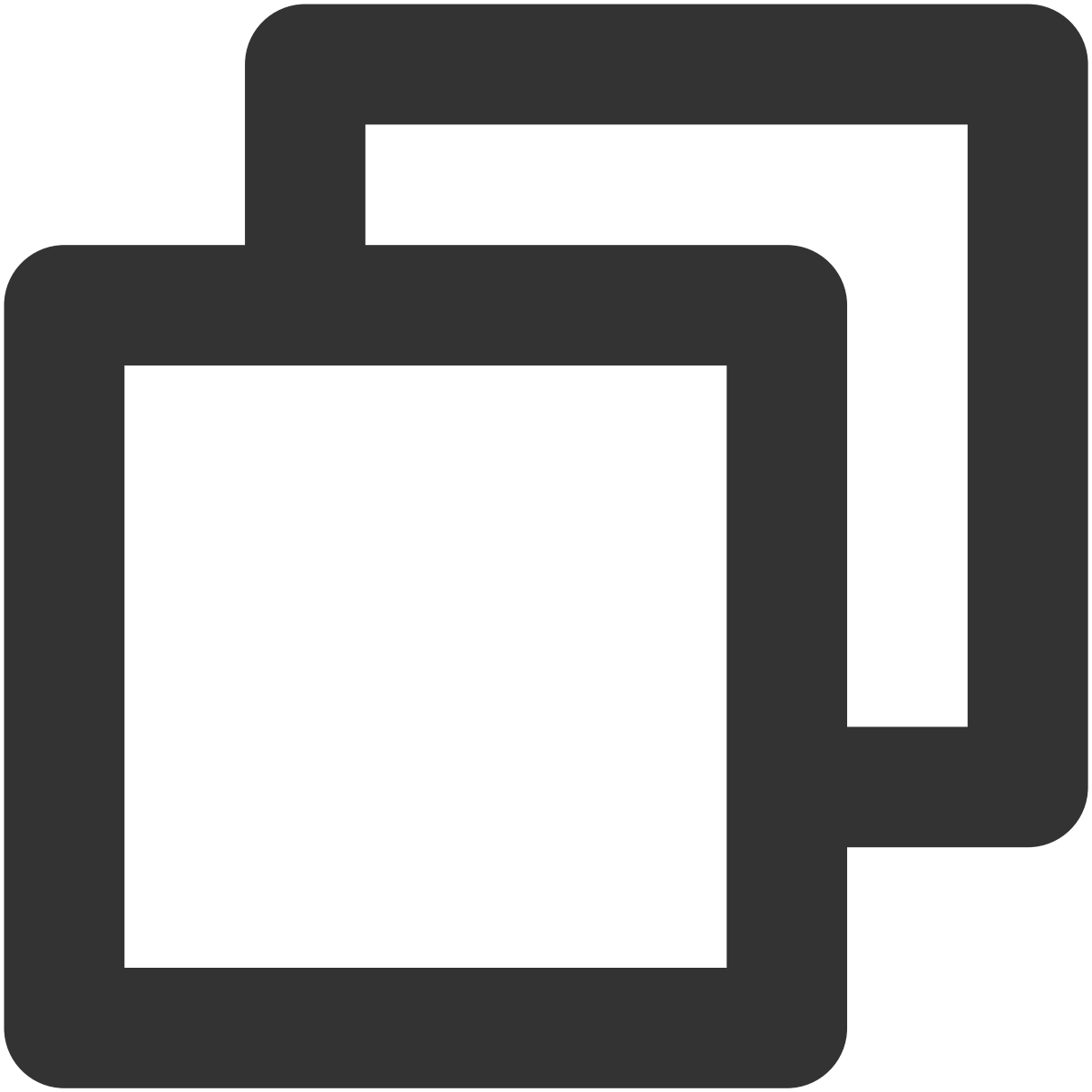
在 NodeLabels 面板中可以看到 normal、cpu 分区的节点个数从0变为1。



在 Scheduler 面板中可以看到，测试系统的两个节点对应的标签已经发生改变。



6. 编辑 `Capacity-Scheduler.xml` 中的配置项，配置集群队列、队列的资源占比和队列的可访问标签。示例如下：

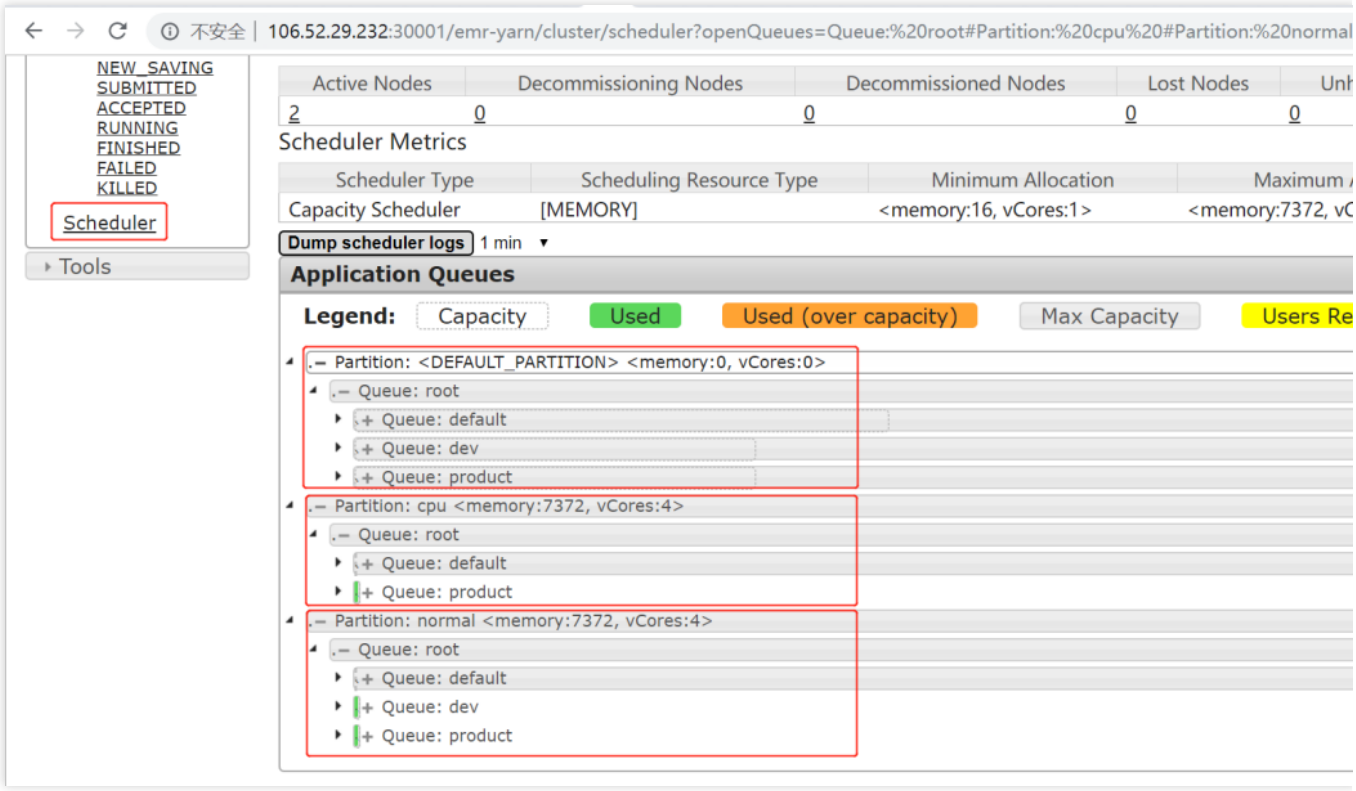


```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration><property>
 <name>yarn.scheduler.capacity.maximum-am-resource-percent</name>
 <value>0.8</value>
</property>
<property>
 <name>yarn.scheduler.capacity.maximum-applications</name>
 <value>1000</value>
</property>
<property>
```

```
 <name>yarn.scheduler.capacity.root.queues</name>
 <value>default,dev,product</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.default.capacity</name>
 <value>20</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.dev.capacity</name>
 <value>40</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.product.capacity</name>
 <value>40</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.accessible-node-labels.cpu.capacity</name>
 <value>100</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.accessible-node-labels.normal.capacity</name>
 <value>100</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.accessible-node-labels</name>
 <value>*</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.dev.accessible-node-labels.normal.capacity</nam
 <value>100</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.product.accessible-node-labels.cpu.capacity</na
 <value>100</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.dev.accessible-node-labels</name>
 <value>normal</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.dev.default-node-label-expression</name>
 <value>normal</value>
</property>
<property>
 <name>yarn.scheduler.capacity.root.product.accessible-node-labels</name>
 <value>cpu</value>
</property>
```

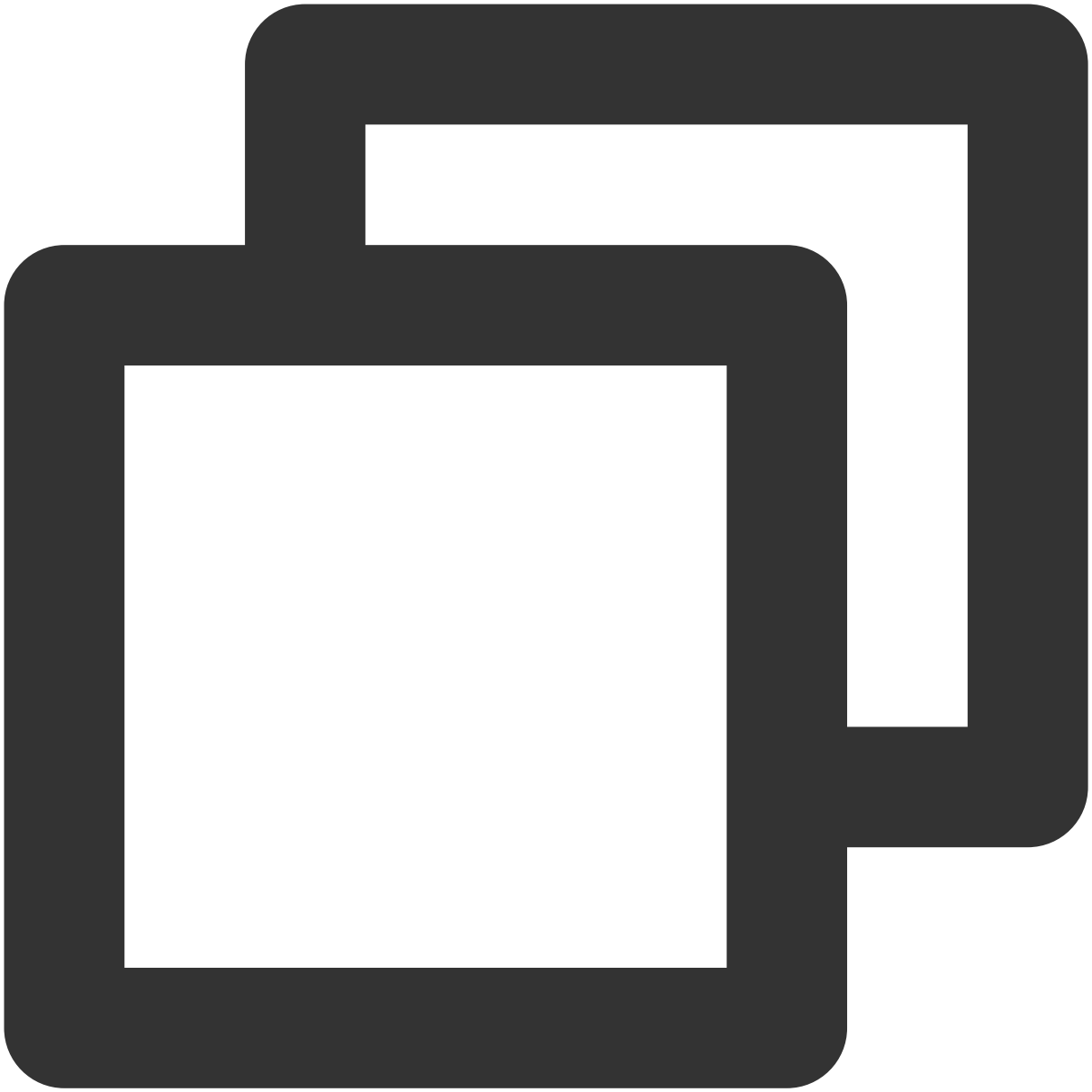
```
<property>
 <name>yarn.scheduler.capacity.root.product.default-node-label-expression</name>
 <value>cpu</value>
</property>
<property>
 <name>yarn.scheduler.capacity.normal.sharable-partitions</name>
 <value>cpu</value>
</property>
<property>
 <name>yarn.scheduler.capacity.normal.require-other-partition-resource</name>
 <value>true</value>
</property>
<property>
 <name>yarn.scheduler.capacity.cpu.sharable-partitions</name>
 <value></value>
</property>
<property>
 <name>yarn.scheduler.capacity.cpu.require-other-partition-resource</name>
 <value>true</value>
</property>
</configuration>
```

在 Scheduler 面板中可以看到，测试集群的3个分区、分区资源分配情况、包含队列情况。在 Application Queues 面板中，共有3个分区：default、normal、cpu，其中 default 分区是默认分区，normal 分区是由带有 normal 标签的节点组成的分区、cpu 分区是由带有 cpu 标签的节点组成的分区。在测试环境中，共有两个节点，这两个节点分别被标记为 normal 和 cpu。单击分区左侧的＋号，能够展开该分区中包含的队列。



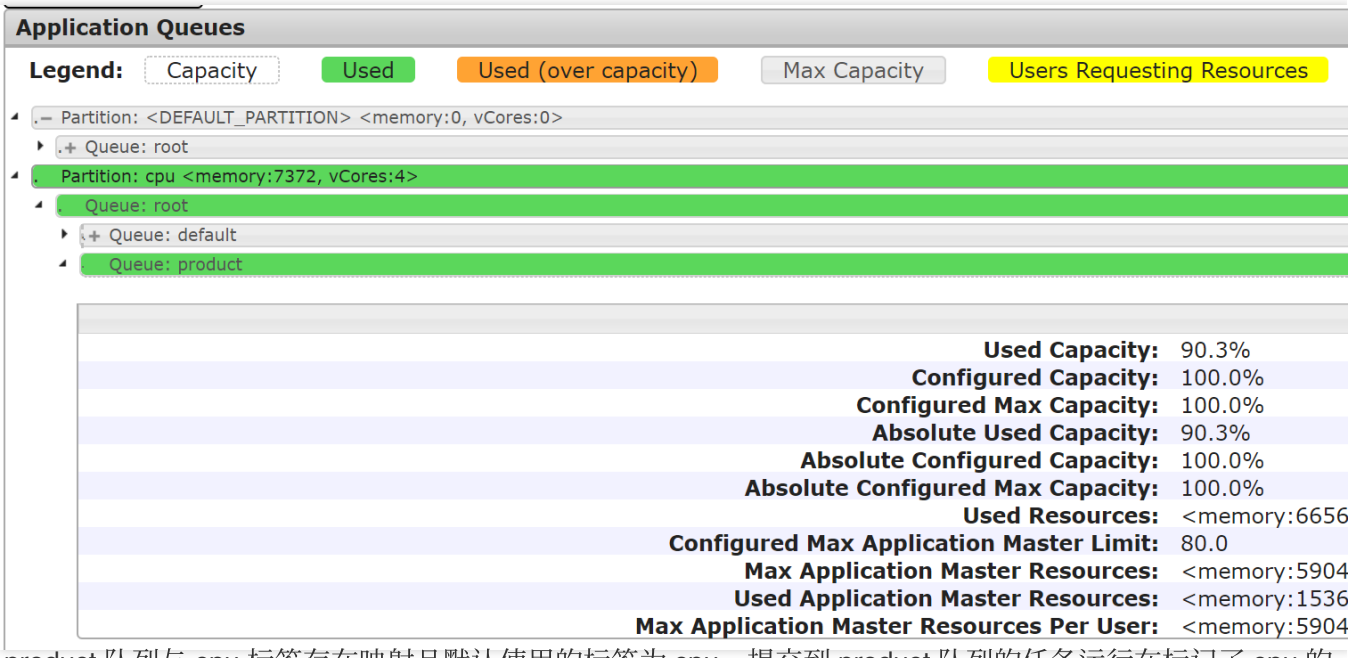
验证标签调度

测试一：将任务提交到 product 队列



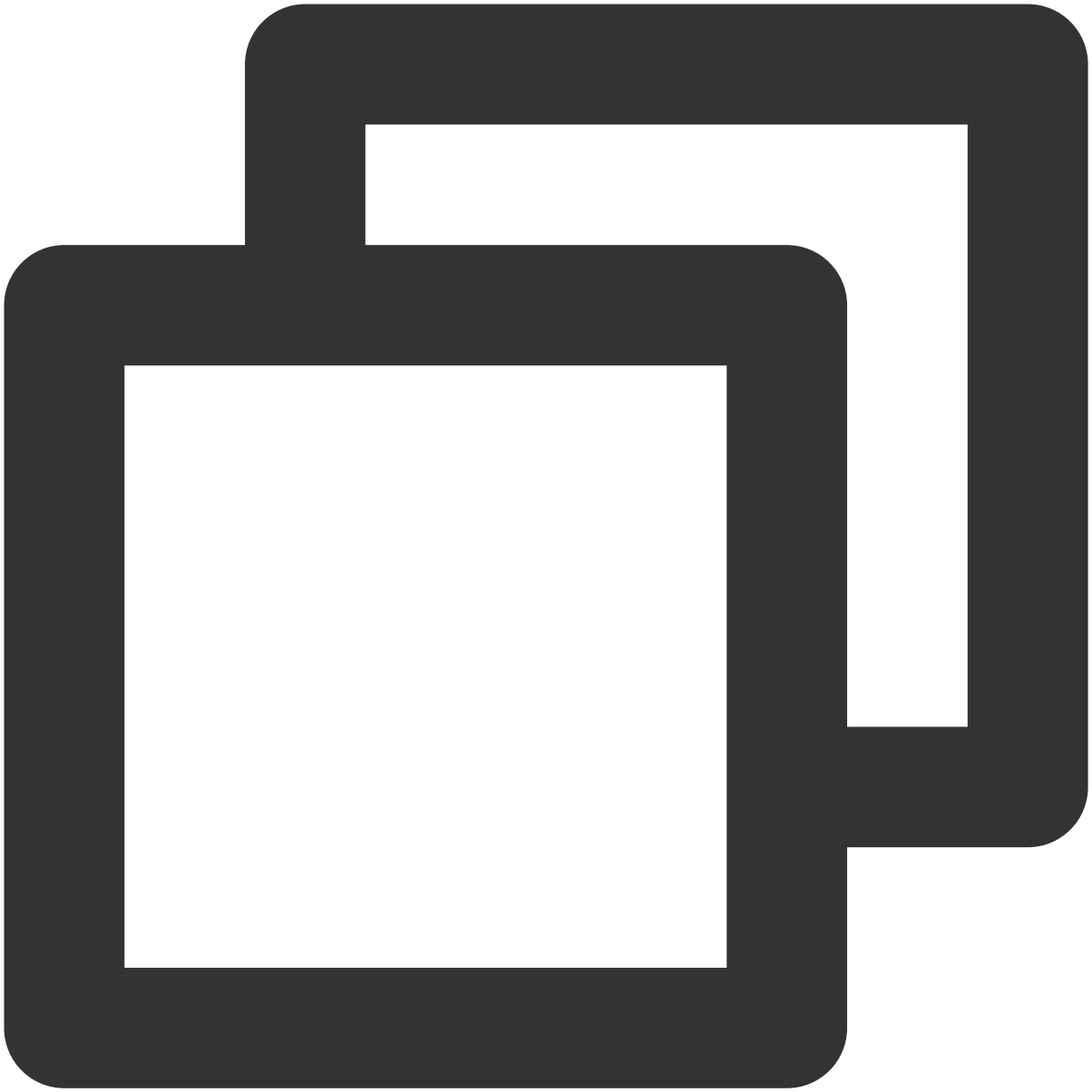
```
[hadoop@172 hadoop]$ cd /usr/local/service/hadoop
[hadoop@172 hadoop]$ yarn jar ./share/hadoop/mapreduce/hadoop-mapreduce-client-jobc
```

任务提交后各分区的队列资源占用情况如下图所示：



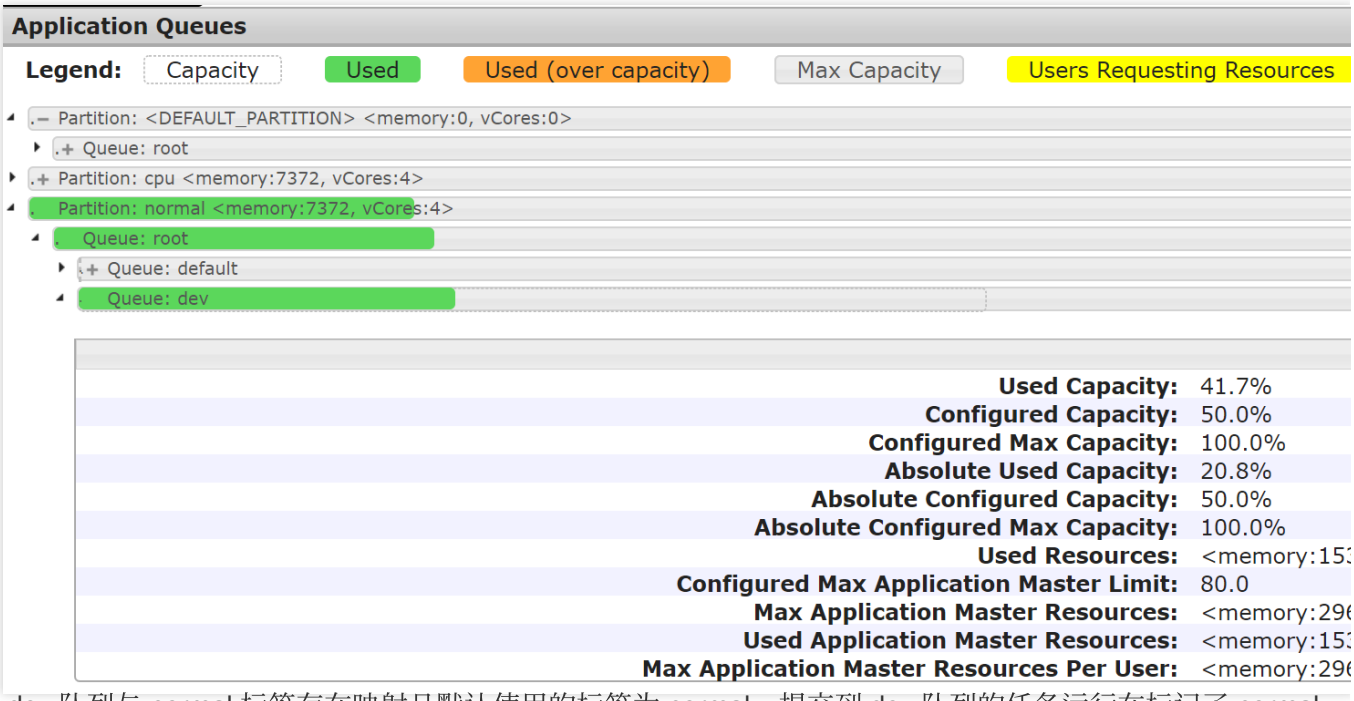
结论：product 队列与 cpu 标签存在映射且默认使用的标签为 cpu，提交到 product 队列的任务运行在标记了 cpu 的节点上。

测试二：将任务提交到 dev 队列



```
[hadoop@172 hadoop]$ cd /usr/local/service/hadoop
[hadoop@172 hadoop]$ yarn jar ./share/hadoop/mapreduce/hadoop-mapreduce-client-jobc
```

任务提交后各分区的队列资源占用情况如下图所示：



结论：dev 队列与 normal 标签存在映射且默认使用的标签为 normal，提交到 dev 队列的任务运行在标记了 normal 的节点上。

Hadoop 最佳实践

最近更新时间：2021-07-13 14:40:56

Hadoop 部分包含了分布式文件系统 HDFS、资源调度框架 YARN 以及迭代式计算框架 MR，腾讯的 Hadoop 版本集成了腾讯云对象存储，让您以 `hadoop fs` 命令行的方式使用对象存储，从而实现计算存储分离，这里的最佳实践包含如下内容。

HDFS

无论您是 HA 集群还是非 HA 集群，**请务必记住不能格式化 namenode**，否则会造成数据丢失，如果是您格式化 namenode 造成的数据丢失腾讯云不承担任何责任。

YARN

腾讯云默认开启的是公平调度，您可以根据您的实际需要修改调度器。

使用 API 分析 HDFS/COS 上的数据

最近更新时间：2023-06-19 17:01:18

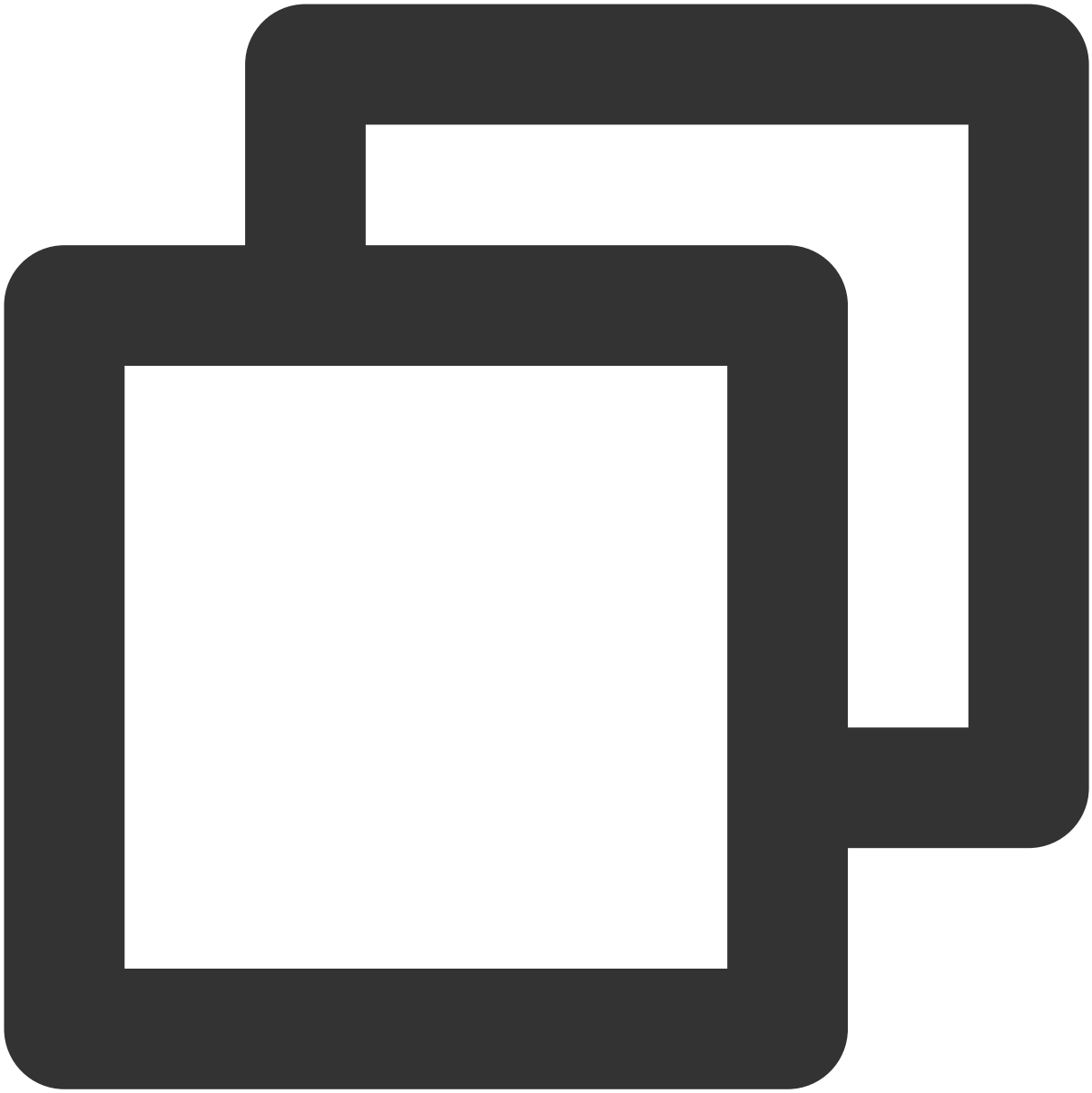
学习 MapReduce 会接触的第一个程序通常都是 WordCount，统计给定文件的单词的词频。本节将会介绍如何自己建立一个工程并编写程序，并且使用编译打包好的程序去统计 HDFS 和腾讯云对象存储 COS 上面的数据，使用的程序基本和 Hadoop 社区的示例程序相同。

1. 开发准备

由于任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建一个存储桶（Bucket）。
确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要选择包含 HDFS 的集群类型，并在基础配置页面开启对象存储的授权。

2. 登录 EMR 服务器

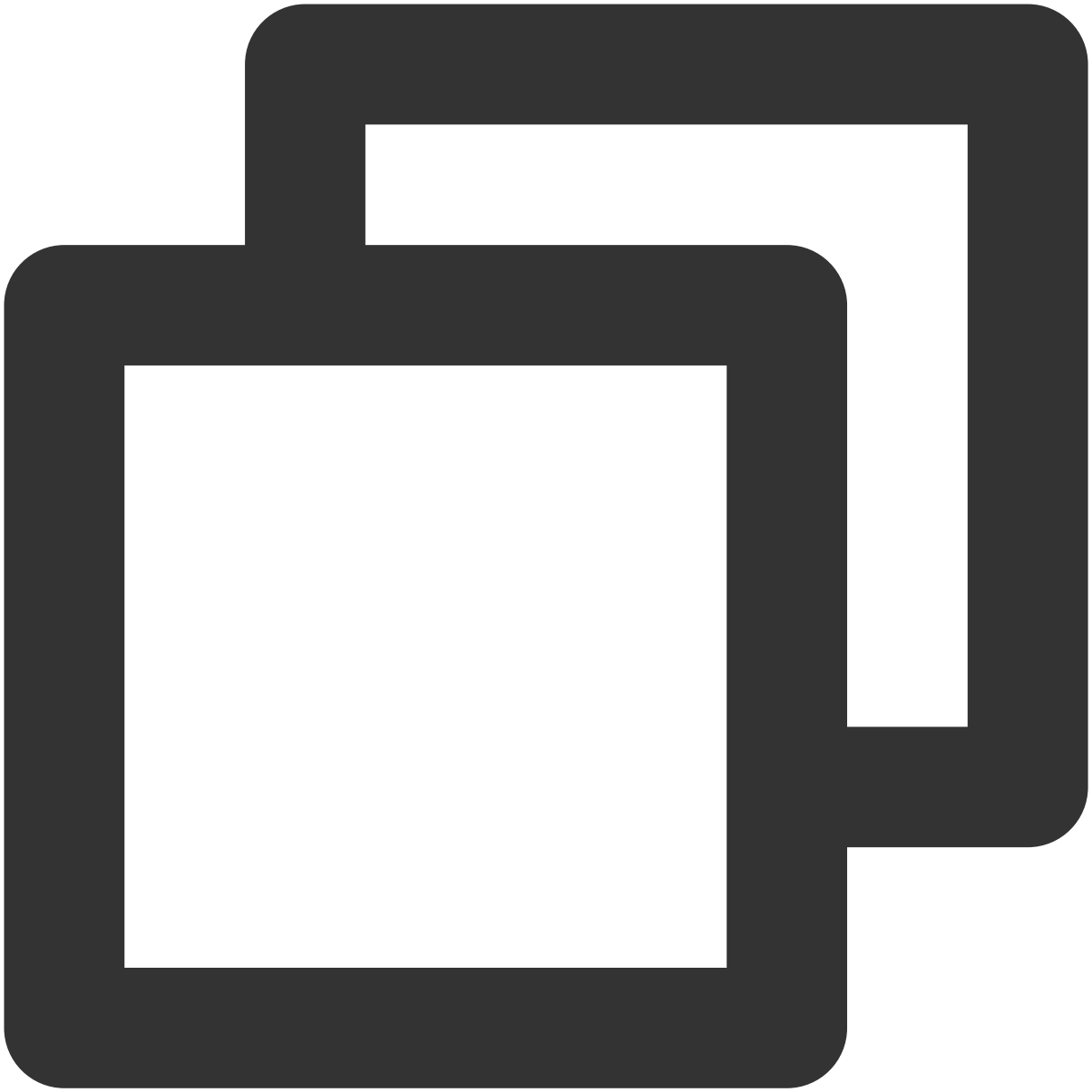
在做相关操作前需要登录到 EMR 集群中的任意一个机器，最好是登录到 Master 节点。EMR 是建立在 Linux 操作系统的腾讯云服务器 CVM 上的，所以在命令行模式下使用 EMR 需要登录 CVM 服务器。
EMR 集群创建后，在控制台中选择弹性 MapReduce，在**详情>集群资源>资源管理>Master 节点**中活跃的 Master 节点的 CVM ID，即可进入云服务器控制台并且找到 EMR 对应的云服务器。
登录 CVM 的方法请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。
输入正确后，即可进入 EMR 集群的命令行界面。所有的 Hadoop 操作都在 Hadoop 用户下，登录 EMR 节点后默认在 root 用户，需要切换到 Hadoop 用户。使用如下命令切换用户，并且进入 Hadoop 文件夹下：



```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/hadoop
[hadoop@172 hadoop]$
```

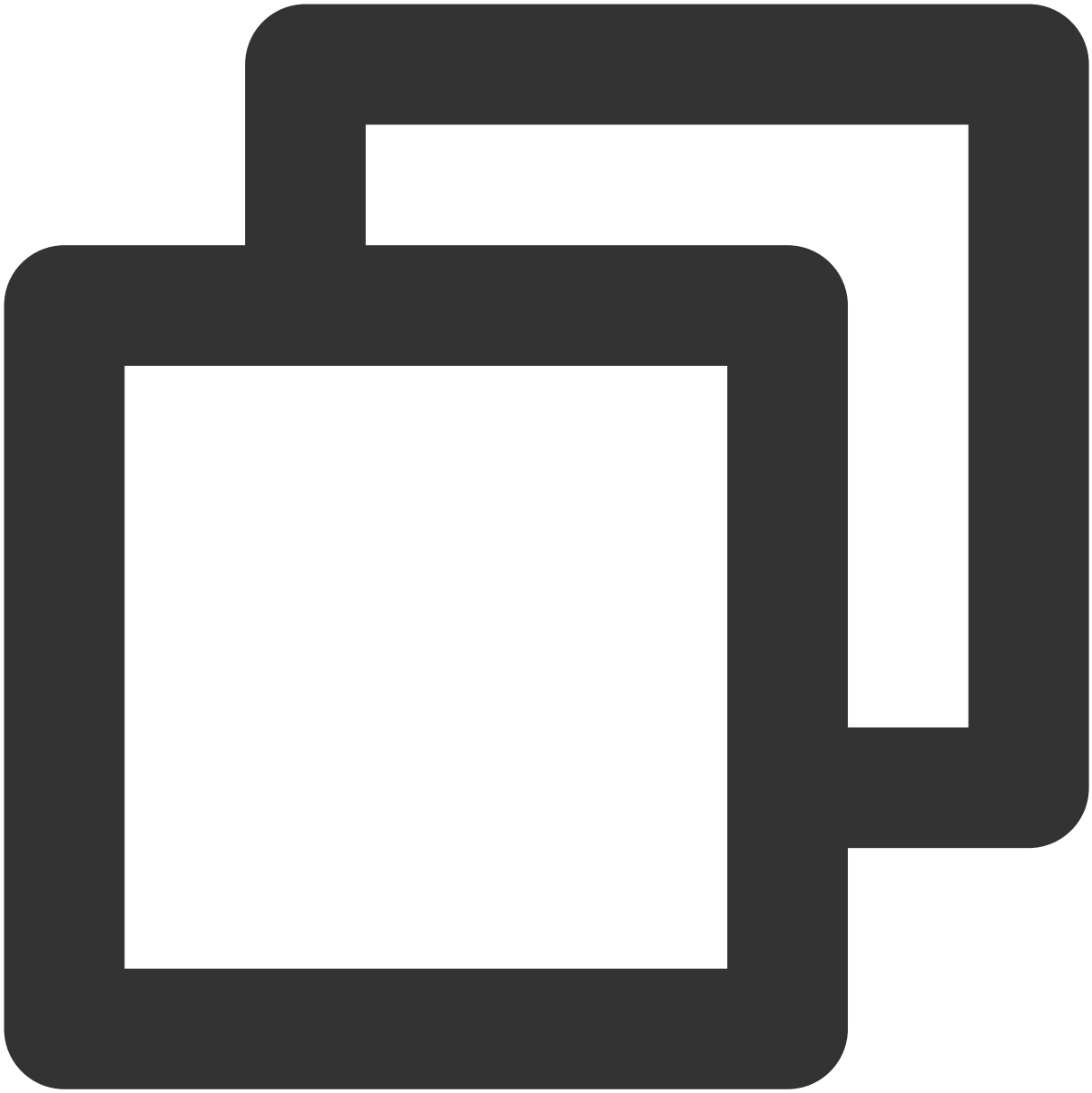
3. 数据准备

您需要准备统计的文本文件。分为两种方式：**将数据存储在 HDFS 集群**和**数据存储在 COS**。
首先在本地新建一个 txt 文件 test.txt，并且添加一些英语单词：



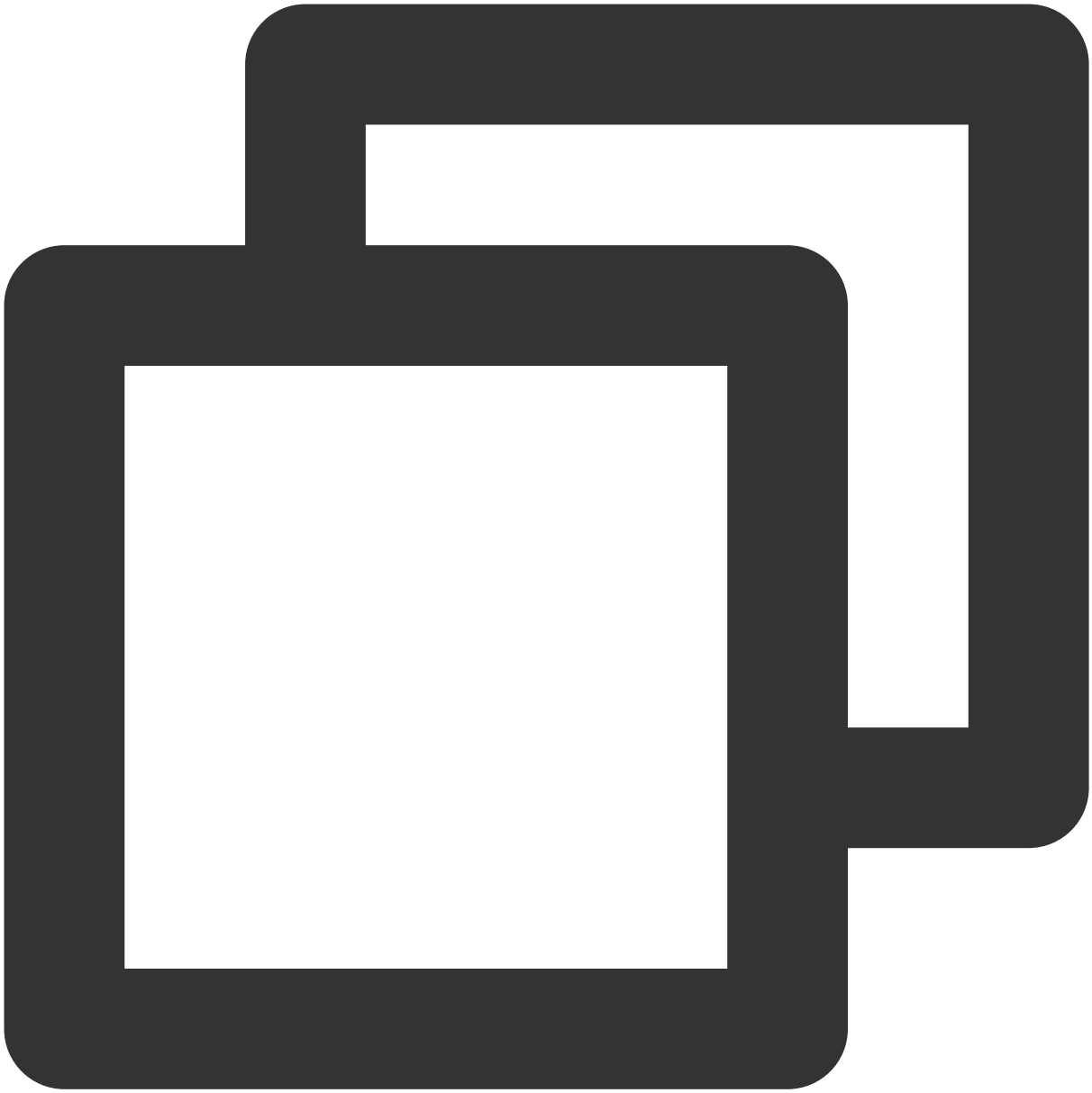
```
Hello World.
this is a message.
this is another message.
Hello world, how are you?
```

把本地的数据上传到云服务器。可以使用 scp 或者 sftp 服务来把本地文件上传到 EMR 集群的云服务器中。在本地 shell 使用：



```
scp $localfile root@公网IP地址:$remotefolder
```

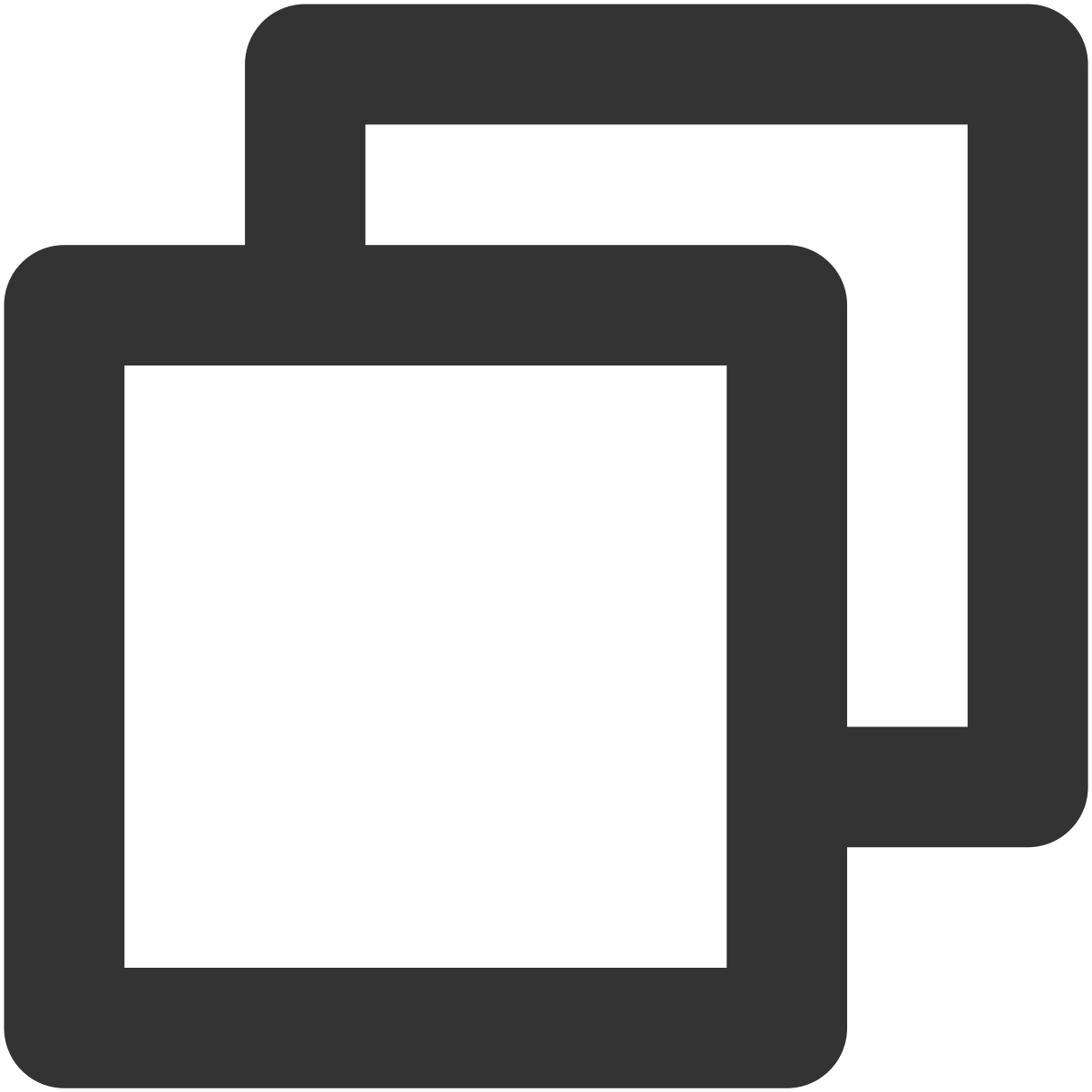
其中，$localfile 是您的本地文件的路径加名称；root 为 CVM 服务器用户名；公网 IP 地址可以在 EMR 控制台的节点信息中或者在云服务器控制台查看；$remotefolder 是您想存放文件的 CVM 服务器路径。
上传完成后，在 EMR 集群命令行中即可查看对应文件夹下是否有相应文件，此处上传到了 EMR 集群的 `/usr/local/service/hadoop` 路径下。



```
[hadoop@172 hadoop]$ ls -l
```

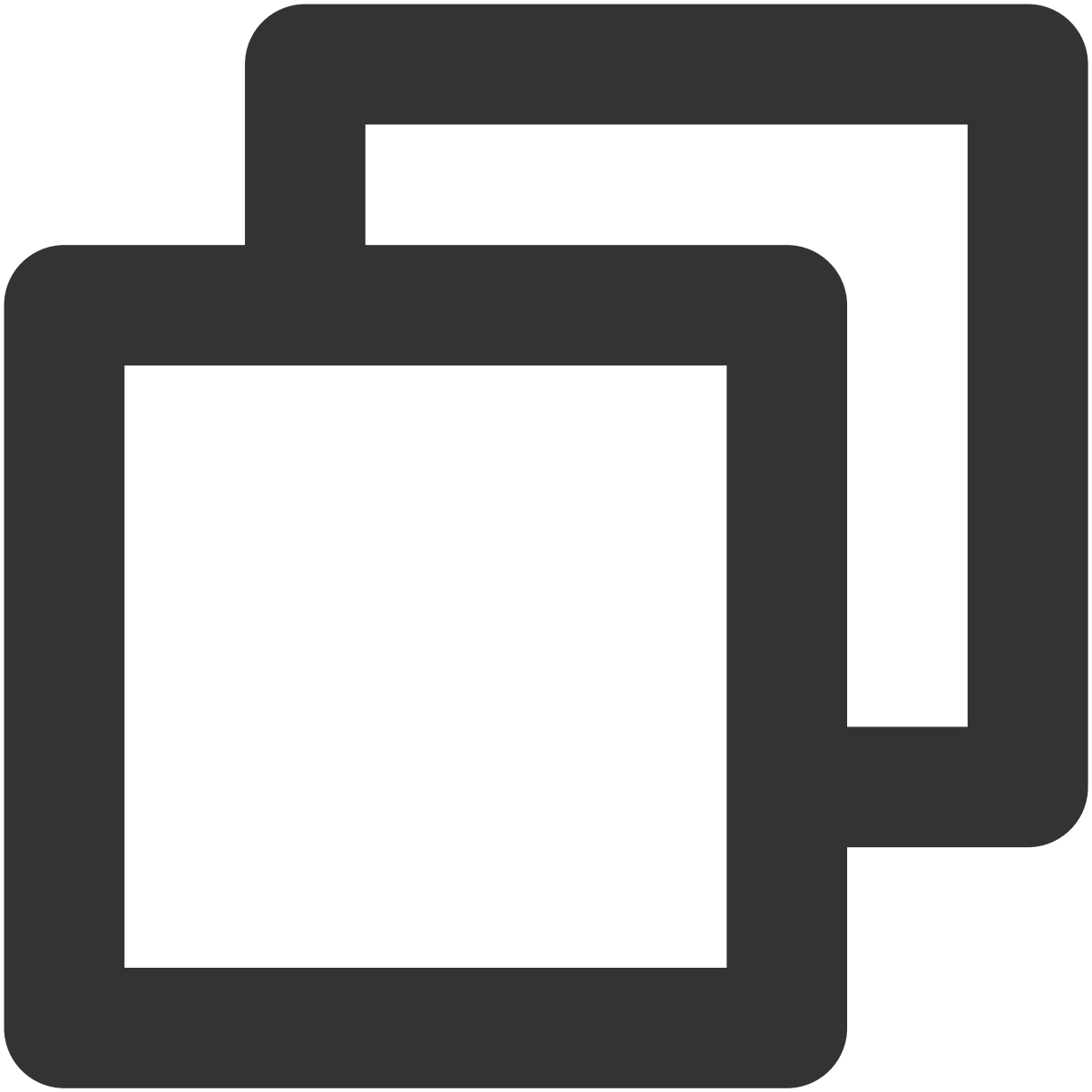
数据存放在 HDFS

将数据上传到腾讯云服务器后，可以把数据拷贝到 HDFS 集群。通过如下指令把文件拷贝到 Hadoop 集群：



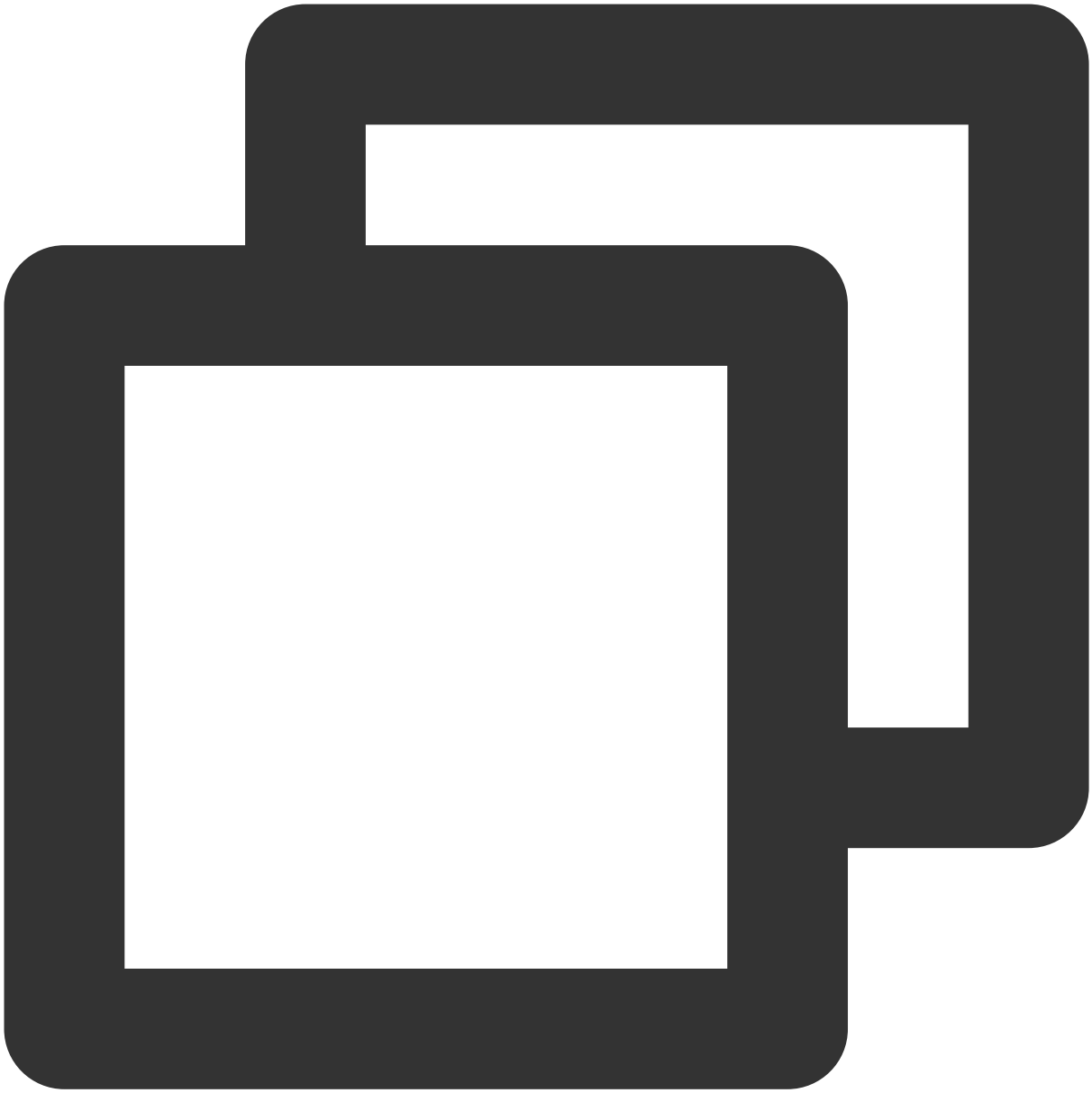
```
[hadoop@172 hadoop]$ hadoop fs -put /usr/local/service/hadoop/test.txt /user/hadoop
```

拷贝完成后，使用以下指令查看拷贝好的文件：



```
[hadoop@172 hadoop]$ hadoop fs -ls /user/hadoop
输出：
-rw-r--r-- 3 hadoop supergroup 85 2018-07-06 11:18 /user/hadoop/test.txt
```

如果 Hadoop 下面没有 `/user/hadoop` 文件夹，用户可以自己创建，指令如下：

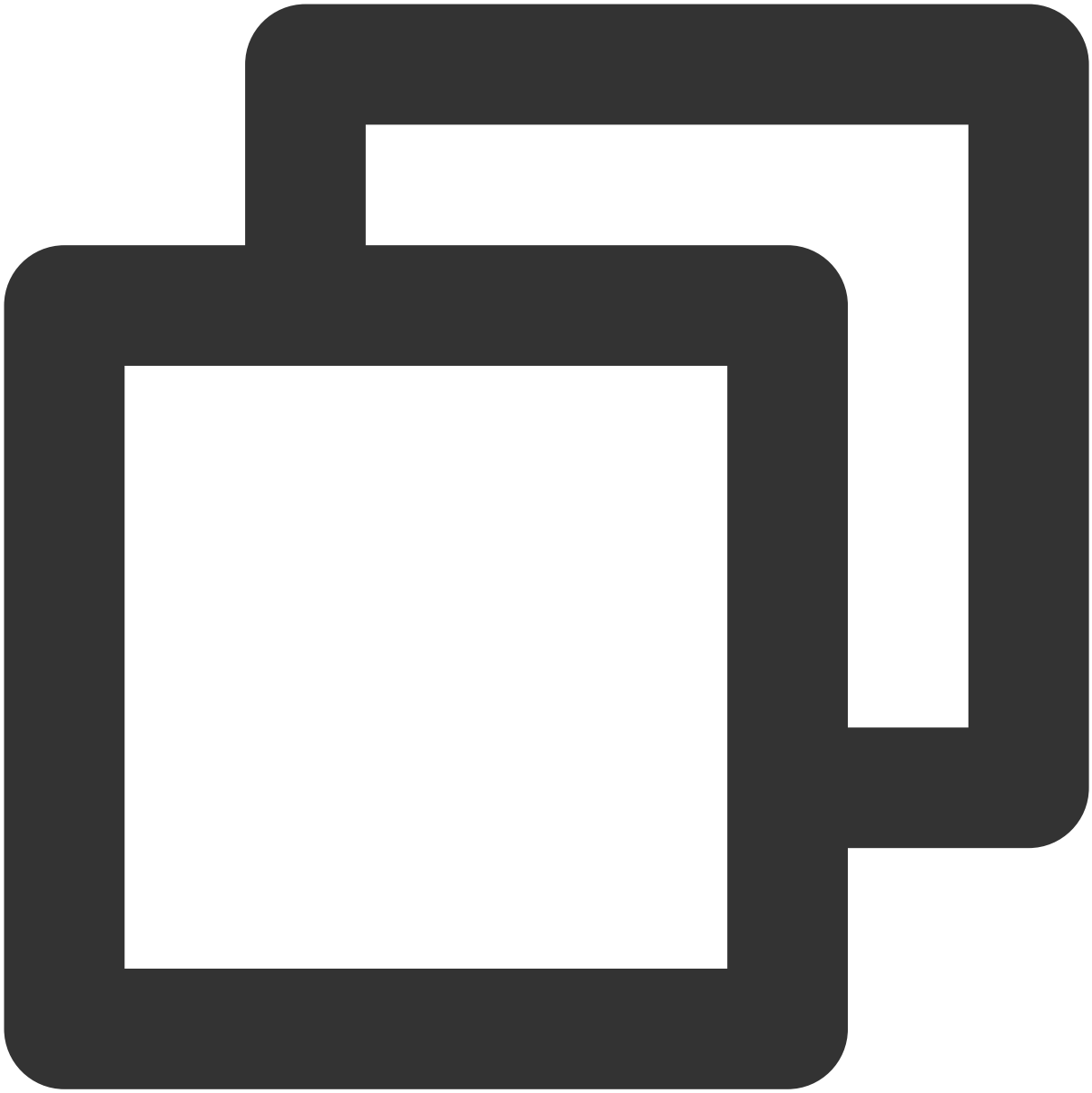


```
[hadoop@172 hadoop]$ hadoop fs -mkdir /user/hadoop
```

更多 hadoop 指令见 HDFS 常见操作。

数据存放在 COS

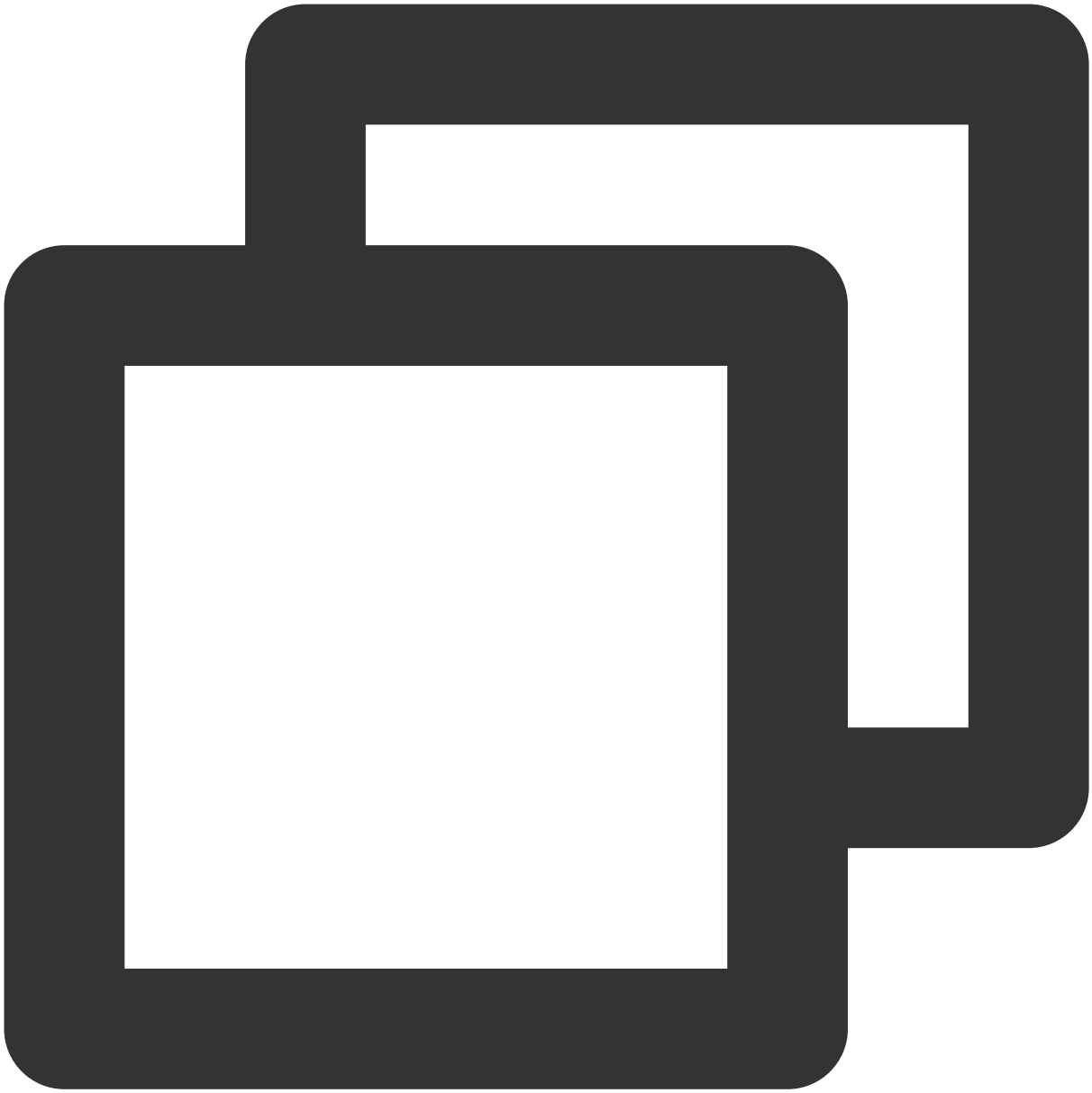
数据存放在 COS 中有两种方式：**从本地直接通过 COS 的控制台上传**和**通过 Hadoop 命令上传**。
从本地直接通过 COS 控制台直接上传，数据文件上传后，可通过如下命令查看：



```
[hadoop@10 hadoop]$ hadoop fs -ls cosn://$bucketname/ test.txt
-rw-rw-rw- 1 hadoop hadoop 1366 2017-03-15 19:09 cosn://$bucketname/test.txt
```

其中 $bucketname 替换成您的储存桶的名字加路径。

通过 Hadoop 命令上传，指令如下：



```
[hadoop@10 hadoop]$ hadoop fs -put test.txt cosn://$bucketname /
[hadoop@10 hadoop]$ hadoop fs -ls cosn:// $bucketname / test.txt
-rw-rw-rw- 1 hadoop hadoop 1366 2017-03-15 19:09 cosn://$bucketname / test.txt
```

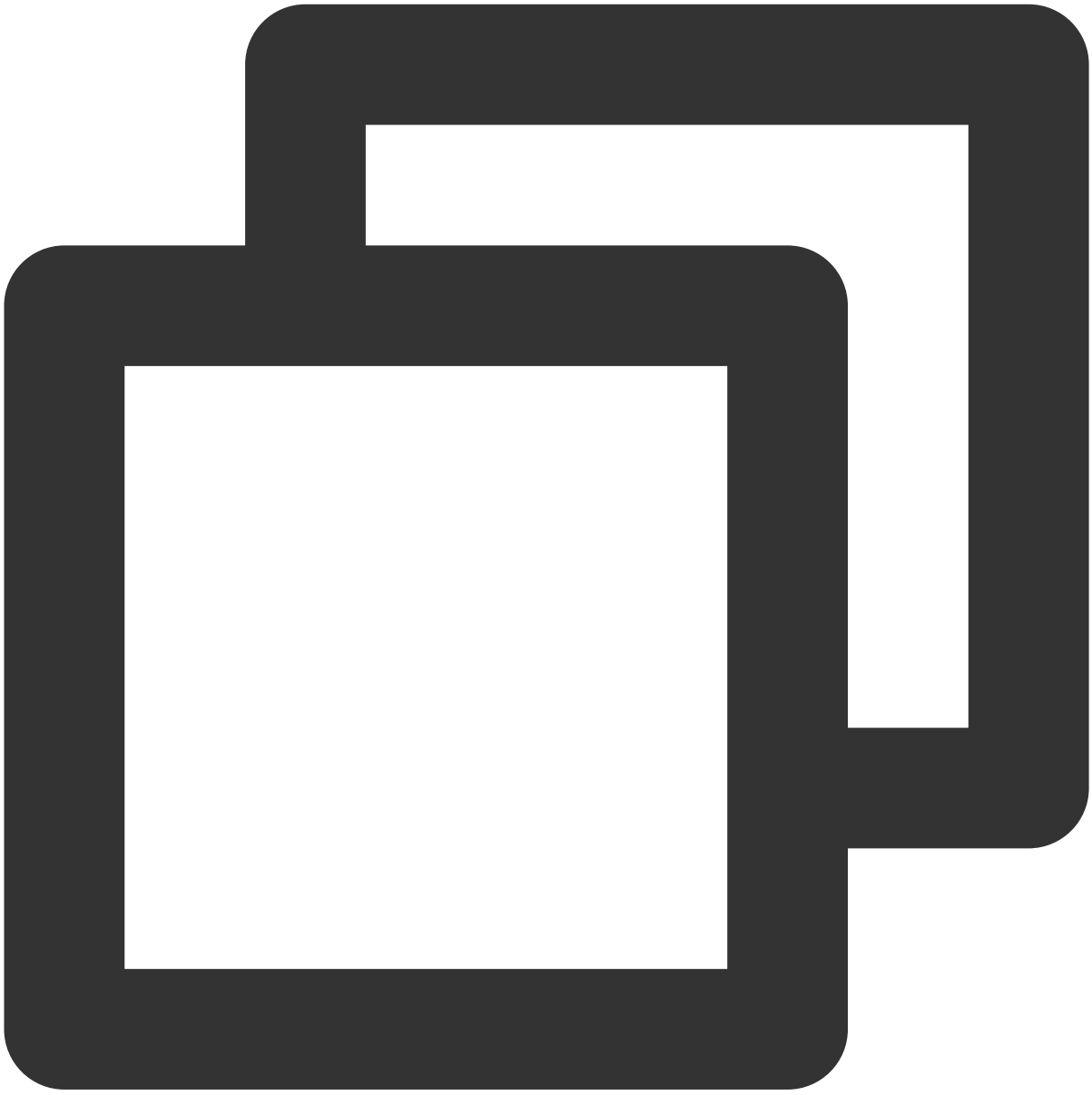
4. 使用 Maven 创建工程

推荐使用 Maven 来管理工程。Maven 是一个项目管理工具，可以方便的管理项目的依赖信息，即它可通过 pom.xml 文件的配置获取 jar 包，而不用手动去添加。

首先下载并安装 Maven，配置好 Maven 的环境变量，如果您使用 IDE，请在 IDE 中设置好 Maven 相关配置。

新建一个 Maven工程

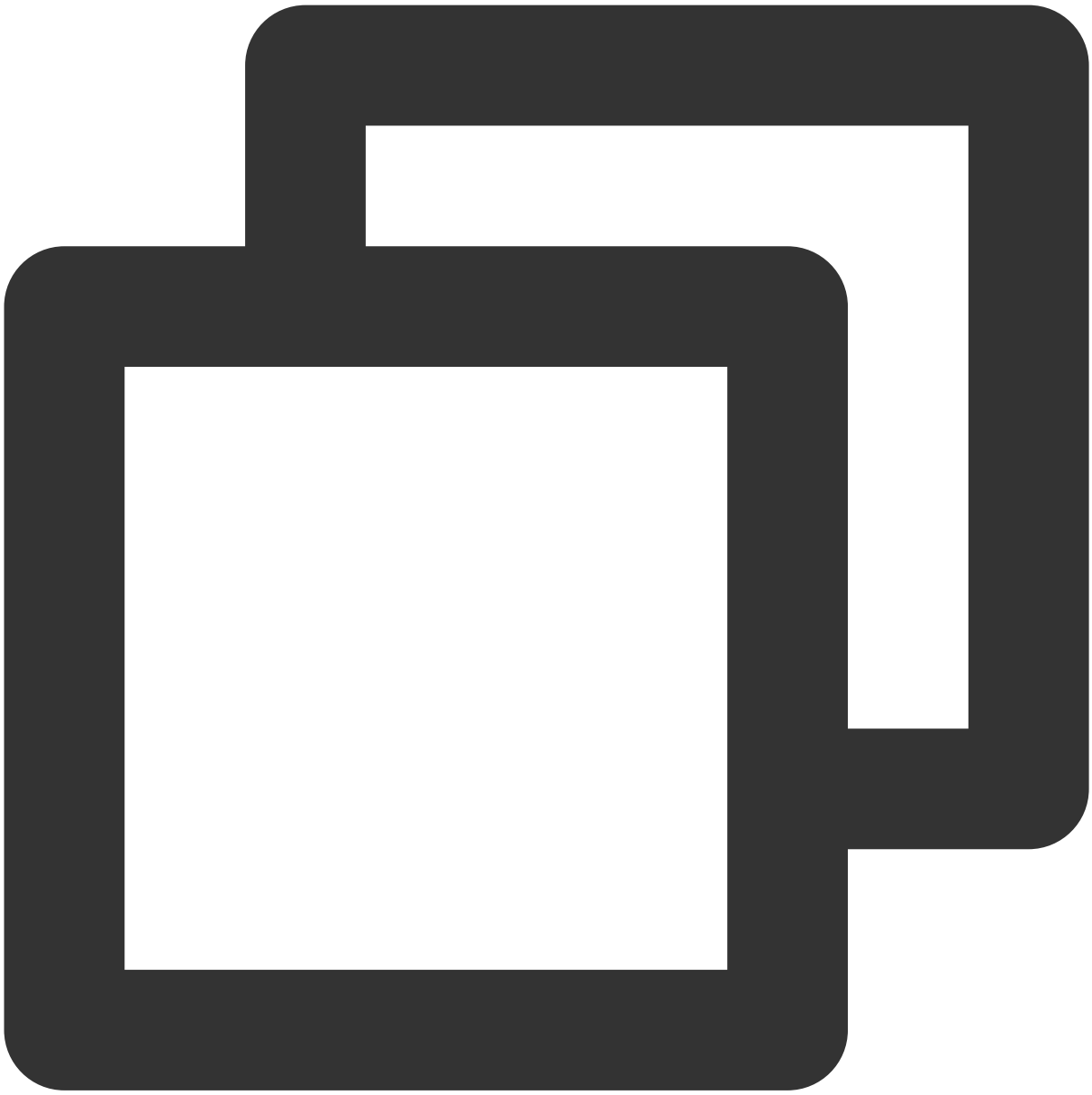
在命令行下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：



```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifac
-DarchetypeArtifactId=maven-archetype-quickstart
```

其中 `$yourgroupID` 即为您的包名； `$yourartifactID` 为您的项目名称； `maven-archetype-quickstart` 表示创建一个 Maven Java 项目，工程创建过程中需要下载一些文件，请保持网络通畅。

创建成功后，在 `D://mavenWorkplace` 目录下就会生成一个名为 `$yourartifactID` 的工程文件夹。其中的文件结构如下所示：

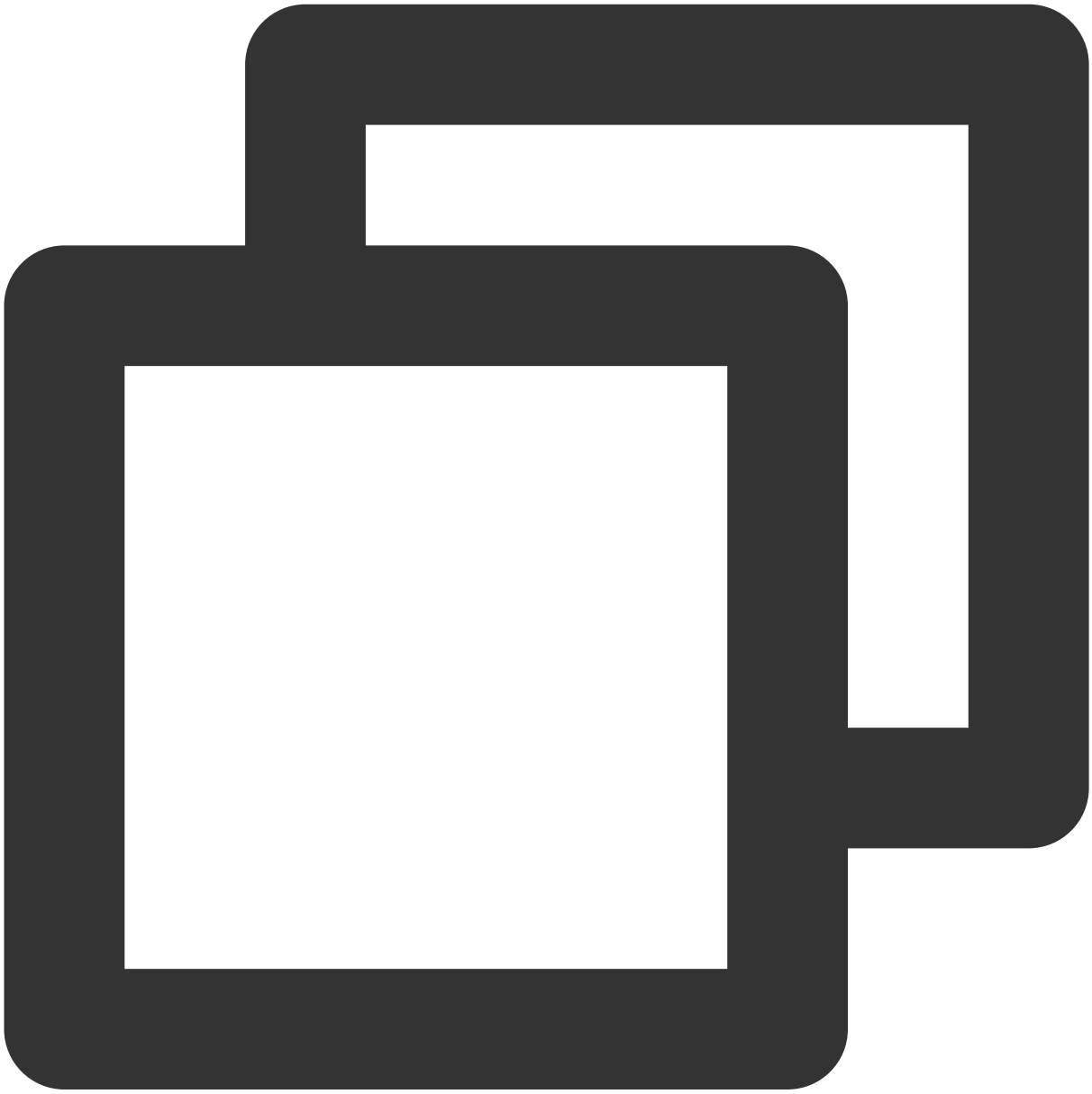


```
simple
     ---pom.xml        核心配置，项目根下
     ---src
        ---main
            ---java        Java 源码目录
           ---resources     Java 配置文件目录
       ---test
           ---java        测试源码目录
```

```
            ---resources      测试配置目录
```

主要关注 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。

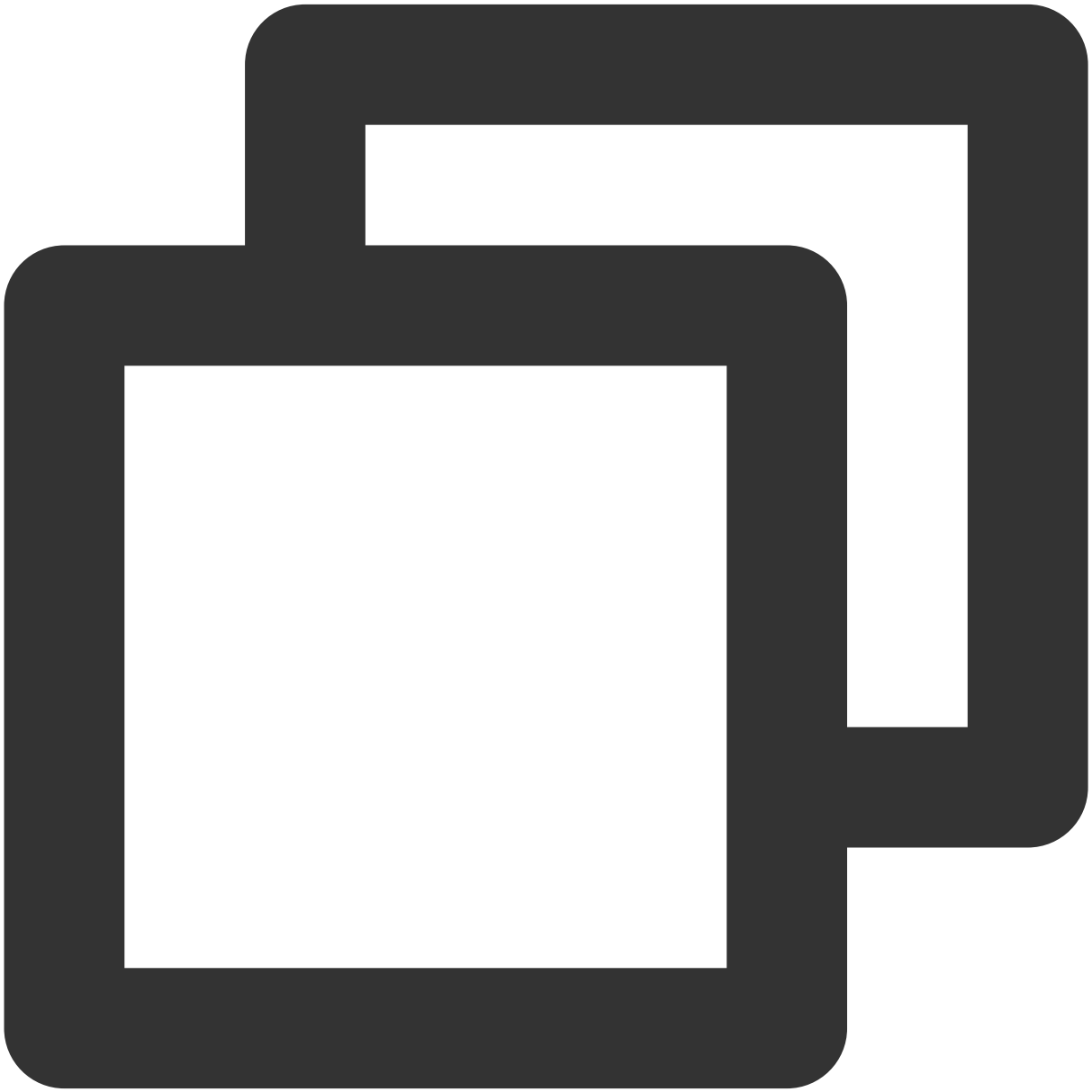
首先在 pom.xml 中添加 Maven 依赖：



```
<dependencies>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-common</artifactId>
            <version>2.7.3</version>
```

```
        </dependency>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-mapreduce-client-core</artifactId>
            <version>2.7.3</version>
        </dependency>
</dependencies>
```

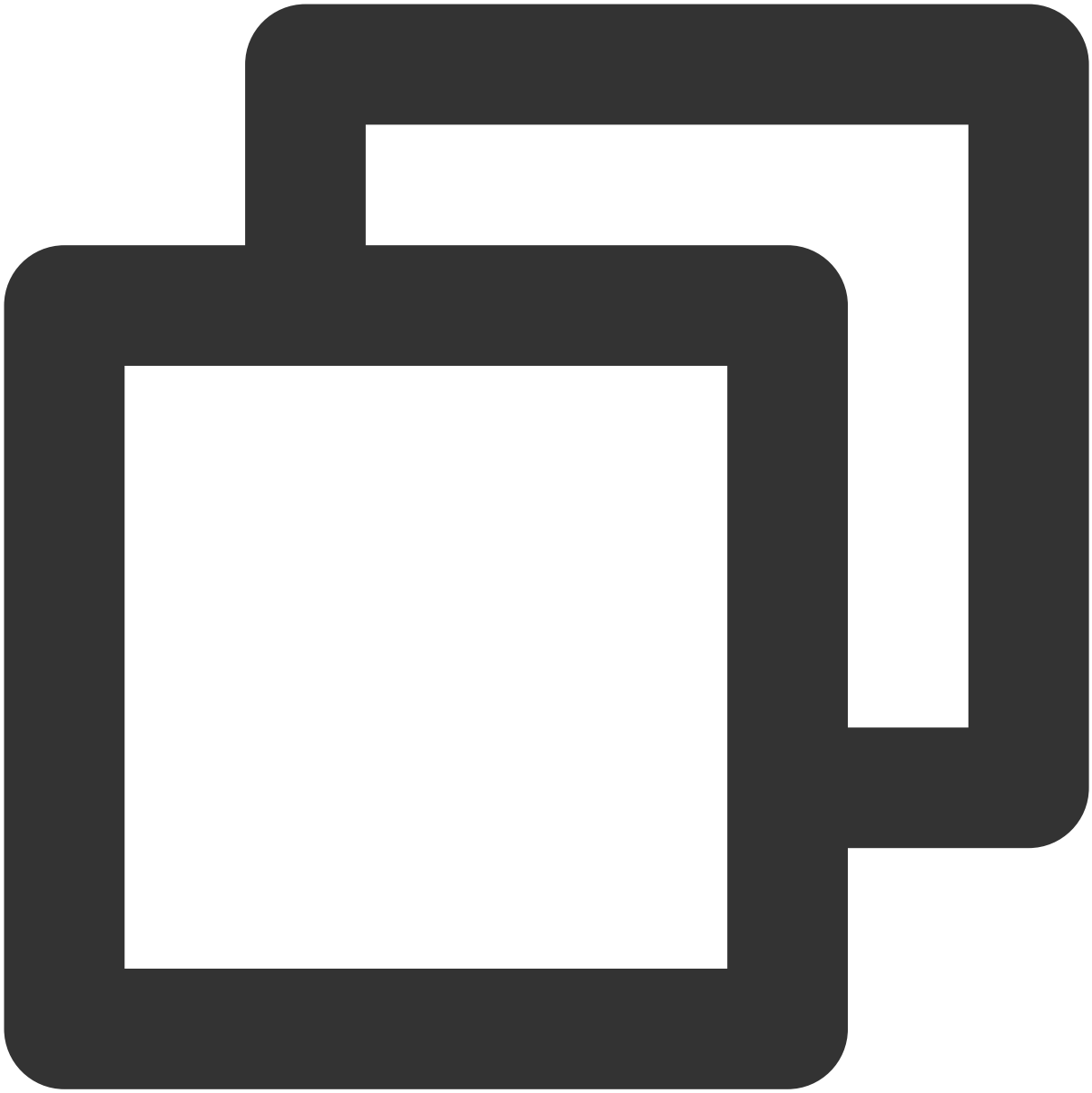
继续在 pom.xml 中添加打包和编译插件：



```
<build>
        <plugins>
```

```
            <plugin>
                <groupId>org.apache.maven.plugins</groupId>
                <artifactId>maven-compiler-plugin</artifactId>
                <configuration>
                    <source>1.8</source>
                    <target>1.8</target>
                    <encoding>utf-8</encoding>
                </configuration>
            </plugin>
            <plugin>
                <artifactId>maven-assembly-plugin</artifactId>
                <configuration>
                    <descriptorRefs>
                        <descriptorRef>jar-with-dependencies</descriptorRef>
                    </descriptorRefs>
                </configuration>
                <executions>
                    <execution>
                        <id>make-assembly</id>
                        <phase>package</phase>
                        <goals>
                            <goal>single</goal>
                        </goals>
                    </execution>
                </executions>
            </plugin>
        </plugins>
    </build>
```

在 src>main>java 下右键新建一个 Java Class，输入 Class 名，这里使用 WordCount，在 Class 添加样例代码：



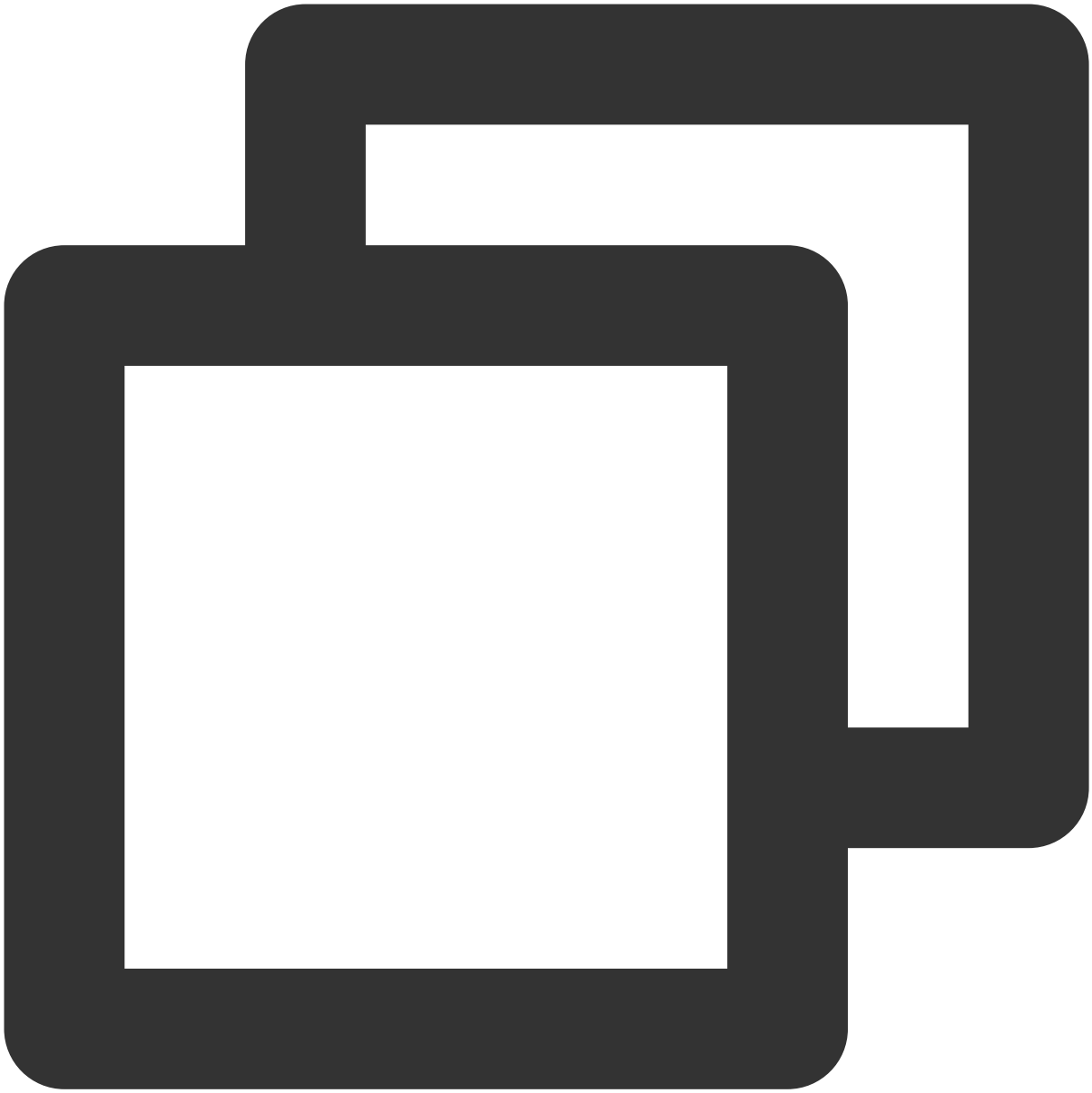
```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.util.GenericOptionsParser;
```

```
import java.io.IOException;
import java.util.StringTokenizer;

/**
 * Created by tencent on 2018/7/6.
 */
public class WordCount {
    public static class TokenizerMapper
            extends Mapper<Object, Text, Text, IntWritable>
    {
        private static final IntWritable one = new IntWritable(1);
        private Text word = new Text();

        public void map(Object key, Text value, Mapper<Object, Text, Text, IntWrita
                throws IOException, InterruptedException
        {
            StringTokenizer itr = new StringTokenizer(value.toString());
            while (itr.hasMoreTokens())
            {
                this.word.set(itr.nextToken());
                context.write(this.word, one);
            }
        }
    }

    public static class IntSumReducer
            extends Reducer<Text, IntWritable, Text, IntWritable>
    {
        private IntWritable result = new IntWritable();

        public void reduce(Text key, Iterable<IntWritable> values, Reducer<Text, In
                throws IOException, InterruptedException
        {
            int sum = 0;
            for (IntWritable val : values) {
                sum += val.get();
            }
            this.result.set(sum);
            context.write(key, this.result);
        }
    }

    public static void main(String[] args)
            throws Exception
    {
        Configuration conf = new Configuration();
        String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs(
```

```
        if (otherArgs.length < 2)
        {
            System.err.println("Usage: wordcount <in> [<in>...] <out>");
            System.exit(2);
        }
        Job job = Job.getInstance(conf, "word count");
        job.setJarByClass(WordCount.class);
        job.setMapperClass(TokenizerMapper.class);
        job.setCombinerClass(IntSumReducer.class);
        job.setReducerClass(IntSumReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(IntWritable.class);
        for (int i = 0; i < otherArgs.length - 1; i++) {
            FileInputFormat.addInputPath(job, new Path(otherArgs[i]));
        }
        FileOutputFormat.setOutputPath(job, new Path(otherArgs[(otherArgs.length -

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

可以看到其中有一个 Map 函数和一个 Reduce 函数。

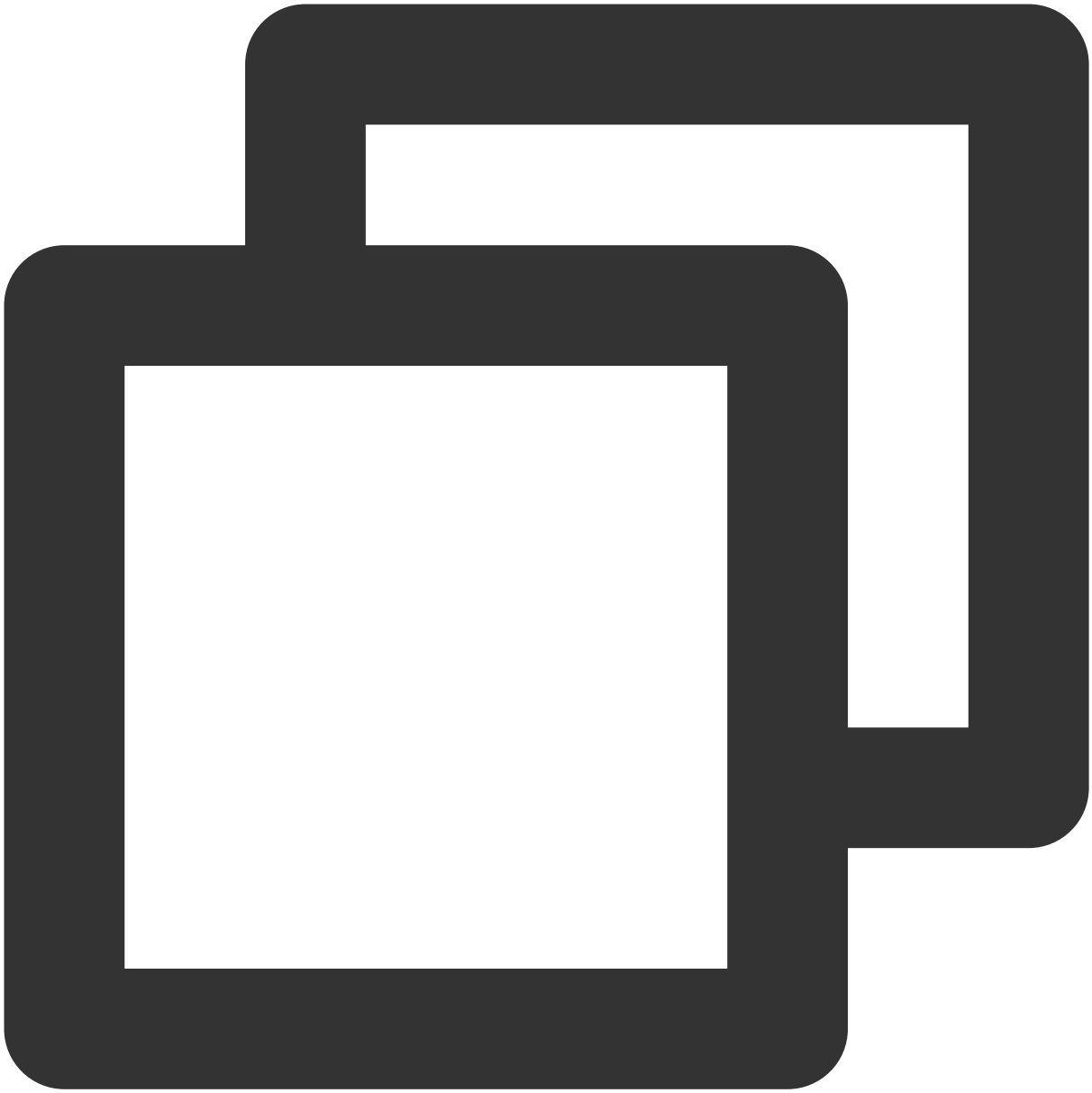
如果您的 Maven 配置正确并且成功的导入了依赖包，那么整个工程即可直接编译。在本地 shell 下进入工程目录，执行下面的命令对整个工程进行打包：



```
mvn package
```

运行过程中可能还需要下载一些文件，直到出现 build success 表示打包成功。然后您可以在工程目录下的 target 文件夹中看到打好的 jar 包。

使用 scp 或者 sftp 服务来把打包好的工程文件上传到 EMR 集群的云服务器中。在本地 shell 使用：

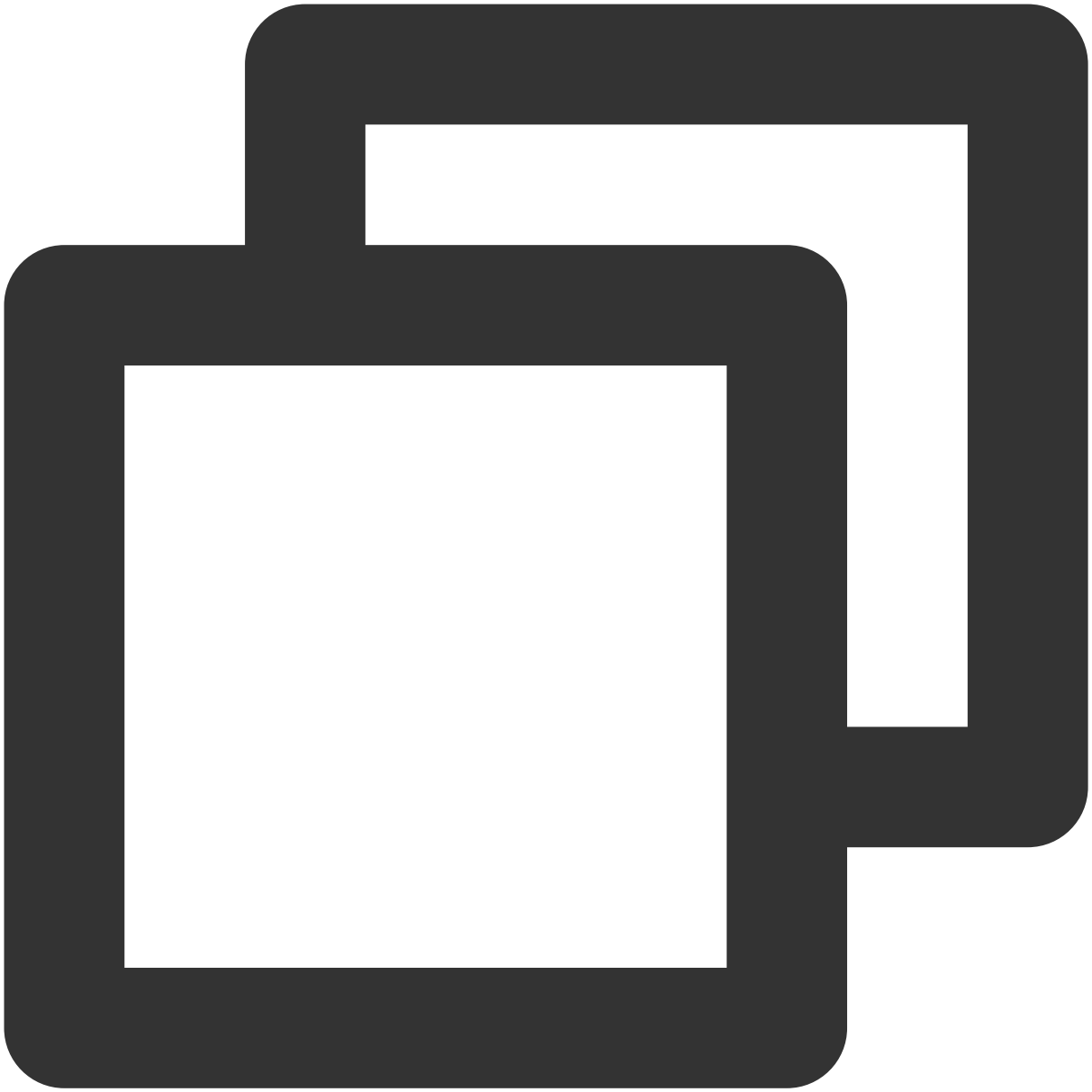


```
scp $jarpackage root@公网IP地址: /usr/local/service/hadoop
```

其中，`$jarpackage` 是您的本地 jar 包的路径加名称；root 为 CVM 服务器用户名；公网 IP 地址可以在 EMR 控制台的节点信息中或者在云服务器控制台查看。这里上传到了 EMR 集群的 `/usr/local/service/hadoop` 文件夹下。

统计 HDFS 中的文本文件

进入 `/usr/local/service/hadoop` 目录，和数据准备中一样。通过如下命令来提交任务：

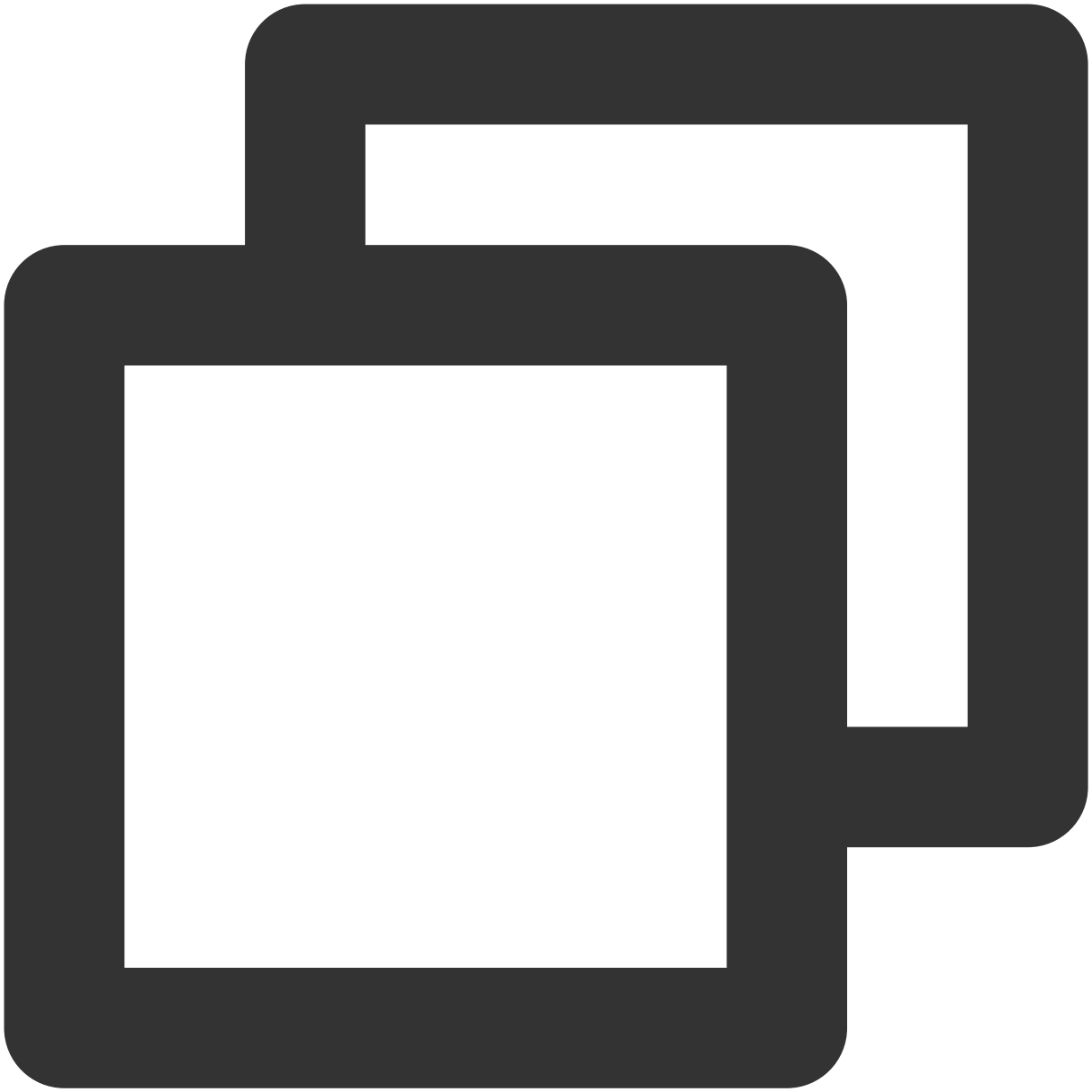


```
[hadoop@10 hadoop]$ bin/hadoop jar
/usr/local/service/hadoop/WordCount-1.0-SNAPSHOT-jar-with-dependencies.jar
WordCount /user/hadoop/test.txt /user/hadoop/WordCount_output
```

注意

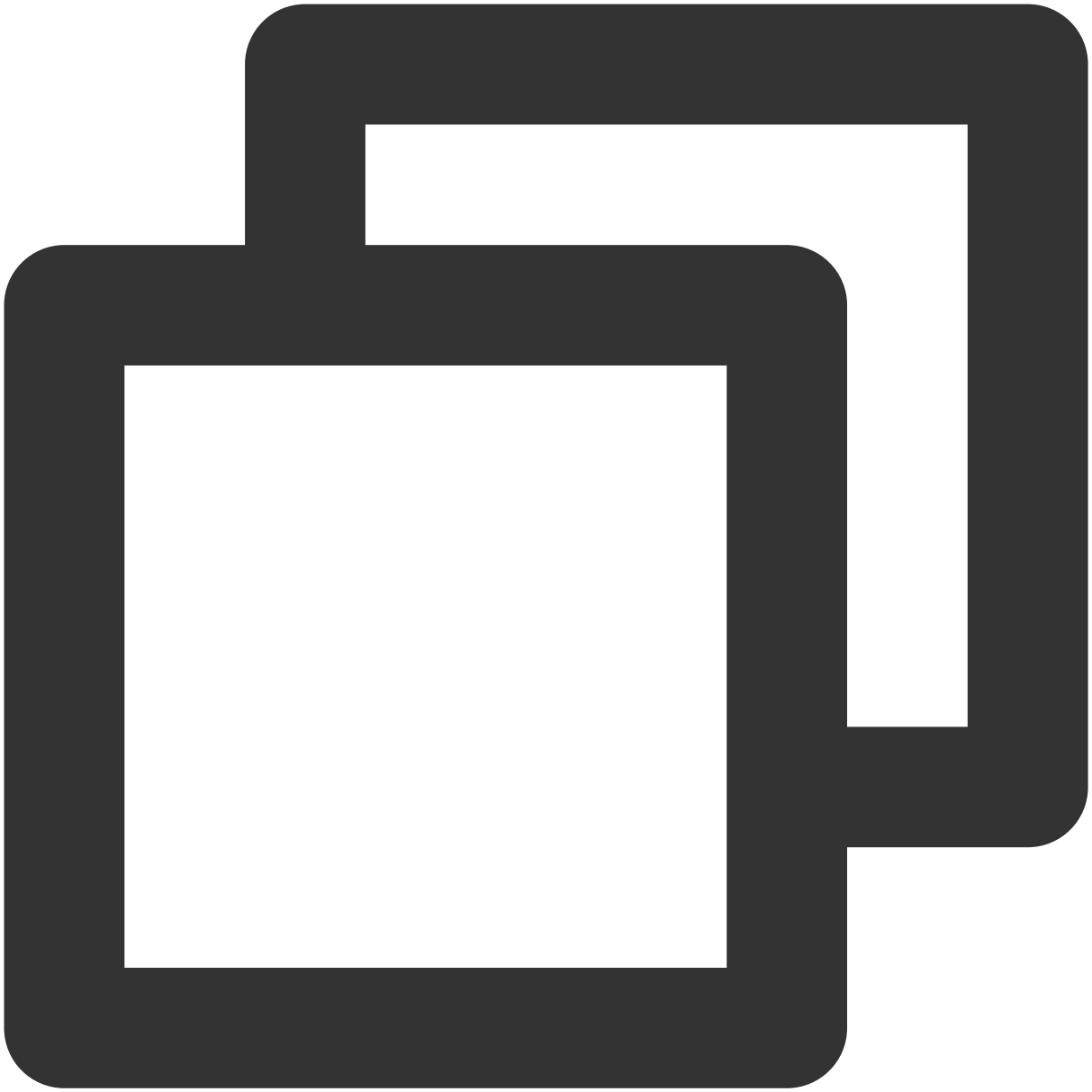
以上整个命令为一条完整的指令，`/user/hadoop/ test.txt` 为输入的待处理文件，`/user/hadoop/WordCount_output` 为输出文件夹，在提交命令前要保证 `WordCount_output` 文件夹尚未创建，否则提交会出错。

执行完成后，通过如下命令查看执行输出文件：



```
[hadoop@172 hadoop]$ hadoop fs -ls /user/hadoop/WordCount_output
Found 2 items
-rw-r--r-- 3 hadoop supergroup 0 2018-07-06 11:35 /user/hadoop/MEWordCount_output/_
-rw-r--r-- 3 hadoop supergroup 82 2018-07-06 11:35 /user/hadoop/MEWordCount_output/
```

通过如下指令查看 part-r-00000 中的统计结果：

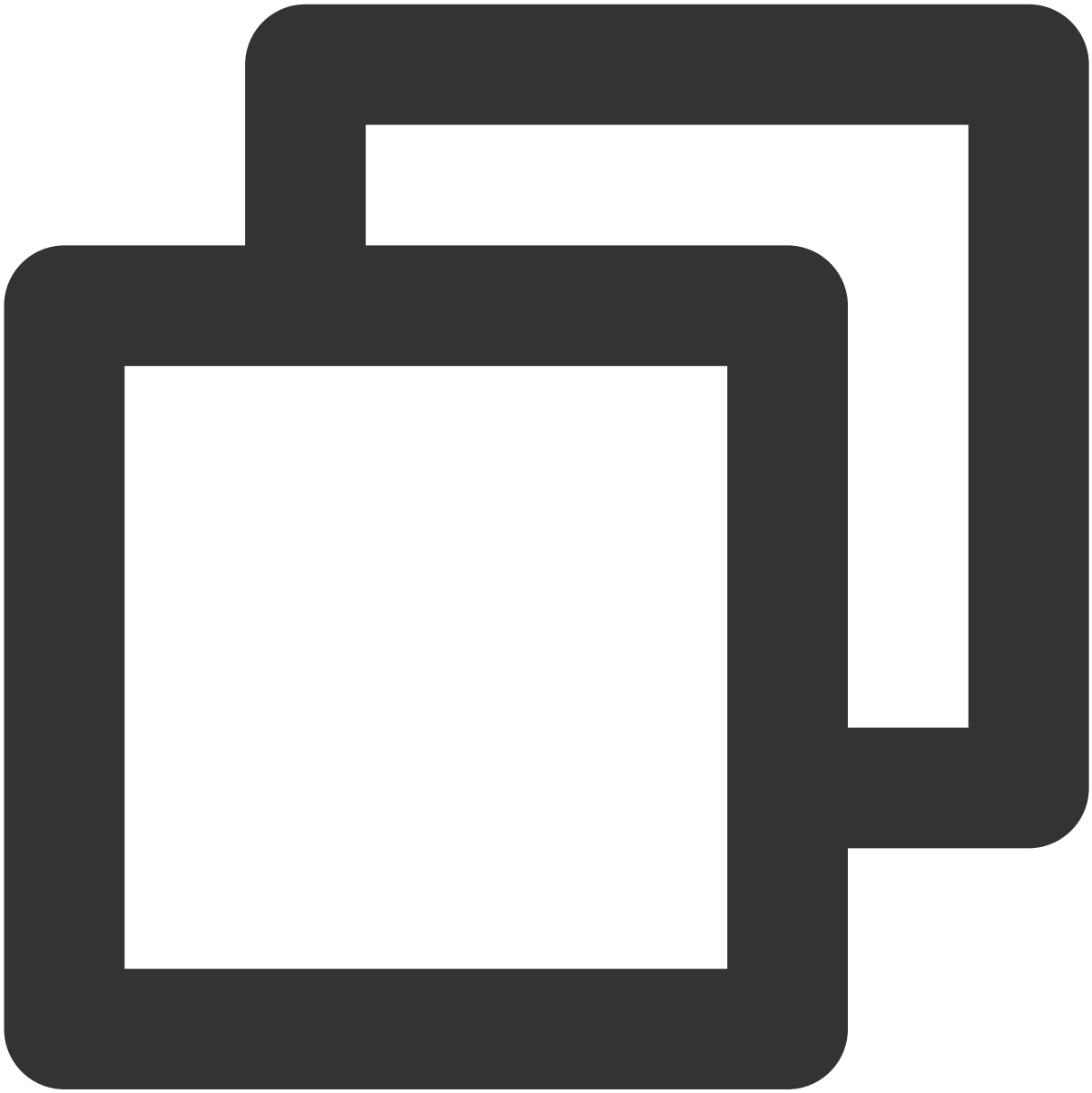


```
[hadoop@172 hadoop]$ hadoop fs -cat /user/hadoop/MEWordCount_output/part-r-00000
Hello   2
World.  1
a       1
another 1
are     1
how     1
is      2
message.        2
this    2
world,  1
```

```
you?    1……
```

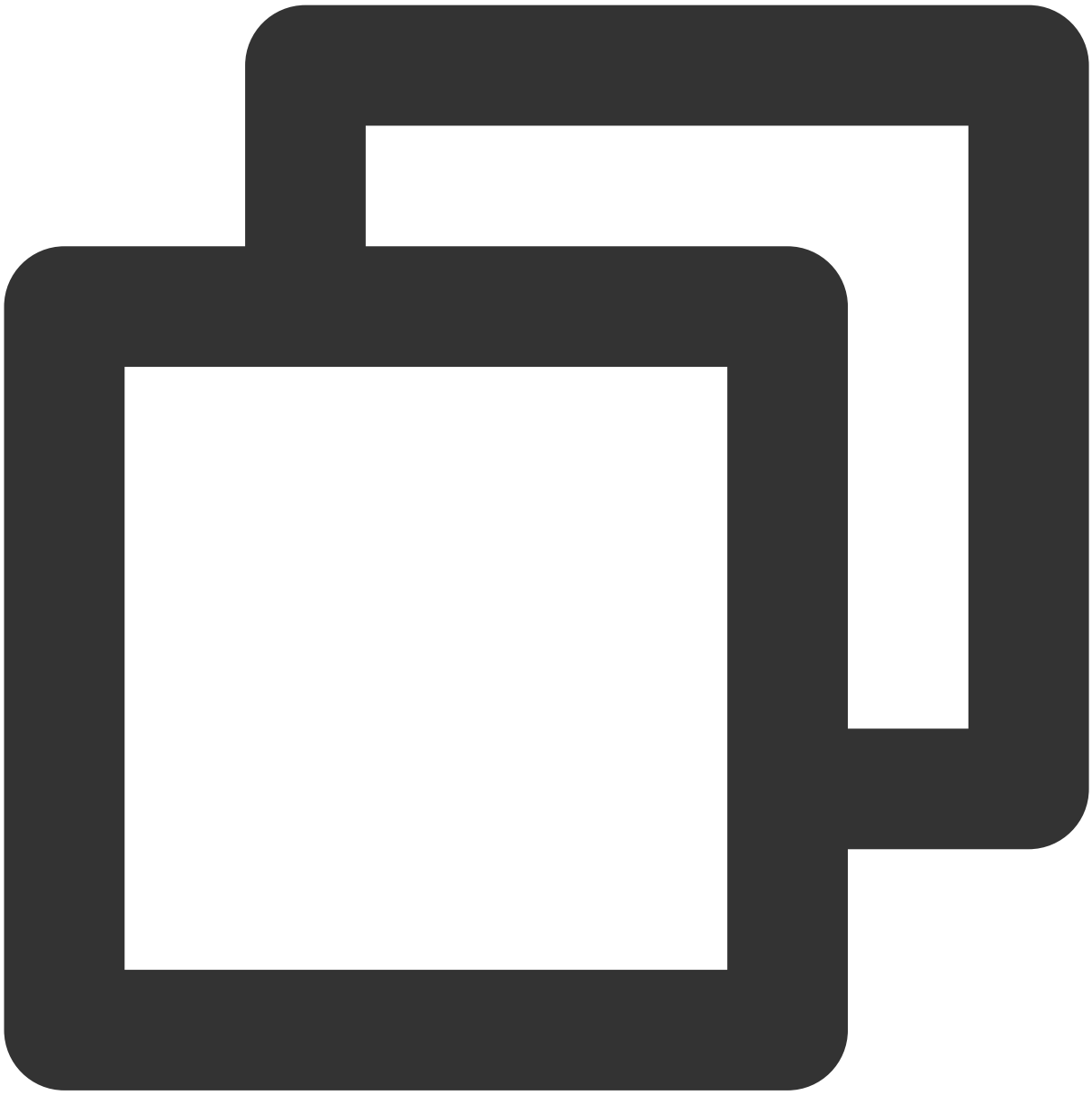
统计 COS 中的文本文件

进入 `/usr/local/service/hadoop` 目录。通过如下命令来提交任务：



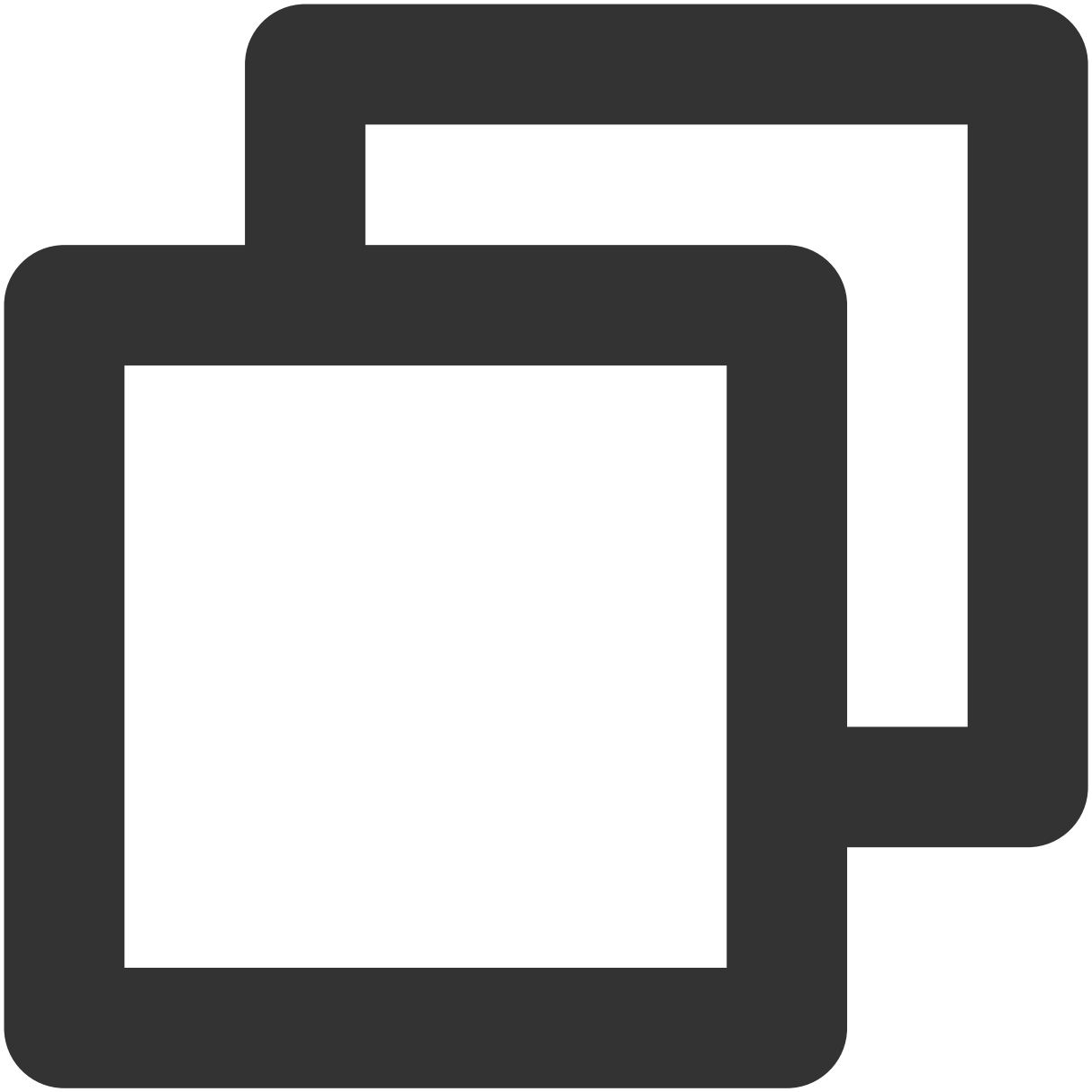
```
[hadoop@10 hadoop]$ hadoop jar
/usr/local/service/hadoop/WordCount-1.0-SNAPSHOT-jar-with-dependencies.jar
WordCount cosn://$bucketname/test.txt cosn://$bucketname /WordCount_output
```

命令的输入文件改为了 `cosn:// $bucketname/ test.txt`，其中 $bucketname 为您的存储桶名字加路径。处理结果同样也输出到 COS 中。使用如下指令查看输出文件：



```
[hadoop@10 hadoop]$ hadoop fs -ls cosn:// $bucketname /WordCount_output
Found 2 items
-rw-rw-rw- 1 hadoop Hadoop 0 2018-07-06 10:34 cosn://$bucketname /WordCount_output/
-rw-rw-rw- 1 hadoop Hadoop 1306 2018-07-06 10:34 cosn://$bucketname /WordCount_outp
```

查看最后输出的结果：



```
[hadoop@10 hadoop]$ hadoop fs -cat cosn:// $bucketname /WordCount_output1/part-r-00
Hello   2
World.  1
a       1
another 1
are     1
how     1
is      2
message.        2
this    2
world,  1
```

```
you?    1
```

YARN 作业日志转存 COS

最近更新时间：2021-07-01 15:33:55

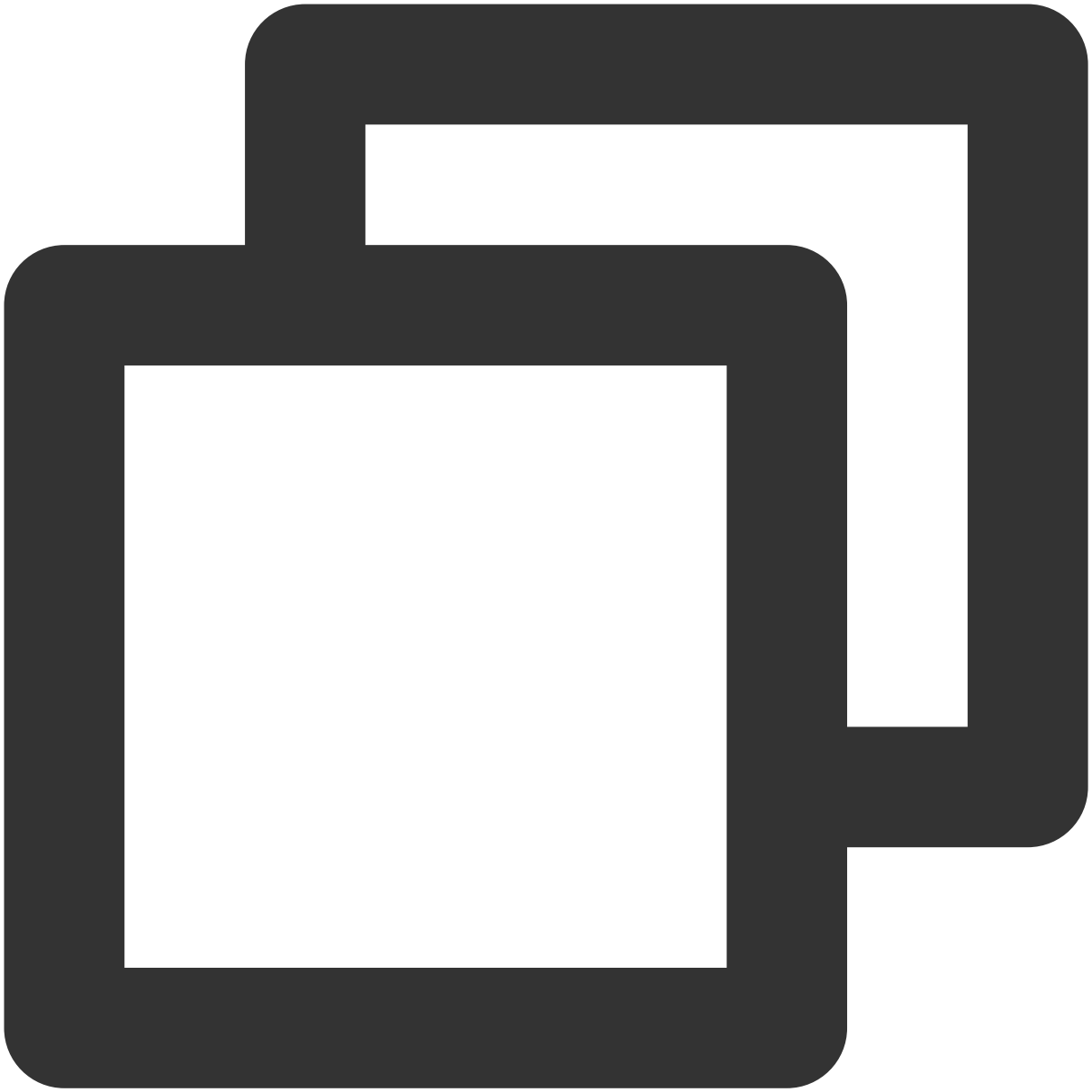
Hadoop 默认将 YARN 的作业日志存储在 hdfs 上，腾讯云 EMR 还提供了将 YARN 作业日志存储在外部存储 COS 上。

前提条件

EMR 集群需要支持 COS，详情可参考 使用 API 分析 HDFS/COS 上的数据。

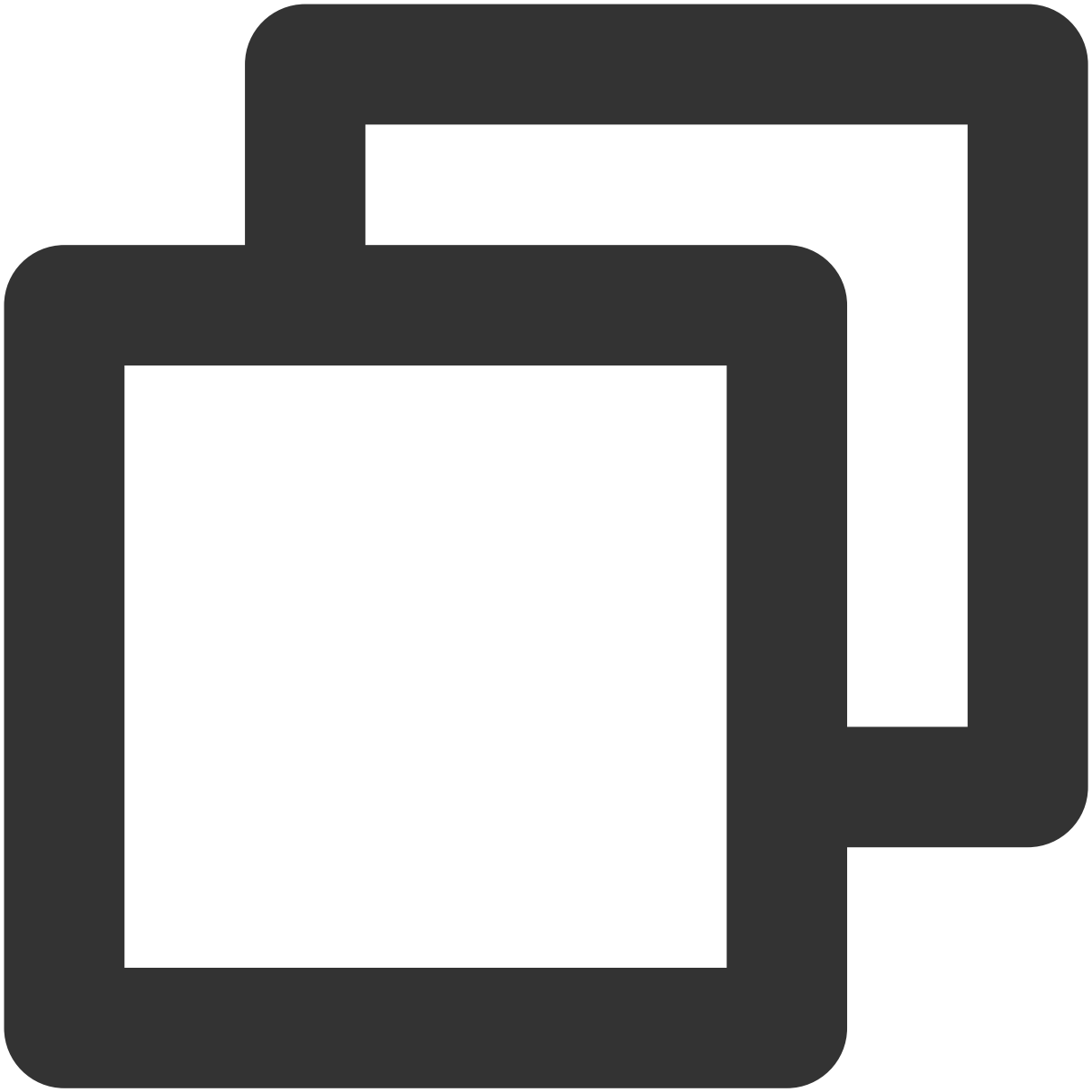
操作步骤

1. 在 `yarn-site.xml` 修改配置，并配置下发所有节点。



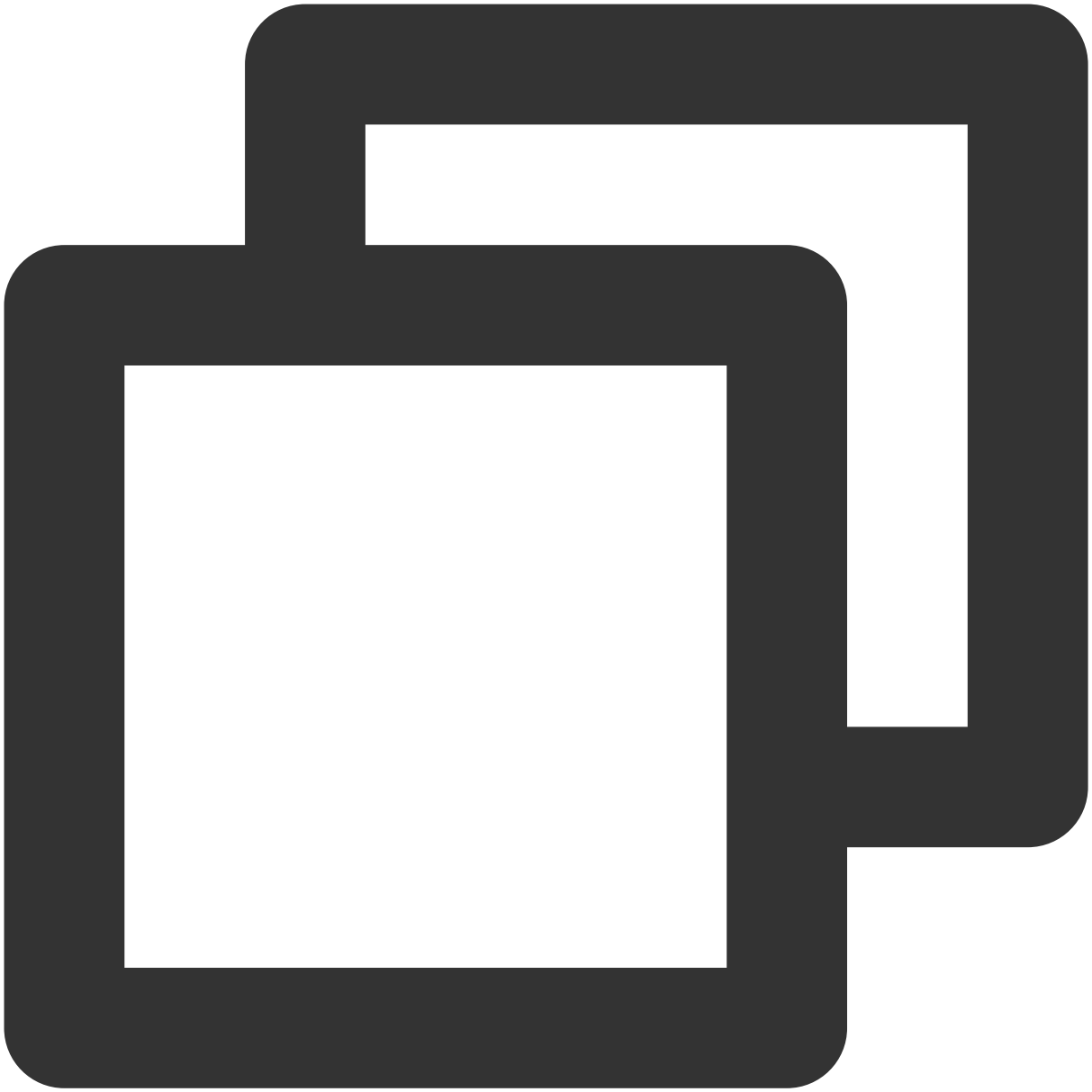
```
yarn.nodemanager.remote-app-log-dir=cosn://[bucket_name]/[logs_dirs]
```

2. 在 `core-site.xml` 新增配置，并配置下发所有节点。



```
fs.AbstractFileSystem.cosn.impl=org.apache.hadoop.fs.cosnative.COS
```

3. 重启集群所有 `nodemanager/datanode` 服务。
4. 运行 `hive/spark` 作业，查看存储在 COS 上的作业日志。



```
hdfs dfs -ls cosn://[bucket_name]/[logs_dirs]
```

Spark 开发指南
Spark 环境信息

最近更新时间：2022-11-25 16:06:47

腾讯云 EMR 提供 Spark 3.x、2.x 多版本支持 ， 软件环境信息如下：

- Spark 默认安装在 master 节点。
- 登录机器后使用命令 `su hadoop` 切换到 Hadoop 用户。
- Spark 软件路径在 `/usr/local/service/spark` 下。
- 相关日志路径在 `/data/emr` 下。

更多详细资料请参考 社区文档，这里主要介绍基于 Spark 访问腾讯云对象存储相关操作。

Spark 分析 COS 上的数据

最近更新时间：2022-03-16 17:52:57

Spark 作为 Apache 高级的开源项目，是一个快速、通用的大规模数据处理引擎，与 Hadoop 的 MapReduce 计算框架类似，但是相对于 MapReduce，Spark 凭借其可伸缩、基于内存计算等特点以及可以直接读写 Hadoop 上任何格式数据的优势，进行批处理时更加高效，并有更低的延迟。实际上，Spark 已经成为轻量级大数据快速处理的统一平台，各种不同的应用，如实时流处理、机器学习、交互式查询等，都可以通过 Spark 建立在不同的存储和运行系统上。

Spark 是基于内存计算的大数据并行计算框架。Spark 基于内存计算，提高了在大数据环境下数据处理的实时性，同时保证了高容错性和高可伸缩性，允许用户将 Spark 部署在大量廉价硬件之上，形成集群。

本教程演示的是提交的任务为 wordcount 任务即统计单词个数，提前需要在集群中上传需要统计的文件。

1. 开发准备

- 因为任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建一个存储桶（Bucket）。
- 确认您已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Spark 组件，并且在**实例信息>基础配置**中开启对象存储的授权。

2. 使用 Maven 创建工程

在本次演示中，不再采用系统自带的演示程序，而是自己建立工程编译打包之后上传到 EMR 集群运行。推荐您使用 Maven 来管理您的工程。Maven 是一个项目管理工具，能够帮助您方便的管理项目的依赖信息，即它可以通过 pom.xml 文件的配置获取 jar 包，而不用去手动添加。

首先下载并安装 Maven，配置 Maven 的环境变量。如果您使用 IDE，请在 IDE 中设置 Maven 相关配置。

新建一个 Maven 工程

在本地 shell 下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID -Darch
etypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名。$yourartifactID 为您的项目名称，而 maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。

创建成功后，在 `D://mavenWorkplace` 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
---pom.xml        核心配置，项目根下
---src
---main
---java          Java 源码目录
---resources     Java 配置文件目录
---test
---java          测试源码目录
---resources     测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。

首先在 pom.xml 中添加 Maven 依赖：

```
<dependencies>
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-core_2.11</artifactId>
<version>2.0.2</version>
</dependency>
</dependencies>
```

继续在 pom.xml 中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
```

```
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

在 src>main>Java 下右键新建一个Java Class，输入您的 Class 名，这里使用 WordCountOnCos，在 Class 添加样例代码：

```
import java.util.Arrays;
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaRDD;
import org.apache.spark.api.java.JavaSparkContext;
import scala.Tuple2;
/**
* Created by tencent on 2018/6/28.
*/
public class WordCountOnCos {
public static void main(String[] args){
SparkConf sc = new SparkConf().setAppName("spark on cos");
JavaSparkContext context = new JavaSparkContext(sc);
JavaRDD<String> lines = context.textFile(args[0]);
lines.flatMap(x -> Arrays.asList(x.split(" ")).iterator())
.mapToPair(x -> new Tuple2<String, Integer>(x, 1))
.reduceByKey((x, y) -> x+y)
.saveAsTextFile(args[1]);
}
}
```

如果您的 Maven 配置正确并且成功的导入了依赖包，那么整个工程应该没有错误可以直接编译。在本地命令行模式下进入工程目录，执行下面的命令对整个工程进行打包：

```
mvn package
```

运行过程中可能还需要下载一些文件，直到出现 build success 表示打包成功。然后您可以在工程目录下的 target 文件夹中看到打好的 jar 包。

数据准备

首先需要把压缩好的 jar 包上传到 EMR 集群中，使用 scp 或者 sftp 工具来进行上传。在本地命令行模式下运行：

```
scp $localfile root@公网IP地址:$remotefolder
```

其中，$localfile 是您的本地文件的路径加名称；root 为 CVM 服务器用户名；公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看；$remotefolder 是您想存放文件的 CVM 服务器路径。上传完成后，在 EMR 命令行中即可查看对应文件夹下是否有相应文件。

需要处理的文件需要事先上传到 COS 中。如果文件在本地则可以通过 COS 控制台直接上传。如果文件在 EMR 集群上，可以使用 Hadoop 命令上传。指令如下：

```
[hadoop@10 hadoop]$ hadoop fs -put $testfile cosn://$bucketname/
```

其中 $testfile 为要统计的文件的完整路径加名字，$bucketname 为您的存储桶名。上传完成后可以在 COS 控制台中查看文件是否已经在 COS 中。

运行样例

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户：

```
[root@172 ~]# su hadoop
```

然后进入您存放 jar 包的文件夹下，执行以下指令：

```
[hadoop@10spark]$ spark-submit --class $WordCountOnCOS --master
yarn-cluster $packagename.jar cosn:// $bucketname /$testfile cosn:// $bucketname
/output
```

其中 $WordCountOnCOS 为您的 Java Class 名字，$packagename 为您新建 Maven 工程中生成的 jar 包名字，$bucketname 为您的存储桶名和路径，$testfile 为您要统计的文件名。最后输出的文件在 output 这个文件夹中，**这个文件夹事先不能被创建，不然运行会失败**。

运行成功后，在指定的存储桶和文件夹下可以看到 wordcount 的结果。

```
[hadoop@172 /]$ hadoop fs -ls cosn:// $bucketname /output
Found 3 items
-rw-rw-rw- 1 hadoop Hadoop 0 2018-06-28 19:20 cosn:// $bucketname /output/_SUCCES
S
-rw-rw-rw- 1 hadoop Hadoop 681 2018-06-28 19:20 cosn:// $bucketname /output/part-
```

```
00000
-rw-rw-rw- 1 hadoop Hadoop 893 2018-06-28 19:20 cosn:// $bucketname /output/part-
00001
[hadoop@172 demo]$ hadoop fs -cat cosn://$bucketname/output/part-00000
18/07/05 17:35:01 INFO cosnative.NativeCosFileSystem: Opening 'cosn:// $bucketnam
e/output/part-00000' for reading
(under,1)
(this,3)
(distribution,2)
(Technology,1)
(country,1)
(is,1)
(Jetty,1)
(currently,1)
(permitted.,1)
(Security,1)
(have,1)
(check,1)
```

通过 Spark Python 分析 COS 上的数据

最近更新时间：2023-06-19 17:25:20

本节主要是通过 Spark Python 来进行 wordcount 的工作。

开发准备

- 因为任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建一个存储桶（Bucket）。
- 确认您已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置页面选择 Spark 组件，并且在基础配置页面开启对象存储的授权。

数据准备

需要处理的文件需要事先上传到 COS 中。如果文件在本地那么就可以通过 COS 控制台直接上传。如果文件在 EMR 集群上，可以使用 Hadoop 命令上传。指令如下：

```
[hadoop@10 hadoop]$ hadoop fs -put $testfile cosn:// $bucketname/
```

其中 $testfile 为要统计的文件的完整路径加名字，$bucketname 为您的存储桶名。上传完成后可以查看文件是否已经在 COS 中。

运行样例

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Spark 安装目录 `/usr/local/service/spark` ：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/spark
```

新建一个 Python 文件 wordcount.py，并添加如下代码：

```
from __future__ import print_function

import sys
```

```
from operator import add
from pyspark.sql import SparkSession

if __name__ == "__main__":
if len(sys.argv) != 3:
print("Usage: wordcount <file>", file=sys.stderr)
exit(-1)

spark = SparkSession\
.builder\
.appName("PythonWordCount")\
.getOrCreate()

sc = spark.sparkContext

lines = spark.read.text(sys.argv[1]).rdd.map(lambda r: r[0])
counts = lines.flatMap(lambda x: x.split(' ')) \
.map(lambda x: (x, 1)) \
.reduceByKey(add)

output = counts.collect()
counts.saveAsTextFile(sys.argv[2])

spark.stop()
```

通过如下指令提交任务：

```
[hadoop@10 spark]$ ./bin/spark-submit --master yarn ./wordcount.py
cosn://$bucketname/$yourtestfile cosn:// $bucketname/$output
```

其中 $bucketname 为您的 COS 存储桶名，$yourtestfile 为您的测试文件在存储桶中的完整路径加名字。$output 为您的输出文件夹。**$output 为一个未创建的文件夹，如果执行指令前该文件夹已经存在，会导致程序运行失败。**

成功后程序自动运行，可以在目标存储桶中查看到输出文件：

```
[hadoop@172 spark]$ hadoop fs -ls cosn:// $bucketname/$output
Found 2 items
-rw-rw-rw- 1 hadoop Hadoop 0 2018-06-29 15:35 cosn:// $bucketname/$output /_SUCCE
SS
-rw-rw-rw- 1 hadoop Hadoop 2102 2018-06-29 15:34 cosn:// $bucketname/$output /par
t-00000
```

最后的结果也可以通过如下指令查看：

```
[hadoop@172 spark]$ hadoop fs -cat cosn:// $bucketname/$output /part-00000
(u'', 27)
```

```
(u'code', 1)
(u'both', 1)
(u'Hadoop', 1)
(u'Bureau', 1)
(u'Department', 1)
```

同样可以把结果输出到 HDFS 中，只需要更改指令中的输出位置即可，如下所示：

```
[hadoop@10spark]$ ./bin/spark-submit ./wordcount.py
cosn://$bucketname/$yourtestfile /user/hadoop/$output
```

其中 `/user/hadoop/` 为 HDFS 中的路径，如果不存在用户可以自己创建。

任务结束后，可以通过如下命令看到 Spark 运行日志：

```
[hadoop@10 spark]$  /usr/local/service/hadoop/bin/yarn logs -applicationId $yourId
```

其中 $yourId 应该替代为您的任务 ID。任务 ID 可以在 YARN 的 WebUI 上面进行查看。

SparkSQL 的使用

最近更新时间：2021-07-13 14:46:41

Spark 为结构化数据处理引入了一个称为 Spark SQL 的编程模块。它提供了一个称为 DataFrame 的编程抽象，并且可以充当分布式 SQL 查询引擎。

1. 开发准备

确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择了 Spark 组件。

2. 使用 SparkSQL 交互式控制台

在使用 SparkSQL 之前请登录 EMR 集群的 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入 EMR 命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入目录 `/usr/local/service/spark` ：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/spark
```

通过如下命令您可以进入 SparkSQL 的交互式控制台：

```
[hadoop@10spark]$ bin/spark-sql --master yarn --num-executors 64 --executor-memor
y 2g
```

其中 --master 表示您的 master URL，--num-executors 表示 executor 数量，--executor-memory 表示 executor 的储存容量。以上参数也可以根据您的实际情况作出修改，您也可以通过 `sbin/start-thriftserver.sh` 或者 `sbin/stop-thriftserver.sh` 来启动或者停止一个 SparkSQLthriftserver。

下面介绍一些 SparkSQL 的基本操作：

- 新建一个数据库并查看：

```
spark-sql> create database sparksql;
Time taken: 0.907 seconds
spark-sql> show databases;
default
sparksql
```

```
test
Time taken: 0.131 seconds, Fetched 5 row(s)
```

- 在新建的数据库中新建一个表，并进行查看：

```
spark-sql> use sparksql;
Time taken: 0.076 seconds
spark-sql> create table sparksql_test(a int,b string);
Time taken: 0.374 seconds
spark-sql> show tables;
sparksql_test false
Time taken: 0.12 seconds, Fetched 1 row(s)
```

- 向表中插入两行数据并查看：

```
spark-sql> insert into sparksql_test values (42,'hello'),(48,'world');
Time taken: 2.641 seconds
spark-sql> select * from sparksql_test;
42 hello
48 world
Time taken: 0.503 seconds, Fetched 2 row(s)
```

更多命令行参数使用教程请参考 社区文档。

3. 使用 Maven 创建工程

首先下载并安装 Maven，配置好 Maven 的环境变量，如果您使用 IDE，请在 IDE 中设置好 Maven 相关配置。

新建一个 Maven 工程

在命令行下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID
-DarchetypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名。$yourartifactID 为您的项目名称， maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。
创建成功之后，在 `D://mavenWorkplace` 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
    ---pom.xml        核心配置，项目根下
    ---src
```

```
        ---main
            ---java         Java 源码目录
      ---resources    Java 配置文件目录
       ---test
          ---java        测试源码目录
          ---resources    测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置， Java 文件夹下放置您的源代码。

添加 Hadoop 依赖和样例代码

首先在 pom.xml 文件中添加 Maven 依赖：

```
<dependencies>
    <dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-core_2.11</artifactId>
    <version>2.0.2</version>
    </dependency>
    <!--spark sql-->
    <dependency>
    <groupId>org.apache.spark</groupId>
    <artifactId>spark-sql_2.11</artifactId>
    <version>2.0.2</version>
    </dependency>
</dependencies>
```

继续在 pom.xml 文件中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
```

```
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

完整的 pom.xml 文件如下所示：

```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xs
d/maven-4.0.0.xsd">
<modelVersion>4.0.0</modelVersion>
<groupId>$yourgroupID </groupId>
<artifactId>$yourartifactID </artifactId>
<version>1.0-SNAPSHOT</version>
<dependencies>
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-core_2.11</artifactId>
<version>2.0.2</version>
</dependency>
<!--spark sql-->
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-sql_2.11</artifactId>
<version>2.0.2</version>
</dependency>
</dependencies>
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
```

```
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
</project>
```

注意：
修改其中的 $yourgroupID 和 $yourartifactID 为您自己的设置。

接下来添加样例代码，在 main>Java 文件夹下新建一个 Java Class 取名为 Demo.java，并将以下代码加入其中：

```
import org.apache.spark.rdd.RDD;
import org.apache.spark.sql.Dataset;
import org.apache.spark.sql.Row;
import org.apache.spark.sql.SparkSession;
/**
* Created by tencent on 2018/6/28.
*/
public class Demo {
public static void main(String[] args){
SparkSession spark = SparkSession
.builder()
.appName("Java Spark Hive Example")
.enableHiveSupport()
.getOrCreate();
```

```
Dataset<Row> df = spark.read().json(args[0]);
RDD<Row> test = df.rdd();

test.saveAsTextFile(args[1]);
}
}
```

编译代码并打包上传

使用本地命令行进入工程目录，执行以下指令对工程进行编译打包：

```
mvn package
```

在显示 build success 表示操作成功，在工程目录下的 target 文件夹中能够看到打包好的文件。
使用 scp 或者 sftp 工具把打包好的文件上传到 EMR 集群。在本地命令行模式下运行：

```
scp $localfile root@公网IP地址:$remotefolder
```

其中，$localfile 是您的本地文件的路径加名称，root 为 CVM 服务器用户名，公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看。$remotefolder 是您想存放文件的 CVM 服务器路径。上传完成后，在 EMR 集群命令行中即可查看对应文件夹下是否有相应文件。

4. 准备数据并运行样例

使用 sparkSQL 来操作存放在 HDFS 上的数据。首先将数据上传到 HDFS 中，这里我们使用自带的文件 people.json，存放在路径 `/usr/local/service/spark/exa-mples/src/main/resources/` 下，使用如下指令把该文件上传到 HDFS 中：

```
[hadoop@10 hadoop]$ hadoop fs -put /usr/local/service/spark/examples/src/main/res
ources/people.json
/user/hadoop
```

测试文件用户也可以另选，这里 `/user/hadoop/` 是 HDFS 下的文件夹，如果没有用户可以自己创建。

执行样例，首先请登录 EMR 集群的 master 节点，并且切换到 Hadoop 用户如使用 SparkSQL 交互式控制台中所示，使用以下命令执行样例：

```
[hadoop@10spark]$ bin/spark-submit --class Demo --master yarn-client $yourjarpack
age /
/user/hadoop/people.json /user/hadoop/$output
```

其中 --class 参数表示要执行的入口类，在本例子中即为 Demo，即在添加 Hadoop 依赖和样例代码中创建的 Java Class 的名字，--master 为集群主要的 URL，$yourjarpackage 是您打包后的包名，$output 为结果输出文件夹（**$output 为一个未创建的文件夹，如果执行指令前该文件夹已经存在，会导致程序运行失败**）。

成功运行后，可以在 `/user/hadoop/$output` 查看结果：

```
[hadoop@172 spark]$ hadoop fs -cat /user/hadoop/$output/part-00000
[null,Michael]
[30,Andy]
[19,Justin]
```

spark-submit 的更多参数，在命令行输入以下命令进行查看，或者请参考 官方文档。

```
[hadoop@10spark]$ spark-submit -h
```

SparkStreaming 对接 Ckafka 服务

最近更新时间：2021-07-13 14:49:56

基于腾讯云的 EMR 服务您可以轻松结合腾讯云的 Ckafka 服务实现以下流式应用：

- 日志信息流式处理
- 用户行为记录流式处理
- 告警信息收集及处理
- 消息系统

1. 开发准备

- 因为任务中需要访问腾讯云消息队列 CKafka，所以需要先创建一个 CKafka 实例，具体见 消息队列 CKafka。
- 确认您已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群时需要在软件配置界面选择 Spark 组件。

2. 在 EMR 集群使用 Kafka 工具包

首先需要查看 CKafka 的内网 IP 与端口号。登录消息队列 CKafka 的控制台，选择您要使用的 CKafka 实例，在基本消息中查看其内网 IP 为 $kafkaIP，而端口号一般默认为9092。在 topic 管理界面新建一个 topic 为 spark_streaming_test。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里可选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入目录 `/usr/local/service/spark` ：

```
[root@172 ~]# su hadoop
[root@172 root]$ cd / usr/local/service/spark
```

从 Kafka 官网 下载安装包，注意选择合适的版本，具体可参考 EMR 各版本 Kafka 与 Spark 版本说明。kafka 客户端版和腾讯云 ckafka 兼容性强，安装对应的 kafka 客户端版本即可。解压压缩包并将解压出来的文件夹移动到 `/opt` 目录下：

```
[hadoop@172 data]$ tar -xzvf kafka_2.10-0.10.2.0.tgz
[hadoop@172 data]$ mv kafka_2.10-0.10.2.0 /opt/
```

解压完成后，Kafka 工具直接能使用。可以使用 `telnet` 命令来测试 EMR 集群是否能够连接到 CKafka 实例：

```
[hadoop@172 kafka_2.10-0.10.2.0]$ telnet $kafkaIP 9092
Trying $kafkaIP...
Connected to $kafkaIP.
```

其中 $kafkaIP 为您创建的 CKafka 实例的内网 IP 地址。

下面可以简单测试 Kafka 工具包，同时用两个 WebShell 登录 EMR 集群并切换到 Hadoop 用户，进入 Kafka 的安装路径：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /opt/kafka_2.10-0.10.2.0/
```

在第一个终端上连接 CKafka，并向其发送消息：

```
[hadoop@172 kafka_2.10-0.10.2.0]$ bin/kafka-console-producer.sh --broker-list $ka
fkaIP:9092
--topic spark_streaming_test
hello world
this is a message
```

在另一个终端上连接 CKafka，并作为消费者获得其中的数据：

```
[hadoop@172 kafka_2.10-0.10.2.0]$ bin/kafka-console-consumer.sh --bootstrap-serve
r
$kafkaIP:9092 --from-beginning --new-consumer --topic spark_streaming_test
hello world
this is a message
```

3. 使用 SparkStreaming 对接 CKafka 服务

在消费者一端，我们利用 Spark Streaming 从 CKafka 中不断拉取数据进行词频统计，即对流数据进行 WordCount 的工作。在生产者一端，也采用程序不断的产生数据，来不断输送给 CKafka。

首先 下载并安装 Maven，配置好 Maven 的环境变量，如果您使用 IDE，请在 IDE 中设置好 Maven 相关配置。

创建 Spark Streamin 消费者工程

在本地命令行下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID
-DarchetypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名。$yourartifactID 为您的项目名称，maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。

创建成功后，在 `D://mavenWorkplace` 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
    ---pom.xml          核心配置，项目根下
    ---src
        ---main
            ---java          Java 源码目录
     ---resources     Java 配置文件目录
      ---test
          ---java          测试源码目录
          ---resources     测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。

首先在 pom.xml 文件中添加 Maven 依赖：

```
<dependencies>
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-core_2.11</artifactId>
<version>2.0.2</version>
</dependency>
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-streaming_2.11</artifactId>
<version>2.0.2</version>
</dependency>
<dependency>
<groupId>org.apache.spark</groupId>
<artifactId>spark-streaming-kafka-0-10_2.11</artifactId>
<version>2.0.2</version>
</dependency>
</dependencies>
```

继续在 pom.xml 文件中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
```

```
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

注意：

修改其中的 $yourgroupID 和 $yourartifactID 为您自己的设置。

接下来添加样例代码，在 main>Java 文件夹下新建一个 Java Class 取名为 KafkaTest.java，并将以下代码加入其中：

```
import org.apache.kafka.clients.consumer.ConsumerRecord;
import org.apache.spark.SparkConf;
import org.apache.spark.api.java.JavaSparkContext;
import org.apache.spark.streaming.Durations;
import org.apache.spark.streaming.api.java.JavaInputDStream;
import org.apache.spark.streaming.api.java.JavaPairDStream;
import org.apache.spark.streaming.api.java.JavaStreamingContext;
import org.apache.spark.streaming.kafka010.ConsumerStrategies;
import org.apache.spark.streaming.kafka010.KafkaUtils;
import org.apache.spark.streaming.kafka010.LocationStrategies;
import scala.Tuple2;
```

```
import java.util.*;
import java.util.concurrent.TimeUnit;
/**
* Created by tencent on 2018/7/3.
*/
public class KafkaTest {
public static void main(String[] args) throws InterruptedException {
String brokers = "$kafkaIP:9092";
String topics = "spark_streaming_test1"; // 订阅的话题，多个话题','分隔
int durationSeconds = 60; // 间隔时间
SparkConf conf = new SparkConf().setAppName("spark streaming word count");
JavaSparkContext sc = new JavaSparkContext(conf);
JavaStreamingContext ssc = new JavaStreamingContext(sc, Durations.seconds(duratio
nSeconds));
Collection<String> topicsSet = new HashSet<>(Arrays.asList(topics.split(",")));
//kafka相关参数
Map<String, Object> kafkaParams = new HashMap<>();
kafkaParams.put("metadata.broker.list", brokers) ;
kafkaParams.put("bootstrap.servers", brokers);
kafkaParams.put("group.id", "group1");
kafkaParams.put("key.serializer", "org.apache.kafka.common.serialization.StringSe
rializer");
kafkaParams.put("key.deserializer", "org.apache.kafka.common.serialization.String
Deserializer");
kafkaParams.put("value.deserializer", "org.apache.kafka.common.serialization.Stri
ngDeserializer");
//创建连接
JavaInputDStream<ConsumerRecord<Object,Object>> lines = KafkaUtils.createDirectSt
ream(
ssc,
LocationStrategies.PreferConsistent(),
ConsumerStrategies.Subscribe(topicsSet, kafkaParams)
);
//wordcount逻辑
JavaPairDStream<String, Integer> counts = lines
.flatMap(x -> Arrays.asList(x.value().toString().split(" ")).iterator())
.mapToPair(x -> new Tuple2<String, Integer>(x, 1))
.reduceByKey((x, y) -> x + y);
// 保存结果
counts.dstream().saveAsTextFiles("$hdfsPath","result");
//
ssc.start();
ssc.awaitTermination();
ssc.close();
}
}
```

代码中要注意以下几点设置：

- brokers 变量要设置为在第二步中查找到的 CKafka 实例的内网 IP。
- topics 变量要设置为自己创建的 topic 的名字，这里为 spark_streaming_test1。
- durationSeconds 为程序去 CKafka 中消费数据的时间间隔，这里为60秒。
- $hdfsPath 为 HDFS 中的路径，结果将会输出到该路径下。

使用本地命令行进入工程目录，执行以下指令对工程进行编译打包：

```
mvn package
```

显示 build success 表示操作成功，在工程目录下的 target 文件夹中能够看到打包好的文件。
使用 scp 或者 sftp 工具来把打包好的文件上传到 EMR 集群，注意一定要上传依赖一起打包的 jar 包：

```
scp $localfile root@公网IP地址:$remotefolder
```

其中，$localfile 是您的本地文件的路径加名称，root 为 CVM 服务器用户名，公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看。$remotefolder 是您想存放文件的 CVM 服务器路径。上传完成后，在 EMR 集群命令行中即可查看对应文件夹下是否有相应文件。

创建 Spark Streaming 生产者工程

在本地命令行下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID
-DarchetypeArtifactId=maven-archetype-quickstart
```

首先在 pom.xml 文件中添加 Maven 依赖：

```
<dependencies>
<dependency>
<groupId>org.apache.kafka</groupId>
<artifactId>kafka_2.11</artifactId>
<version>0.10.1.0</version>
</dependency>
<dependency>
<groupId>org.apache.kafka</groupId>
<artifactId>kafka-clients</artifactId>
<version>0.10.1.0</version>
</dependency>
<dependency>
<groupId>org.apache.kafka</groupId>
<artifactId>kafka-streams</artifactId>
```

```
<version>0.10.1.0</version>
</dependency>
</dependencies>
```

继续在 pom.xml 文件中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

> 注意：
> 修改其中的 $yourgroupID 和 $yourartifactID 为您自己的设置。

接下来添加样例代码，在 main>Java 文件夹下新建一个 Java Class 取名为 SendData.java，并将以下代码加入其中：

```
import org.apache.kafka.clients.producer.KafkaProducer;
import org.apache.kafka.clients.producer.ProducerRecord;
import java.util.Properties;

/**
* Created by tencent on 2018/7/4.
*/
public class SendData {
public static void main(String[] args) {
Properties props = new Properties();
props.put("bootstrap.servers", "$kafkaIP:9092");
props.put("acks", "all");
props.put("retries", 0);
props.put("batch.size", 16384);
props.put("linger.ms", 1);
props.put("buffer.memory", 33554432);
props.put("key.serializer", "org.apache.kafka.common.serialization.StringSerializ
er");
props.put("value.serializer", "org.apache.kafka.common.serialization.StringSerial
izer");
//生产者发送消息
String topic = "spark_streaming_test1";
org.apache.kafka.clients.producer.Producer<String, String> procuder = new KafkaPr
oducer<String,String>(props);
while(true){
int num = (int)((Math.random())*10);
for (int i = 0; i <= 10; i++) {
int tmp = (num+i)%10;
String value = "value_" + tmp;
ProducerRecord<String, String> msg = new ProducerRecord<String, String>(topic, va
lue);
procuder.send(msg);
}
try {Thread.sleep(1000*10);}
catch (InterruptedException e) {}
}
}
}
```

修改其中的 $kafkaIP 为您的 CKafka 的内网 IP 地址。

这个程序每10秒向 CKafka 发送10条消息从 value_0 到 value_9，其开始的顺序随机。程序中的参数信息参考消费者程序。

使用本地命令行进入工程目录，执行以下指令对工程进行编译打包：

```
mvn package
```

显示 build success 表示操作成功，在工程目录下的 target 文件夹中能够看到打包好的文件。
使用 scp 或者 sftp 工具来把打包好的文件上传到 EMR 集群，注意一定要上传依赖一起打包的 jar 包：

```
scp $localfile root@公网IP地址:$remotefolder
```

使用程序消费 CKafka 的数据

使用两个界面分别登录 EMR 集群的 Web Shell。
第一个界面：登录 EMR 集群的 Master 节点，并且切换到 Hadoop 用户如2节中所示，使用以下命令执行样例：

```
[hadoop@172 ~]$ bin/spark-submit --class KafkaTest --master yarn-cluster $consume
rpackage
```

其中参数如下：

- --class 参数表示要执行的入口类，在本例中即为 KafkaTest。
- --master 为集群主要的 URL。
- $ consumerpackage 是您的消费者打包后的包名。

程序开始执行后，将会在 yarn 集群上一直运行，使用以下指令可以查看到程序运行的状态：

```
[hadoop@172 ~]$ yarn application -list
```

第二个界面：登录 EMR 的 Web Shell，然后运行生产者程序，以便 Spark Streaming 能够从中取数据消费。

```
[hadoop@172 spark]$ bin/spark-submit --class SendData $producerpackage
```

其中 $producerpackage 为您的生产者打包后的包名。等待一段时间后，会在指定的 HDFS 文件夹中输出 wordcount 的结果，可以到 HDFS 中查看 Spark Streaming 消费 CKafka 数据后输出的结果：

```
[hadoop@172 root]$ hdfs dfs -ls /user
Found 9 items
drwxr-xr-x - hadoop supergroup 0 2018-07-03 16:37 /user/hadoop
drwxr-xr-x - hadoop supergroup 0 2018-06-19 10:10 /user/hive
-rw-r--r-- 3 hadoop supergroup 0 2018-06-29 10:19 /user/pythontest.txt
drwxr-xr-x - hadoop supergroup 0 2018-07-05 20:25 /user/sparkstreamingtest-153079
3500000.result
[hadoop@172 root]$ hdfs dfs -cat /user/sparkstreamingtest-1530793500000.result/*
(value_6,16)
(value_7,22)
(value_8,18)
(value_0,18)
(value_9,17)
```

```
(value_1,18)
(value_2,17)
(value_3,17)
(value_4,16)
(value_5,17)
```

最后需要退出 yarn 集群中的 KafkaTest 程序：

```
[hadoop@172 ~]$ yarn application -kill $Application-Id
```

其中 $Application-Id 为使用 `yarn application -list` 命令查找到的 ID。

更多 Kafka 的相关信息请查看 官方文档。

Spark 资源动态调度实践

最近更新时间：2021-10-08 11:20:24

开发准备

确定您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候，需要在软件配置界面选择 spark_hadoop 组件。

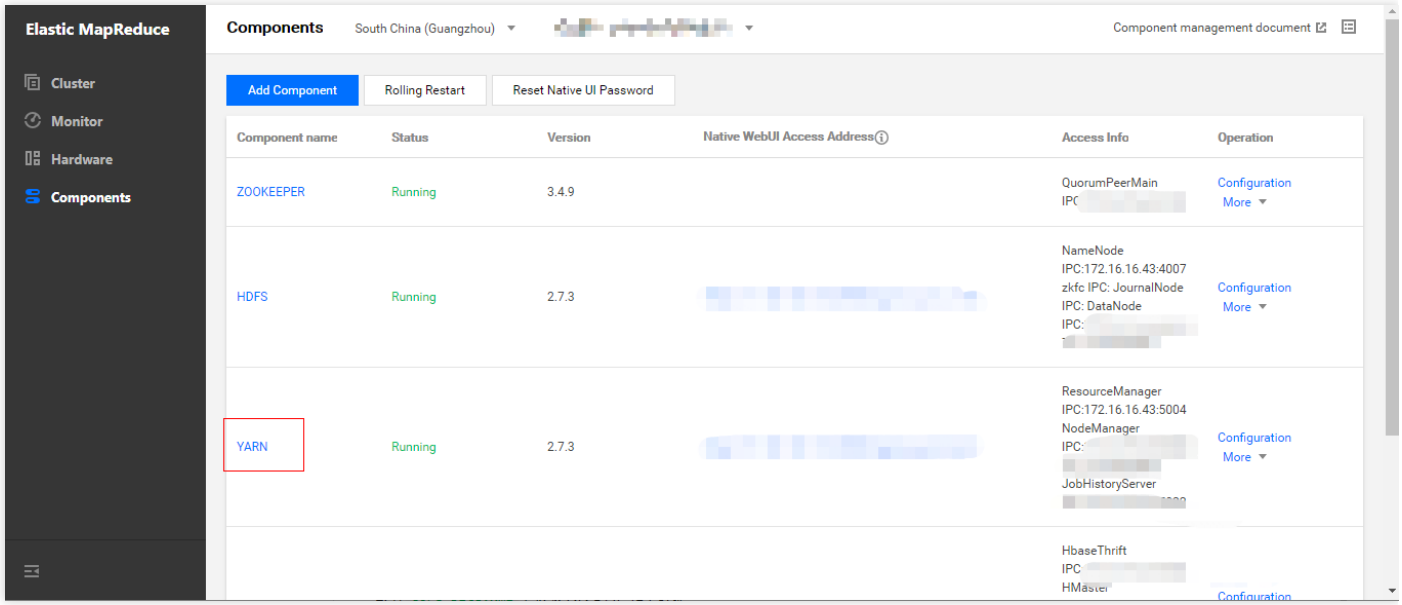
Spark 安装在 EMR 云服务器的 `/usr/local/service` 路径下（`/usr/local/service/spark`）。

拷贝 jar 包

需要将 `spark-<version>-yarn-shuffle.jar` 拷贝到集群各个节点的 `/usr/local/service/hadoop/share/hadoop/yarn/lib` 目录路径下。

方法一：SSH 控制台操作

1. 在**集群服务>YARN**组件中，选择**操作>角色管理**，确定 NodeManager 所在节点 IP。



2. 依次登录每个 NodeManager 所在节点。
   - 首先，需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例，这里我们可以使用 XShell 登录 Master 节点。
   - 使用 SSH 登录到其他 NodeManager 所在节点。指令为 `ssh $user@$ip`，$user 为登录的用户名，$ip 为远程节点 IP（即步骤1中确定的 IP 地址）。

```
[root@172 ~]# ssh root@17
The authenticity of host '                (1          )' can't be established.
ECDSA key fingerprint is b9:                                                :2d.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added '1          3' (ECDSA) to the list of known hosts.
root@1          's password:
```

- 验证已经成功切换。

```
[root@172 ~]# ifconfig
eth0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 1          netmask 255.255.240.0  broadcast 1
        ether 52:54:00:0d:ad:e3  txqueuelen 1000  (Ethernet)
        RX packets 212406  bytes 110895121 (105.7 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 178536  bytes 26768271 (25.5 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 1          netmask 255.0.0.0
        loop  txqueuelen 1  (Local Loopback)
        RX packets 378979  bytes 167242376 (159.4 MiB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 378979  bytes 167242376 (159.4 MiB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

3. 搜索 `spark-<version>-yarn-shuffle.jar` 文件路径。

```
[root@172 ~]# find / -name *shuffle.jar
/usr/local/service/hadoop/share/hadoop/yarn/spark-2.3.2-yarn-shuffle.jar
/usr/local/service/spark/yarn/spark-2.3.2-yarn-shuffle.jar
/usr/local/service/hive/spark/yarn/spark-2.0.2-yarn-shuffle.jar
/usr/local/service/apps/kylin-2.5.2/spark/yarn/spark-2.1.2-yarn-shuffle.jar
```

4. 将 `spark-<version>-yarn-shuffle.jar` 拷贝到 `/usr/local/service/hadoop/share/hadoop/yarn/lib` 下。

```
[root@172 ~]# cd /usr/local/service/hadoop/share/hadoop/yarn/lib/
[root@172 lib]# cp /usr/local/service/spark/yarn/spark-2.3.2-yarn-shuffle.jar spark-2.3.2-yarn-shuffle.jar
[root@172 lib]# ls
activation-1.1.jar              hadoop-lzo-0.4.20.jar          jetty-util-6.1.26.jar
aopalliance-1.0.jar             jackson-core-asl-1.9.13.jar    json-io-2.5.1.jar
asm-3.2.jar                     jackson-jaxrs-1.9.13.jar       jsr305-3.0.0.jar
commons-cli-1.2.jar             jackson-mapper-asl-1.9.13.jar  leveldbjni-all-1.8.jar
commons-codec-1.4.jar           jackson-xc-1.9.13.jar          log4j-1.2.17.jar
commons-collections-3.2.2.jar   javassist-3.18.1-GA.jar        netty-3.6.2.Final.jar
commons-compress-1.4.1.jar      java-util-1.9.0.jar            protobuf-java-2.5.0.jar
commons-io-2.4.jar              javax.inject-1.jar             ranger-plugin-classloader-0.7.1.jar
commons-lang-2.6.jar            jaxb-api-2.2.2.jar             ranger-yarn-plugin-impl
commons-logging-1.1.3.jar       jaxb-impl-2.2.3-1.jar          ranger-yarn-plugin-shim-0.7.1.jar
commons-math-2.2.jar            jersey-client-1.9.jar          servlet-api-2.5.jar
curator-client-2.7.1.jar        jersey-core-1.9.jar            spark-2.3.2-yarn-shuffle.jar
curator-test-2.7.1.jar          jersey-guice-1.9.jar           stax-api-1.0-2.jar
fst-2.50.jar                    jersey-json-1.9.jar            xz-1.0.jar
guava-11.0.2.jar                jersey-server-1.9.jar          zookeeper-3.4.6.jar
guice-3.0.jar                   jettison-1.1.jar               zookeeper-3.4.6-tests.jar
guice-servlet-3.0.jar           jetty-6.1.26.jar
[root@172 lib]# ls | grep spark-*
spark-2.3.2-yarn-shuffle.jar
```

5. 退出登录，并切换其余节点。

```
[root@172 lib]# exit
logout
Connection to [redacted] closed.
```

方法二：批量部署脚本

首先，需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例，这里我们可以使用 XShell 登录 Master 节点。

编写如下批量传输文件的 Shell 脚本。当集群节点很多时，为了避免多次输入密码，可以使用 sshpass 工具传输。sshpass 的优势在于可以免密传输避免多次输入，但其缺点在于密码是明文容易暴露，可以使用 history 命令找到。

1. 免密，安装 sshpass。

```
[root@172 ~]# yum install sshpass
```

编写如下脚本：

```
#!/bin/bash
nodes=(ip1 ip2 … ipn) #集群各节点 IP 列表，空格分隔
len=${#nodes[@]}
password=<your password>
file=" spark-2.3.2-yarn-shuffle.jar "
source_dir="/usr/local/service/spark/yarn"
target_dir="/usr/local/service/hadoop/share/hadoop/yarn/lib"
```

```
echo $len
for node in ${nodes[*]}
do
echo $node;
sshpass -p $password scp "$source_dir/$file"root@$node:"$target_dir";
done
```

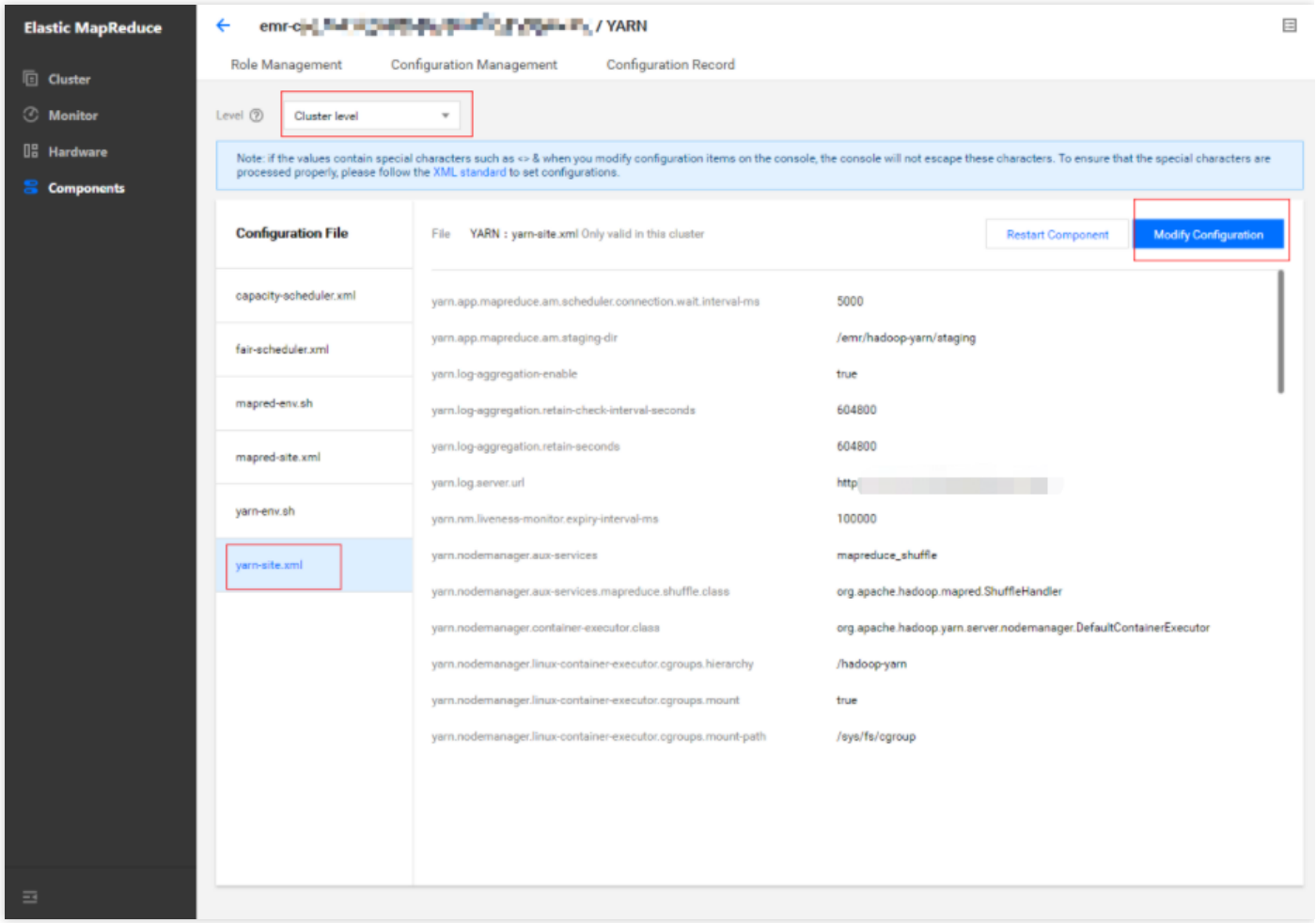
2. 非免密。

   编写如下脚本：

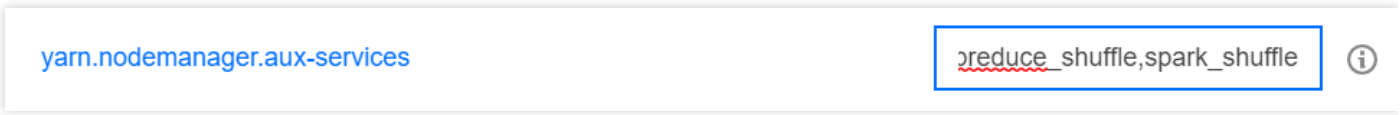
```
#!/bin/bash
nodes=(ip1 ip2 … ipn) #集群各节点 IP 列表，空格分隔
len=${#nodes[@]}
password=<your password>
file=" spark-2.3.2-yarn-shuffle.jar "
source_dir="/usr/local/service/spark/yarn"
target_dir="/usr/local/service/hadoop/share/hadoop/yarn/lib"
echo $len
for node in ${nodes[*]}
do
echo $node;
scp "$source_dir/$file" root@$node:"$target_dir";
done
```

修改 Yarn 配置

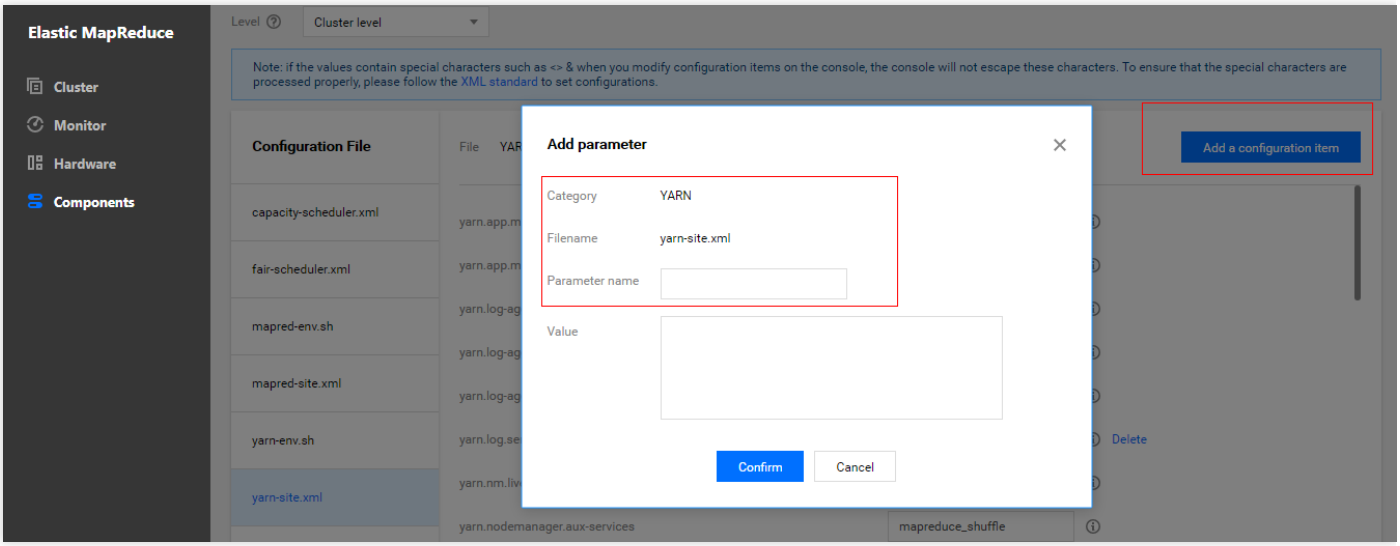
1. 在**集群服务>YARN**组件中，选择**操作>配置管理**。选中配置文件 `yarn-site.xml` ，**维度范围**选择“集群维度”（集群维度的配置项修改将应用到所有节点），然后单击**修改配置**。



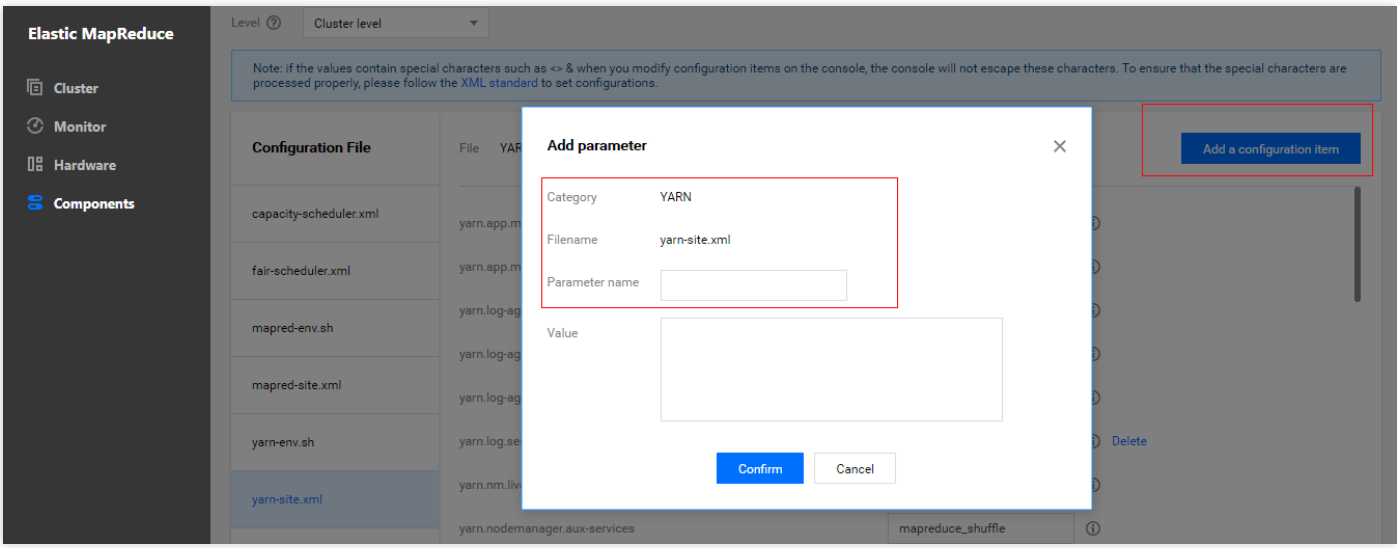
2. 修改配置项 `yarn.nodemanager.aux-services` ，添加 spark_shuffle。



3. 新增配置项 `yarn.nodemanager.aux-services.spark_shuffle.class` ，该配置项的值设置为 `org.apache.spark.network.yarn.YarnShuffleService` 。



4. 新增配置项 `spark.yarn.shuffle.stopOnFailure` ，该配置项的值设置为 false。

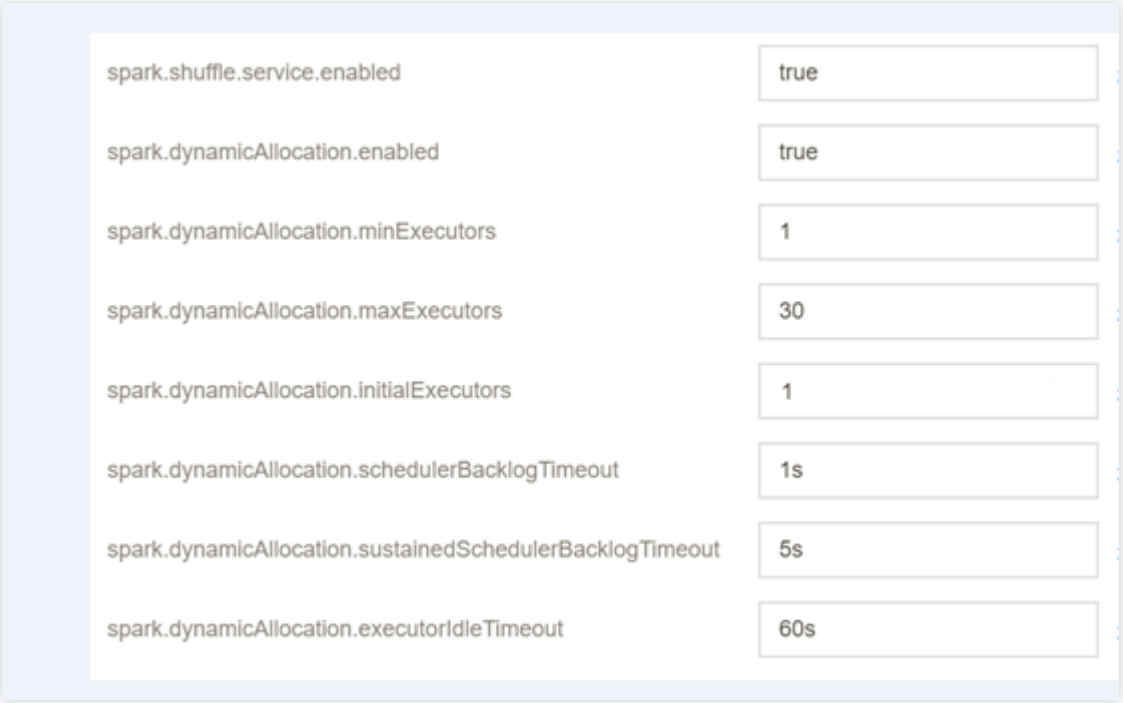


5. 保存设置并下发，重启 YARN 组件使得配置生效。

修改 Spark 配置

1. 在**集群服务>SPARK**组件中，选择**操作>配置管理**。

2. 选中配置文件**spark-defaults.conf**，单击**修改配置**。新建配置项如下：



配置项	值	备注
spark.shuffle.service.enabled	true	启动 shuffle 服务。
spark.dynamicAllocation.enabled	true	启动动态资源分配。
spark.dynamicAllocation.minExecutors	1	每个 Application 最小分配的 executor 数。
spark.dynamicAllocation.maxExecutors	30	每个 Application 最大分配的 executor 数。
spark.dynamicAllocation.initialExecutors	1	一般情况下与 spark.dynamicAllocation.minExecutors 值相同。
spark.dynamicAllocation.schedulerBacklogTimeout	1s	已有挂起的任务积压超过此持续事件，则将请求新的执行程序。
spark.dynamicAllocation.sustainedSchedulerBacklogTimeout	5s	带处理任务队列依然存在，则此后每隔几秒再次出发，每轮请求的 executor 数目与上轮相比呈指数增长。
spark.dynamicAllocation.executorIdleTimeout	60s	Application 在空闲超过几秒钟时会删除 executor。

3. 保存配置、下发并重启组件。

测试 Spark 资源动态调整

1. 测试环境资源配置说明

测试环境下，共两个部署 NodeManager 服务的节点，每个节点资源配置为4核 CPU、8GB内存，集群总资源为8核 CPU、16GB内存。

2. 测试任务说明

测试一：

- 在 EMR 控制台中，进入 `/usr/local/service/spark` 目录，切换 hadoop 用户，使用 `spark-submit` 提交一个任务，数据需存储在 hdfs 上。

```
[root@172 ~]# cd /usr/local/service/spark/
[root@172 spark]# su hadoop
[hadoop@172 spark]$ hadoop fs -put ./README.md /
[hadoop@172 spark]$ spark-submit --class org.apache.spark.examples.JavaWordCount --master yarn-client --num-executors 10 --driver-memory 4g --executor-memory 4g --executor-cores 2 ./examples/jars/spark-examples_2.11-2.3.2.jar /README.md /output
```

- 在 YARN 组件的 WebUI 界面 Application 面板中，可以观察到配置前后容器和 CPU 分配情况。

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blackl Node
application_1562242321020_0001	hadoop	spark word count	SPARK	default	0	Thu Jul 4 20:17:09 +0800 2019	N/A	RUNNING	UNDEFINED	3	3	9920	168.2	67.3		ApplicationMaster	0

- 未设置资源动态调度前，CPU 最多分配个数为3。

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklis Nodes
application_1562243747104_0001	hadoop	spark word count	SPARK	root.default	0	Thu Jul 4 20:40:57 +0800 2019	N/A	RUNNING	UNDEFINED	3	5	11264	76.4	76.4		ApplicationMaster	0

- 设置资源动态调度后，CPU 最多分配个数为5。

结论：配置资源动态调度后，调度器会根据应用程序的需要动态的增加分配的资源。

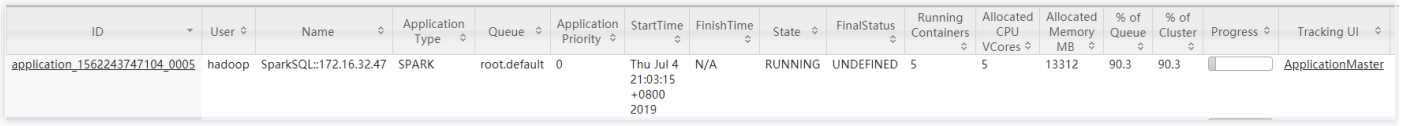
测试二：

- 在 EMR 控制台中，进入 `/usr/local/service/spark` 目录，切换 hadoop 用户，使用 `spark-sql` 启动 SparkSQL 交互式控制台。交互式控制台被设置成占用测试集群的大部分资源，观察设置资源动态调度前后资源分配情况。

```
[root@172 ~]# cd /usr/local/service/spark/
[root@172 spark]# su hadoop
[hadoop@172 spark]$ spark-sql --master yarn-client --num-executors 5 --driver-m
emory 4g --executor-memory 2g --executor-cores 1
```

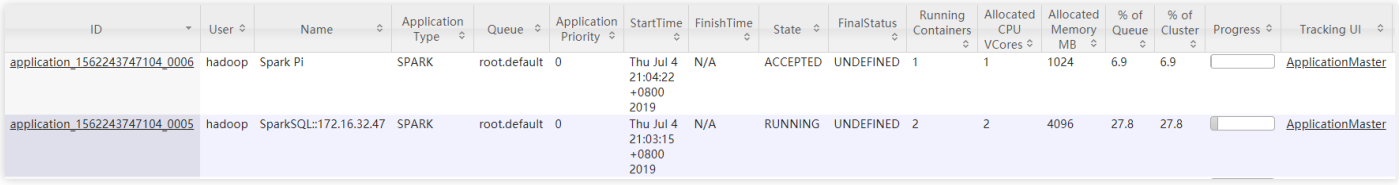
- 使用 spark2.3.0 自带的计算圆周率的 example 作为测试任务，提交任务时将任务的 executor 数设置为5，driver 内存设置为4g，executor 内存设置为4g，executor 核数设置为2。

```
[root@172 ~]# cd /usr/local/service/spark/
[root@172 spark]# su hadoop
[hadoop@172 spark]$ spark-submit --class org.apache.spark.examples.SparkPi --ma
ster yarn-client --num-executors 5 --driver-memory 4g --executor-memory 4g --ex
ecutor-cores 2 examples/jars/spark-examples_2.11-2.3.2.jar 500
```



ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI
application_1562243747104_0005	hadoop	SparkSQL::172.16.32.47	SPARK	root.default	0	Thu Jul 4 21:03:15 +0800 2019	N/A	RUNNING	UNDEFINED	5	5	13312	90.3	90.3		ApplicationMaster

- 只运行 SparkSQL 任务时资源占用率90.3%。



ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	% of Queue	% of Cluster	Progress	Tracking UI
application_1562243747104_0006	hadoop	Spark Pi	SPARK	root.default	0	Thu Jul 4 21:04:22 +0800 2019	N/A	ACCEPTED	UNDEFINED	1	1	1024	6.9	6.9		ApplicationMaster
application_1562243747104_0005	hadoop	SparkSQL::172.16.32.47	SPARK	root.default	0	Thu Jul 4 21:03:15 +0800 2019	N/A	RUNNING	UNDEFINED	2	2	4096	27.8	27.8		ApplicationMaster

- 提交 SparkPi 任务后，SparkSQL 资源占用率27.8%。

结论：SparkSQL 任务虽然在提交时申请了大量资源，但并未执行任何分析任务，因此实际上有大量空闲的资源。当超过 `spark.dynamicAllocation.executorIdleTimeout` 设置时间时，空闲的 executor 被释放，其他任务获得资源。在本次测试中 SparkSQL 任务的集群资源占用率从90%降至28%，空闲资源分配给圆周率计算任务，自动调度有效。

> 说明：
> 配置项 `spark.dynamicAllocation.executorIdleTimeout` 的值将影响资源动态调度的快慢，测试发现资源调度用时基本与该值相等，建议您根据实际需求调整该配置项的值以获得最佳性能。

Spark 集成 Kafka 说明

最近更新时间：2022-11-25 16:15:15

依赖关系

Spark 从2.3 起不再支持 Kafka0.8.2，EMR 现网版本已集成 Spark2.4.3及以上版本，需要集成 kafka 0.10.0及更高版本。

查找方法

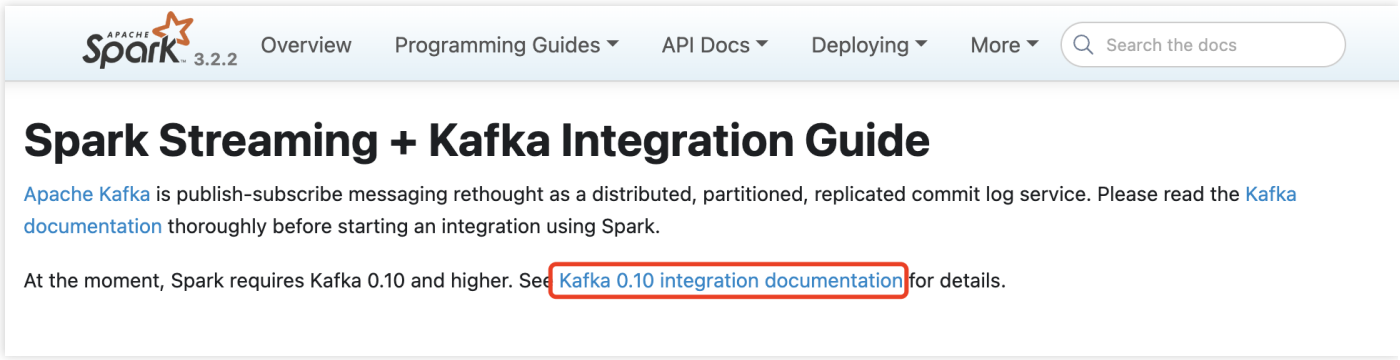
1. 访问官网链接，输入版本号链接模板：

```
https://spark.apache.org/docs/{spark.version}/streaming-kafka-integration.html
```

将 `{spark.version}` 替换为对应的 Spark 版本。例如查看3.2.2版本的依赖关系，访问链接如下：

```
https://spark.apache.org/docs/3.2.2/streaming-kafka-integration.html
```

2. 单击相应链接可看到详细集成说明。



EMR 各版本 Spark 相关依赖说明

最近更新时间：2022-11-25 16:06:47

依赖关系

Spark	scala	Python	R	Java
2.4.3	2.12.x	2.7+/3.4+	3.1+	8+
3.0.0	2.12.x	2.7+/3.4+	3.1+	8/11
3.0.2	2.12.x	2.7+/3.4+	3.5+	8/11
3.2.1	2.12.x/2.13.x	3.6+	3.5+	8/11
3.2.2	2.12.x/2.13.x	3.6+	3.5+	8/11

查找方法

1. 访问官网链接，输入版本号链接模板：

```
https://spark.apache.org/docs/{spark.version}/index.html
```

将 `{spark.version}` 替换为对应的 spark 版本，例如查看3.2.2版本的依赖关系，访问链接如下：

```
https://spark.apache.org/docs/3.2.2/index.html
```

2. 查看依赖

Note: Kafka 0.8 support is deprecated as of Spark 2.3.0.

	spark-streaming-kafka-0-8	spark-streaming-kafka-0-10
Broker Version	0.8.2.1 or higher	0.10.0 or higher
API Maturity	Deprecated	Stable
Language Support	Scala, Java, Python	Scala, Java
Receiver DStream	Yes	No
Direct DStream	Yes	Yes
SSL / TLS Support	No	Yes
Offset Commit API	No	Yes
Dynamic Topic Subscription	No	Yes

HBASE开发指南
通过 API 使用 Hbase

最近更新时间：2022-11-07 16:24:33

HBase 是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，是 Google BigTable 的开源实现。HBase 利用 Hadoop HDFS 作为其文件存储系统；Hadoop MapReduce 来处理 HBase 中的海量数据；Zookeeper 来做协同服务。

Hbase 主要由 Zookeeper、HMaster 和 HRegionServer 组成。其中 ：

- ZooKeeper 可避免 Hmaster 的单点故障，其 Master 选举机制可保证一个 Master 提供服务。
- Hmaster 管理用户对表的增删改查操作，管理 HRegionServer 的负载均衡。并可调整 Region 的分布，在 HRegionServer 退出时迁移其内的 HRegion 到其他 HRegionServer 上。
- HRegionServer 是 Hbase 中最核心的模块，其主要负责响应用户的 I/O 请求，向 HDFS 文件系统中读写数据。HRegionServer 内部管理了一系列 HRegion 对象，每个 HRegion 对应一个 Region，HRegion 中由多个 Store 组成。每个 Store 对应了 Column Family 的存储。

本开发指南将从技术人员的角度帮助用户使用 EMR 集群开发。考虑用户数据安全，EMR 中当前只支持 VPC 网络访问。

1. 开发准备

确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择了 Hbase 组件和 Zookeeper 组件。

2. 使用 Hbase Shell

在使用 Hbase Shell 之前请登录 EMR 集群的 Master 节点。登录 EMR 的方式可参考 登录 Linux 实例。这里可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入 EMR 命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入目录 `/usr/local/service/hbase` ：

```
[root@172 ~]# su hadoop
[hadoop@10root]$ cd /usr/local/service/hbase
```

通过如下命令您可以进入 Hbase Shell ：

```
[hadoop@10hbase]$ bin/hbase shell
```

在 hbase shell 下输入 help 可以查看基本的使用信息和示例的指令。接下来我们使用以下指令建立一个新表：

```
hbase(main):001:0> create 'test', 'cf'
```

表格建立后，可以使用 `list` 指令来查看您建立的表是否已经存在。

```
hbase(main):002:0> list 'test'
TABLE
test
1 row(s) in 0.0030 seconds
=> ["test"]
```

使用 `put` 指令来为您创建的表加入元素：

```
hbase(main):003:0> put 'test', 'row1', 'cf:a', 'value1'
0 row(s) in 0.0850 seconds
hbase(main):004:0> put 'test', 'row2', 'cf:b', 'value2'
0 row(s) in 0.0110 seconds
hbase(main):005:0> put 'test', 'row3', 'cf:c', 'value3'
0 row(s) in 0.0100 seconds
```

我们在创建的表中加入了三个值，第一次在“row1”行“cf:a”列插入了一个值“value1”，以此类推。

使用 `scan` 指令来遍历整个表：

```
hbase(main):006:0> scan 'test'
ROW COLUMN+CELL
row1 column=cf:a, timestamp=1530276759697, value=value1
row2 column=cf:b, timestamp=1530276777806, value=value2
row3 column=cf:c, timestamp=1530276792839, value=value3
3 row(s) in 0.2110 seconds
```

使用 `get` 指令来取得表中指定行的值：

```
hbase(main):007:0> get 'test', 'row1'
COLUMN CELL
cf:a timestamp=1530276759697, value=value
1 row(s) in 0.0790 seconds
```

使用 `drop` 指令来删除一个表，在删除表之前需要先使用 `disable` 指令来禁用一个表：

```
hbase(main):010:0> disable 'test'
hbase(main):011:0> drop 'test'
```

最后可以使用 `quit` 指令来关闭 hbase shell。

更多的 Hbase shell 指令请查看 官方文档。

3. 通过 API 使用 Hbase

首先 下载并安装 Maven，配置 Maven 的环境变量，如果您使用 IDE，请在 IDE 中设置 Maven 相关配置。

新建一个 Maven 工程

在命令行下进入您想要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID
-DarchetypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名。$yourartifactID 为您的项目名称，而 maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。

创建成功后，在 `D://mavenWorkplace` 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
    ---pom.xml          核心配置，项目根下
    ---src
        ---main
            ---java          Java 源码目录
     ---resources     Java 配置文件目录
      ---test
          ---java          测试源码目录
          ---resources     测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。

添加 Hadoop 依赖和样例代码

首先在 pom.xml 文件中添加 Maven 依赖：

```
<dependencies>
    <dependency>
    <groupId>org.apache.hbase</groupId>
    <artifactId>hbase-client</artifactId>
    <version>1.2.4</version>
```

```
</dependency>
</dependencies>
```

然后在 pom.xml 文件中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

在添加样例代码之前，需要用户获取 Hbase 集群的 zookeeper 地址。登录 EMR 任意一台 Master 节点或者 Core 节点，进入 `/usr/local/service/hbase/conf` 目录，查看 hbase-site.xml 的 hbase.zookeeper.quorum 配置获得 zookeeper 的 IP 地址 $quorum，hbase.zookeeper.property.clientPort 配置获得 zookeeper 的端口号 $clientPort。

接下来添加样例代码，在 main>java 文件夹下新建一个 Java Class 取名为 PutExample.java，并将以下代码加入其中：

```
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.hbase.*;
import org.apache.hadoop.hbase.client.*;
```

```
import org.apache.hadoop.hbase.util.Bytes;
import org.apache.hadoop.hbase.io.compress.Compression.Algorithm;
import java.io.IOException;
/**
* Created by tencent on 2018/6/30.
*/
public class PutExample {
public static void main(String[] args) throws IOException {
Configuration conf = HBaseConfiguration.create();
conf.set("hbase.zookeeper.quorum","$quorum");
conf.set("hbase.zookeeper.property.clientPort","$clientPort");
conf.set("zookeeper.znode.parent", "$znodePath");
Connection connection = ConnectionFactory.createConnection(conf);
Admin admin = connection.getAdmin();
HTableDescriptor table = new HTableDescriptor(TableName.valueOf("test1"));
table.addFamily(new HColumnDescriptor("cf").setCompressionType(Algorithm.NONE));
System.out.print("Creating table. ");
if (admin.tableExists(table.getTableName())) {
admin.disableTable(table.getTableName());
admin.deleteTable(table.getTableName());
}
admin.createTable(table);
Table table1 = connection.getTable(TableName.valueOf("test1"));
Put put1 = new Put(Bytes.toBytes("row1"));
put1.addColumn(Bytes.toBytes("cf"), Bytes.toBytes("a"),
Bytes.toBytes("value1"));
table1.put(put1);
Put put2 = new Put(Bytes.toBytes("row2"));
put2.addColumn(Bytes.toBytes("cf"), Bytes.toBytes("b"),
Bytes.toBytes("value2"));
table1.put(put2);
Put put3 = new Put(Bytes.toBytes("row3"));
put3.addColumn(Bytes.toBytes("cf"), Bytes.toBytes("c"),
Bytes.toBytes("value3"));
table1.put(put3);
System.out.println(" Done.");
}
}
```

编译代码并打包上传

使用本地命令行进入工程目录，执行以下指令对工程进行编译打包：

```
mvn package
```

显示 build success 表示操作成功，在工程目录下的 target 文件夹中能够看到打包好的文件。

使用 scp 或者 sftp 工具把打包好的文件上传到 EMR 集群。**这里一定要上传把依赖一起进行打包的 jar 包**。在本地命令行模式下运行：

```
scp $localfile root@公网IP地址:$remotefolder
```

其中，$localfile 是您的本地文件的路径加名称，root 为 CVM 服务器用户名，公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看。$remotefolder 是您想存放文件的 CVM 服务器路径。上传完成后，在 EMR 集群命令行中即可查看对应文件夹下是否有相应文件。

4. 运行样例

登录 EMR 集群的 Master 节点，并且切换到 hadoop 用户。使用如下命令执行样例：

```
[hadoop@10 hadoop]$ java -jar $package.jar
```

在控制台输出“Done”后，说明所有的操作已完成。可切换到 hbase shell 中使用 `list` 命令来查看使用 API 创建的 Hbase 表是否成功。如果成功可使用 `scan` 命令来查看表的具体内容。

```
[hadoop@10hbase]$ bin/hbase shell
hbase(main):002:0> list 'test1'
TABLE
Test1
1 row(s) in 0.0030 seconds
=> ["test1"]
hbase(main):006:0> scan 'test1'
ROW COLUMN+CELL
row1 column=cf:a, timestamp=1530276759697, value=value1
row2 column=cf:b, timestamp=1530276777806, value=value2
row3 column=cf:c, timestamp=1530276792839, value=value3
3 row(s) in 0.2110 seconds
```

更多的 API 使用说明请参见 官方文档。

通过 Thrift 使用 Hbase

最近更新时间：2021-08-12 10:25:16

Apache Thrift 是一个跨平台、跨语言的开发框架，提供多语言的编译功能，并提供多种服务器工作模式。用户通过 Thrift 的 IDL（接口描述语言）来描述接口函数及数据类型，然后通过 Thrift 的编译环境生成各种语言类型的接口文件，用来进行可扩展且跨语言的服务的开发。

它结合了功能强大的软件堆栈和代码生成引擎，以构建在 C++、Java、Go、Python、PHP、Ruby、Erlang、Perl、Haskell、C#、Cocoa、JavaScript、Node.js、Smalltalk 和 OCaml 编程语言间无缝结合的、高效的服务。

Thrift server 是 HBase 中的一种服务，主要用于对多语言 API 的支持。基于 Apache Thrift 开发。Thrift API 依赖于客户端和服务器进程。本节将以 Python 为例子，说明如何通过 Thrift 利用 Python 编程来使用 Hbase。

1. 开发准备

确认您已开通腾讯云，并且创建了一个 EMR 集群。创建 EMR 集群时需要在软件配置界面选择 Hbase 组件。

2. 通过 Python API 使用 Hbase

EMR 集群中 Hbase 默认集成了 Thrift，并在 Master1（外网 IP 节点）节点上启动了 Thrift Server。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户并进入 Hbase 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/hbase/
[hadoop@172 hbase]$
```

在 Hbase 的配置文件中查看 thrift 的 IP 地址和端口号：

```
[hadoop@172 hbase]$ vim conf/hbase-site.xml
<property>
<name>hbase.master.hostname</name>
<value>$thriftIP</value>
</property>
<property>
<name>hbase.regionserver.thrift.port</name>
```

```
<value>$port</value>
</property>
```

其中 $port 为 ThriftServer 的端口号。

因为 EMR 集群的 Hbase 默认集成了 Thrift，所以不需要再进行安装配置，使用以下命令查看 Thrift Server 是否已经启动：

```
[hadoop@172 hbase]$ jps
4711 ThriftServer
```

可见 Thrift Server 已经在后台运行。我们可以直接使用 Python 编程来操作 Hbase。

负载均衡

HA 集群有两个 master 节点，两个节点默认都启动了 Thrift Server。若需要实现负载均衡，客户端代码需要自定义策略将请求分散到两台 Thrift Server 上，这两台 Thrift Server 是完全独立的，之间没通信。

准备数据

使用 Hbase Shell 在 Hbase 中新建一个表，如果您使用过 EMR 的 Hbase 并且创建过自己的表，那么该步骤可以略过：

```
[hadoop@172 hbase]$ hbase shell
hbase(main):001:0> create 'thrift_test', 'cf'
hbase(main):005:0> list
thrift_test
1 row(s) in 0.2270 seconds
hbase(main):001:0> quit
```

使用 Python 查看 Hbase 中的表

首先需要安装 Python 依赖包，切换到 root 用户下，密码即为创建 EMR 集群时您设置的密码，先安装 python-pip 工具再安装依赖包：

```
[hadoop@172 hbase]$ su
Password: ********
[root@172 hbase]# yum install python-pip
[root@172 hbase]# pip install hbase-thrift
```

然后切换回 Hadoop 用户并新建一个 Python 文件 Hbase_client.py，在其中加入以下代码：

```
#! /usr/bin/env python
#coding=utf-8
from thrift.transport import TSocket,TTransport
```

```
from thrift.protocol import TBinaryProtocol
from hbase import Hbase
socket = TSocket.TSocket('$thriftIP ', $port)
socket.setTimeout(5000)
transport = TTransport.TBufferedTransport(socket)
protocol = TBinaryProtocol.TBinaryProtocol(transport)
client = Hbase.Client(protocol)
transport.open()
print client.getTableNames()
```

注意：
其中 $thriftIP 为 Master 节点在内网的 IP 地址，$port 为 ThriftService 的端口号，下同。

保存后直接运行程序，会直接在控制台输出 Hbase 中的存在的表：

```
[hadoop@172 hbase]$ python Hbase_client.py
['thrift_test']
```

使用 Python 创建一个 Hbase 表

新建一个 Python 文件 Create_table.py，把以下代码加入其中：

```
#! /usr/bin/env python
#coding=utf-8
from thrift import Thrift
from thrift.transport import TSocket,TTransport
from thrift.protocol import TBinaryProtocol
from hbase import Hbase
from hbase.ttypes import ColumnDescriptor,Mutation,BatchMutation,TRegionInfo
from hbase.ttypes import IOError,AlreadyExists
socket = TSocket.TSocket('$thriftIP ',$port)
socket.setTimeout(5000)
transport = TTransport.TBufferedTransport(socket)
protocol = TBinaryProtocol.TBinaryProtocol(transport)
client = Hbase.Client(protocol)
transport.open()
new_table = ColumnDescriptor(name = 'cf:',maxVersions = 1)
client.createTable('thrift_test_1',[new_table])
tables = client.getTableNames()
socket.close()
print tables
```

该程序会在 Hbase 中添加一个名为 thrift_test_1 的新表，并且输出所有存在的表，运行效果如下：

```
[hadoop@172 hbase]$ python Create_table.py
['thrift_test', 'thrift_test_1']
```

使用 Python 在 Hbase 表中插入数据

新建一个 Python 文件 Insert.py，把以下代码加入其中：

```
#! /usr/bin/env python
#coding=utf-8
from thrift import Thrift
from thrift.transport import TSocket,TTransport
from thrift.protocol import TBinaryProtocol
from hbase import Hbase
from hbase.ttypes import ColumnDescriptor,Mutation,BatchMutation,TRegionInfo
from hbase.ttypes import IOError,AlreadyExists
socket = TSocket.TSocket('$thriftIP ', $port)
socket.setTimeout(5000)
transport = TTransport.TBufferedTransport(socket)
protocol = TBinaryProtocol.TBinaryProtocol(transport)
client = Hbase.Client(protocol)
transport.open()
mutation1 = [Mutation(column = "cf:a",value = "value1")]
client.mutateRow('thrift_test_1',"row1",mutation1)
mutation2 = [Mutation(column = "cf:b",value = "value2")]
client.mutateRow('thrift_test_1',"row1",mutation2)
mutation1 = [Mutation(column = "cf:a",value = "value3")]
client.mutateRow('thrift_test_1',"row2",mutation1)
mutation2 = [Mutation(column = "cf:b",value = "value4")]
client.mutateRow('thrift_test_1',"row2",mutation2)
socket.close()
```

该程序会在 Hbase 的 thrift_test_1 表中添加两行数据，每行分别有两个数据，可以在 Hbase Shell 中查看插入的数据：

```
hbase(main):005:0> scan 'thrift_test_1'
ROW COLUMN+CELL
row1 column=cf:a, timestamp=1530697238581, value=value1
row1 column=cf:b, timestamp=1530697238587, value=value2
row2 column=cf:a, timestamp=1530704886969, value=value3
row2 column=cf:b, timestamp=1530704886975, value=value4
2 row(s) in 0.0190 seconds
```

使用 Python 查看 Hbase 表中的数据

有两种查看方式，一种是查看一行，一种是使用 Scan 来查看全部数据，新建一个 Python 文件 Scan_table.py，加入以下代码：

```
#! /usr/bin/env python
#coding=utf-8
from thrift import Thrift
from thrift.transport import TSocket,TTransport
from thrift.protocol import TBinaryProtocol
from hbase import Hbase
from hbase.ttypes import ColumnDescriptor,Mutation,BatchMutation,TRegionInfo
from hbase.ttypes import IOError,AlreadyExists
socket = TSocket.TSocket('$thriftIP ', $port)
socket.setTimeout(5000)
transport = TTransport.TBufferedTransport(socket)
protocol = TBinaryProtocol.TBinaryProtocol(transport)
client = Hbase.Client(protocol)
transport.open()
result1 = client.getRow("thrift_test_1","row1")
print result1
for r in result1:
print 'the rowname is ',r.row
print 'the frist value is ',r.columns.get('cf:a').value
print 'the second value is ',r.columns.get('cf:b').value
scanId = client.scannerOpen('thrift_test_1',"",["cf"])
result2 = client.scannerGetList(scanId,10)
print result2
client.scannerClose(scanId)
socket.close()
```

其中使用 GetRow 来取得一行的数据，使用 scannerGetList 来得到整个表格中的数据，运行该程序后输出如下：

```
[hadoop@172 hbase]$ python Scan_table.py
[TRowResult(columns={'cf:a': TCell(timestamp=1530697238581, value='value1'), 'cf:
b': TCell(timestamp=1530697238587, value='value2')}, row='row1')]
the rowname is row1
the frist value is value1
the second value is value2
[TRowResult(columns={'cf:a': TCell(timestamp=1530697238581, value='value1'), 'cf:
b': TCell(timestamp=1530697238587, value='value2')}, row='row1'), TRowResult(colu
mns={'cf:a': TCell(timestamp=1530704886969, value='value3'), 'cf:b': TCell(timest
amp=1530704886975, value='value4')}, row='row2')]
```

分别输出了第一行的数据和整个表中的数据。

使用 Python 删除 Hbase 中的数据

新建一个 Python 文件 Delete_row.py，把以下代码加入其中：

```
#! /usr/bin/env python
#coding=utf-8
from thrift import Thrift
from thrift.transport import TSocket,TTransport
from thrift.protocol import TBinaryProtocol
from hbase import Hbase
from hbase.ttypes import *
socket = TSocket.TSocket('$thriftIP ',$port)
socket.setTimeout(5000)
transport = TTransport.TBufferedTransport(socket)
protocol = TBinaryProtocol.TBinaryProtocol(transport)
client = Hbase.Client(protocol)
transport.open()
client.deleteAllRow("thrift_test_1","row2")
socket.close()
```

该程序会删除测试表中的第二行数据，运行后可以在 Hbase Shell 中查看该表中的内容：

```
[hadoop@172 hbase]$ python Delete_row.py
[hadoop@172 hbase]$ hbase shell
hbase(main):004:0> scan 'thrift_test_1'
ROW COLUMN+CELL
row1 column=cf:a, timestamp=1530697238581, value=value1
row1 column=cf:b, timestamp=1530697238587, value=value2
1 row(s) in 0.2050 seconds
```

此时表中只剩第一行中的数据。

更多关于 Thrift 的操作详见 如何使用 Thrift。

Spark On Hbase

最近更新时间：2021-07-13 15:09:19

关于 Spark on Hbase 的详细操作，具体可查看 Github 主页 Spark-HBase Connector。

MapReduce On Hbase

最近更新时间：2021-07-13 15:03:33

关于 MapReduce on Hbase 的读写操作等内容，具体可参见 使用示例。

Phoenix on Hbase 开发指南
Phoenix 客户端使用

最近更新时间：2022-11-25 16:06:47

Phoenix 查询引擎支持使用 SQL 进行 HBase 数据的查询，会将 SQL 查询转换为一个或多个 HBase API，协同处理器与自定义过滤器的实现，并编排执行。使用 Phoenix 进行简单查询，其性能量级是毫秒，对于百万级别的行数来说，其性能量级是秒。EMR 中选择 HBase 组件的集群，默认集成 phoenix 客户端。

1. 启动客户端

   切换成 hadoop 用户，进入/usr/local/service/hbase/phoenix-client/bin 目录，使用 Phoenix 的 Python 命令行工具：

```
./sqlline.py
```

执行成功后显示：

```
[hadoop@172 /usr/local/service/hbase/phoenix-client/bin]$ ./sqlline.py
Setting property: [incremental, false]
Setting property: [isolation, TRANSACTION_READ_COMMITTED]
issuing: !connect -p driver org.apache.phoenix.jdbc.PhoenixDriver -p user "none" -p password "none" "jdbc:phoenix:"
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/service/hbase/phoenix-client/phoenix-client-hbase-2.4-5.1.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/service/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Connecting to jdbc:phoenix:
Connected to: Phoenix (version 5.1)
Driver: PhoenixEmbeddedDriver (version 5.1)
Autocommit status: true
Transaction isolation: TRANSACTION_READ_COMMITTED
sqlline version 1.9.0
0: jdbc:phoenix:>
```

2. Phoenix 引擎支持使用 SQL 进数据查询，一些常见操作如下：

- 创建表

```
0: jdbc:phoenix:> CREATE TABLE IF NOT EXISTS TEST (
host char(50) not null,
txn_count bigint
CONSTRAINT pk PRIMARY KEY (host)
);
```

- 插入数据

```
0: jdbc:phoenix:>UPSERT INTO TEST(host,txn_count) VALUES('192.168.1.1',1);
0: jdbc:phoenix:>UPSERT INTO TEST(host,txn_count) VALUES('192.168.1.2',2);
```

- 查询数据

```
0: jdbc:phoenix:>SELECT * FROM TEST;
```

- 删除数据表

```
0: jdbc:phoenix:>DROP TABLE IF EXISTS TEST;
```

更多操作及说明，可参考 社区文档。

Phoenix JDBC 使用

最近更新时间：2022-11-25 16:06:47

添加 maven 依赖

```
<dependency>
<groupId>org.apache.phoenix</groupId>
<artifactId>phoenix-core</artifactId>
<version>${phoenix.version}</version>
</dependency>
```

其中，phoenix.version 与集群中 phoenix 版本保持一致。

创建 JDBC 连接对象

```
Class.forName("org.apache.phoenix.jdbc.PhoenixDriver");
// Connect to the database
connection = DriverManager.getConnection("jdbc:phoenix:10.0.0.3:2181,10.0.0.5:218
1,10.0.0.8:2181");
```

执行查询

```
private static void instertPhoenix(Connection connection)throws Exception{
String sql="upsert into album_subscribe_log(id,album_id,user_id,op_time,sub_flag,
is_optimize,type_parent_id,type_id,host_id,is_pay,user_type,identtity_typ)"
+" values(?,?,?,?,?,?,?,?,?,?,?,?)";
PreparedStatement ps=connection.prepareStatement(sql);
ps.setLong(0,1);
ps.setLong(1,3);
ps.setLong(2,1);
ps.setString(3,"2017-09-05 14:00:00");
ps.setInt(4,1);
ps.setString(5,"1");
ps.setInt(6,3);
ps.setInt(7,5);
ps.setInt(8,6);
ps.setInt(9,7);
```

```
ps.setString(10,"1");
ps.setString(11,"1");
ps.setString(12,"1");
ps.executeUpdate();
ps.close();
connection.commit();
}
```

Phoenix 最佳实践

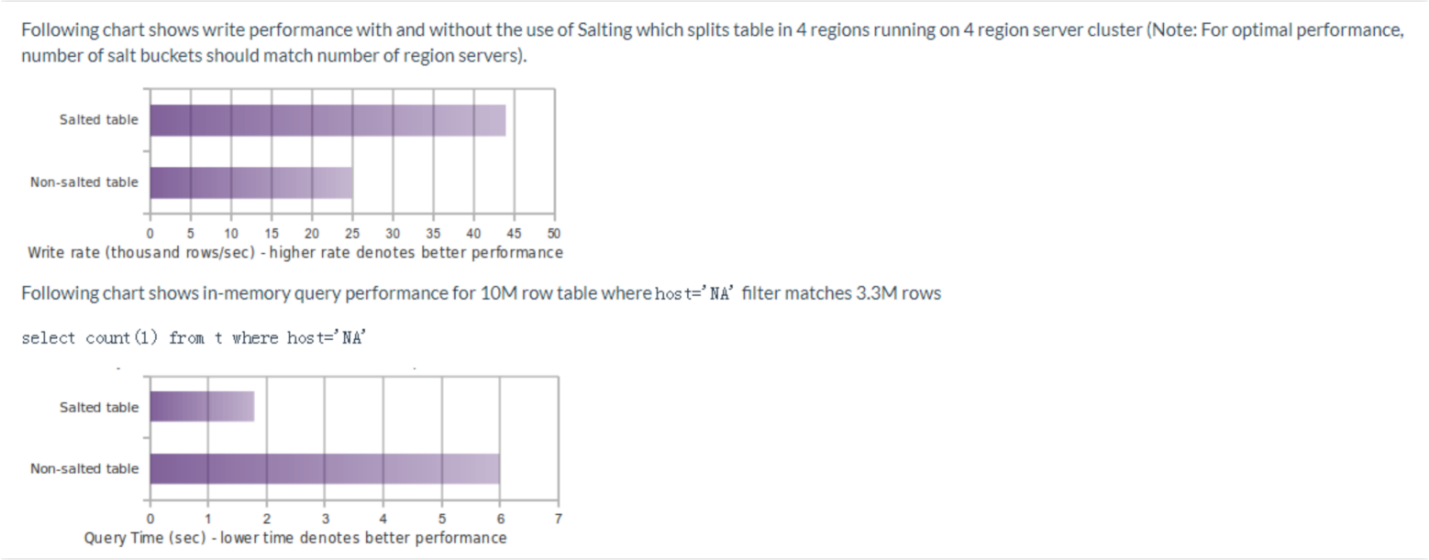
最近更新时间：2021-08-12 10:25:16

使用 Phoenix salted 表

Hbase 的 Row Key 假如没有经过精心设计，如果它又是自增长的，那么顺序写很可能会导致热点问题。PHoenix 提供了一种透明的方法，关联 salt 和 RowKey 到一张指定表的方案。只需要在创建表的时候添加 SALT_BUCKETS 关键字，这个值的范围是1到256。例如：

```
0: jdbc:phoenix:>CREATE TABLE table (a_key VARCHAR PRIMARY KEY, a_col VARCHAR) SA
LT_BUCKETS = 20;
```

可以免除使用 Hbase API 时需要精心设计 RowKey 的麻烦。如果用户对 Hbase Row Key 的设计理解不够，建议使用 salted 表。Phoenix 官方提供的写和查询，salted 表和非 salted 表性能对比：



更多 salte 性能或者操作说明，可查看 Phoenix salted 表 社区文档。

Phoenix 二级索引

对于 HBase 而言，如果想精确地定位到某行记录，唯一的办法是通过 rowkey 来查询。如果不通过 rowkey 来查找数据，就必须逐行地比较每一列的值，即全表扫瞄。对于较大的表，全表扫瞄的代价是不可接受的。但是，很多情况下，需要从多个角度查询数据。这需要二级索引（secondary index）来完成这件事。二级索引的原理很简单，但是如果自己维护的话则会麻烦一些。

Phoenix 二级索引配置

EMR 中 Phoenix 直接可支持 Phoenix 的二级索引。如果需要使用非事务性的，可变的索引只需按照以下步骤配置即可。打开 EMR 组件管理页面，单击【Hbase】，选择【配置】>【配置管理】，增加 hbase-site.xml 的 `hbase.regionserver.wal.codec` 、 `hbase.region.server.rpc.scheduler.factory.class` 和 `hbase.rpc.controllerfactory.class` 三个配置项即可，详细配置如下：

```
<property>
<name>hbase.regionserver.wal.codec</name>
<value>org.apache.hadoop.hbase.regionserver.wal.IndexedWALEditCodec</value>
</property>
<property>
<name>hbase.region.server.rpc.scheduler.factory.class</name>
<value>org.apache.hadoop.hbase.ipc.PhoenixRpcSchedulerFactory</value>
<description>Factory to create the Phoenix RPC Scheduler that uses separate queue
s for index and metadata updates</description>
</property>
<property>
<name>hbase.rpc.controllerfactory.class</name>
<value>org.apache.hadoop.hbase.ipc.controller.ServerRpcControllerFactory</value>
<description>Factory to create the Phoenix RPC Scheduler that uses separate queue
s for index and metadata updates</description>
</property>
```

Phoenix 二级索引使用

创建二级索引，命令如下：

```
0: jdbc:phoenix:>CREATE INDEX my_index ON my_table (v1) INCLUDE (v2)；
```

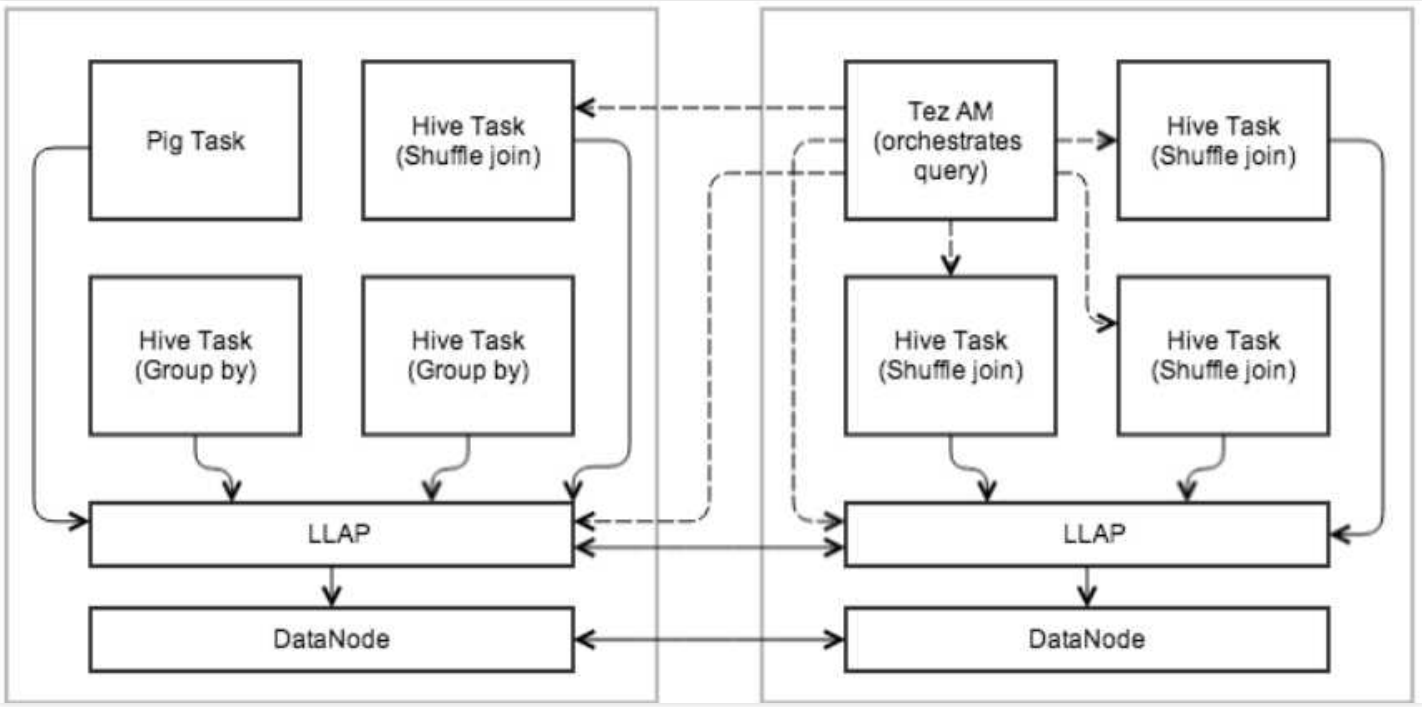
更多二级索引操作说明，可查看 Phoenix 二级索引社区文档。

Hive 开发指南
Hive 支持 LLAP

最近更新时间：2022-04-18 15:37:37

apache hive 从 Hive 2.0 版本引入了 LLAP（Live Long And Process），2.1版本进行了比较大的优化，可以说 hive 已经走向了内存计算。目前 hortonworks 测试 llap + tez 比 hive1.x 快了25倍。LLAP 提供了混合执行 model，由一个 long-lived 守护进程组成，取代了与 HDFS DataNode 的直接交互和一个紧密集成的 DAG-based framework。例如，缓存 pre-fetching，部分查询处理和访问控制之类的功能被移动到守护进程中。Small/short 查询由此守护程序直接处理，其他较为复杂的查询将在标准 YARN 容器中执行。

hive-llap 架构图



- 持久守护进程
  为了便于缓存和 JIT 优化，并消除大部分启动成本，守护程序在 cluster 上的 worker 节点上运行。守护程序处理 I/O、缓存和查询片段执行。
- 执行引擎
  LLAP 在现有的 process-based Hive 执行中工作，以保持 Hive 的可伸缩性和多功能性。它不会替换现有的执行 model，而是会增强它。

- 查询片段执行
  LLAP 节点执行“查询片段”，例如过滤器、投影、数据变换、部分聚合、排序、分组、散列 `joins/semi-joins` 等。
- I/O
  守护进程 off-loads I/O 并从压缩格式转换为单独的线程。数据在准备就绪时传递给执行，因此可以在准备下一批数据时处理以前的批次。数据以简单的 RLE-encoded 柱状格式传递给执行，可以进行矢量化处理；这也是缓存格式，旨在最小化 I/O、缓存和执行之间的复制。
- 缓存
  守护进程缓存输入 files 的元数据以及数据。即使对于当前未缓存的数据，也可以缓存元数据和索引信息。

安装 hive-llap

1. 进入 EMR 购买页。

2. 选择产品版本：EMR-V2.3.0。

3. 在【可选组件】列表中，选择【TEZ 0.9.2】后就会默认安装 hive-llap，安装目录位于 `/usr/local/service/slider` 。

使用 hive-llap

- 修改 `/usr/local/service/slider/conf/slider-client.xml` ，增加配置项：

```
<property>
<name>hadoop.registry.zk.quorum</name>
<value>zk_ip:zk_port</value>
</property>
```

这里的 value，如果是非高可用集群 IP 是 master 节点，如果是高可用集群，执行如下命令查看 zookeeper 组件（即 zk）。

```
cat /usr/local/service/hadoop/etc/hadoop/hdfs-site.xml |grep 2181
```

- 修改 hive-site.xml（通过控制台下发）

```
<property>
<name>hive.execution.engine</name>
<value>tez</value>
```

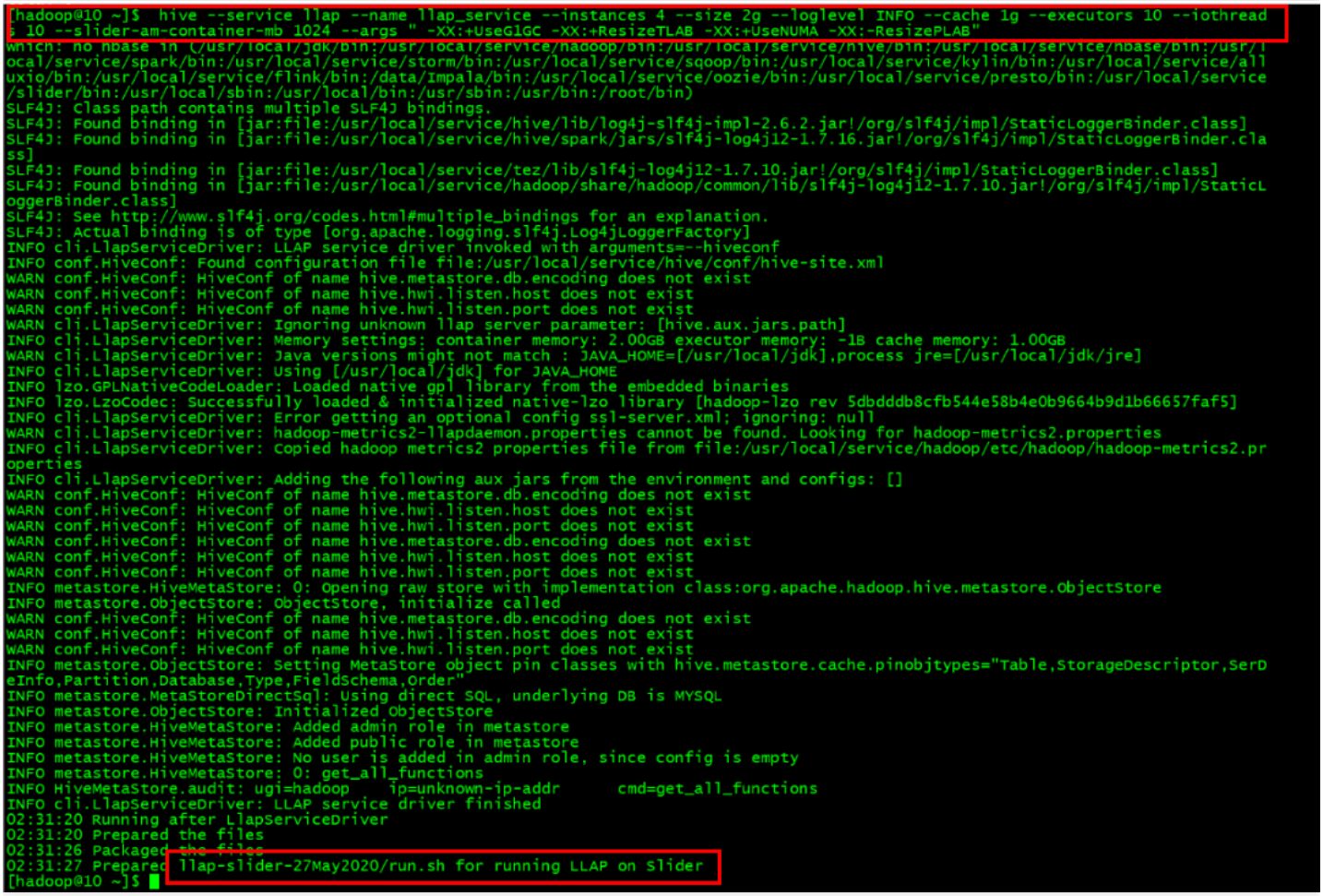
```
</property>
<property>
<name>hive.llap.execution.mode</name>
<value>all</value>
</property>
<property>
<name>hive.execution.mode</name>
<value>llap</value>
</property>
<property>
<name>hive.llap.daemon.service.hosts</name>
<value>@llap_service</value>
</property>
<property>
<name>hive.zookeeper.quorum</name>
<value>zk_ip:zk_port</value>
</property>
<property>
<name>hive.llap.daemon.memory.per.instance.mb</name>
<value>4000</value>
</property>
<property>
<name>hive.llap.daemon.num.executors</name>
<value>2</value>
</property>
<property>
<name>hive.server2.tez.default.queues</name>
<value>root.default</value>
</property>
<property>
<name>hive.server2.tez.initialize.default.sessions</name>
<value>true</value>
</property>
<property>
<name>hive.server2.tez.sessions.per.default.queue</name>
<value>2</value>
</property>
```

注意：
这里的 `hive.zookeeper.quorum` 配置项需要填写实际的 zookeeper 地址和端口。

- 重启 hive 所有服务
- 生成 llap 启动文件和命令

```
hive --service llap --name llap_service --instances 2 --size 2g --loglevel INFO
--cache 1g --executors 2 --iothreads 5 --slider-am-container-mb 1024 --args " -
XX:+UseG1GC -XX:+ResizeTLAB -XX:+UseNUMA -XX:-ResizePLAB"
```

这里会在当前目录下生成 llap 的运行和配置文件，根据提示执行 `run.sh` 脚本，如：




```
llap-slider-27May2020/run.sh
```

执行上面这个 `run.sh` 文件，这里会在 yarn 上提交一个常驻 application，等待几分钟会在 `yarn-ui` 上看到这个application。

> 说明：
> 这里 LLAP 初始化会需要几分钟，等待几分钟后再去执行数据操作。

- 验证 llap

  进入 hive cli

```
create table t1(id int, name string);
insert into t1 values('1','test1'),('2', 'test2');
select id, count(1) from t1 group by id;
```

预期结果：

```
hive> insert into t1 values('1','test1'),('2', 'test2');
Query ID = hadoop_20200527114442_1d4a9647-b10f-435f-9b78-5fbdacd78f1f
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1590545548684_0014)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........      llap     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 1.16 s
--------------------------------------------------------------------------------
Loading data to table default.t1
OK
Time taken: 2.732 seconds
hive> select id,count(1) from t1 group by id;
Query ID = hadoop_20200527115110_847ad7c3-634b-44b4-af5f-f1c0e9e8b43b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1590545548684_0015)

--------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........      llap     SUCCEEDED      2          2        0        0       0       0
Reducer 2 ......      llap     SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 0.77 s
--------------------------------------------------------------------------------
OK
1       5
2       5
Time taken: 6.925 seconds, Fetched: 2 row(s)
hive>
```

Hive 基础操作

最近更新时间：2021-07-09 10:40:22

Hive 是一个建立在 Hadoop 文件系统上的数据仓库架构，它为数据仓库的管理提供了许多功能，包括数据 ETL（抽取、转换和加载）工具、数据存储管理和大型数据集的查询和分析能力。同时 Hive 还定义了类 SQL 的语言 Hive-SQL。Hive-SQL 允许用户进行和 SQL 相似的操作，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能。还允许开发人员方便地使用 Mapper 和 Reducer 操作，可以将 SQL 语句转换为 MapReduce 任务运行，这对 MapReduce 框架来说是一个强有力的支持。其优点是学习成本低，可通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库统计分析。

Hive 使用 Hadoop 的 HDFS 作为文件的存储系统，很容易扩展自己的存储能力和计算能力，可达到 Hadoop 所能达到的横向扩展能力，数千台服务器的集群已不难做到，是为海量数据做数据挖掘而设计，不过实时性比较差。

Hive 的内部表和外部表：

- **内部表：**实际上是将 hdfs 的文件映射成 table，然后 Hive 的数据仓库会生成对应的目录，EMR 中默认的仓库路径为 `usr/hive/warehouse/$tablename`，**这个路径在 hdfs 上面**，其中 `$tablename` 是您创建的表名。这时只要将符合 table 定义的文件加载到此目录中，即可通过 Hql（Hive-SQL）对整个目录的文件进行查询。
- **外部表：**Hive 中的外部表和表很类似，但是其数据不是放在自己表所属的目录中，而是存放到其他地方。这样的好处是如果您要删除这个外部表，此外部表所指向的数据是不会被删除的，它只会删除外部表对应的元数据。而如果您要删除内部表，该表对应的所有数据包括元数据都会被删除。

Hive 基础操作演示了如何在 EMR 集群上创建表以及通过 Hive 查询表。

1. 开发准备

- 任务中要访问腾讯云对象存储 COS，所以需要先在 COS 中创建一个存储桶（Bucket），具体可参考 创建存储桶。
- 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群时，需要在软件配置界面选择 Hive 组件，并且在基础配置页面勾选“开启 COS”，在下方填写自己的 SecretId 和 SecretKey。SecretId 和 SecretKey 可在 API 密钥管理界面 查看。如果还没有密钥，可单击【新建密钥】创建一个新的密钥。
- Hive 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 准备数据

首先登录 EMR 集群中的任意机器，推荐登录到 Master 节点。登录 EMR 的方式可参考 登录 Linux 实例，这里可选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面。用户名默认为 root，密码为创建 EMR 时用

户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行使用以下命令切换到 Hadoop 用户，并进入 Hive 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hive
```

新建一个 bash 脚本文件 gen_data.sh，在其中添加以下代码：

```
#!/bin/bash
MAXROW=1000000 #指定生成数据行数
for((i = 0; i < $MAXROW; i++))
do
echo $RANDOM, \"$RANDOM\"
done
```

并按如下方式执行：

```
[hadoop@172 hive]$ chmod +x 脚本名称
[hadoop@172 hive]$ ./gen_data.sh > hive_test.data
```

这个脚本文件会生成1000000个随机数对，并且保存到文件 hive_test.data 中。

- 使用如下命令把生成的测试数据先上传到 HDFS 中，其中 $hdfspath 为 HDFS 上的您存放文件的路径。

  ```
  [hadoop@172 hive]$ hdfs dfs -put ./hive_test.data /$hdfspath
  ```

- 也可以使用 COS 上面的数据。将数据上传到 COS 中，如果数据在本地，那么可以使用 COS 控制台来上传数据。如果数据在 EMR 集群，那么使用如下命令来上传数据，其中 $bucketname 为您创建的 COS 桶名。

  ```
  [hadoop@172 hive]$ hdfs dfs -put ./hive_test.data cosn://$bucketname/
  ```

3. Hive 基础操作

连接 Hive

登录 EMR 集群的 Master 节点，切换到 Hadoop 用户并且进入 Hive 目录，并连接 Hive：

```
[hadoop@172 hive]$ su hadoop
[hadoop@172 hive]$ cd /usr/local/service/hive/bin
[hadoop@172 bin]$ hive
```

用户也可以使用 `-h` 参数来获取 Hive 指令的基本信息。也可以使用 beeline 模式连接数据库，同样需要登录 EMR 的Master 节点，切换到 Hadoop 用户并且进入 Hive 目录，在 `conf/hive-site.xml` 配置文件中，获得 hive

server2 的连接端口 $port 和 host 地址 $host：

```
<property>
<name>hive.server2.thrift.bind.host</name>
<value>$host</value>
</property>
<property>
<name>hive.server2.thrift.port</name>
<value>$port</value>
</property>
```

在 bin 目录下，执行下列语句连接 Hive：

```
[hadoop@172 hive]$ cd bin
[hadoop@172 bin]$ ./beeline -u "jdbc:hive2:// $host: $port " -n hadoop -p hadoop
```

创建 Hive 表

无论以 Hive 模式还是 beeline 模式成功连接到 Hive 数据库后，Hive-SQL 的执行语句都是一样的，现在以 Hive 模式执行 Hive-SQL。在 Hive 下执行如下命令查看数据库：

```
hive> show databases;
OK
default
Time taken: 0.26 seconds, Fetched: 1 row(s)
```

使用 `create` 指令创建一个数据库：

```
hive> create database test; #创建数据库 test
OK
Time taken: 0.176 seconds
```

使用 `use` 指令转到刚刚创建的 test 数据库下：

```
hive> use test;
OK
Time taken: 0.176 seconds
```

使用 `create` 指令在 test 数据库下创建一个新的名为 hive_test 的内部表：

```
hive> create table hive_test (a int, b string)
hive> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
#创建数据表 hive_test, 并指定列分割符为','
```

```
OK
Time taken: 0.204 seconds
```

这里只有一条指令，如果不输入分号";"，Hive-SQL 可以把一条指令放在多行输入。最后可用以下指令查看表是否创建成功：

```
hive> show tables;
OK
hive_test
Time taken: 0.176 seconds, Fetched: 1 row(s)
```

将数据导入表中

对于存放在 HDFS 中的数据，使用如下指令来将其导入表中：

```
hive> load data inpath "/$hdfspath/hive_test.data" into table hive_test;
```

其中 $hdfspath 为 HDFS 上的您存放文件的路径。导入完成后，HDFS 上导入路径上的源数据文件将会被删除。

对于存放在 COS 中的数据，使用如下指令来将其导入表中：

```
hive> load data inpath "cosn://$bucketname/hive_test.data" into table hive_test;
```

其中 $bucketname 为您的 COS 桶名加桶内存放数据的路径。

同样在导入完成后，COS 导入路径上的源数据文件将会被删除。也可以将存放在 EMR 集群本地的数据导入到 Hive 中，使用如下指令：

```
hive>load data local inpath "/$localpath/hive_test.data" into table hive_test;
```

其中 $localpath 为您的 EMR 集群本地存放数据的路径。导入完成后，源数据会被删除。

执行查询

使用 `select` 指令来执行查询操作，统计表中一共有多少行数据：

```
hive> select count(*) from hive_test;
Query ID = hadoop_20170316142922_967b5f0e-1f89-4464-bfa3-b6ed53273fc2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
set hive.exec.reducers.bytes.per.reducer=
In order to limit the maximum number of reducers:
set hive.exec.reducers.max=
```

```
In order to set a constant number of reducers:
set mapreduce.job.reduces=
Starting Job = job_1489458311206_9869, Tracking URL =
http://10.0.1.125:5004/proxy/application_1489458311206_9869/
Kill Command = /usr/local/service/hadoop/bin/hadoop job -kill job_1489458311206_9
869
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-03-16 14:29:29,023 Stage-1 map = 0%, reduce = 0%
2017-03-16 14:29:34,208 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.87 sec
2017-03-16 14:29:40,404 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.79 se
c
MapReduce Total cumulative CPU time: 5 seconds 790 msec
Ended Job = job_1489458311206_9869
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.79 sec
HDFS Read: 40974623 HDFS Write: 107 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 790 msec
OK
1000000
Time taken: 18.504 seconds, Fetched: 1 row(s)
```

最后输出结果为1000000。

使用 `select` 指令来查询表中的前10个元素：

```
hive> select * from hive_test limit 10;
OK
30847 "31583"
14887 "32053"
19741 "16590"
8104 "20321"
29030 "32724"
27274 "5231"
10028 "22594"
924 "32569"
10603 "27927"
4018 "30518"
Time taken: 2.133 seconds, Fetched: 10 row(s)
```

删除 Hive 表

使用 `drop` 指令来删除 Hive 表：

```
hive> drop table hive_test;
Moved: 'hdfs://HDFS/usr/hive/warehouse/hive_test' to trash at: hdfs://HDFS/user/h
adoop/.Trash/Current
```

```
OK
Time taken: 2.327 seconds
```

更多关于 Hive 的操作，详见 官方文档。

通过 Java 连接 Hive

最近更新时间：2021-06-30 15:27:55

Hive 中集成了 Thrift 服务。Thrift 是 Facebook 开发的一个软件框架，它用来进行可扩展且跨语言的服务的开发。Hive 的 HiveServer2 就是基于 Thrift 的，所以能让不同的语言如 Java、Python 来调用 Hive 的接口。对于 Java，Hive 提供了 jdbc 驱动，用户可以使用 Java 代码来连接 Hive 并进行一系列操作。

本节将演示如何使用 Java 代码来连接 HiveServer2。

1. 开发准备

- 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Hive 组件。
- Hive 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 使用 Maven 来创建您的工程

查看参数

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器机右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 安装文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/hive/
[hadoop@172 hive]$
```

查看在程序中需要使用的参数：

```
[hadoop@172 hive]$ vim conf/hive-site.xml
<property>
<name>hive.server2.thrift.bind.host</name>
<value>$hs2host</value>
</property>
<property>
<name>hive.server2.thrift.port</name>
<value>$hs2port</value>
</property>
```

其中 $hs2host 为您的 HiveServer2 的 hostID，$hs2port 为您的 HiveServer2 的端口号。

新建一个 Maven 工程

推荐使用 Maven 来管理您的工程。Maven 是一个项目管理工具，能够帮助您方便的管理项目的依赖信息，即它可以通过 pom.xml 文件的配置获取 jar 包，而不用去手动添加。

首先在本地下载并安装 Maven，配置好 Maven 的环境变量，如果您使用 IDE，请在 IDE 中设置好 Maven 相关配置。

在本地 shell 下进入要新建工程的目录，例如 `D://mavenWorkplace` 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID -Darch
etypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名；$yourartifactID 为您的项目名称；maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。

创建成功后，在 `D://mavenWorkplace` 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
---pom.xml          核心配置，项目根下
---src
---main
---java          Java 源码目录
---resources     Java 配置文件目录
---test
---java          测试源码目录
---resources     测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。

首先在 pom.xml 中添加 Maven 依赖：

```
<dependencies>
<dependency>
<groupId>org.apache.hive</groupId>
<artifactId>hive-jdbc</artifactId>
<version>2.1.1</version>
</dependency>
<dependency>
<groupId>org.apache.hadoop</groupId>
<artifactId>hadoop-common</artifactId>
<version>2.7.3</version>
```

```
</dependency>
</dependencies>
```

继续在 pom.xml 中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

在 src>mai>Java 下右键新建一个 Java Class，输入您的 Class 名，这里使用 HiveTest.java，在 Class 添加样例代码：

```
import java.sql.*;
/**
* Created by tencent on 2018/7/6.
*/
public class HiveTest {
private static String driverName =
"org.apache.hive.jdbc.HiveDriver";
```

```
public static void main(String[] args)
throws SQLException {
try {
Class.forName(driverName);
} catch (ClassNotFoundException e) {
e.printStackTrace();
System.exit(1);
}
Connection con = DriverManager.getConnection(
"jdbc:hive2://$hs2host:$hs2port/default", "hadoop", "");
Statement stmt = con.createStatement();
String tableName = "HiveTestByJava";
stmt.execute("drop table if exists " + tableName);
stmt.execute("create table " + tableName +
" (key int, value string)");
System.out.println("Create table success!");
// show tables
String sql = "show tables '" + tableName + "'";
System.out.println("Running: " + sql);
ResultSet res = stmt.executeQuery(sql);
if (res.next()) {
System.out.println(res.getString(1));
}
// describe table
sql = "describe " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(res.getString(1) + "\t" + res.getString(2));
}
sql = "insert into " + tableName + " values (42,\"hello\"),(48,\"world\")";
stmt.execute(sql);
sql = "select * from " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(String.valueOf(res.getInt(1)) + "\t"
+ res.getString(2));
}
sql = "select count(1) from " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(res.getString(1));
}
}
}
```

> 注意：
> 将程序中的参数 $hs2host 和 $hs2port 分别修改为您查到的 HiveServer2 的 hostID 和端口号的值。

整个程序会先连接 HiveServer2 服务，然后在 default 数据库中建立一个名为 HiveTestByJave 的表。之后在该表中插入两个元素，并输出整个表的内容。

如果您的 Maven 配置正确并且成功的导入了依赖包，那么整个工程即可直接编译。在本地 shell 下进入工程目录，执行下面的命令对整个工程进行打包：

```
mvn package
```

运行过程中可能还需要下载一些文件，直到出现 build success 表示打包成功。然后您可以在工程目录下的 target 文件夹中看到打好的 jar 包。

3. 上传并运行程序

首先需要把压缩好的 jar 包上传到 EMR 集群中，使用 scp 或者 sftp 工具来进行上传。在本地 shell 下运行：

```
scp $localfile root@公网IP地址:/usr/local/service/hive
```

其中，$localfile 是您的本地文件的路径加名称，root 为 CVM 服务器用户名，公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看。将打好的 jar 包上传到 EMR 集群的 `/usr/local/service/hive` 目录下。上传完成后，在 EMR 命令行中即可查看对应文件夹下是否有相应文件。**一定要上传具有依赖的 jar 包。**

登录 EMR 集群切换到 Hadoop 用户并且进入目录 `/usr/local/service/hive` 。接下来可以执行程序：

```
[hadoop@172 hive]$ yarn jar $package.jar HiveTest
```

其中 $package.jar 为您的 jar 包的路径 + 名字，HiveTest 为之前的 Java Class 的名字。运行结果如下：

```
Create table success!
Running: show tables 'HiveTestByJava'
hivetestbyjava
Running: describe HiveTestByJava
key int
value string
Running: select * from HiveTestByJava
42 hello
48 world
```

```
Running: select count(1) from HiveTestByJava
2
```

通过 Python 连接 Hive

最近更新时间：2021-04-28 16:18:07

Hive 中集成了 Thrift 服务。Thrift 是 Facebook 开发的一个软件框架，它用来进行可扩展且跨语言的服务的开发。Hive 的 HiveServer2 就是基于 Thrift 的，所以能让不同的语言如 Java、Python 来调用 Hive 的接口。

本节将演示如何使用 Python 代码来连接 HiveServer2。

1. 开发准备

- 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Hive 组件。
- Hive 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 查看参数

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例，这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 安装文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/hive/
[hadoop@172 hive]$
```

查看在程序中需要使用的参数：

```
[hadoop@172 hive]$ vim conf/hive-site.xml

<property>
<name>hive.server2.thrift.bind.host</name>
<value>$hs2host</value>
</property>
<property>
<name>hive.server2.thrift.port</name>
<value>$hs2port</value>
</property>
```

其中 $hs2host 为您的 HiveServer2 的 hostID，$hs2port 为您的 HiveServer2 的端口号。

3. 使用 Python 进行 Hive 操作

使用 Python 程序操作 Hive 需要安装 pip：

```
[hadoop@172 hive]$ su
[root@172 hive]# pip install pyhs2
```

安装完成后切换回 Hadoop 用户。然后在 `/usr/local/service/hive/` 目录下新建一个 Python 文件 hivetest.py，并且添加以下代码：

```
#!/usr/bin/env python

import pyhs2
import sys

default_encoding = 'utf-8'

conn = pyhs2.connect(host='$hs2host',
port=$hs2port,
authMechanism='PLAIN',
user='hadoop',
password='',
database='default',)


tablename = 'HiveByPython'
cur = conn.cursor()
print 'show the databases: '
print cur.getDatabases()

print "\n"
print 'show the tables in default: '
cur.execute('show tables')
for i in cur.fetch():
print i

cur.execute('drop table if exists ' + tablename)
cur.execute('create table ' + tablename + ' (key int,value string)')

print "\n"
print 'show the new table: '
cur.execute('show tables ' +"'" +tablename+"'")
for i in cur.fetch():
print i
```

```
print "\n"
print "contents from " + tablename + ":";
cur.execute('insert into ' + tablename + ' values (42,"hello"),(48,"world")')
cur.execute('select * from ' + tablename)
for i in cur.fetch():
print i
```

⚠ 注意：

将程序中的参数 $hs2host 和 $hs2port 分别修改为您查到的 HiveServer2 的 hostID 和端口号的值。

该程序连接 HiveServer2 后，首先输出所有的数据库，然后显示“default”数据库中的表。创建一个名为“hivebypython”的表，在表中插入两个数据并输出。运行该程序：

```
[hadoop@172 hive]$ python hivetest.py
show the databases:
[['default', ''], ['hue_test', ''], ['test', '']]


show the tables in default:
['dd']
['ext_table']
['hive_test']
['hivebypython']


show the new table:
['hivebypython']


contents from HiveByPython:
[42, 'hello']
[48, 'world']
```

如何映射 Hbase表

最近更新时间：2021-07-09 10:45:12

使用 Hive 来映射 Hbase 表，可以使用 Hive 来读取 Hbase 上的数据，使用 Hive-SQL 语句在 Hbase 表上进行查询、插入等操作。

开发准备

- 确认已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群时需要在软件配置界面选择 Hive、Hbase 组件。
- Hive 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

创建一个 Hbase 表

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，进入 Hbase 文件夹并进入 Hbase shell：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hbase
[hadoop@10hbase]$ bin/hbase shell
```

在 Hbase 中建立一个新表，如下所示：

```
hbase(main):001:0> create 'test', 'cf'
hbase(main):003:0> put 'test', 'row1', 'cf:a', 'value1'
hbase(main):004:0> put 'test', 'row1', 'cf:b', 'value2'
hbase(main):005:0> put 'test', 'row1', 'cf:c', 'value3'
```

更多在 Hbase 中的操作详见 Hbase 操作指南，或者查看 官方文档。

创建完成后，可使用 `list` 和 `scan` 操作来查看新建的表。

```
hbase(main):001:0> list 'test'
TABLE
test
1 row(s) in 0.0030 seconds
```

```
=> ["test"]
hbase(main):002:0> scan 'test'
ROW COLUMN+CELL
row1 column=cf:a, timestamp=1530276759697, value=value1
row2 column=cf:b, timestamp=1530276777806, value=value2
row3 column=cf:c, timestamp=1530276792839, value=value3
3 row(s) in 0.2110 seconds
```

映射 Hive 表

切换到 Hive 文件夹下，并且连接到 Hive 上：

```
[hadoop@172 hive]$ cd /usr/local/service/hive/
[hadoop@172 hive]$ bin/hive
```

接下来创建一个 Hive 外部表让它映射到第二步中创建的 Hbase 表上：

```
hive> CREATE EXTERNAL TABLE hive_test (
> rowkey string,
> a string,
> b string,
> c string
> ) STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' WITH
> SERDEPROPERTIES("hbase.columns.mapping" = ":key,cf:a,cf:b,cf:c")
> TBLPROPERTIES("hbase.table.name" = "test");
OK
Time taken: 2.086 seconds
```

这样就建立了一个 Hive 表到 Hbase 表的映射。可使用以下指令来查看 Hive 表中的元素：

```
hive> select * from hive_test;
OK
row1 value1 value2 value3
Time taken: 0.305 seconds, Fetched: 1 row(s)
```

Hive 最佳实践

最近更新时间：2021-07-09 10:42:51

执行引擎设置

腾讯云 EMR 中的 Hive 目前支持三种执行引擎 MR、TEZ 和 Spark。如果需要 TEZ 那么在初始购买集群的时候需要勾选 TEZ，在普通情况下建议执行引擎为 TEZ，这样您会获得更好的计算效率。

存储选择

腾讯云存储介质目前支持本地数据盘、普通云硬盘、SSD 云硬盘以及 COS 对象存储，如果您对成本敏感，那么基于 COS 的数据仓库方式是一个不错的选择。

数据格式

腾讯云压缩支持 snappy、lzo 等压缩算法，如果使用 Hive 建议您的数据文件格式使用 ORC 或者 parquet 的格式，这样您会更节省空间以及会获得更好的计算效率。

查询引擎如何选择

腾讯云 EMR 目前支持的查询引擎有 Presto、SparkSQL、Hive，如果您想实现多种数据源耦合查询建议您使用 Presto，如果普通数据仓库建议您使用 Hive+TEZ 的模式，如果您对时延比较敏感可以考虑 SparkSQL。

数据安全

如果您是使用 COS 作为底层存储，建议您使用外部表的方式以免误删数据；如果是存储在 HDFS 那么建议您开启 HDFS 回收站来避免数据误删除。

基于对象存储 COS 的数据仓库

最近更新时间：2021-08-12 10:25:16

基于对象存储 COS 的数据仓库有两种方式：

方式一：将整个数据库建立在 COS 上

整个数据库建立在 COS 上可通过如下语句实现。

```
create database hivewithcos location 'cosn://huadong/hive';
```

其中 huadong 是 bucket 名称，而 hive 是路径，可根据您的实际需要进行设置。库创建完成后，创建一张数据表。然后向表中 load 数据，其使用方式和 HDFS 相同。

```
create table record(id int, name string) row format delimited fields terminated by ',' stored as textfile;
```

方式二：将指定表放在 COS 上

首先需要在 Hive 中选择一个数据库或者创建一个数据库，然后通过如下语句实现单表存储在 COS 上。

```
create table record(id int, name string) row format delimited fields terminated by ',' stored as textfile location 'cosn://huadong/hive/cos';
```

然后向表中 load 数据，单个表在 COS 上即设定表的存储位置在对象存储中。

Hive 加载 json 数据实践

最近更新时间：2020-12-24 10:31:39

1. 连接 Hive

登录 EMR 集群的 Master 节点，切换到 hadoop 用户并且进入 hive 目录：

```
[root@10 ~]# su hadoop
[hadoop@10 root]$ cd /usr/local/service/hive
```

2. 准备数据

创建数据文件（JSON 格式），编译以下内容并保存：

```
vim test.data
{"name":"Mary","age":12,"course":[{"name":"math","location":"b208"},{"name":"engl
ish","location":"b702"}],"grade":[99,98,95]}
{"name":"Bob","age":20,"course":[{"name":"music","location":"b108"},{"name":"hist
ory","location":"b711"}],"grade":[91,92,93]}
```

将数据文件存储在 hdfs 上：

```
hadoop fs -put ./test.data /
```

3. 创建表格

连接 Hive：

```
[hadoop@10 hive]$ hive
```

根据映射关系创建表格：

```
hive> CREATE TABLE test (name string, age int, course array<map<string,string>>,
grade array<int>) ROW FORMAT SERDE 'org.apache.hive.hcatalog.data.JsonSerDe' STOR
ED AS TEXTFILE;
```

4. 导入数据

```
hive>LOAD DATA INPATH '/test.data' into table test;
```

5. 检查数据是否导入成功

查询所有数据：

```
hive> select * from test;
OK
Mary 12 [{"name":"math","location":"b208"},{"name":"english","location":"b702"}]
[99,98,95]
Bob 20 [{"name":"music","location":"b108"},{"name":"history","location":"b711"}]
[91,92,93]
Time taken: 0.153 seconds, Fetched: 2 row(s)
```

查询每条记录的第一个得分：

```
hive> select grade[0] from test;
OK
99
91
Time taken: 0.374 seconds, Fetched: 2 row(s)
```

查询每条记录的第一个课程的名称和地点：

```
hive> select course[0]['name'], course[0]['location'] from test;
OK
math b208
music b108
Time taken: 0.162 seconds, Fetched: 2 row(s)
```

Hive 元数据管理

最近更新时间：2023-07-14 10:57:58

当选择部署 Hive 组件时，Hive 元数据库提供了两种存储方式：第一种集群默认，Hive 元数据存储于集群独立购买 MetaDB；第二种是关联外部 Hive 元数据库，可选择关联 EMR-MetaDB 或自建 MySQL 数据库，元数据将存储于关联的数据库中，不随集群销毁而销毁。

集群默认是独立自动购买一个 MetaDB 云数据库实例存储单元作为元数据存储地，与其余组件元数据一起存储，并随集群销毁而销毁 MetaDB 云数据库，若需保存元数据，需提前在云数据库中手动保存元数据。

注意：

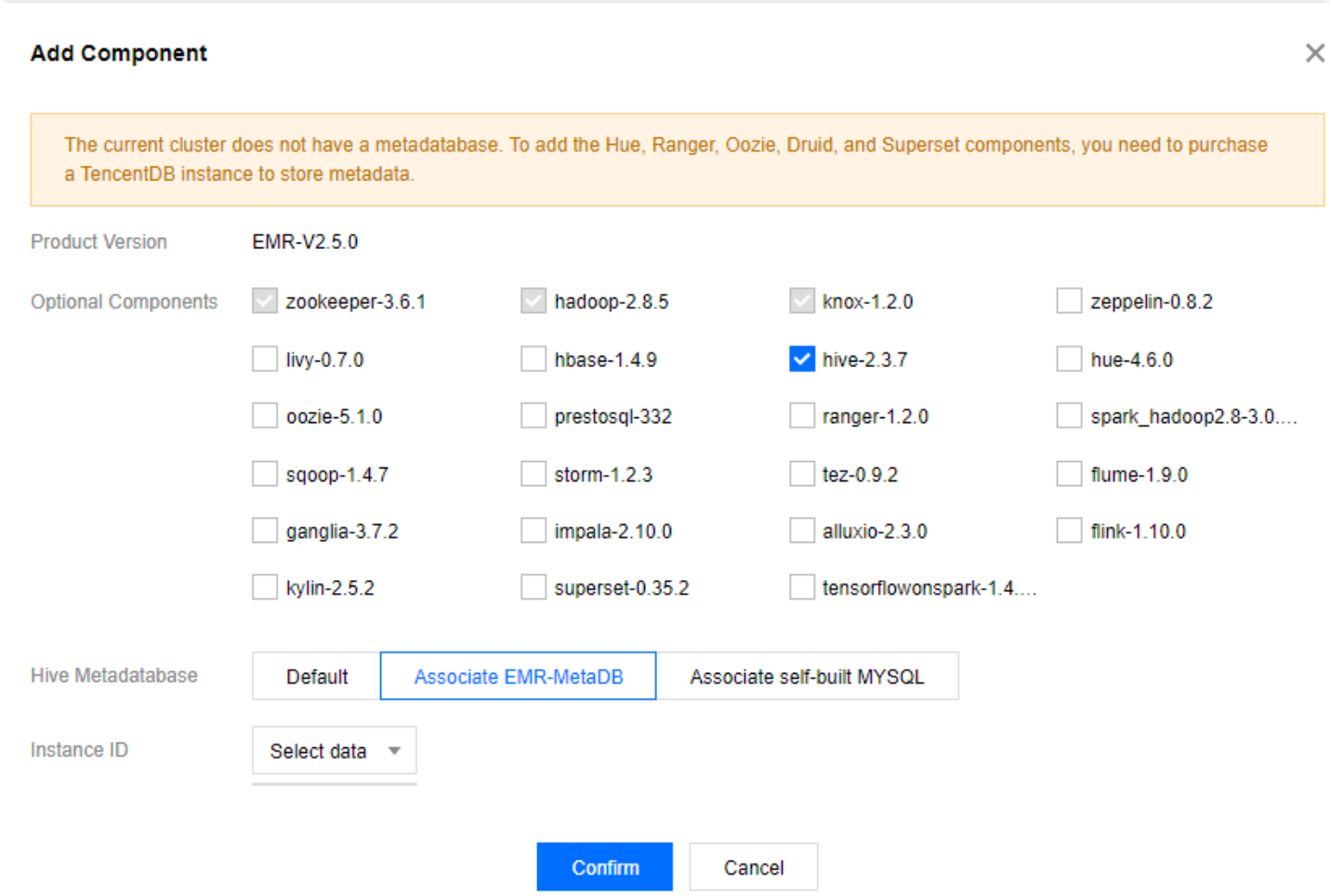
1. Hive 元数据与 Druid、Superset、Hue、Ranger、Oozie、Presto 组件元数据一起存储。
2. 集群需要单独购买一个 MetaDB 作为元数据存储单元。
3. MetaDB 随集群销毁而销毁，即元数据随集群而销毁。

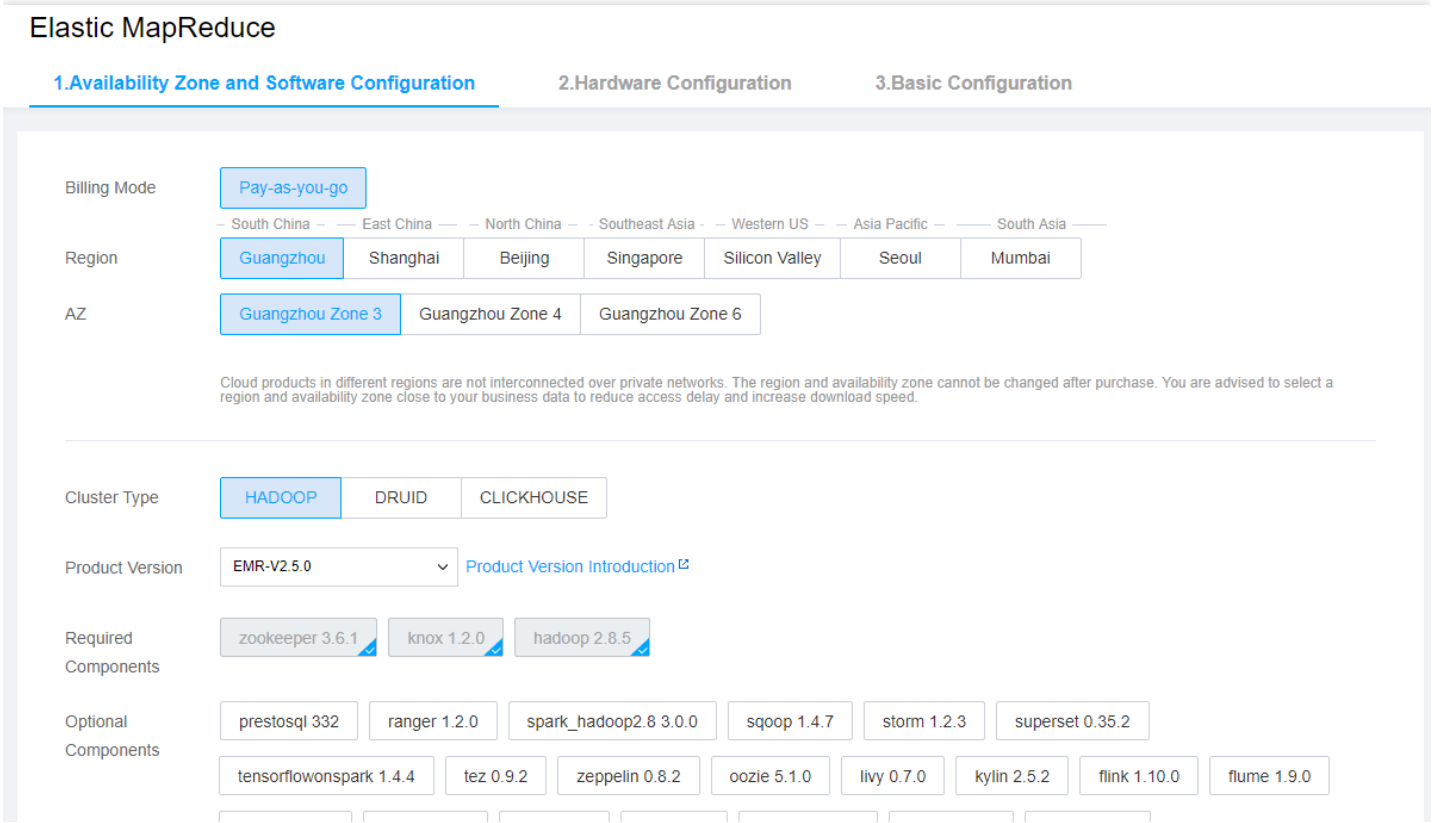
关联 EMR-MetaDB 共享 Hive 元数据

集群创建时系统会拉取云上可用的 MetaDB，用于新集群 Hive 组件存储元数据，无需单独购买 MetaDB 存储 Hive 元数据节约成本；并且 Hive 元数据不会随当前集群的销毁而销毁。

注意：

1. 可用 MetaDB 实例 ID 为同一账号下 EMR 集群中已有的 MetaDB。
2. 当选择 Hue、Ranger、Oozie、Druid、Superset 一个或多个组件时系统会自动购买一个 MetaDB 用于除 Hive 外的组件元数据存储。
3. 要销毁关联的 EMR-MetaDB 需前往云数据库销毁，销毁后 Hive 元数据库将无法恢复。
4. 需保持关联的 EMR-MetaDB 网络与当前新建集群在同一网络环境下。


Add Component
The current cluster does not have a metadatabase. To add the Hue, Ranger, Oozie, Druid, and Superset components, you need to purchase a TencentDB instance to store metadata.
Product Version
EMR-V2.5.0
Optional Components
zookeeper-3.6.1
hadoop-2.8.5
knox-1.2.0
zeppelin-0.8.2
livy-0.7.0
hbase-1.4.9
hive-2.3.7
hue-4.6.0
oozie-5.1.0
prestosql-332
ranger-1.2.0
spark_hadoop2.8-3.0....
sqoop-1.4.7
storm-1.2.3
tez-0.9.2
flume-1.9.0
ganglia-3.7.2
impala-2.10.0
alluxio-2.3.0
flink-1.10.0
kylin-2.5.2
superset-0.35.2
tensorflowonspark-1.4....
Hive Metadatabase
Default
Associate EMR-MetaDB
Associate self-built MYSQL
Instance ID
Select data
Confirm
Cancel

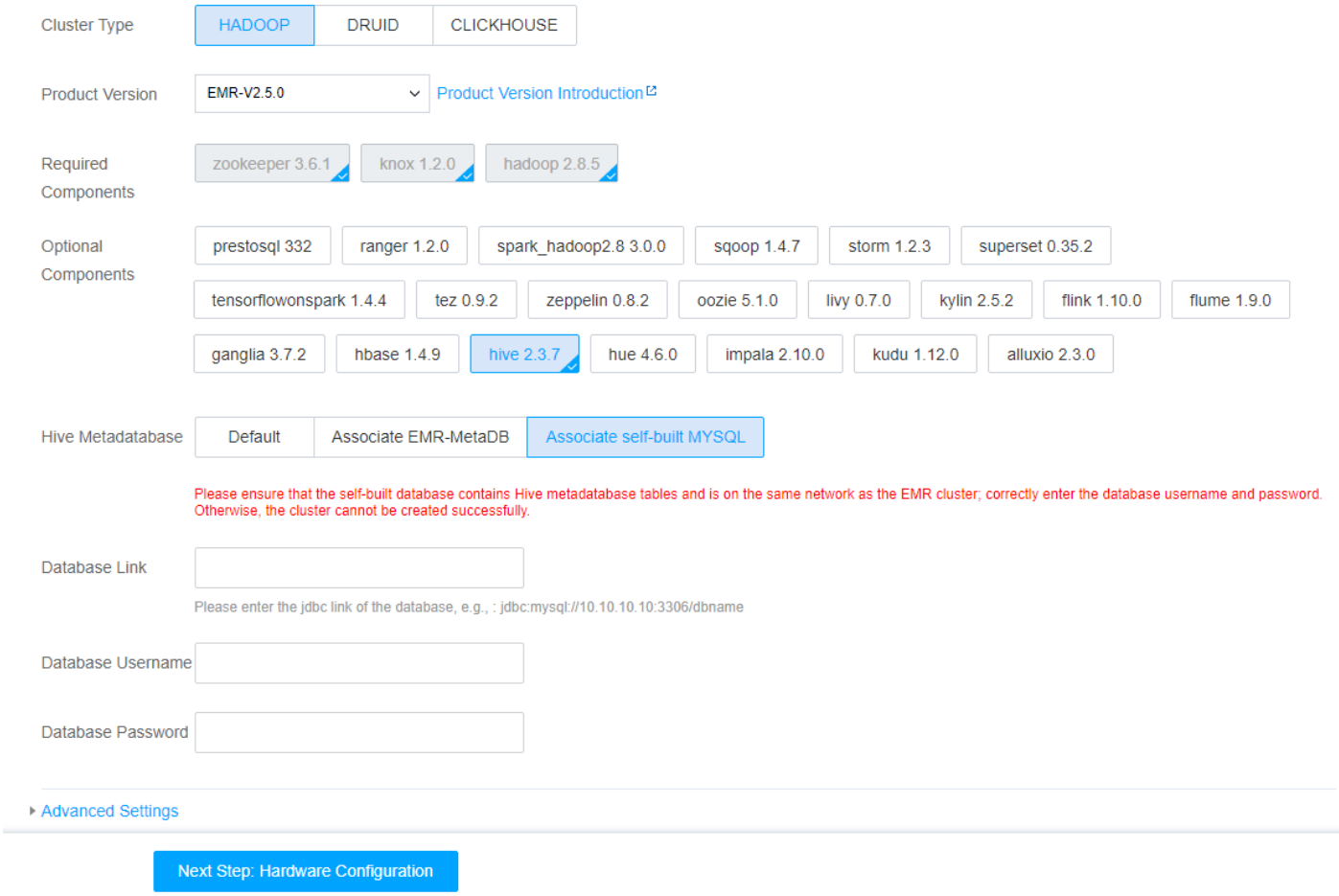

Elastic MapReduce
1.Availability Zone and Software Configuration
2.Hardware Configuration
3.Basic Configuration
Billing Mode
Pay-as-you-go
South China
East China
North China
Southeast Asia
Western US
Asia Pacific
South Asia
Region
Guangzhou
Shanghai
Beijing
Singapore
Silicon Valley
Seoul
Mumbai
AZ
Guangzhou Zone 3
Guangzhou Zone 4
Guangzhou Zone 6
Cloud products in different regions are not interconnected over private networks. The region and availability zone cannot be changed after purchase. You are advised to select a region and availability zone close to your business data to reduce access delay and increase download speed.
Cluster Type
HADOOP
DRUID
CLICKHOUSE
Product Version
EMR-V2.5.0
Product Version Introduction
Required Components
zookeeper 3.6.1
knox 1.2.0
hadoop 2.8.5
Optional Components
prestosql 332
ranger 1.2.0
spark_hadoop2.8 3.0.0
sqoop 1.4.7
storm 1.2.3
superset 0.35.2
tensorflowonspark 1.4.4
tez 0.9.2
zeppelin 0.8.2
oozie 5.1.0
livy 0.7.0
kylin 2.5.2
flink 1.10.0
flume 1.9.0

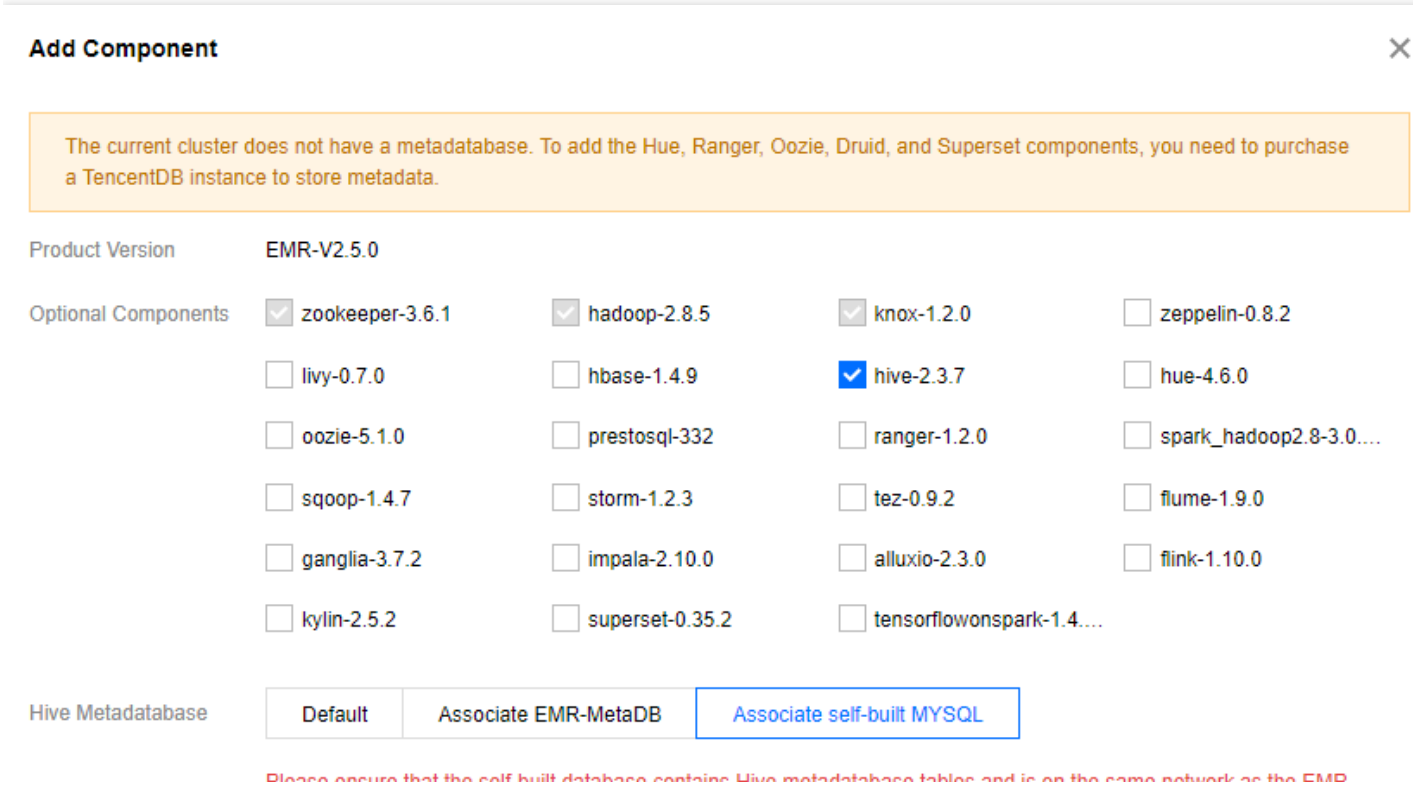
关联自建 MySQL 共享 Hive 元数据

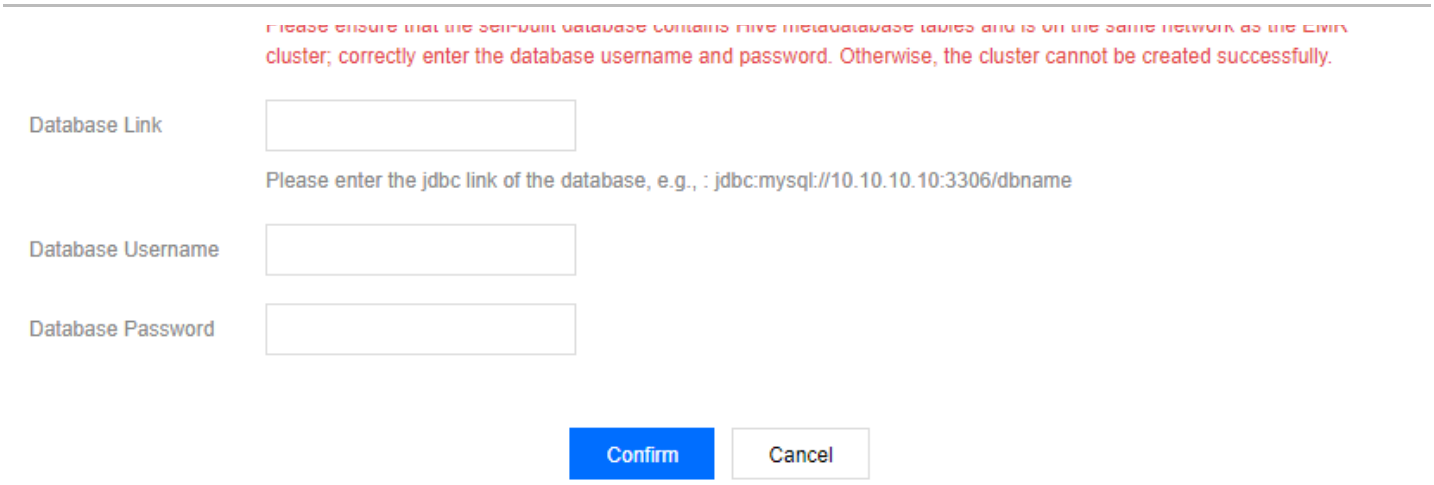
关联自己本地自建 MySQL 数据库作为 Hive 元数据存储，也无需单独购买 MetaDB 存储 Hive 元数据节约成本，需准确填写输入以“jdbc:mysql://开头”的本地址、数据库名字、数据库登录密码，并确保网络与当前集群网络打通。

注意：

1. 请确保自建数据库与 EMR 集群在同一网络下。
2. 准确填写数据库用户名和数据库密码。
3. 当选择 Hue、Ranger、Oozie、Druid、Superset 一个或多个组件时系统会自动购买一个 MetaDB 用于除 Hive 外的元数据存储。
4. 需保证自定义数据库中的 Hive 元数据版本大于等于新集群中的 Hive 版本。


Cluster Type
HADOOP
DRUID
CLICKHOUSE
Product Version
EMR-V2.5.0
Product Version Introduction
Required Components
zookeeper 3.6.1
knox 1.2.0
hadoop 2.8.5
Optional Components
prestosql 332
ranger 1.2.0
spark_hadoop2.8 3.0.0
sqoop 1.4.7
storm 1.2.3
superset 0.35.2
tensorflowonspark 1.4.4
tez 0.9.2
zeppelin 0.8.2
oozie 5.1.0
livy 0.7.0
kylin 2.5.2
flink 1.10.0
flume 1.9.0
ganglia 3.7.2
hbase 1.4.9
hive 2.3.7
hue 4.6.0
impala 2.10.0
kudu 1.12.0
alluxio 2.3.0
Hive Metadatabase
Default
Associate EMR-MetaDB
Associate self-built MYSQL
Please ensure that the self-built database contains Hive metadatabase tables and is on the same network as the EMR cluster; correctly enter the database username and password. Otherwise, the cluster cannot be created successfully.
Database Link
Please enter the jdbc link of the database, e.g., : jdbc:mysql://10.10.10.10:3306/dbname
Database Username
Database Password
Advanced Settings
Next Step: Hardware Configuration

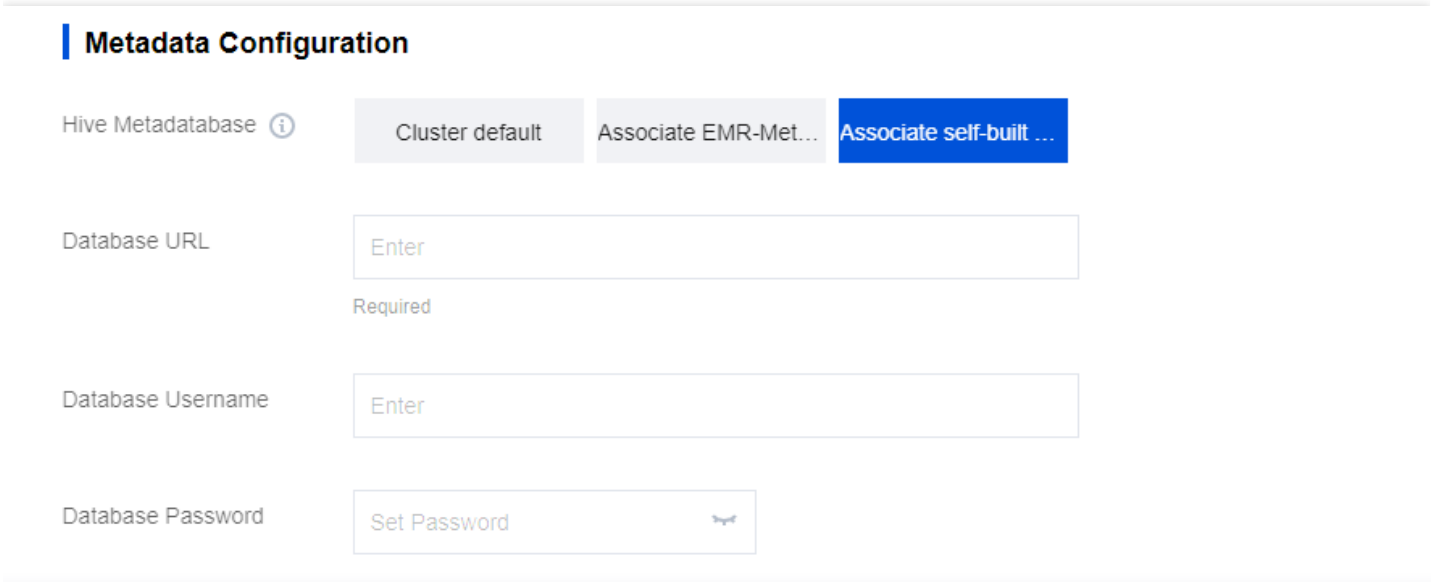

Add Component
The current cluster does not have a metadatabase. To add the Hue, Ranger, Oozie, Druid, and Superset components, you need to purchase a TencentDB instance to store metadata.
Product Version
EMR-V2.5.0
Optional Components
zookeeper-3.6.1
hadoop-2.8.5
knox-1.2.0
zeppelin-0.8.2
livy-0.7.0
hbase-1.4.9
hive-2.3.7
hue-4.6.0
oozie-5.1.0
prestosql-332
ranger-1.2.0
spark_hadoop2.8-3.0....
sqoop-1.4.7
storm-1.2.3
tez-0.9.2
flume-1.9.0
ganglia-3.7.2
impala-2.10.0
alluxio-2.3.0
flink-1.10.0
kylin-2.5.2
superset-0.35.2
tensorflowonspark-1.4....
Hive Metadatabase
Default
Associate EMR-MetaDB
Associate self-built MYSQL



HIVE 关联自建元数据异常修复方法

由于在创建 EMR 集群时选用了关联自建 MySQL，且自建 MySQL 无 HIVE 的元数据，这会导致 HIVE 进程异常。

问题复现



解决方案

对于无数据的 hive 元数据操作步骤如下：

> 说明：
> 操作时替换成用户实际的 ${ip}、${port}、${database}。

1. 控制台将 HIVE 的 hs2 和 metastore 停掉。

2. hive 组件修改 hive-site.xml proto-hive-site.xml 下发。
   配置项：javax.jdo.option.ConnectionURL

```
jdbc:mysql://${ip}:${port}/${database}?useSSL=false&createDatabaseIfNotExist=tr
ue&characterEncodin
g=UTF-8
```

3. CDB 数据库里删除库操作：

```
drop database ${database};
```

4. hadoop 用户执行如下命令：

```
/usr/local/service/hive/bin/schematool -dbType mysql -initSchema
```

5. 控制台 HIVE 启动 hs2和 metastore。

6. 访问 hive 是否异常。
   如果有字符异常，CDB 再执行如下命令：

```
alter database ${database} character set latin1；
flush privileges；
```

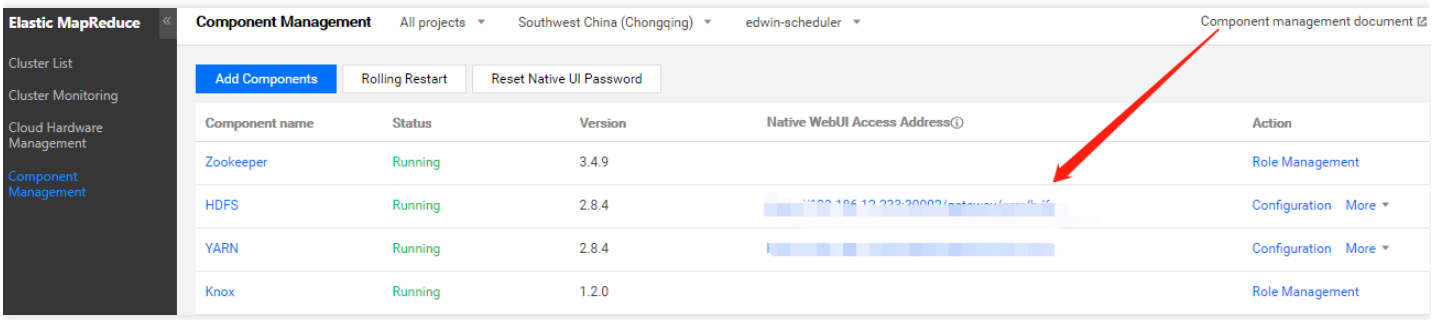
Presto开发指南

Presto 服务 UI

最近更新时间：2021-07-09 10:47:08

Presto 是一个开源的分布式 SQL 查询引擎，适用于交互式分析查询，数据量支持 GB 到 PB 字节。Presto 的设计和编写是为了解决大规模的甚至超大规模商业数据仓库的交互式分析和处理速度的问题。
Presto 支持在线数据查询，包括 Hive、Cassandra、关系数据库以及专有数据存储。一条 Presto 查询可以将这些多个数据源的数据进行合并，可以跨越整个组织进行分析。

EMR 代理了 Presto 原生 Web UI，可以直接在 EMR 控制台查看。登录 EMR 控制台，单击集群实例 ID，进入实例管理页面。单击左侧菜单栏【集群服务】，即可看到【WebUI地址】页面快捷入口，单击 Presto 的入口即可。登录用户名为 root，密码为创建集群时设置的密码。如下图：

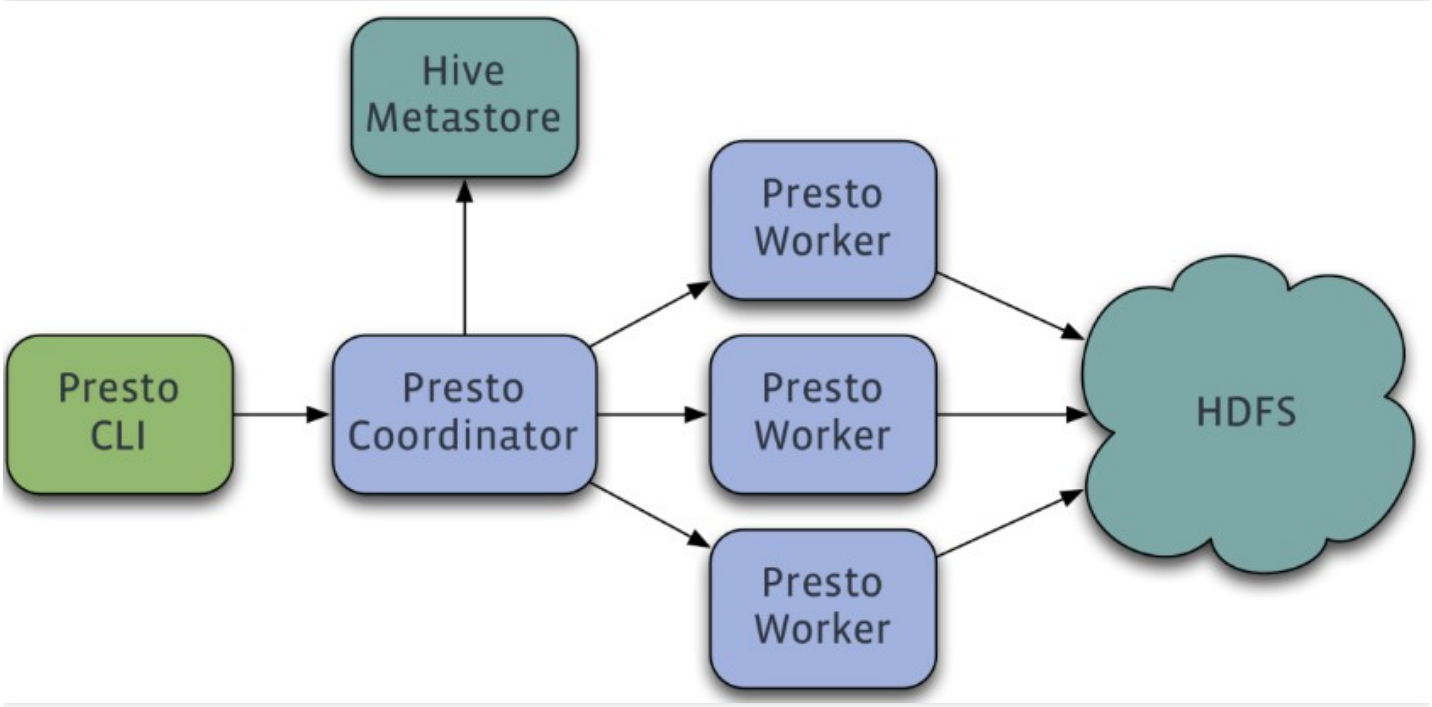


访问地址需要进行身份验证，用户名为 root，默认密码为创建集群时输入的密码，如需修改密码，可在该页面中单击【重置原生UI密码】进行修改。

连接器

最近更新时间：2021-07-09 11:08:46

Presto 是由Facebook 开发的一个分布式 SQL 查询引擎，被设计为用来专门进行高速、实时的数据分析，适用于交互式分析查询，数据量支持 GB 到 PB 字节。支持标准的 ANSI SQL，包括复杂查询、聚合（aggregation）、连接（join）和窗口函数（window functions）。采用 Java 实现。Presto 的数据源包括 Hive、HBase、关系数据库，甚至专有数据存储。其架构图如下所示：



Presto 是一个运行在多台服务器上的分布式系统，采用了主从（Master-Slave）架构，包括一个主节点 Coordinator 和多个从节点 Worker。客户端 Presto CLI 负责提交查询到 Coordinator 节点；Coordinator 节点负责解析 SQL 语句、生成查询执行计划、管理 Worker 节点等；Worker 节点负责实际执行查询任务。

EMR 中 Presto 组件预置了 Hive、Mysql 和 Kafka 等连接器，本节将以 Hive 连接器为例说明 Presto 读取 Hive 的表信息进行查询的使用，EMR 集群机器配置了 presto-client 的相关环境变量，可直接切换 Hadoop 用户并使用 Presto 客户端工具。

1. 开发准备

- 确认您已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Presto 组件。
- Presto 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 使用连接器操作 Hive

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Presto 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/presto
```

在 `etc/config.properties` 配置文件中查看 uri 的值：

```
[hadoop@172 presto]$ vim etc/config.properties
http-server.http.port=$port
discovery.uri=http://$host:$port
```

其中 $host 为您的 host 地址；$port 为您的端口号。然后切换到 presto-client 文件夹中，并且使用 Presto 连接 Hive：

```
[hadoop@172 presto]# cd /usr/local/service/presto-client
[hadoop@172 presto-client]$ ./presto --server $host:$port --catalog hive --schema
default
```

其中 --catalog 参数表示要操纵的数据库类型，--schema 表示数据库名，这里进入的是默认的 default 数据库。更多的参数信息，可以通过命令 `presto -h` 来查看，或者查看 官方文档。

执行成功后即可进入 Presto 的界面，并且直接进入指定的数据库。可以使用 Hive-SQL 来查看 Hive 数据库中的表

```
presto:default> show tables;
Table
---------------
hive_from_cos
test
(2 rows)
Query 20180702_140619_00006_c4qzg, FINISHED, 2 nodes
Splits: 2 total, 2 done (100.00%)
0:00 [3 rows, 86B] [17 rows/s, 508B/s]
```

其中表 hive_from_cos 是在 Hive 开发指南中建立的表。
更多 Presto 操作请查看 官方文档。

分析 COS 上的数据

最近更新时间：2021-07-08 10:43:44

本节将基于腾讯云对象存储 COS 展示 Hive 连接器更多使用方法，数据来源于直接插入数据、COS 数据和 lzo 压缩数据。

1. 开发准备

- 因为任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建一个存储桶（Bucket）。
- 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Presto 组件，并且在基础配置页面开启对象存储的授权。
- Presto 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 数据准备

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hive
```

新建文件 cos.txt，并添加数据如下：

```
5,cos_patrick
6,cos_stone
```

使用 HDFS 指令把文件上传到 COS 中。其中 $bucketname 为您创建的存储桶的名字和路径。

```
[hadoop@172 hive]# hdfs dfs -put cosn://$bucketname/
```

再新建文件 lzo.txt，并添加数据如下：

```
10,lzo_pop
11,lzo_tim
```

将其压缩为 .lzo 文件：

```
[hadoop@172 hive]$ lzop -v lzo.txt
compressing hive_test.data into lzo.txt.lzo
```

注意：
压缩 lzo 文件需要先安装 lzo 和 lzop，安装执行命令：`yum -y install lzo lzop`。

3. 新建 Hive 表并使用 Presto 查询

这里使用了一个脚本文件来生产进行 Hive 数据库和表的创建。新建一个脚本文件 presto_on_cos_test.sql，并添加以下程序：

```
create database if not exists test;
use test;
create external table if not exists presto_on_cos (id int,name string) ROW FORMAT
DELIMITED FIELDS TERMINATED BY ',';
insert into presto_on_cos values (12,'hello'),(13,'world');
load data inpath "cosn://$bucketname/cos.txt" into table presto_on_cos;
load data local inpath "/$yourpath/lzo.txt.lzo" into table presto_on_cos;
```

其中 $bucketname 为您的 COS 存储桶名加路径，$yourpath 为您放置 lzo.txt.lzo 文件的路径。

脚本文件首先新建一个数据库“test”，在新建的数据库中新建一个表“presto_on_cos”。分三步进行了数据的插入操作，首先采用直接插入的方法。然后插入了 COS 中数据，最后插入 lzo 压缩包中的数据。

建议如示例一样，使用外部表进行 Hive 测试，以免删除重要数据。使用 hive-cli 执行这个脚本：

```
[hadoop@172 hive]$ hive -f "presto_on_cos_test.sql"
```

执行完成之后，就可以进入 Presto 查看表中的数据。使用上一节的方法进入 Presto，不过需要改动 schema 参数。

```
[hadoop@172 presto-client]$ ./presto --server $host:$port --catalog hive --schema
test
```

对刚刚创建的 Hive 表进行查询：

```
presto:test> select * from presto_on_cos ;
id | name
----+-------------
5 | cos_patrick
6 | cos_stone
```

```
10 | lzo_pop
11 | lzo_tim
12 | hello
13 | world
(6 rows)
Query 20180702_150000_00011_c4qzg, FINISHED, 3 nodes
Splits: 4 total, 4 done (100.00%)
0:03 [6 rows, 127B] [1 rows/s, 37B/s]
```

更多 Presto 操作请查看 官方文档。

Sqoop 开发指南
关系型数据库和 HDFS 的导入导出

最近更新时间：2021-06-30 15:37:05

Sqoop 是一款开源的工具，主要用于在 Hadoop 和传统数据库（MySQL、PostgreSQL 等）之间进行数据传递，可以将一个关系型数据库（例如 MySQL、Oracle、Postgres 等）中的数据导入到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入到关系型数据库中。Sqoop 中一大亮点就是可以通过 Hadoop 的 MapReduce 把数据从关系型数据库中导入数据到 HDFS。

本文介绍了使用腾讯云 Sqoop 服务将数据在 MySQL 和 HDFS 之间导入/导出的使用方法。

1. 开发准备

- 确认已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Sqoop 组件。
- Sqoop 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 新建一个 MySQL 表

首先要连接已经创建好的 MySQL 数据库，进入 EMR 控制台，复制目标集群的实例 ID，即集群的名字。然后进入关系型数据库控制台，使用 Ctrl+F 进行搜索，找到集群对应的 MySQL 数据库，查看该数据库的内网地址 $mysqlIP。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Sqoop 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/sqoop
```

连接 MySQL 数据库：

```
[hadoop@172 sqoop]$ mysql -h $mysqlIP -p
Enter password:
```

密码为您创建 EMR 集群的时候设置的密码。

MySQL 数据库连接后，进入 test 数据库并且新建一个表，用户也可以自己选择目标数据库：

```
mysql> use test;
Database changed
mysql> create table sqoop_test(id int not null primary key auto_increment, title
varchar(64), time timestamp, content varchar(255));
Query ok , 0 rows affected(0.00 sec)
```

该指令创建了一个 MySQL 表，它的主键为 ID，然后还有三列分别为 title、time 和 content。向该表中插入数据如下：

```
mysql> insert into sqoop_test values(null, 'first', now(), 'hdfs');
Query ok, 1 row affected(0.00 sec)
mysql> insert into sqoop_test values(null, 'second', now(), 'mr');
Query ok, 1 row affected (0.00 sec)
mysql> insert into sqoop_test values(null, 'third', now(), 'yarn');
Query ok, 1 row affected(0.00 sec)
```

使用如下指令可以查看表中的数据：

```
Mysql> select * from sqoop_test;
+----+--------+---------------------+---------+
| id | title | time | content |
+----+--------+---------------------+---------+
| 1 | first | 2018-07-03 15:29:37 | hdfs |
| 2 | second | 2018-07-03 15:30:57 | mr |
| 3 | third | 2018-07-03 15:31:07 | yarn |
+----+--------+---------------------+---------+
3 rows in set (0.00 sec)
```

退出 MySQL 数据库：

```
Mysql> exit;
```

3. 将 MySQL 的数据导入到 HDFS 中

使用 sqoop-import 把上一步中创建的 sqoop_test 表中数据导入到 HDFS 中：

```
[hadoop@172 sqoop]$ bin/sqoop-import --connect jdbc:mysql://$mysqlIP/test --usern
ame root
-P --table sqoop_test --target-dir /sqoop
```

其中 --connect 用于连接 MySQL 数据库，test 也可以换成您的数据库名字，-P 表示之后需要输入密码，--table 为您想要导出的数据库的名字，--target-dir 为导出到 HDFS 中的路径。 **`/sqoop` 文件夹在执行命令之前并未创建，如果文件夹已经存在则会出错。**

回车后需要您输入密码，密码为您创建 EMR 时设置的密码。

执行成功后，可以在 HDFS 的相应路径下查看导入的数据：

```
[hadoop@172 sqoop]$ hadoop fs -cat /sqoop/*
1, first, 2018-07-03 15:29:37.0,hdfs
2, second, 2018-07-03 15:30:57.0,mr
3, third, 2018-07-03 15:31:07.0,yarn
```

4. 将 HDFS 的数据导入到 MySQL 中

首先需要在 MySQL 新建一个表准备存放 HDFS 中的数据：

```
[hadoop@172 sqoop]$ mysql -h $mysqlIP -p
Enter password:
mysql> use test;
Database changed
mysql> create table sqoop_test_back(id int not null primary key auto_increment, t
itle varchar(64), time timestamp, content varchar(255));
Query ok , 0 rows affected(0.00 sec)
```

查看表是否创建成功后退出 MySQL：

```
mysql> show tables;
+-----------------+
| Tables_in_test |
+-----------------+
| sqoop_test |
| sqoop_test_back |
+-----------------+
2 rows in set (0.00 sec)
mysql> exit;
```

使用 sqoop-export 把上一步导入 HDFS 中的数据再一次导入到 MySQL 中：

```
[hadoop@172 sqoop]$ bin/sqoop-export --connect jdbc:mysql://$mysqlIP/test --usern
ame
root -P --table sqoop_test_back --export-dir /sqoop
```

参数和 sqoop-import 类似，只不过变成了 --export-dir，该参数为 HDFS 中的存放数据的路径。回车后也需要输入密码。

执行成功，后即可验证数据库 sqoop_test_back 中的数据：

```
[hadoop@172 sqoop]$ mysql -h $mysqlIP -p
Enter password:
mysql> use test;
Database changed
mysql> select * from sqoop_test_back;
+----+---------+---------------------+---------+
| id | title | time | content |
+----+---------+---------------------+---------+
| 1 | first | 2018-07-03 15:29:37 | hdfs |
| 2 | second | 2018-07-03 15:30:57 | mr |
| 3 | third | 2018-07-03 15:31:07 | yarn |
+----+---------+---------------------+---------+
3 rows in set (0.00 sec)
```

更多的 Sqoop 操作可以查看 官方文档。

增量 DB 数据到 HDFS

最近更新时间：2021-07-09 11:12:55

Sqoop 是一款开源的工具，主要用于在 Hadoop 和传统数据库（MySQL、PostgreSQL 等）之间进行数据传递，可以将一个关系型数据库（例如 MySQL、Oracle、Postgres 等）中的数据导入到 Hadoop 的 HDFS 中，也可以将 HDFS 的数据导入到关系型数据库中。Sqoop 中一大亮点就是可以通过 Hadoop 的 MapReduce 把数据从关系型数据库中导入数据到 HDFS。

本文介绍了 Sqoop 的增量导入操作，即在数据库中的数据增加或更新后，把数据库的改动同步到导入 HDFS 的数据中。其中分为 append 模式和 lastmodified 模式，append 模式只能用在数据库的数据增加但不更新的场景，lastmodified 模式用在数据增加并且更新的场景。

1. 开发准备

- 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Sqoop 组件。
- Sqoop 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 使用 append 模式

本节将继续使用上一节的用例。
进入 EMR 控制台，复制目标集群的实例 ID，即集群的名字。再进入关系型数据库控制台，使用 Ctrl+F 进行搜索，找到集群对应的 MySQL 数据库，查看该数据库的内网地址 $mysqlIP。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Sqoop 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/sqoop
```

连接 MySQL 数据库：

```
[hadoop@172 sqoop]$ mysql -h $mysqlIP -p
Enter password:
```

密码为您创建 EMR 集群的时候设置的密码。

连接了 MySQL 数据库之后，在表 sqoop_test 中新增一条数据，如下：

```
mysql> use test;
Database changed
mysql> insert into sqoop_test values(null, 'forth', now(), 'hbase');
Query ok, 1 row affected(0.00 sec)
```

查看表中的数据：

```
Mysql> select * from sqoop_test;
+----+--------+---------------------+---------+
| id | title | time | content |
+----+--------+---------------------+---------+
| 1 | first | 2018-07-03 15:29:37 | hdfs |
| 2 | second | 2018-07-03 15:30:57 | mr |
| 3 | third | 2018-07-03 15:31:07 | yarn |
| 4 | forth | 2018-07-03 15:39:38 | hbase |
+----+--------+---------------------+---------+
4 rows in set (0.00 sec)
```

使用 append 模式将新增的数据同步到上一节中储存数据的 HDFS 路径中：

```
[hadoop@172 sqoop]$ bin/sqoop-import --connect jdbc:mysql://$mysqlIP/test --usern
ame
root -P --table sqoop_test --check-column id --incremental append --last-value 3
--target-dir
/sqoop
```

其中 $mysqlIP 为您的 MySQL 数据库的内网地址。

执行命令会需要您输入数据库的密码，默认为您创建 EMR 集群时设置的密码。比普通的 sqoop-import 命令多出一些参数，其中 --check-column 为导入时参照的数据，--incremental 为导入的模式，在此例中为 append，--last-value 为参考数据的参考值，比该值更新的数据都会导入到 HDFS 中。

执行成功后，可以查看 HDFS 相应目录下更新后的数据：

```
[hadoop@172 sqoop]$ hadoop fs -cat /sqoop/*
1, first, 2018-07-03 15:29:37.0,hdfs
2, second, 2018-07-03 15:30:57.0,mr
3, third, 2018-07-03 15:31:07.0,yarn
4,forth,2018-07-03 15:39:38.0,hbase
```

使用 Sqoop job

使用 append 同步 HDFS 中的数据每次需要手动输入 --last-value，也可以使用 sqoop job 的方式，Sqoop 会自动保存上次导入成功的 last-value 值。如果要使用 sqoop job。需要启动 sqoop-metastore 进程，操作步骤如下：

首先在 conf/sqoop-site.xml 中启动 sqoop-metastore 进程：

```
<property>
<name>sqoop.metastore.client.enable.autoconnect</name>
<value>true</value>
</property>
```

然后在 bin 目录下启动 sqoop-metastore 服务：

```
./sqoop-metastore &
```

使用如下指令创建 Sqoop job：

> 说明：
> 此命令适用于 Sqoop 1.4.6 版本。

```
[hadoop@172 sqoop]$ bin/sqoop job --create job1 -- import --connect
jdbc:mysql://$mysqlIP/test --username root -P --table sqoop_test --check-column i
d
--incremental append --last-value 4 --target-dir /sqoop
```

其中 $mysqlIP 为您的 MySQL 的内网地址。使用该命令就成功创建了一个 Sqoop job，每一次执行，会自动从上次更新的 last-value 值自动更新。

为 MySQL 中的 sqoop_test 表格新增一条记录：

```
mysql> insert into sqoop_test values(null, 'fifth', now(), 'hive');
Query ok, 1 row affected(0.00 sec)
Mysql> select * from sqoop_test;
+----+--------+---------------------+---------+
| id | title | time | content |
+----+--------+---------------------+---------+
| 1 | first | 2018-07-03 15:29:37 | hdfs |
| 2 | second | 2018-07-03 15:30:57 | mr |
| 3 | third | 2018-07-03 15:31:07 | yarn |
| 4 | forth | 2018-07-03 15:39:38 | hbase |
| 5 | fifth | 2018-07-03 16:02:29 | hive |
+----+--------+---------------------+---------+
5 rows in set (0.00 sec)
```

然后执行 Sqoop job：

```
[hadoop@172 sqoop]$ bin/sqoop job --exec job1
```

执行该命令会让您输入 MySQL 的密码。执行成功后，可以查看 HDFS 相应目录下更新后的数据：

```
[hadoop@172 sqoop]$ hadoop fs -cat /sqoop/*
1, first, 2018-07-03 15:29:37.0,hdfs
2, second, 2018-07-03 15:30:57.0,mr
3, third, 2018-07-03 15:31:07.0,yarn
4,forth,2018-07-03 15:39:38.0,hbase
5,fifth,2018-07-03 16:02:29.0,hive
```

3. 使用 lastmodified 模式

直接创建一个 sqoop-import 的 lastmodified 模式的 Sqoop job，首先查询 sqoop_test 中最后更新的时间：

```
mysql> select max(time) from sqoop_test;
```

创建一个 Sqoop job：

```
[hadoop@172 sqoop]$ bin/sqoop job --create job2 -- import --connect jdbc:mysql://
$mysqlIP/test --username root -P --table sqoop_test --check-column time --increme
ntal lastmodified --merge-key id --last-value '2018-07-03 16:02:29' --target-dir
/sqoop
```

参数说明：

- $mysqlIP 为您的 MySQL 的内网地址。
- --check-column 必须使用 timestamp。
- --incremental 模式选择 lastmodified。
- --merge-key 选择 ID。
- --last-value 为我们查询到的表中的最后更新时间。在此时间后做出的更新都会被同步到 HDFS 中，而 Sqoop job 每次会自动保存和更新该值。

对 MySQL 中的 sqoop_test 表添加数据并做出更改：

```
mysql> insert into sqoop_test values(null, 'sixth', now(), 'sqoop');
Query ok, 1 row affected(0.00 sec)
mysql> update sqoop_test set time=now(), content='spark' where id = 1;
Query OK, 1 row affected (0.00 sec)
```

```
Rows matched: 1 changed: 1 warnings: 0
Mysql> select * from sqoop_test;
+----+--------+---------------------+---------+
| id | title | time | content |
+----+--------+---------------------+---------+
| 1 | first | 2018-07-03 16:07:46 | spark |
| 2 | second | 2018-07-03 15:30:57 | mr |
| 3 | third | 2018-07-03 15:31:07 | yarn |
| 4 | forth | 2018-07-03 15:39:38 | hbase |
| 5 | fifth | 2018-07-03 16:02:29 | hive |
| 6 | fifth | 2018-07-03 16:09:58 | sqoop |
+----+--------+---------------------+---------+
6 rows in set (0.00 sec)
```

执行 Sqoop job：

```
[hadoop@172 sqoop]$ bin/sqoop job --exec job2
```

执行该命令会让您输入 MySQL 的密码。执行成功后，可以查看 HDFS 相应目录下更新后的数据：

```
[hadoop@172 sqoop]$ hdfs dfs -cat /sqoop/*
1,first,2018-07-03 16:07:46.0,spark
2,second,2018-07-03 15:30:57.0,mr
3,third,2018-07-03 15:31:07.0,yarn
4,forth,2018-07-03 15:39:38.0,hbase
5,fifth,2018-07-03 16:02:29.0,hive
6,sixth,2018-07-03 16:09:58.0,sqoop
```

更多的 Sqoop 操作可以查看 官方文档。

Hive 存储格式和关系型数据库之间进行导入导出

最近更新时间：2021-10-29 10:09:03

本文介绍了使用腾讯云 Sqoop 服务将数据在 MySQL 和 Hive 之间相互导入导出的方法。

1. 开发准备

- 确认已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Sqoop、Hive 组件。
- Sqoop 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

2. 将关系型数据库导入到 Hive 中

本节将继续使用上一节的用例。

进入 弹性 MapReduce 控制台，复制目标集群的实例 ID，即集群的名字。再进入关系型数据库控制台，使用 Ctrl+F 进行搜索，找到集群对应的 MySQL 数据库，查看该数据库的内网地址 $mysqlIP。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hive
```

新建一个 Hive 数据库：

```
[hadoop@172 hive]$ hive
hive> create database hive_from_sqoop;
OK
Time taken: 0.167 seconds
```

使用 sqoop-import 命令把上一节中创建的 MySQL 数据库导入到 Hive 中：

```
[hadoop@172 hive]# cd /usr/local/service/sqoop
[hadoop@172 sqoop]$ bin/sqoop-import --connect jdbc:mysql://$mysqlIP/test --usern
```

```
ame
root -P --table sqoop_test_back --hive-database hive_from_sqoop --hive-import --h
ive-table hive_from_sqoop
```

- $mysqlIP：腾讯云关系型数据库（CDB）的内网地址。
- test：MySQL 数据库名称。
- --table：要导出的 MySQL 表名。
- --hive-database：Hive 数据库名。
- --hive-table：导入的 Hive 表名。

执行指令需要输入您的 MySQL 密码，默认为您创建 EMR 集群时设置的密码。执行成功后，可以在 Hive 中查看导入的数据库：

```
hive> select * from hive_from_sqoop;
OK
1 first 2018-07-03 16:07:46.0 spark
2 second 2018-07-03 15:30:57.0 mr
3 third 2018-07-03 15:31:07.0 yarn
4 forth 2018-07-03 15:39:38.0 hbase
5 fifth 2018-07-03 16:02:29.0 hive
6 sixth 2018-07-03 16:09:58.0 sqoop
Time taken: 1.245 seconds, Fetched: 6 row(s)
```

3. 将 Hive 导入到关系型数据库中

Sqoop 支持将 Hive 表中的数据导入到关系型数据库中。先在 Hive 中创建新表并导入数据。

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hive
```

新建一个 bash 脚本文件 gen_data.sh，在其中添加以下代码：

```
#!/bin/bash
MAXROW=1000000 #指定生成数据行数
for((i = 0; i < $MAXROW; i++))
do
    echo $RANDOM, \"$RANDOM\"
done
```

并按如下方式执行：

```
[hadoop@172 hive]$ ./gen_data.sh > hive_test.data
```

这个脚本文件会生成1,000,000个随机数对，并且保存到文件 hive_test.data 中。

使用如下指令把生成的测试数据先上传到 HDFS 中：

```
[hadoop@172 hive]$ hdfs dfs -put ./hive_test.data /$hdfspath
```

其中 $hdfspath 为 HDFS 上的您存放文件的路径。

连接 Hive 并创建测试表：

```
[hadoop@172 hive]$ bin/hive
hive> create database hive_to_sqoop; #创建数据库 hive_to_sqoop
OK
Time taken: 0.176 seconds
hive> use hive_to_sqoop; #切换数据库
OK
Time taken: 0.176 seconds
hive> create table hive_test (a int, b string)
hive> ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
                        #创建数据表 hive_test, 并指定列分割符为',,'
OK
Time taken: 0.204 seconds
hive> load data inpath "/$hdfspath/hive_test.data" into table hive_test; #导入数据
```

$hdfspath 为 HDFS 上的您存放文件的路径。

成功后可使用 `quit` 命令退出 Hive 数据仓库。连接关系型数据库并创建对应的表格：

```
[hadoop@172 hive]$ mysql -h $mysqlIP -p
Enter password:
```

其中 $mysqlIP 为该数据库的内网地址，密码为您创建集群时设置的密码。

在 MySQL 中创建一个名为 test 的表格，**MySQL 中的表字段名字和 Hive 中的表字段名字必须完全一致**：

```
mysql> create table table_from_hive (a int,b varchar(255));
```

成功创建表格后即可退出 MySQL。

使用 Sqoop 把 Hive 数据仓库中的数据导入到关系型数据库中有两种方法，可以直接使用 HDFS 存储的 Hive 数据，也可以使用 Hcatalog 来进行数据的导入。

使用 HDFS 中的 Hive 数据

切换进入 Sqoop 文件夹，然后使用以下指令把 Hive 数据库中的数据导出到关系型数据库中：

```
[hadoop@172 hive]$ cd ../sqoop/bin
[hadoop@172 bin]$ ./sqoop-export --connect jdbc:mysql://$mysqlIP/test --username
root -P
--table table_from_hive --export-dir /usr/hive/warehouse/hive_to_sqoop.db/hive_te
st
```

其中 $mysqlIP 为您的关系型数据库的内网 IP 地址，test 为关系型数据库中的数据库名，--table 后跟的参数为您的关系型数据库的表名，--export-dir 后跟的参数为 Hive 表中的数据在 HDFS 中存储的位置。

使用 Hcatalog 进行导入

切换进入 Sqoop 文件夹，然后使用以下指令把 Hive 数据库中的数据导出到关系型数据库中：

```
[hadoop@172 hive]$ cd ../sqoop/bin
[hadoop@172 bin]$ ./sqoop-export --connect jdbc:mysql://$mysqlIP/test --username
root -P
--table table_from_hive --hcatalog-database hive_to_sqoop --hcatalog-table hive_t
est
```

其中 $mysqlIP 为您的关系型数据库的内网 IP 地址，test 为关系型数据库中的数据库名，--table 后跟的参数为您的关系型数据库的表名，--hcatalog-database 后面跟的参数是要导出的 Hive 表所在的数据库的名称，--hcatalog-table 后面跟的参数是要 Hive 中要导出的表的名称。

操作完成后可以进入关系型数据库查看是否导入成功：

```
[hadoop@172 hive]$ mysql -h $mysqlIP -p #连接 MySQL
Enter password:
mysql> use test;
Database changed
mysql> select count(*) from table_from_hive; #现在表中有1000000条数据
+----------+
| count(*) |
+----------+
| 1000000 |
+----------+
1 row in set (0.03 sec)
mysql> select * from table_from_hive limit 10; #查看表中前10条记录
+-------+----------+
| a | b |
+-------+----------+
| 28523 | "3394" |
| 31065 | "24583" |
```

```
| 399 | "23629" |
| 18779 | "8377" |
| 25376 | "30798" |
| 20234 | "22048" |
| 30744 | "32753" |
| 21423 | "6117" |
| 26867 | "16787" |
| 18526 | "5856" |
+-------+----------+
10 rows in set (0.00 sec)
```

更多关于 sqoop-export 命令的参数可以通过如下命令查看：

```
[hadoop@172 bin]$ ./sqoop-export --help
```

4. 将 orc 格式的 Hive 表格导入到关系型数据库中

orc 是按列存储的一种文件存储格式，使用该格式能够极大的提升 Hive 的性能。本节介绍了如何创建一个 orc 格式的表并载入数据，然后使用腾讯云 Sqoop 服务把 Hive 中以 orc 格式进行存储的数据导出到关系型数据库。

> 注意：
> 将 orc 存储格式的 Hive 表格导入到关系型数据库中不能直接使用 HDFS 中存储的数据，只能使用 Hcatalog 进行操作。

本节将继续使用上一节的用例。

登录 EMR 集群的 Master 节点后，在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Hive 文件夹：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]# cd /usr/local/service/hive
```

在上一节中创建的 hive_from_sqoop 数据库中创建一个新表格：

```
[hadoop@172 hive]$ hive
hive> use hive_to_sqoop;
OK
Time taken: 0.013 seconds
hive> create table if not exists orc_test(a int,b string) ROW FORMAT DELIMITED FI
ELDS TERMINATED BY ',' stored as orc;
```

可以通过如下指令来查看表格中数据的存储格式：

```
hive> show create table orc_test;
OK
CREATE TABLE `orc_test`(
`a` int,
`b` string)
ROW FORMAT SERDE
'org.apache.hadoop.hive.ql.io.orc.OrcSerde'
WITH SERDEPROPERTIES (
'field.delim'=',',
'serialization.format'=',')
STORED AS INPUTFORMAT
'org.apache.hadoop.hive.ql.io.orc.OrcInputFormat'
OUTPUTFORMAT
'org.apache.hadoop.hive.ql.io.orc.OrcOutputFormat'
LOCATION
'hdfs://HDFS2789/usr/hive/warehouse/hive_to_sqoop.db/orc_test'
TBLPROPERTIES (
'COLUMN_STATS_ACCURATE'='{\"BASIC_STATS\":\"true\"}',
'numFiles'='0',
'numRows'='0',
'rawDataSize'='0',
'totalSize'='0',
'transient_lastDdlTime'='1533563293')
Time taken: 0.041 seconds, Fetched: 21 row(s)
```

由返回的数据可以看出该表格中的数据存储格式为 orc。

有多种方式可以向 orc 格式的 Hive 表格导入数据，下面主要介绍通过创建临时的存储格式为 text 的 Hive 表格来向 orc 存储格式的表格导入数据，这里我们使用上一节中创建的 hive_test 表格作为临时表格，使用以下指令来导入数据：

```
hive> insert into table orc_test select * from hive_test;
```

导入成功后可通过 `select` 指令查看表格中的数据。

然后使用 Sqoop 把 orc 格式的 Hive 表格导出到 MySQL中。连接关系型数据库并创建对应的表格，连接关系型数据库的具体方式见上文：

```
[hadoop@172 hive]$ mysql -h $mysqlIP -p
Enter password:
```

其中 $mysqlIP 为该数据库的内网地址，密码为您创建集群时设置的密码。

在 MySQL 中创建一个名为 test 的表格，**MySQL 中的表字段名字和 Hive 中的表字段名字必须完全一致**：

```
mysql> create table table_from_orc (a int,b varchar(255));
```

成功创建表格后即可退出 MySQL。

切换进入 Sqoop 文件夹，然后使用以下指令把 Hive 数据库中以 orc 格式存储的数据导出到关系型数据库中：

```
[hadoop@172 hive]$ cd ../sqoop/bin
[hadoop@172 bin]$ ./sqoop-export --connect jdbc:mysql://$mysqlIP/test --username
root -P
--table table_from_orc --hcatalog-database hive_to_sqoop --hcatalog-table orc_tes
t
```

其中 $mysqlIP 为您的关系型数据库的内网 IP 地址，test 为关系型数据库中的数据库名，--table 后跟的参数为您的关系型数据库的表名，--hcatalog-database 后面跟的参数是要导出的 Hive 表所在的数据库的名称，--hcatalog-table 后面跟的参数是要 Hive 中要导出的表的名称。

导入成功后可以在 MySQL 中查看相应表中的数据：

```
mysql> select count(*) from table_from_orc;
+----------+
| count(*) |
+----------+
| 1000000 |
+----------+
1 row in set (0.24 sec)
mysql> select * from table_from_orc limit 10;
+-------+----------+
| a | b |
+-------+----------+
| 28523 | "3394" |
| 31065 | "24583" |
| 399 | "23629" |
| 18779 | "8377" |
| 25376 | "30798" |
| 20234 | "22048" |
| 30744 | "32753" |
| 21423 | "6117" |
| 26867 | "16787" |
| 18526 | "5856" |
+-------+----------+
10 rows in set (0.00 sec)
```

更多的 Sqoop 操作可以查看 官方文档。

Hue 开发指南

Hue 简介

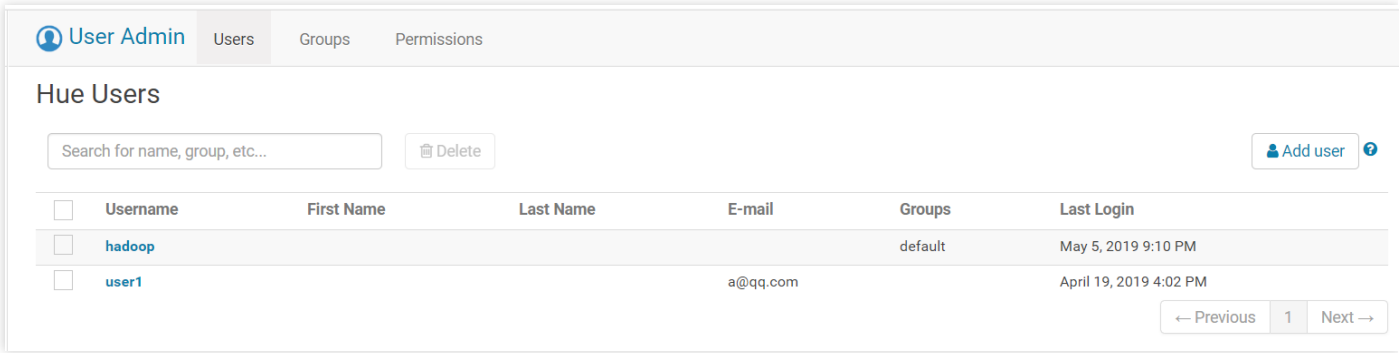
最近更新时间：2023-06-19 17:18:18

Hue 是一个开源的 Apache Hadoop UI 系统，由 Cloudera Desktop 演化而来，最后 Cloudera 公司将其贡献给 Apache 基金会的 Hadoop 社区，它是基于 Python Web 框架 Django 实现的。通过使用 Hue 我们可以在浏览器端的 Web 控制台上与 Hadoop 集群进行交互来分析处理数据，例如操作 HDFS 上的数据、运行 MapReduce Job、执行 Hive 的 SQL 语句和浏览 HBase 数据库等。

访问 Hue WebUI

使用 Hue 组件管理工作流时，请先登录 Hue 控制台页面，具体步骤如下：

1. 登录 EMR 控制台，单击对应集群 **ID/名称**，进入集群详情页面，然后单击**集群服务**。
2. 在列表页找到 Hue 组件，单击 **WebUI 访问地址**进入 Hue 页面。
3. 首次登录 Hue 控制台页面，请使用 hadoop 帐号，密码为创建集群时提供的密码。



> 注意：
> EMR-V2.5.0及以前版本、EMR-V3.1.0及以前版本未集成 OpenLDAP ，需要在首次以 root 帐号登录 Hue 控制台，参考 社区官方文档 于 WebUI 新建帐号。EMR 产品的组件启动帐号为 hadoop，历史版本建议首次登录 Hue 控制台后，新建 hadoop 帐号，后续可以通过 hadoop 帐号来提交作业。

用户权限管理

1. 添加用户。

i. 登录 EMR 控制台，使用 用户管理 功能添加新用户。

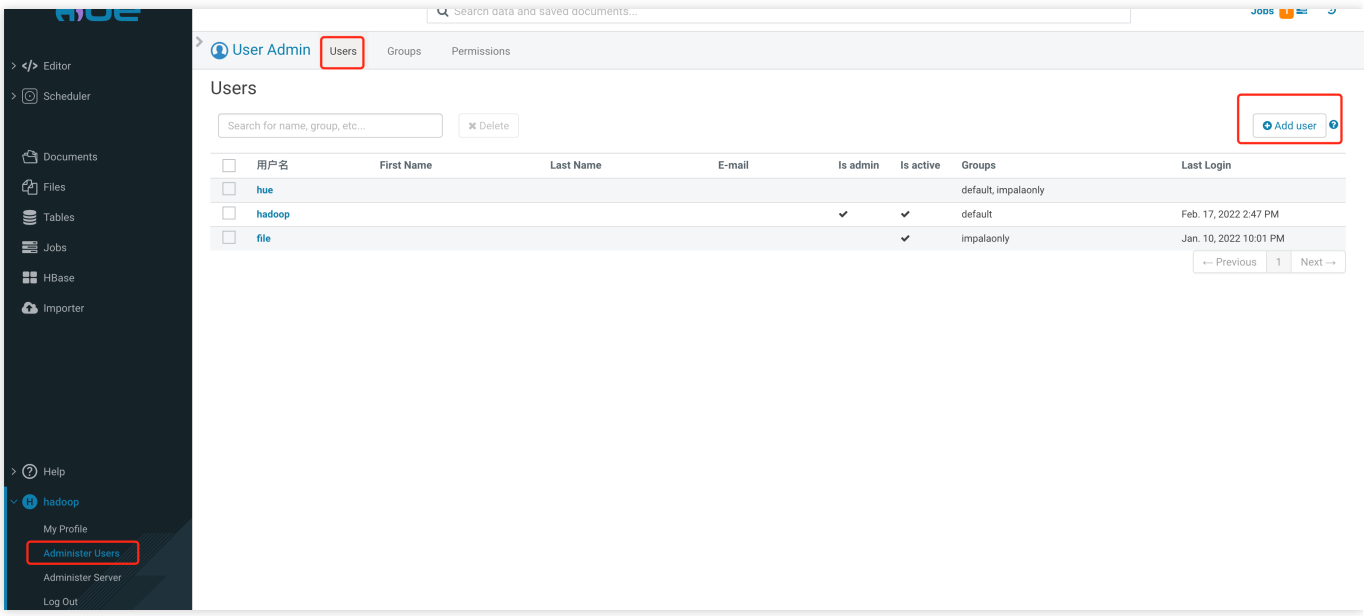
ii. 如果您的集群部署了 Ranger，添加新用户后，需要手动触发 ranger-ugsync-site.xml 的配置下发，重启 EnableUnixAuth 服务进行用户同步，具体操作步骤可参考 用户管理。然后进入 Ranger WebUI 设置新用户访问权限。

iii. 在列表页找到 Hue 组件，单击 WebUI 访问地址进入 Hue 页面，完成新用户登录及使用。

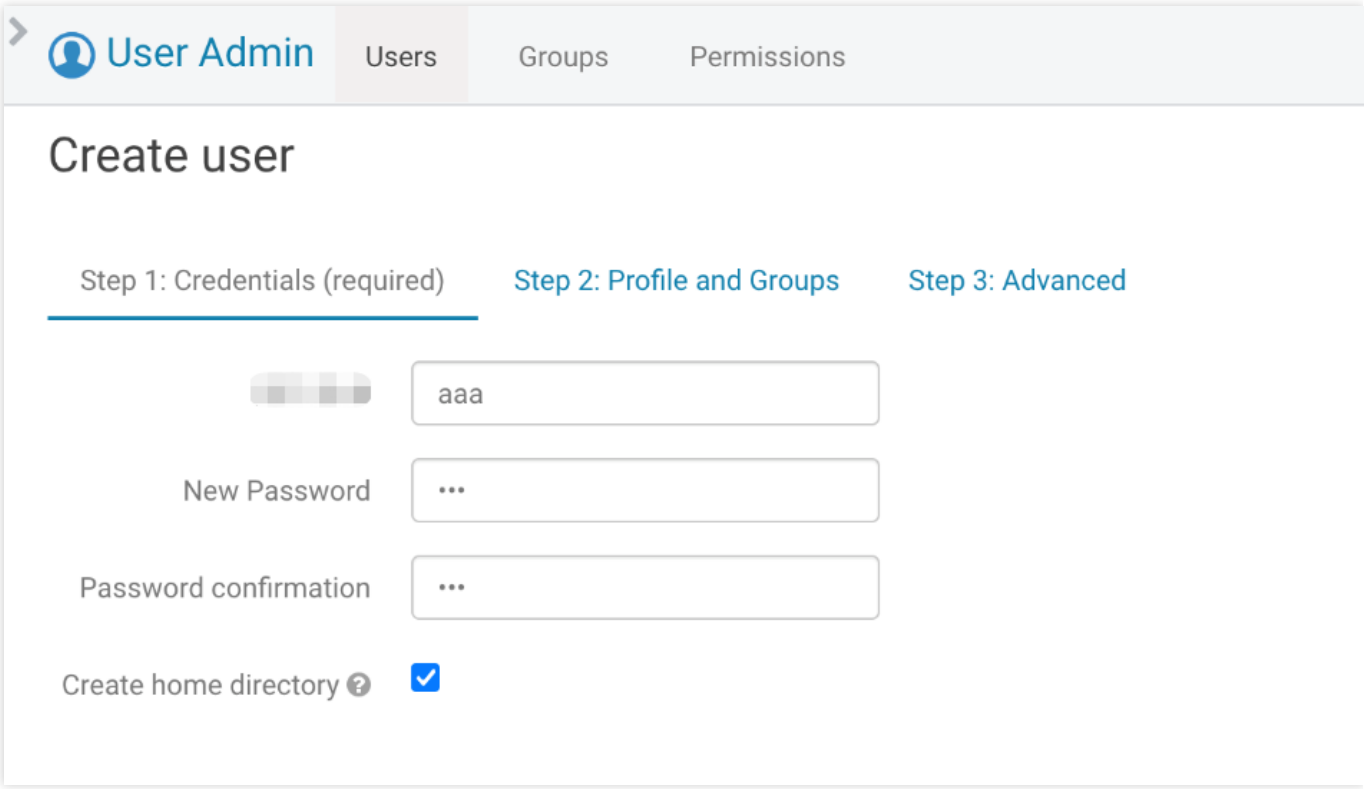
2. 权限控制。

hue 通过将不同的权限添加到组，用户通过加入不同的组获得对应权限。

i. 单击用户管理页面上方的 **Groups**，然后单击右侧的 **Add group**。

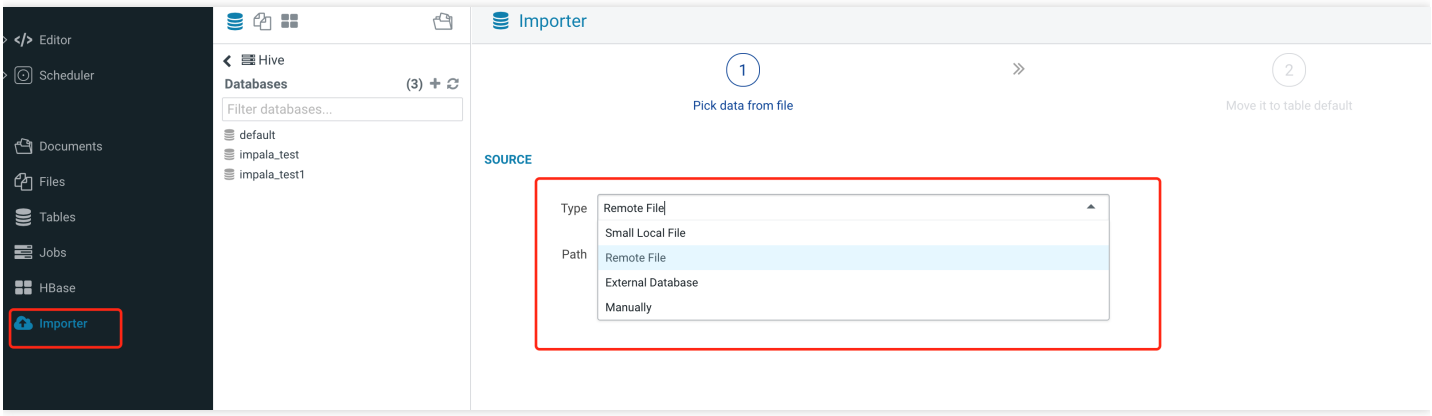


ii. 填写用户组信息，可勾选目标用户加入此组，并勾选此用户组的权限，单击下方的 **Add Group**。



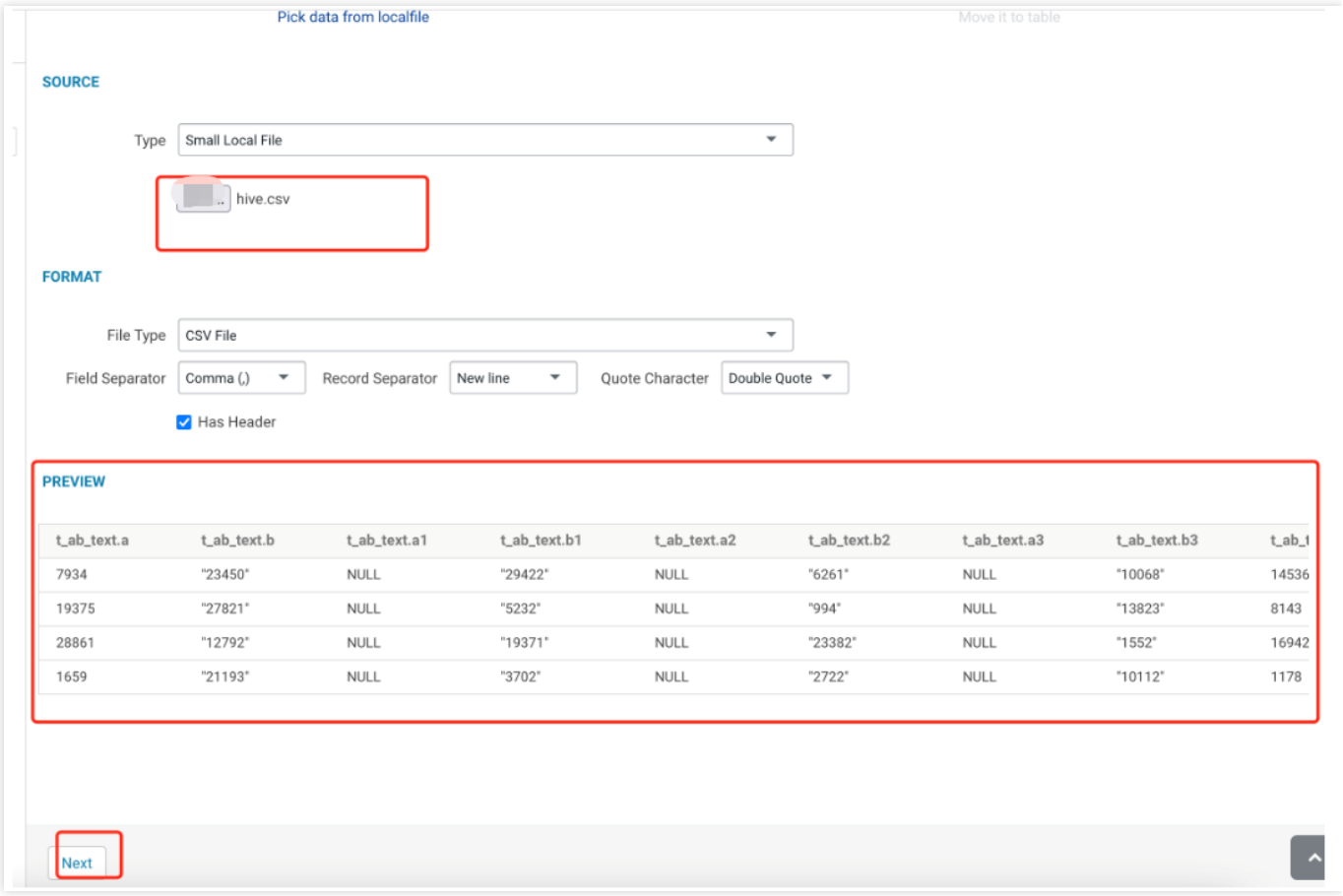
数据导入

Hue 支持4种导入方式：本地文件、HDFS 上的文件、外部数据库以及人工导入。

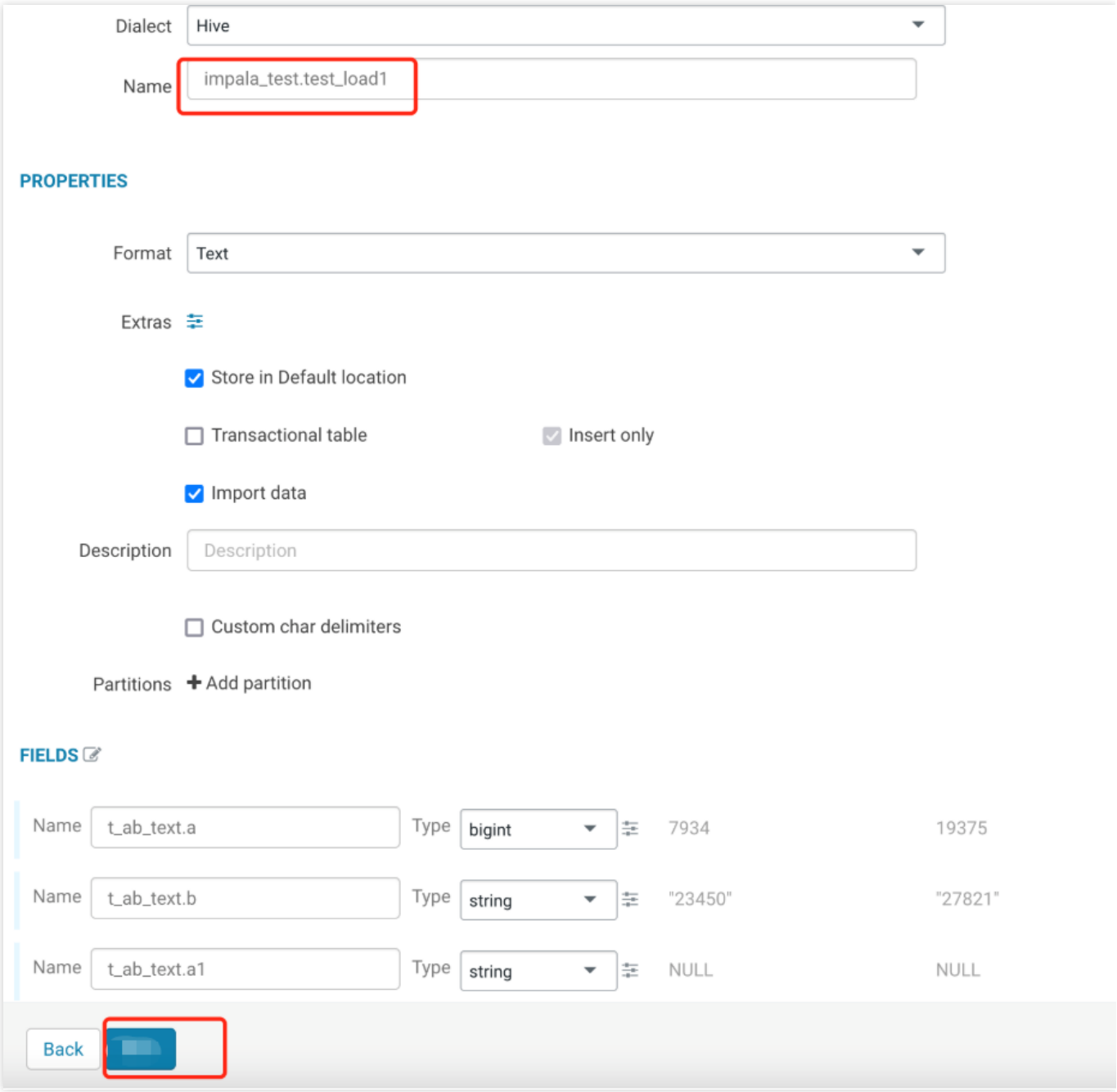


1. 本地文件导入。

i. 单击浏览选择 csv 文件，hue 会自动识别出分隔符并生成预览，单击 Next 导入到表。

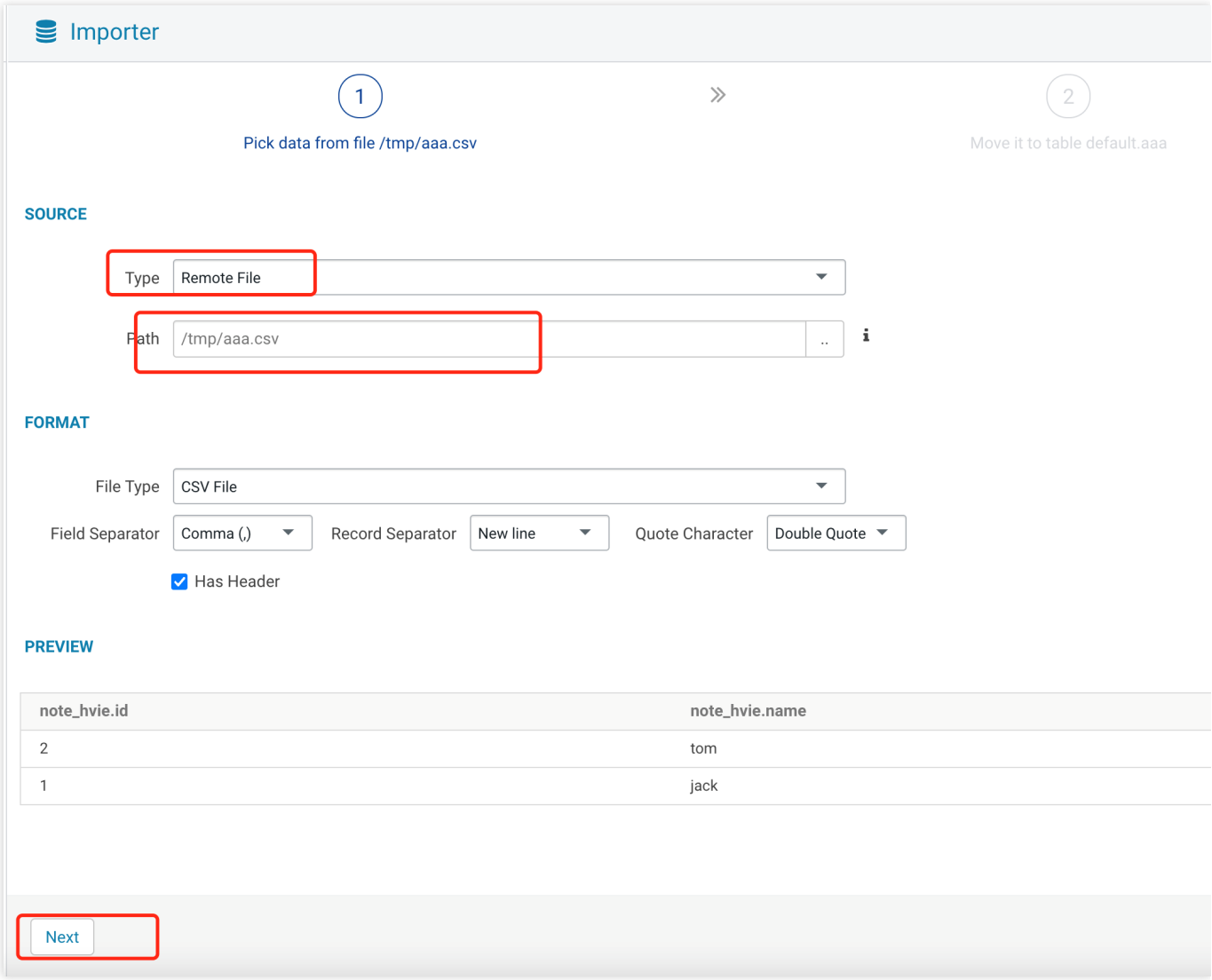


ii. 填写需要导入的表信息等，单击保存。

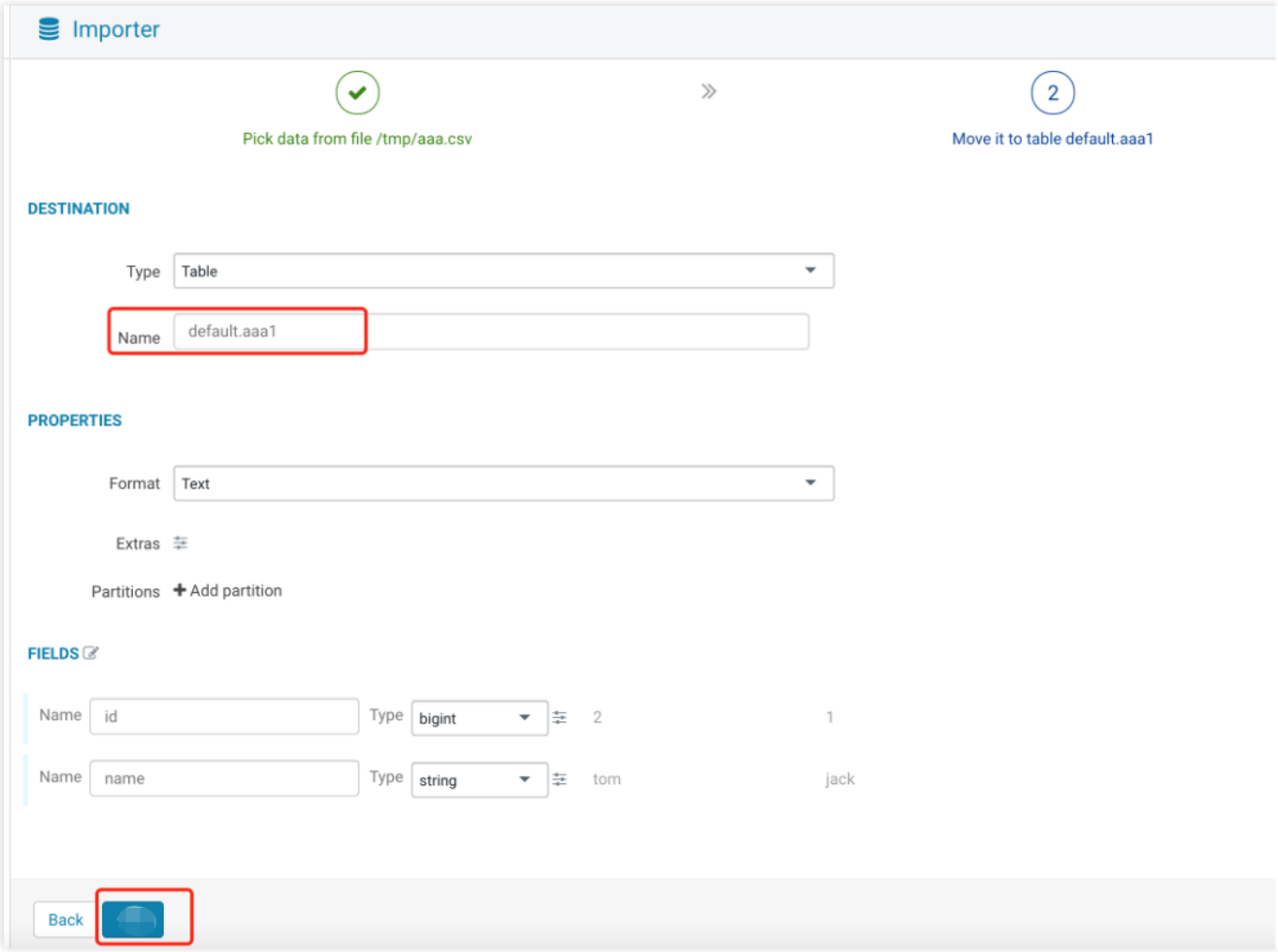


2. HDFS 文件导入。

i. 选择 HDFS 上的 csv 文件。

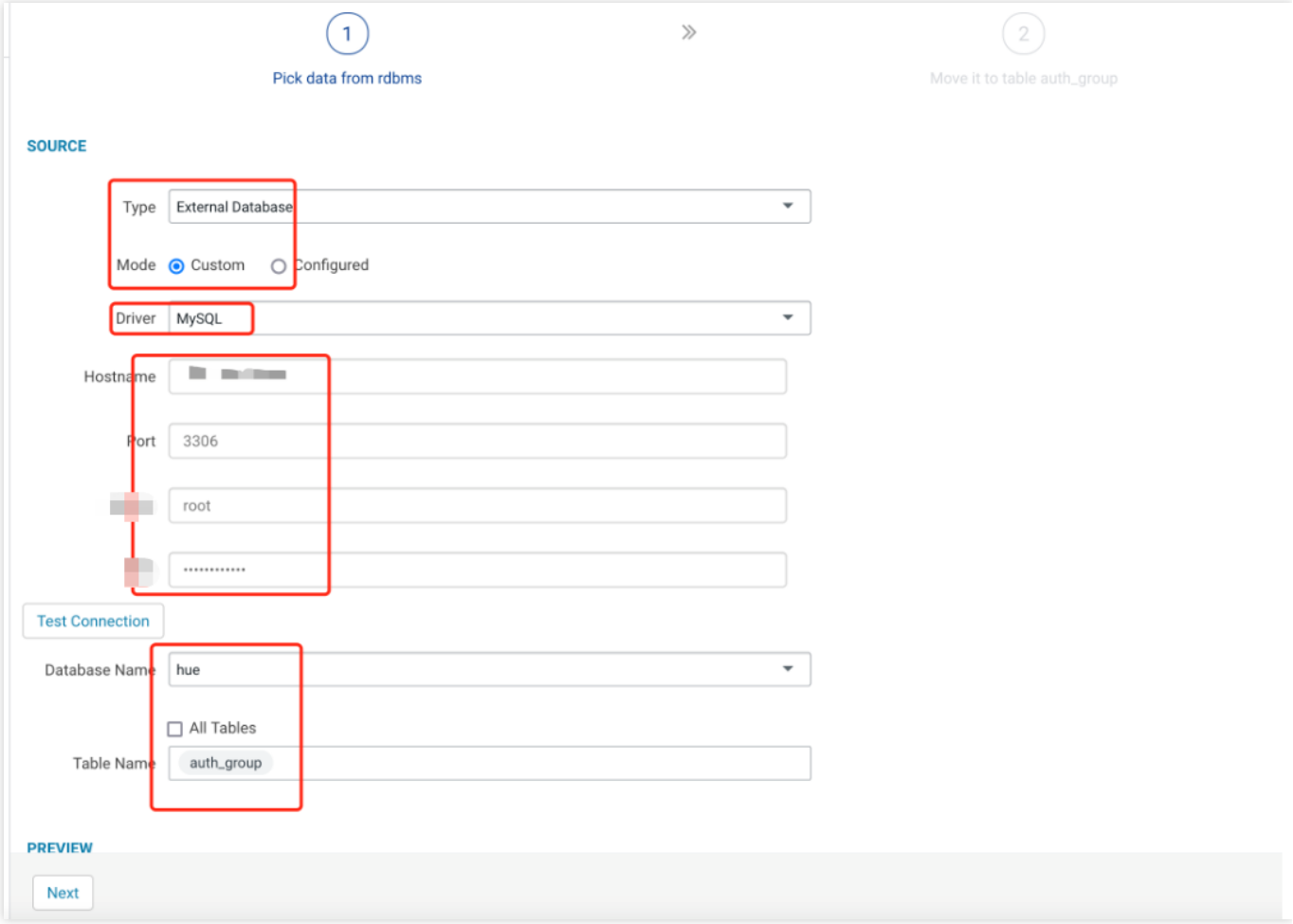


ii. 填写需要导入的表信息等，单击保存。

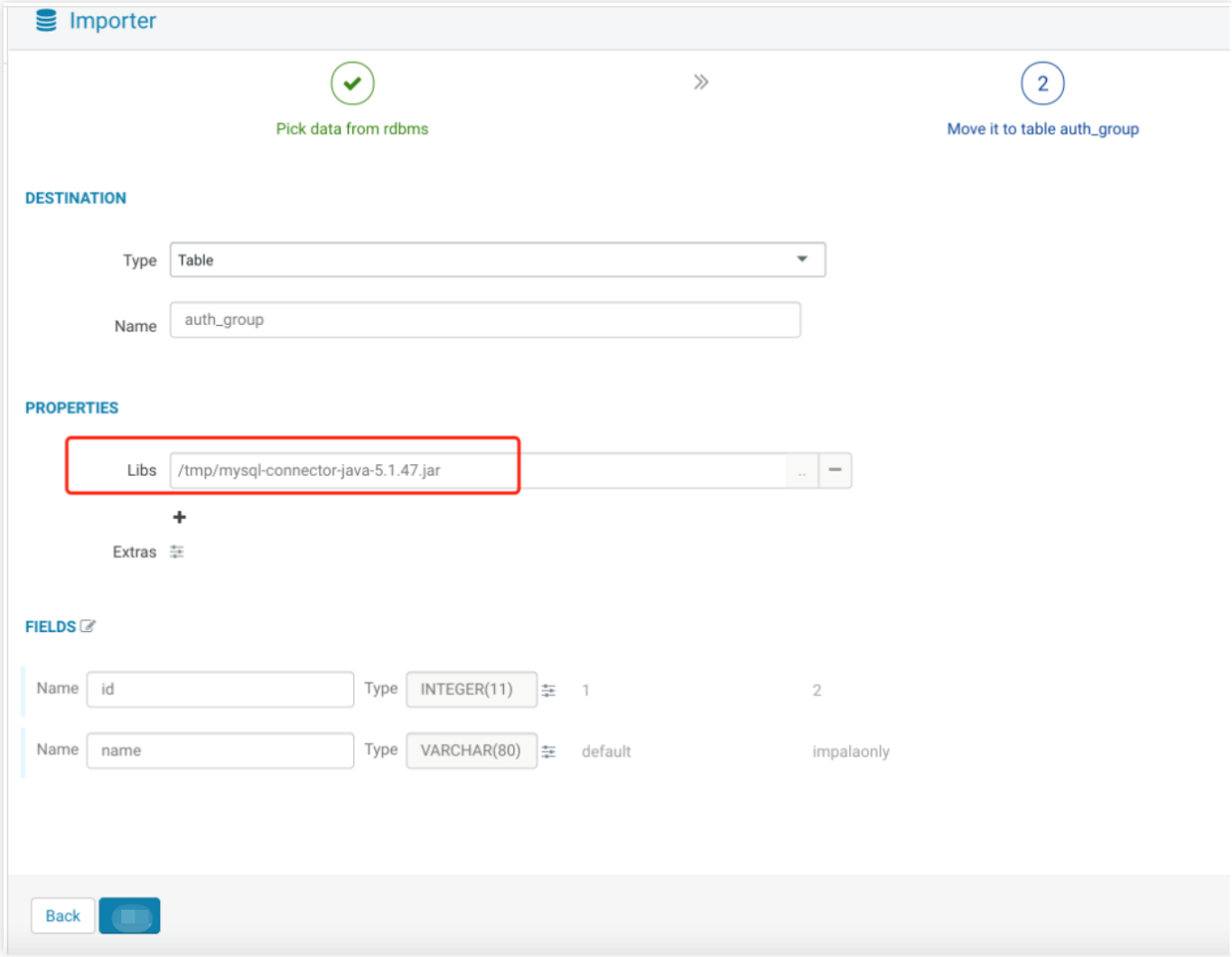


3. 外部数据库 External Database.

i. 填写外部数据库信息，单击 **Test Connection** 获取到数据库信息，选择库和表后单击 Next。

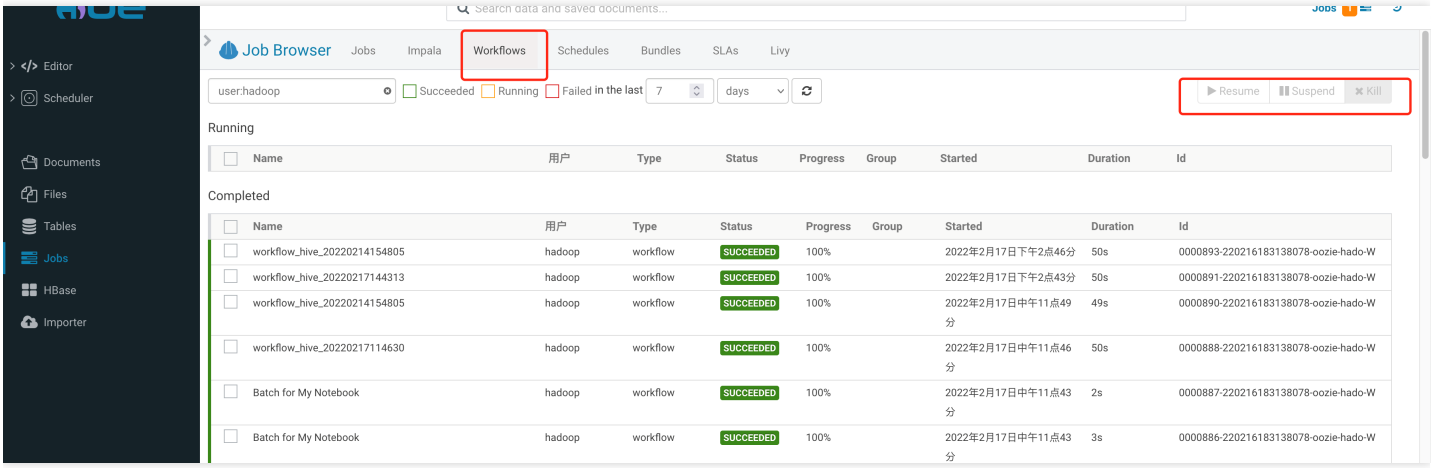


ii. 填写需要导入的目的表信息，并单击 lib 选择 mysql 驱动，然后单击保存。



Job 管理

单击右侧的 **Jobs**标签，即可进入任务管理页面，单击上方的各个任务类型标签，可进行查看管理。



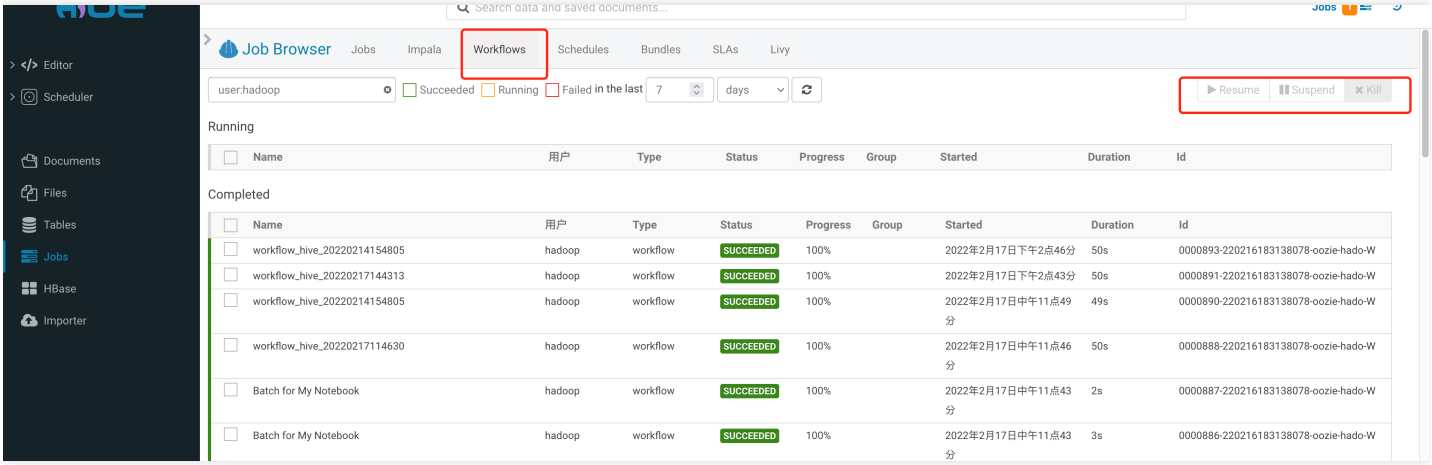
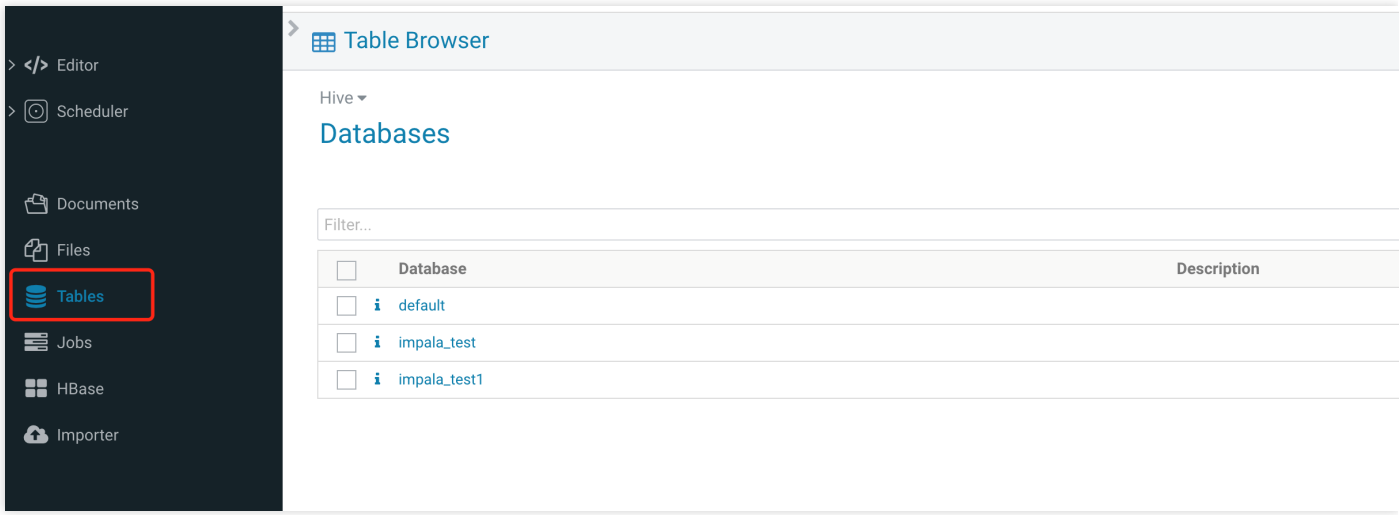
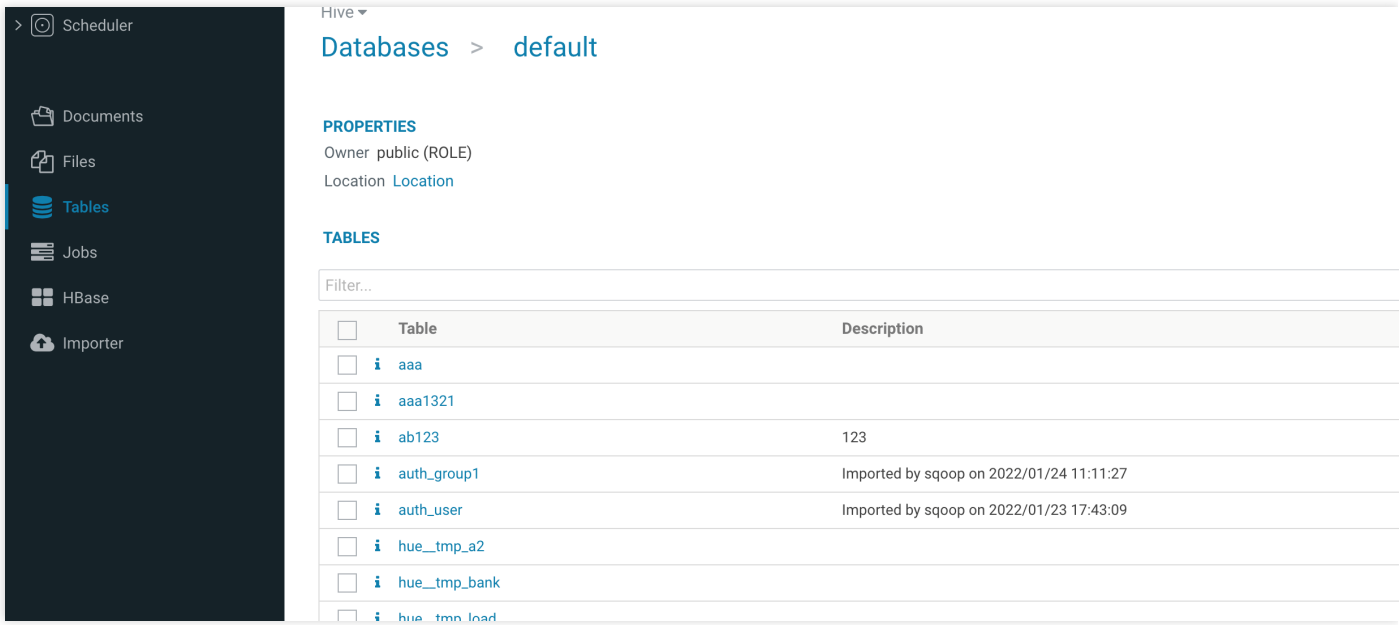


Table 管理

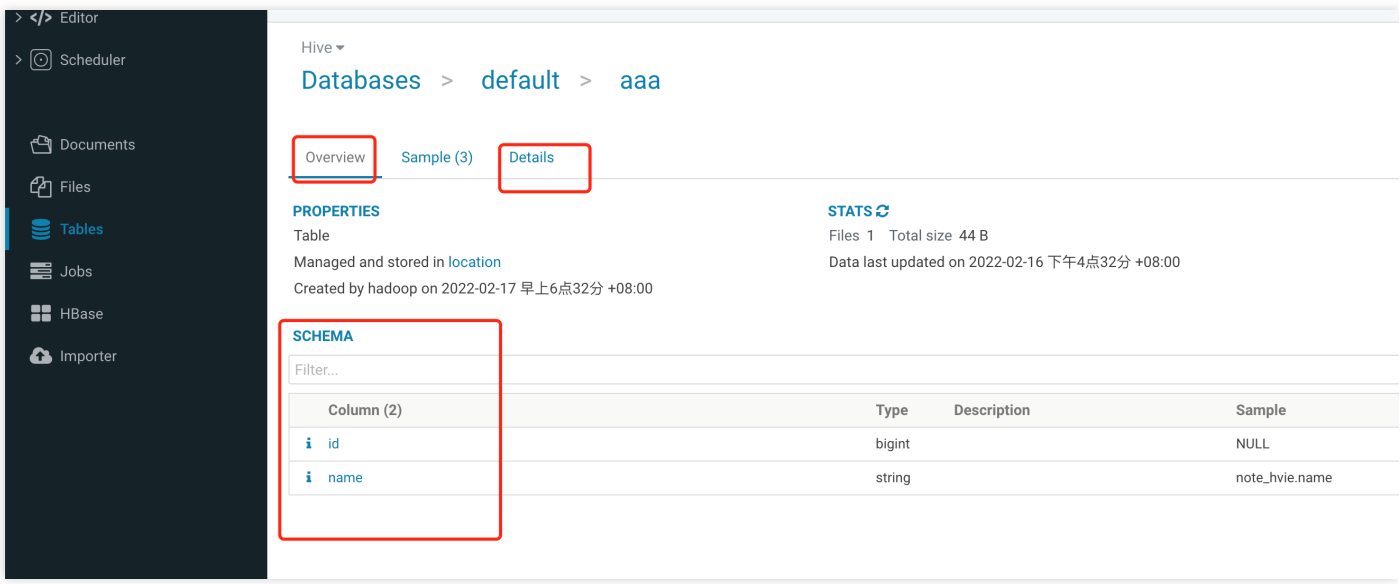
1. 单击右侧的 **Tables** 进入到 Table 管理页面，可以查看到基本的数据库信息。



2. 单击其中一个数据库，可查看此数据库的表。



3. 单击各个表，可以查看表的具体详细信息。



Hue 最佳实践

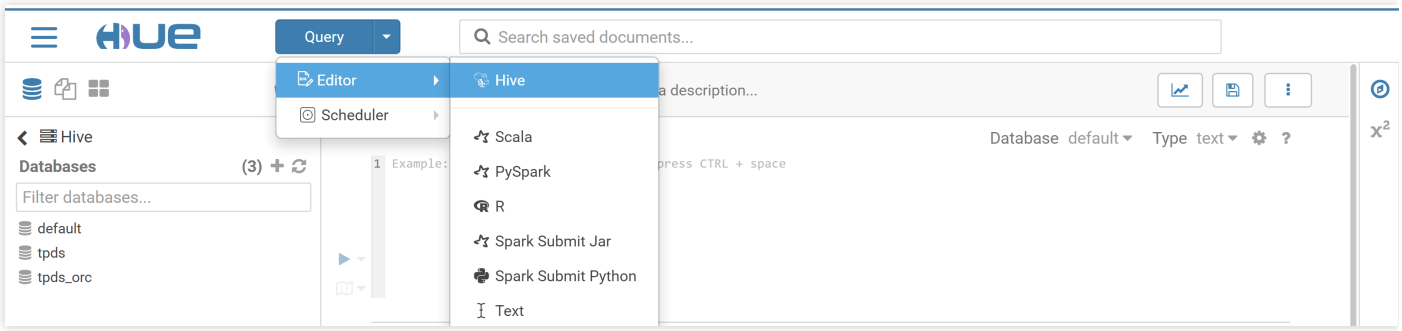
最近更新时间：2022-05-24 16:54:42

本文主要介绍 Hue 的实践用法。

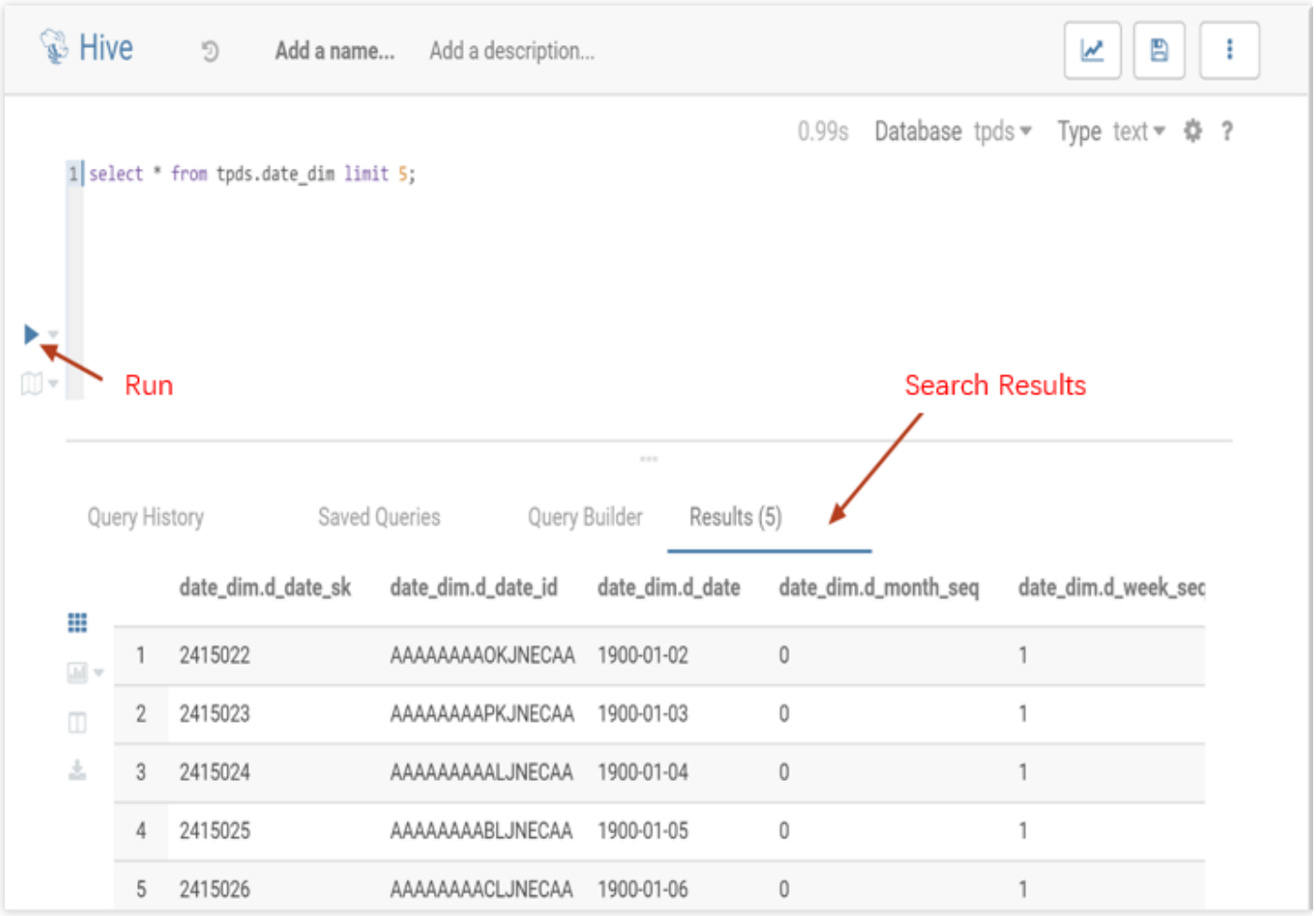
Hive SQL 查询

Hue 的 beeswax App 提供了友好方便的 Hive 查询功能，可以选择不同的 Hive 数据库、编写 HQL 语句、提交查询任务、查看结果。

1. 在 Hue 控制台上方，选择 Query > Editor > Hive。

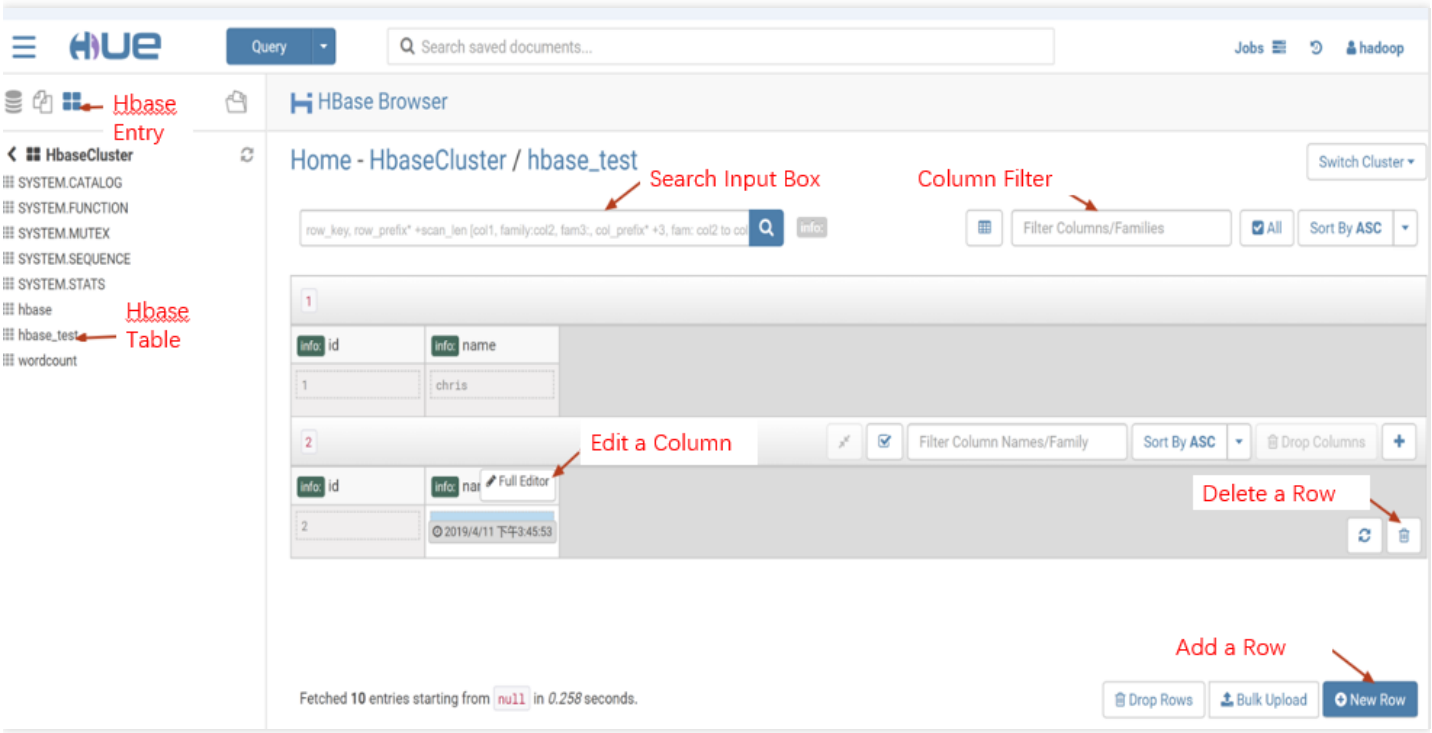


2. 在语句输入框中输入要执行语句，然后单击**执行**，执行语句。



Hbase 数据查询和修改、数据展示

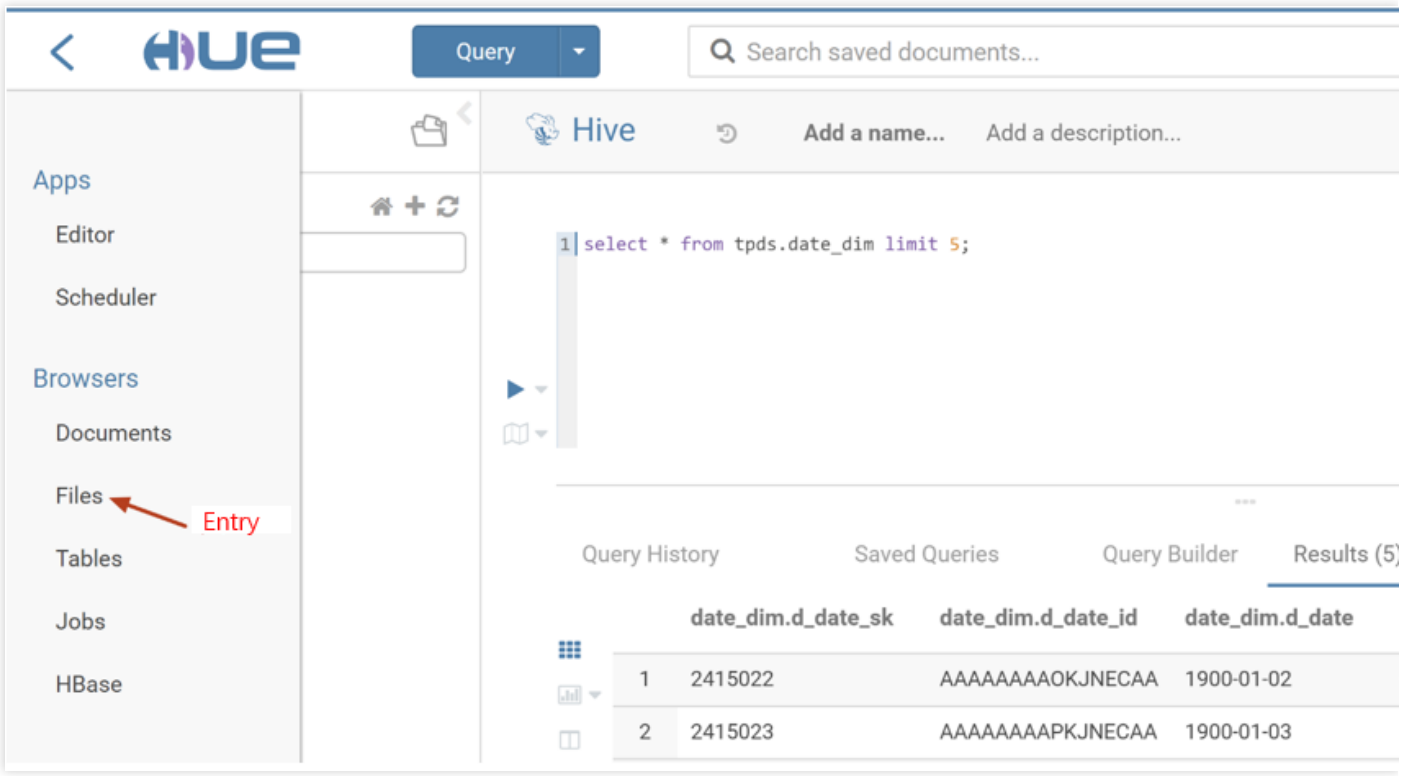
使用 Hbase Browser 可以查询、修改、展示 Hbase 集群中表的数据。



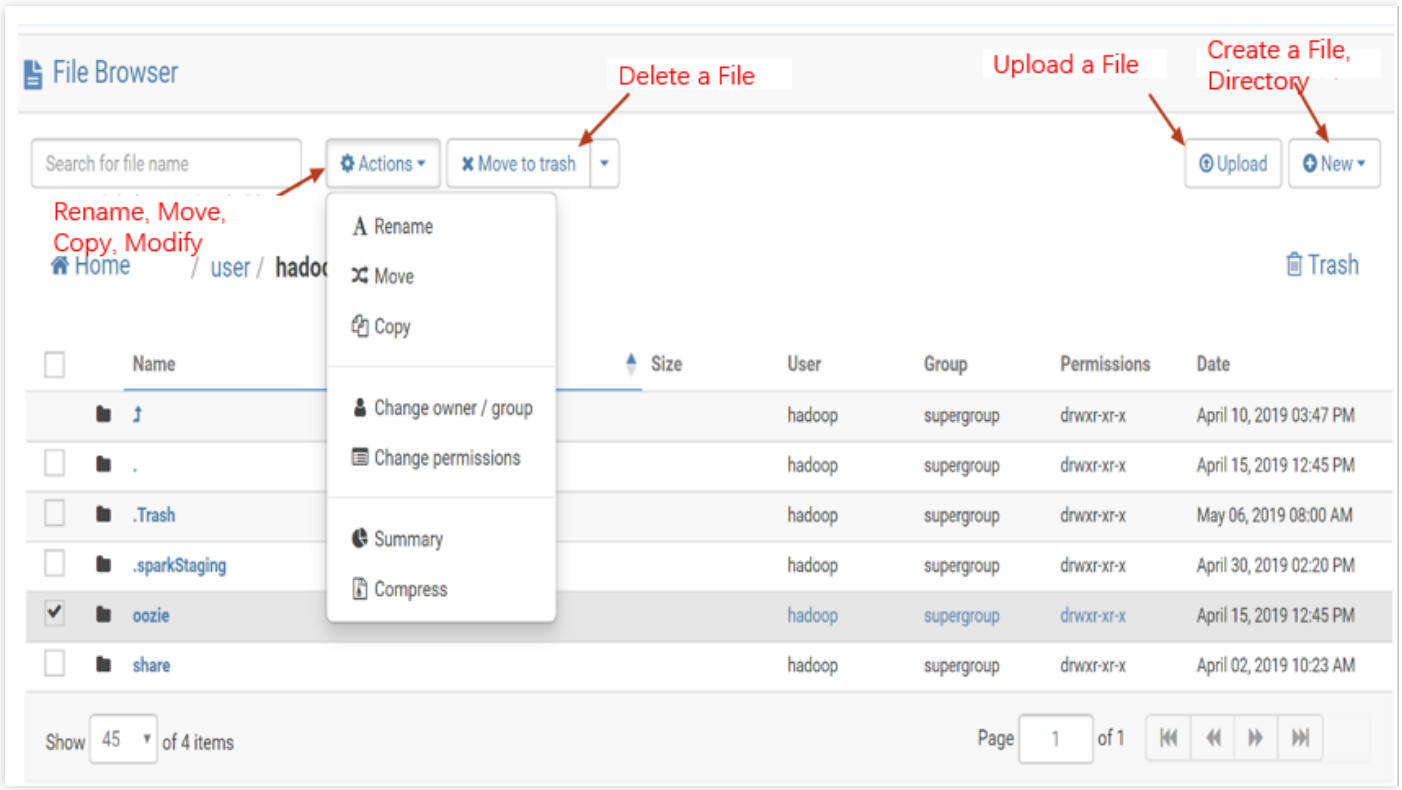
访问 HDFS 和文件浏览

通过 Hue 的 Web 页面可方便查看 HDFS 中的文件和文件夹，并对其进行创建、下载、上传、复制、修改和删除等操作。

1. 在 Hue 控制台左侧，选择 Browsers > Files 进入 HDFS 文件浏览。



2. 在 Hue 控制台左侧，选择 Browsers > Files 进入 HDFS 文件浏览。



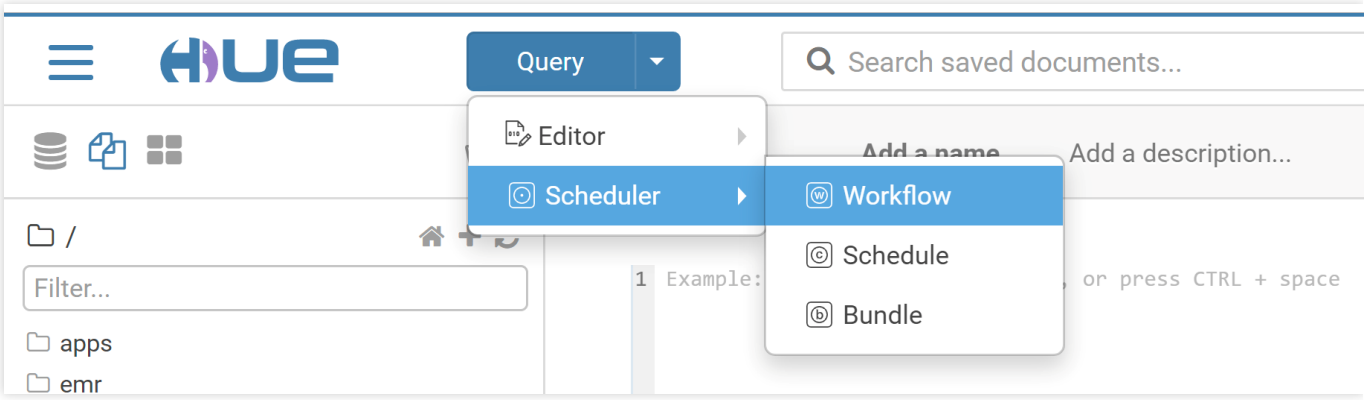
Oozie 任务的开发

1. 准备工作流数：Hue 的任务调度基于工作流，先创建一个包含 Hive script 脚本的工作流，Hive script 脚本的内容如下：

```
create database if not exists hive_sample;
show databases;
use hive_sample;
show tables;
create table if not exists hive_sample (a int, b string);
show tables;
insert into hive_sample select 1, "a";
select * from hive_sample;
```

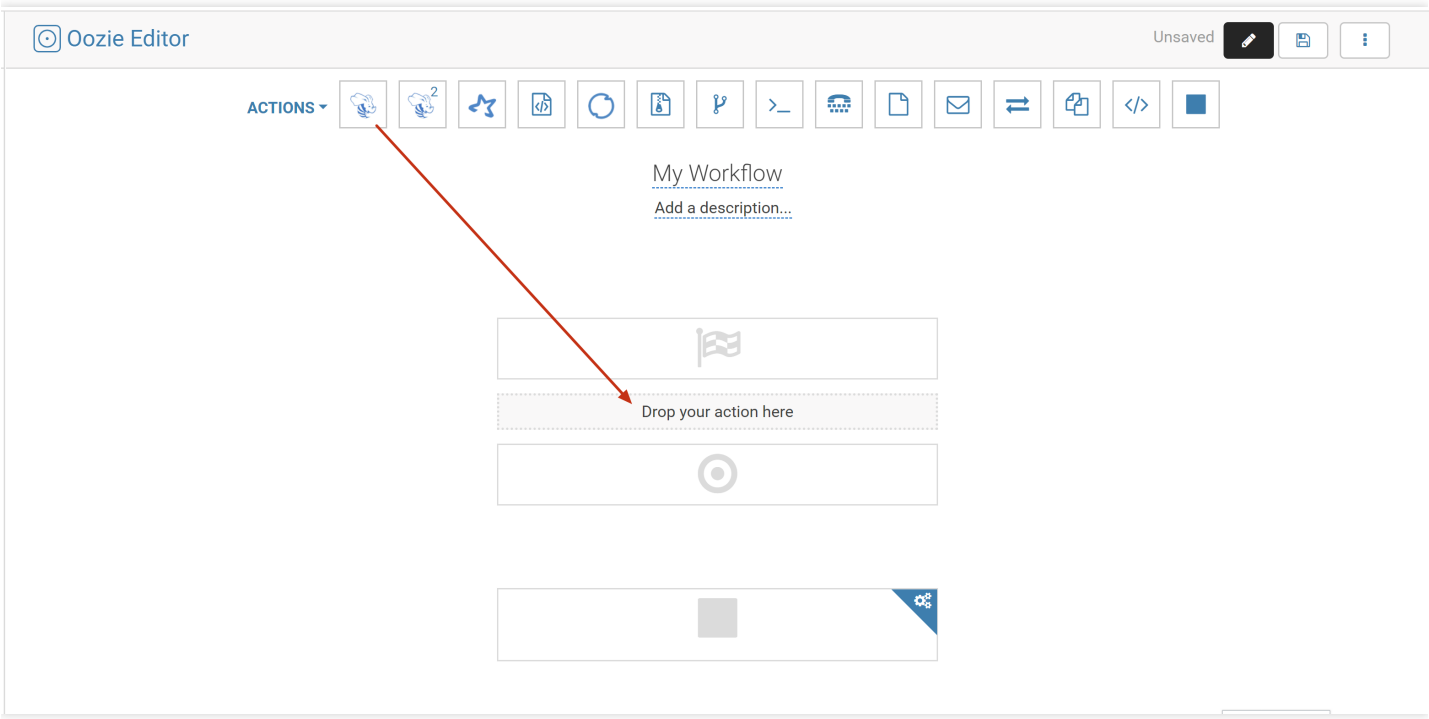
将以上内容保存为 hive_sample.sql 文件。Hive 工作流还需要一个 hive-site.xml 配置文件，此配置文件可以在集群中安装了 Hive 组件的节点上找到。具体路径：/usr/local/service/hive/conf/hive-site.xml，复制一个 hive-site.xml 文件。然后上传 Hive script 文件和 hive-site.xml 到 hdfs 的目录，例如：/user/hadoop。

2. 创建工作流。
   i. 切换到 hadoop 用户，在 Hue 页面上方，选择 Query > Scheduler > Workflow。

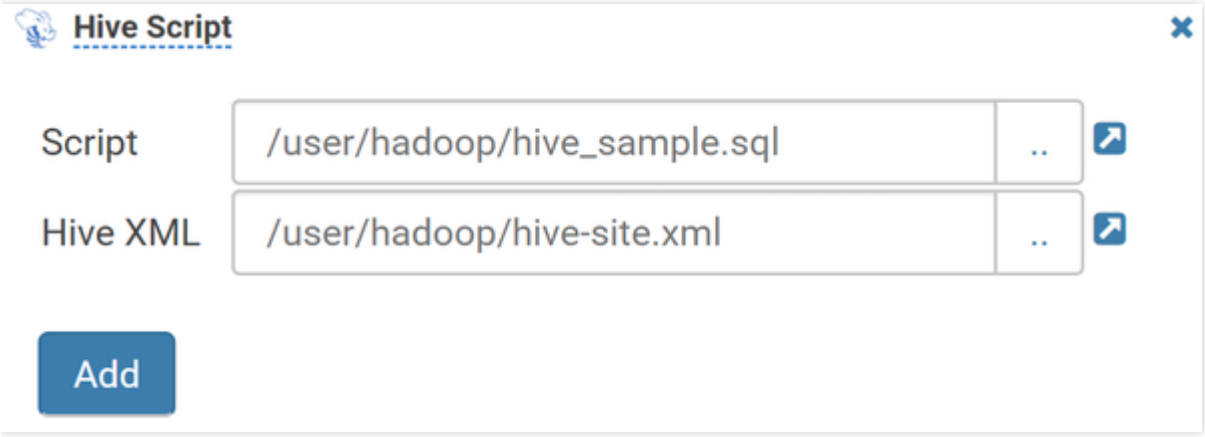


   ii. 在工作流编辑页面中拖一个 Hive Script。

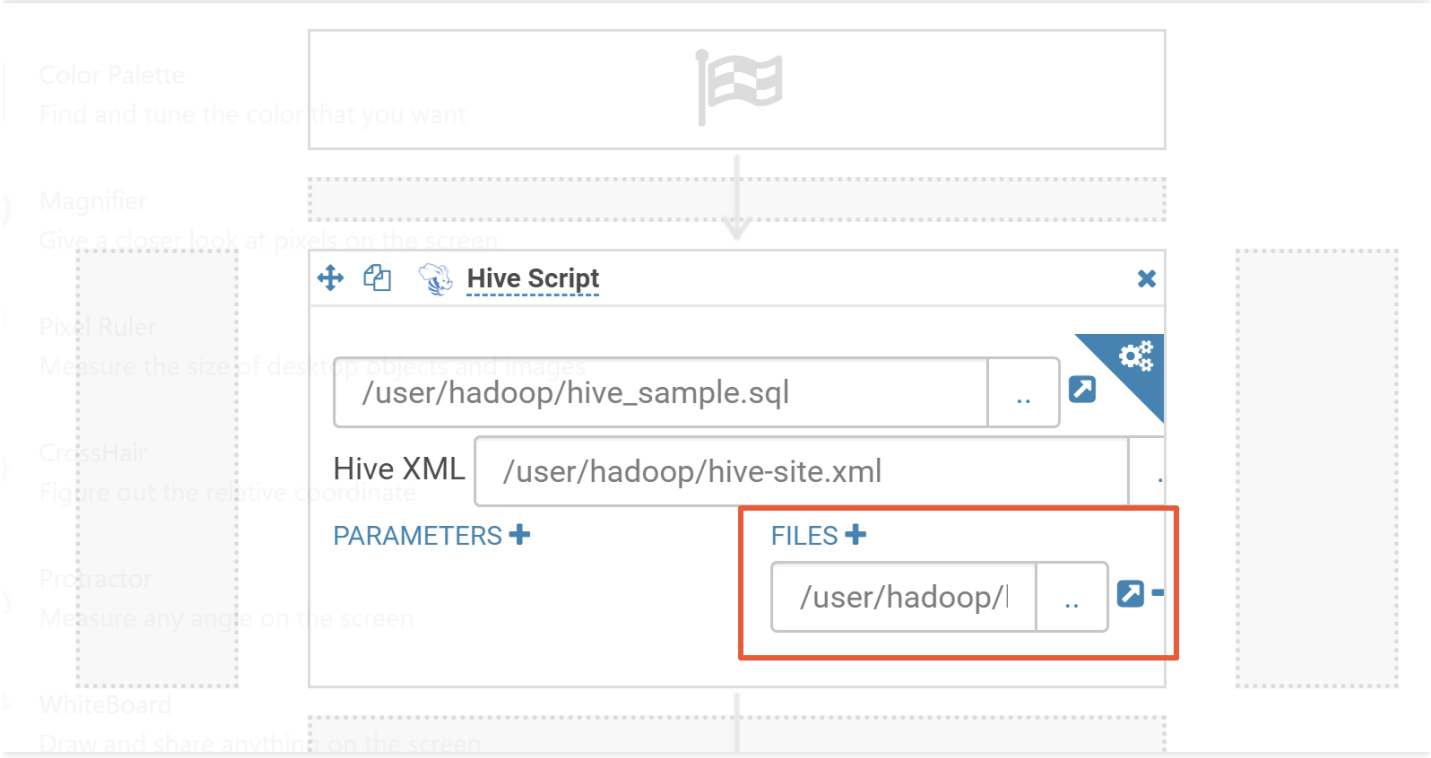
> 注意：
> 本文以安装 Hive 版本为 Hive1 为例，配置参数为 HiveServer1。与其他 Hive 版本混合部署时（即配置其他版本的配置参数时），会报错。



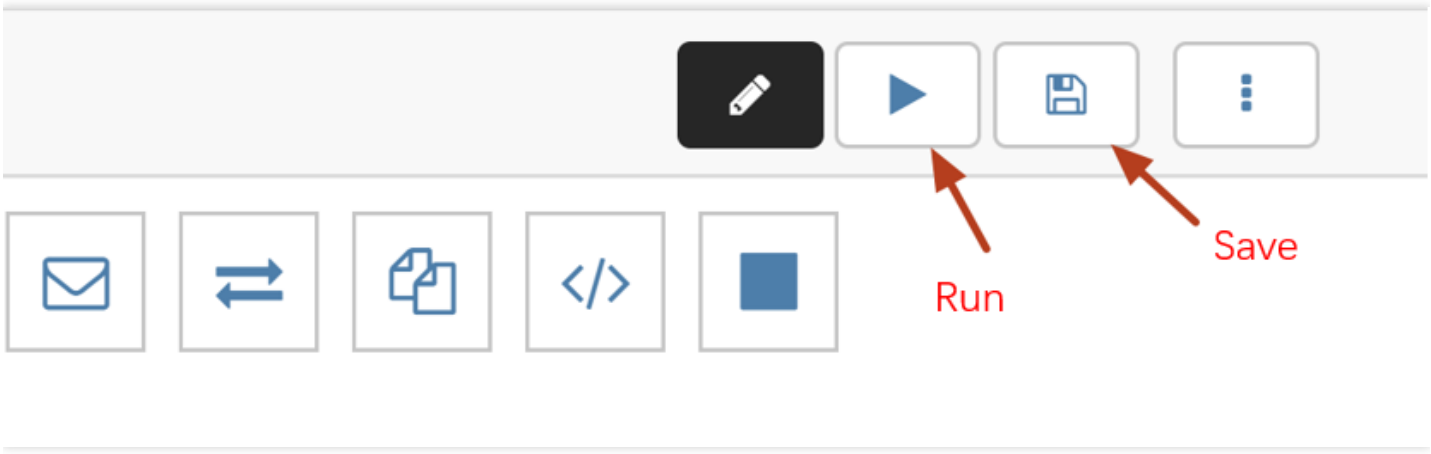
3. 选择刚上传的 Hive scipt 文件和 hive-site.xml 文件。



4. 单击 **Add** 后，还需在 FILES 中指定 hive script 文件。



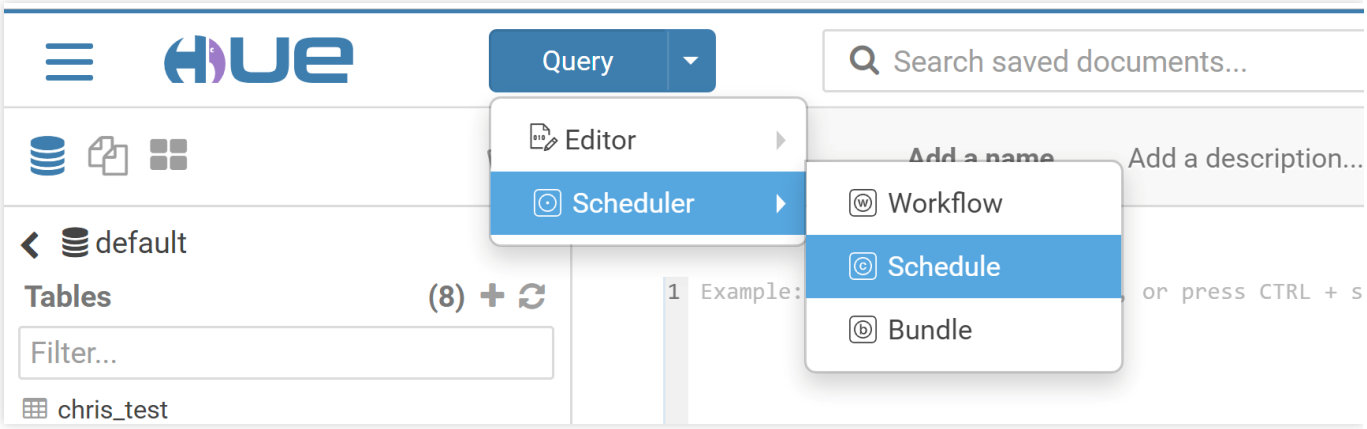
5. 单击右上角**保存**，然后单击**执行**，运行 workflow。



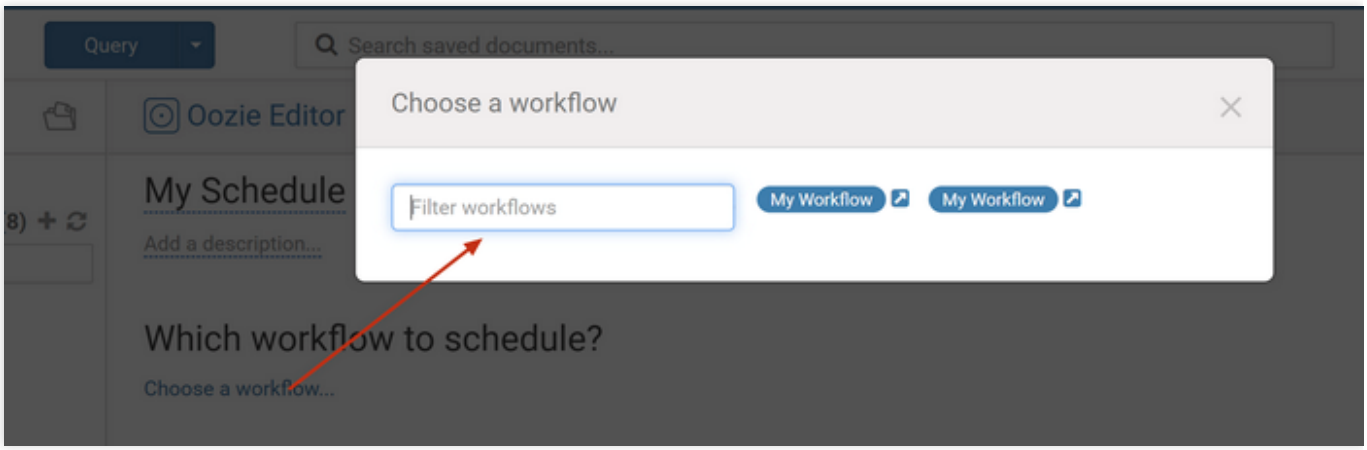
3. 创建定时调度任务。

Hue 的定时调度任务是 schedule，类似于 Linux 的 crontab，支持的调度粒度可以到分钟级别。

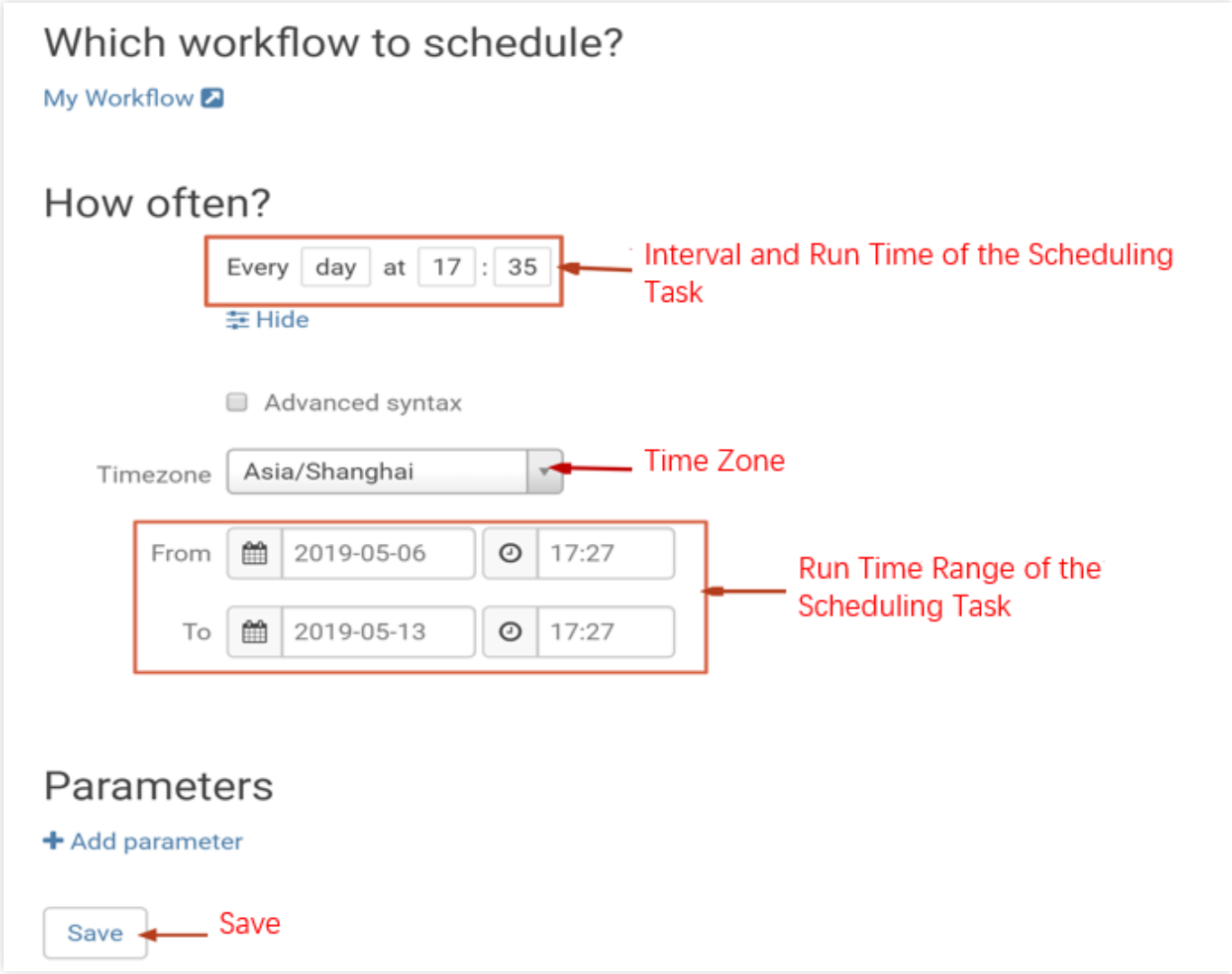
i. 选择 Query > Scheduler > Schedule，创建 Schedule。



ii. 单击 **Choose a workflow**，选择一个创建好的工作流。

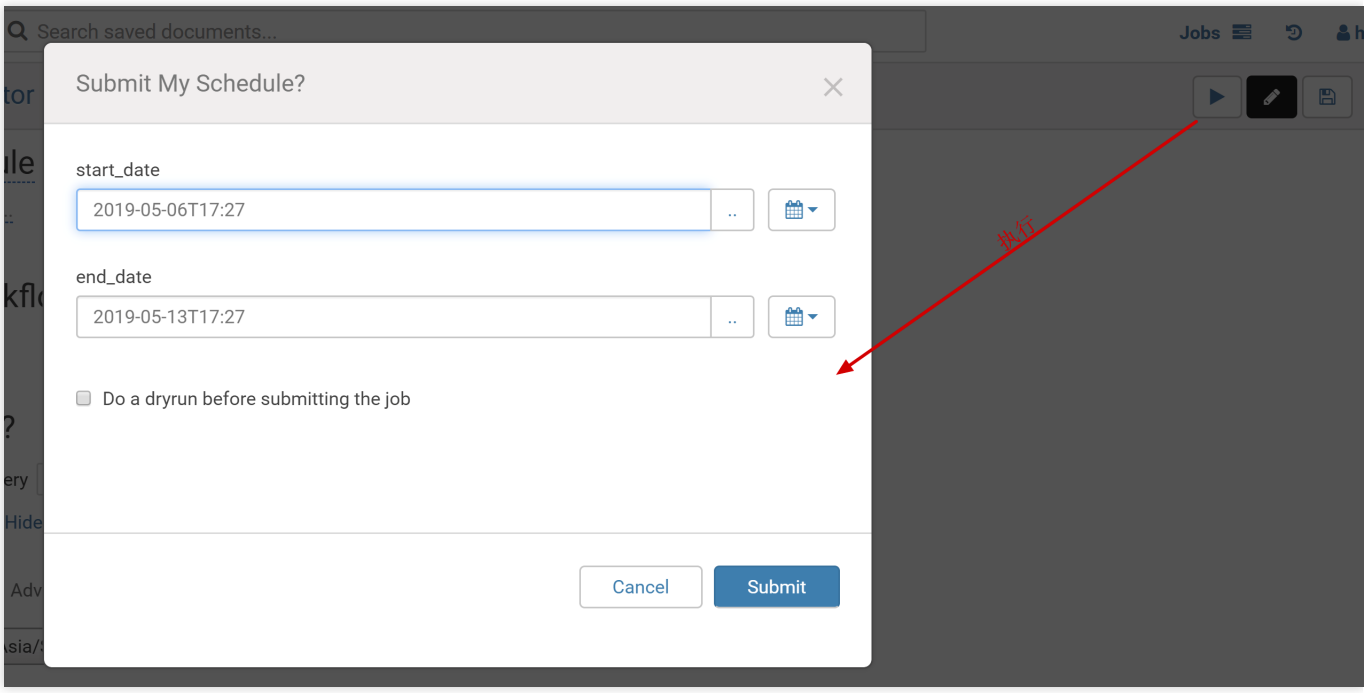


iii. 选择需要调度的时间点和时间间隔、时区、调度任务的开始时间和结束时间，然后单击 **Save** 保存。

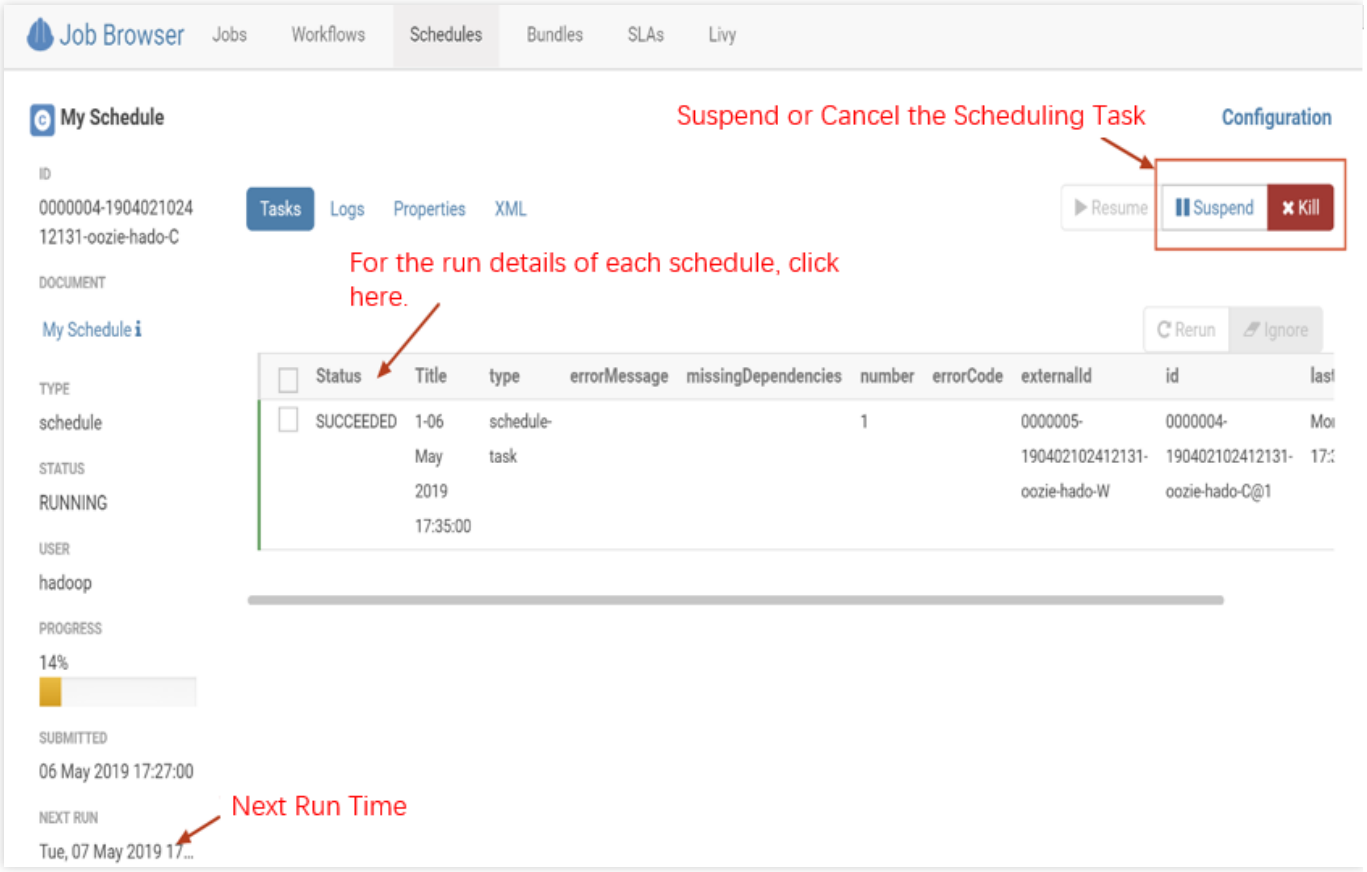


4. 创建定时调度任务。

i. 单击右上角的**提交**，提交调度任务。



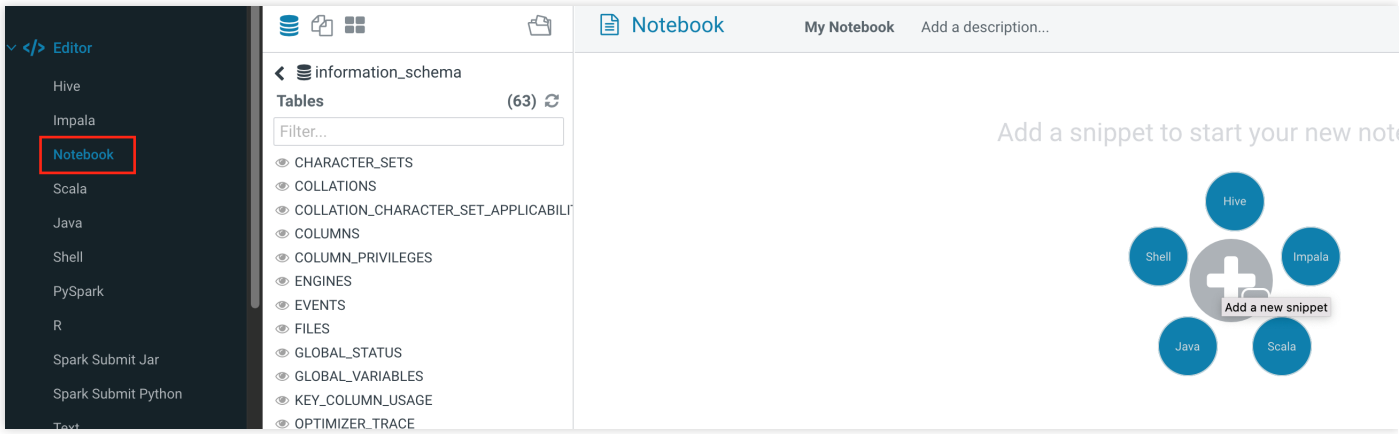
ii. 在 schedulers 的监控页面可以查看任务调度情况。



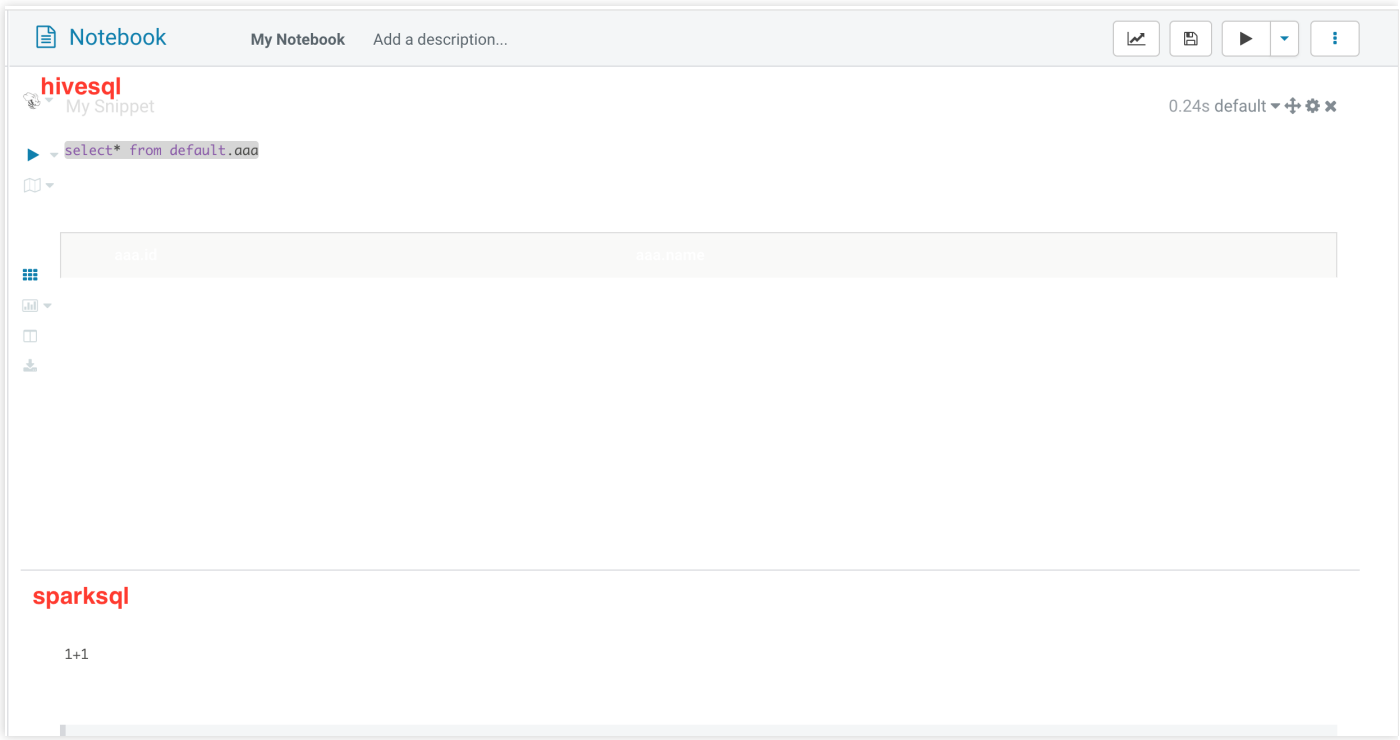
Notebook 查询对比分析

notebook 可以快速的构建间的查询，将查询结果放到一起做对比分析，支持5种类型：hive，impala，spark，java，shell。

1. 依次单击 **Editor**，**Notebook**，单击"+" 号添加需要的查询。



2. 依次单击**保存**可保存添加的notebook，单击**执行**可以执行整个 notebook。



Oozie 开发指南

最近更新时间：2022-05-16 12:52:25

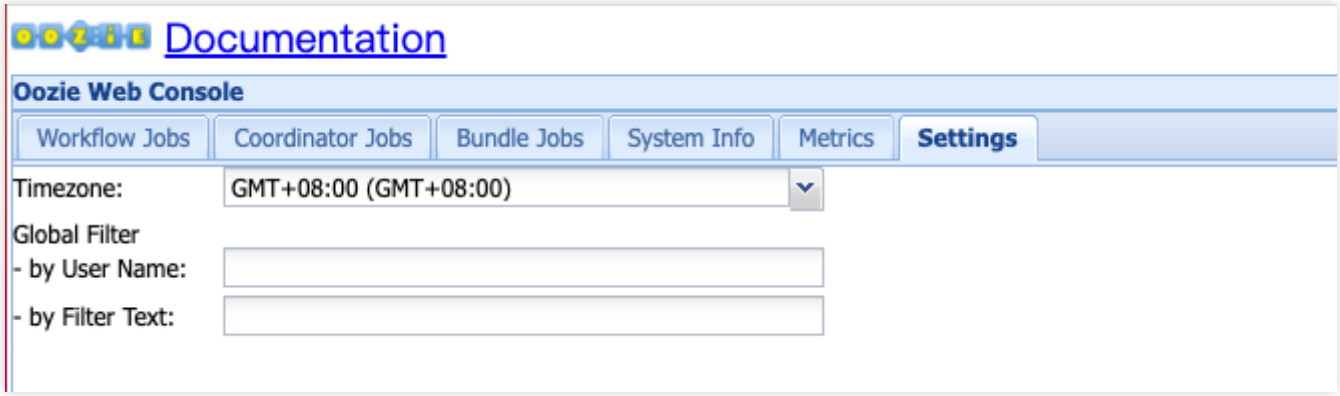
Apache Oozie 是一个开源的工作流引擎，被设计将 hadoop 生态组件的任务编排成 Workflow，然后对其进行调度、执行、监控。本文简单介绍如何在 EMR 上使用 Oozie，详细的使用文档请参考官网,另外这里建议用户通过 Hue 的图像化界面来使用 Oozie，使用文档请移步 Hue 开发文档。

前提条件

已创建弹性 MapReduce（简称EMR）的 Hadoop 集群，并选择了 Oozie 服务，详情请参见 创建 EMR 集群。

访问 Oozie WebUI

- 如果您在购买集群时勾选了开启集群节点外网，就可以在 EMR 控制台上通过单击 WebUI 链接来访问。
- 对于国内用户，建议将 WebUI 时区设置为 GMT+08:00。



sharelib 的更新

在 EMR 集群中，已安装了 sharelib，所以您使用 Oozie 提交 Workflow 作业时，不需要再安装 sharelib。当然您也可以对 sharelib 进行编辑与更新，操作步骤如下：

```
cd /usr/local/service/oozie
tar -xf oozie-sharelib.tar.gz添加jar包到解压出的share目录下要支持的action对应的目录下bin
/oozie-setup.sh sharelib create -fs hdfs://active-namenode-ip:4007 -locallib shar
eoozie admin --oozie http://oozie-server-ip:12000/oozie -sharelibupdate
```

在非 Kerberos 环境下提交 Workflow

在 oozie 的安装目录/usr/local/service/oozie，对文件 oozie-examples.tar.gz 进行解压，里面有 Oozie 支持的组件的 Workflow 示例：

```
tar -xf oozie-examples.tar.gz
```

这里以 action hive2来进行举例：

- su hadoop。
- cd examples/apps/hive2/。
- 修改 job.properties。
  - namenode 设置为 core-site.xml 下 `fs.defaultFS` 的值。
  - **resourceManager** 的值在 HA 模式下设置为 yarn-site.xml 下 `yarn.resourcemanager.ha.rm-ids` 的值，非 HA 模式下为 `yarn.resourcemanager.address` 的值。
  - **jdbcURL** 的值为 `jdbc:hive2://hive2-server:7001/default` 。
- hadoop fs -put examples。
- oozie job -debug -oozie http://oozie-server-ip:12000/oozie -config examples/apps/hive2/job.properties -run。
- oozie job -info 上一步返回的Job ID（或者通过WebUI查看） 。

在 Kerberos 环境下提交 Workflow

仍然以 action hive2 来进行举例，其它的注意事项请查看 hive2目录下的 README，此处不再赘述。

- kinit -kt /var/krb5kdc/emr.keytab hadoop 的 principal && su hadoop。
- cd examples/apps/hive2/。
- mv job.properties.security job.properties && mv workflow.xml.security workflow.xml。
- 修改 job.properties：
  - namenode 设置为 core-site.xml 下 `fs.defaultFS` 的值。
  - resourceManager 的值在 HA 模式下设置为 yarn-site.xml 下 `yarn.resourcemanager.ha.rm-ids` 的值，非 HA 模式下为 `yarn.resourcemanager.address` 的值。
  - jdbcURL 的值为 `jdbc:hive2://hive2-server:7001/default` 。
  - jdbcPrincipal 的值为 `hive.server2.authentication.kerberos.principal` 的值。
- hadoop fs -put examples。
- oozie job -debug -oozie http://oozie-server-ip:12000/oozie -config examples/apps/hive2/job.properties -run。
- oozie job -info 上一步返回的 Job ID（或者通过 WebUI 查看）。

Flume 开发指南

Flume 简介

最近更新时间：2021-09-30 16:37:10

Flume 简介

Apache Flume 是可以收集例如日志、事件等数据资源，并将这些数量庞大的数据从各项数据资源中集中起来存储的工具/服务。Flume 具有高可用、分布式、配置工具等特性，其设计原理也是将数据流（例如日志数据）从各种网站服务器上汇集起来存储到 HDFS、HBase 等集中存储器中。

Flume 架构

一个 Flume 事件被定义为一个数据流单元。Flume agent 其实是一个 JVM 进程，该进程中包含完成任务所需要的各个组件，其中最核心的三个组件是 Source、Channel 以及 Sink。

- **Source**
  消费外部源（例如 Web 服务器或者其他 Source）传递给它的事件，并将其保存到 Channel（一个或多个）中。
- **Channel**
  Channel 位于 Source 和 Sink 之间，用于缓存进来的 events，当 Sink 成功的将 events 发送到下一跳的 Channel 或最终目的，events 从 Channel 移除。
- **Sink**
  Sink 负责将 events 传输到下一跳或最终目的，成功完成后将 events 从 Channel 移除。

使用指南

使用准备

- 已创建一个 EMR 集群。创建 EMR 集群 时需要在软件配置界面选择 flume 组件。

- flume 安装在 EMR 云服务器（core 节点和 task 节点）的 `/usr/local/service/flume` 路径下；master 节点的安装路径是 `/usr/local/service/apps/` 。

配置 Flume

进入 `/usr/local/service/flume` 文件夹，并创建 example.conf 文件。

```
[hadoop@10 /usr/local/service/flume]$ cd /usr/local/service/flume/
[hadoop@10 /usr/local/service/flume]$ vim example.conf
[hadoop@10 /usr/local/service/flume]$
```

```
# example.conf: A single-node Flume configuration

# Name the components on this agent
a1.sources = r1
a1.sinks = k1
a1.channels = c1

# Describe/configure the source
a1.sources.r1.type = netcat
a1.sources.r1.bind = localhost
a1.sources.r1.port = 44444

# Describe the sink
a1.sinks.k1.type = logger

# Use a channel which buffers events in memory
a1.channels.c1.type = memory
a1.channels.c1.capacity = 1000
a1.channels.c1.transactionCapacity = 100

# Bind the source and sink to the channel
a1.sources.r1.channels = c1
a1.sinks.k1.channel = c1
```

启动 Flume

```
bin/flume-ng agent --conf conf --conf-file example.conf --name a1 -Dflume.root.logger=INFO,console
```

配置测试样例

配置后将会看到之前启动的 Flume Agent 向终端打印。

```
telnet localhost 44444
Trying 127.0.0.1...
Connected to localhost.localdomain (127.0.0.1).
Escape character is '^]'.
Hello world! <ENTER>
OK
```

Kafka 数据通过 Flume 存储到 Hive

最近更新时间：2023-07-12 10:47:30

场景说明

将 Kafka 中的数据通过 Flume 收集并存储到 Hive。

开发准备

- 因为任务中需要访问腾讯云消息队列 CKafka，所以需要先创建一个 CKafka 实例，具体见 消息队列 CKafka。
- 确认您已开通腾讯云，且已创建一个 EMR 集群。创建 EMR 集群时，需要在软件配置界面选择 Flume 组件。

在 EMR 集群使用 Kafka 工具包

首先需要查看 CKafka 的内网 IP 与端口号。登录消息队列 CKafka 的控制台，选择您要使用的 CKafka 实例，在基本消息中查看其内网 IP 为 $kafkaIP，而端口号一般默认为9092。在 topic 管理界面新建一个 topic 为 kafka_test。

配置 flume

1. 创建 flume 的配置文件 `hive_kafka.properties`

```
vim hive_kafka.properties
agent.sources = kafka_source
agent.channels = mem_channel
agent.sinks = hive_sink
# 以下配置 source
agent.sources.kafka_source.type = org.apache.flume.source.kafka.KafkaSource
agent.sources.kafka_source.channels = mem_channel
agent.sources.kafka_source.batchSize = 5000
agent.sources.kafka_source.kafka.bootstrap.servers = $kafkaIP:9092
agent.sources.kafka_source.kafka.topics = kafka_test
# 以下配置 sink
agent.sinks.hive_sink.channel = mem_channel
agent.sinks.hive_sink.type = hive
agent.sinks.hive_sink.hive.metastore = thrift://172.16.32.51:7004
agent.sinks.hive_sink.hive.database = default
agent.sinks.hive_sink.hive.table = weblogs
```

```
agent.sinks.hive_sink.hive.partition = asia,india,%y-%m-%d-%H-%M
agent.sinks.hive_sink.useLocalTimeStamp = true
agent.sinks.hive_sink.round = true
agent.sinks.hive_sink.roundValue = 10
agent.sinks.hive_sink.roundUnit = minute
agent.sinks.hive_sink.serializer = DELIMITED
agent.sinks.hive_sink.serializer.delimiter = ","
agent.sinks.hive_sink.serializer.serdeSeparator = ','
agent.sinks.hive_sink.serializer.fieldnames =id,msg
# 以下配置 channel
agent.channels.mem_channel.type = memory
agent.channels.mem_channel.capacity = 100000
agent.channels.mem_channel.transactionCapacity = 100000
```

其中 hive.metastore 可以通过以下方式确认：

```
grep "hive.metastore.uris" -C 2 /usr/local/service/hive/conf/hive-site.xml
```

```
<property>
<name>hive.metastore.uris</name>
<value>thrift://172.16.32.51:7004</value>
</property>
```

2. 创建 hive 表

```
create table weblogs ( id int , msg string )
partitioned by (continent string, country string, time string)
clustered by (id) into 5 buckets
stored as orc TBLPROPERTIES ('transactional'='true');
```

注意：
一定要是分区且分桶的表，存储为 orc 且设置 TBLPROPERTIES ('transactional'='true')，以上条件缺一不可。
3. 开启 hive 事务
在控制台给 `hive-site.xml` 添加以下配置项。

```
<property>
<name>hive.support.concurrency</name>
<value>true</value>
</property>
<property>
```

```
<name>hive.exec.dynamic.partition.mode</name>
<value>nonstrict</value>
</property>
<property>
<name>hive.txn.manager</name>
<value>org.apache.hadoop.hive.ql.lockmgr.DbTxnManager</value>
</property>
<property>
<name>hive.compactor.initiator.on</name>
<value>true</value>
</property>
<property>
<name>hive.compactor.worker.threads</name>
<value>1</value>
</property>
<property>
<name>hive.enforce.bucketing</name>
<value>true</value>
</property>
```

注意：
配置下发并重启后，在 `hadoop-hive` 日志中会提示 metastore 无法连接，请忽略该错误。由于进程启动顺序导致，需先启动 metastore 再启动 hiveserver2。

4. 复制 hive 的 `hive-hcatalog-streaming-xxx.jar` 到 flume 的 lib 目录

```
cp -ra /usr/local/service/hive/hcatalog/share/hcatalog/hive-hcatalog-streaming-2.3.3.jar /usr/local/service/flume/lib/
```

5. 运行 flume

```
./bin/flume-ng agent --conf ./conf/ -f hive_kafka.properties -n agent -Dflume.root.logger=INFO,console
```

6. 运行 kafka producer

```
[hadoop@172 kafka]$ ./bin/kafka-console-producer.sh --broker-list $kafkaIP:9092 --topic kafka_test
1,hello
2,hi
```

测试

- 在 kafka 生产者客户端输入信息并回车。
- 观察 hive 表中是否有相应数据。

参考文档

- hive-sink 配置说明
- hive 日志配置说明

Kafka 数据通过 Flume 存储到 HDFS 或 COS

最近更新时间：2023-07-12 10:30:31

场景说明

将 Kafka 中的数据通过 Flume 收集并存储到 HDFS 或 COS。

开发准备

- 因为任务中需要访问腾讯云消息队列 CKafka，所以需要先创建一个 CKafka 实例，具体见 消息队列 CKafka。
- 确认您已开通腾讯云，且已创建一个 EMR 集群。创建 EMR 集群时，需要在软件配置界面选择 Flume 组件，并且在基础配置页面开启对象存储的授权。

在 EMR 集群使用 Kafka 工具包

首先需要查看 CKafka 的内网 IP 与端口号。登录消息队列 CKafka 的控制台，选择您要使用的 CKafka 实例，在基本消息中查看其内网 IP 为 $kafkaIP，而端口号一般默认为9092。在 topic 管理界面新建一个 topic 为 kafka_test。

配置 flume

1. 创建 flume 的配置文件 `kafka.properties`

```
vim kafka.properties
agent.sources = kafka_source
agent.channels = mem_channel
agent.sinks = hdfs_sink
# 以下配置 source
agent.sources.kafka_source.type = org.apache.flume.source.kafka.KafkaSource
agent.sources.kafka_source.channels = mem_channel
agent.sources.kafka_source.batchSize = 5000
agent.sources.kafka_source.kafka.bootstrap.servers = $kafkaIP:9092
agent.sources.kafka_source.kafka.topics = kafka_test
# 以下配置 sink
agent.sinks.hdfs_sink.type = hdfs
agent.sinks.hdfs_sink.channel = mem_channel
agent.sinks.hdfs_sink.hdfs.path = /data/flume/kafka/%Y%m%d（或cosn://bucket/xx
```

```
x)
agent.sinks.hdfs_sink.hdfs.rollSize = 0
agent.sinks.hdfs_sink.hdfs.rollCount = 0
agent.sinks.hdfs_sink.hdfs.rollInterval = 3600
agent.sinks.hdfs_sink.hdfs.threadsPoolSize = 30
agent.sinks.hdfs_sink.hdfs.fileType=DataStream
agent.sinks.hdfs_sink.hdfs.useLocalTimeStamp=true
agent.sinks.hdfs_sink.hdfs.writeFormat=Text
# 以下配置 channel
agent.channels.mem_channel.type = memory
agent.channels.mem_channel.capacity = 100000
agent.channels.mem_channel.transactionCapacity = 10000
```

2. 运行 flume

```
./bin/flume-ng agent --conf ./conf/ -f kafka.properties -n agent -Dflume.root.l
ogger=INFO,console
```

3. 运行 kafka producer

```
[hadoop@172 kafka]$ ./bin/kafka-console-producer.sh --broker-list $kafkaIP:9092
--topic kafka_test
test
hello
```

测试

- 在 kafka 生产者客户端输入信息并回车。
- 观察 hdfs 是否生成相应目录和文件 `hadoop fs -ls /data/flume/kafka/` 。

参考文档

kafka-source 配置说明

Kafka 数据通过 Flume 存储到 Hbase

最近更新时间：2023-07-12 10:32:04

场景说明

将 Kafka 中的数据通过 Flume 收集并存储到 Hbase。

开发准备

- 因为任务中需要访问腾讯云消息队列 CKafka，所以需要先创建一个 CKafka 实例，具体见 消息队列 CKafka。
- 确认您已开通腾讯云，且已创建一个 EMR 集群。创建 EMR 集群时，需要在软件配置界面选择 Flume 组件。

在 EMR 集群使用 Kafka 工具包

首先需要查看 CKafka 的内网 IP 和端口号。登录消息队列 CKafka 的控制台，选择您要使用的 CKafka 实例，在基本消息中查看其内网 IP 为 $kafkaIP，而端口号一般默认为9092。在 topic 管理界面新建一个 topic 为 kafka_test。

配置 flume

1. 创建 flume 的配置文件 `hbase_kafka.properties`

```
vim hbase_kafka.properties
agent.sources = kafka_source
agent.channels = mem_channel
agent.sinks = hbase_sink
# 以下配置 source
agent.sources.kafka_source.type = org.apache.flume.source.kafka.KafkaSource
agent.sources.kafka_source.channels = mem_channel
agent.sources.kafka_source.batchSize = 5000
agent.sources.kafka_source.kafka.bootstrap.servers = $kafkaIP:9092
agent.sources.kafka_source.kafka.topics = kafka_test
# 以下配置 sink
agent.sinks.hbase_sink.channel = mem_channel
agent.sinks.hbase_sink.table = foo_table
agent.sinks.hbase_sink.columnFamily = cf
agent.sinks.hbase_sink.serializer = org.apache.flume.sink.hbase.RegexHbaseEvent
Serializer
```

```
# 以下配置 channel
agent.channels.mem_channel.type = memory
agent.channels.mem_channel.capacity = 100000
agent.channels.mem_channel.transactionCapacity = 10000
```

2. 创建 hbase 表

```
hbase shell
create 'foo_table','cf'
```

3. 运行 flume

```
./bin/flume-ng agent --conf ./conf/ -f hbase_kafka.properties -n agent -Dflume.root.logger=INFO,console
```

4. 运行 kafka producer

```
[hadoop@172 kafka]$ ./bin/kafka-console-producer.sh --broker-list $kafkaIP:9092 --topic kafka_test
hello
hbase_test
```

测试

- 在 kafka 生产者客户端输入信息并回车。
- 观察 hbase 表中是否有相应数据。

参考文档

hbase-sink 配置说明

Kerberos 开发指南
Kerberos 支持高可用

最近更新时间：2021-07-01 14:41:37

对于一个启用了 Kerberos 的正式生产系统，还需要考虑 KDC 的高可用。而 Kerberos 服务是支持配置为主备模式的，数据同步是通过 kprop 服务将主节点的数据同步到备节点。在购买了腾讯云 EMR 高可用安全集群后，Kerberos 默认是高可用的，用户无需任何配置。

本文主要介绍 Kerberos 服务高可用的相关配置和使用。

前提条件

- 已购买腾讯云 EMR 高可用集群。
- 购买集群时，已选择开启 Kerberos 认证。

KDC 高可用配置介绍

配置 `/etc/krb5.conf` 文件，设置如下：

> 注意：
> 示例中配置了两个 KDC 地址，active kdc 和 backup kdc，这样能保证当其中任意一个 KDC 服务异常时，仍可以对集群提供正常的 KDC 服务。

```
[libdefaults]
dns_lookup_realm = false
ticket_lifetime = 24h
renew_lifetime = 7d
forwardable = true
rdns = false
default_realm = REALM
default_tgs_enctypes = des3-cbc-sha1
default_tkt_enctypes = des3-cbc-sha1
permitted_enctypes = des3-cbc-sha1
[realms]
REALM = {
kdc = active_kdc:88
```

```
admin_server = active_kdc
kdc = backup_kdc:88
admin_server = backup_kdc
}
[domain_realm]
# .example.com = EXAMPLE.COM
```

KDC 数据同步

- kprop 配置

  在两个 kdc server 上默认有 `/var/kerberos/krb5kdc/kpropd.acl` 配置文件。

```
host/active_kdc@REALM
host/backup_kdc@REALM
```

2. 执行 `/var/kerberos/krb5kdc/kpropd.acl` 文件

   执行默认配置文件 `/var/kerberos/krb5kdc/kpropd.acl` 后，两个 kdc server 上会生成 `/etc/krb5.keyta` 文件。同时，两个 kdc server 也会启动 kprop 服务。

```
[root@10 krb5kdc]# service kprop status
Redirecting to /bin/systemctl status kprop.service
kprop.service - Kerberos 5 Propagation
Loaded: loaded (/usr/lib/systemd/system/kprop.service; disabled; vendor preset:
disabled)
Active: active (running) since Thu 2020-05-07 15:33:35 CST; 1h 9min ago
Process: 3752 ExecStart=/usr/sbin/_kpropd $KPROPD_ARGS (code=exited, status=0/S
UCCESS)
Main PID: 3753 (kpropd)
CGroup: /system.slice/kprop.service
└─3753 /usr/sbin/kpropd
```

3. 周期性同步 kdc 数据指令

   EMR agent 会周期性（默认5分钟）将 active kdc 的 principals 同步到 backup kdc，保证两个 kdc 上的数据是同步的。会在 active kdc 周期性执行如下同步指令：

```
kdb5_util dump /var/kerberos/krb5kdc/master.dump & kprop -f /var/kerberos/krb5k
dc/master.dump -d -P 754 backup_kdc
```

结果验证

1. 在 active kdc 上执行如下命令。

```
kadmin.local "-q addprinc -randkey test"
```

2. 在 backup kdc 上可以看到新增的 principal，说明 kdc 之间已完成数据同步，这里需要等待一段时间（默认5分钟）。

```
kadmin.local "-q listprincs"|grep test
```

Kerberos 简介

最近更新时间：2020-04-03 15:05:13

当前仅 EMR-V2.1.0 版本支持创建安全类型集群，即集群中的开源组件以 Kerberos 的安全模式启动，在这种安全环境下只有经过认证的客户端（Client）才能访问集群的服务（Service，例如 HDFS）。

当前 EMR-V2.1.0 版本支持的 Kerderos 的组件列表如下所示：

组件名称	组件版本
Hadoop	2.8.4
Hbase	1.3.1
Hive	2.3.3
Hue	4.4.0
Ooize	4.3.1
Presto	0.7.1
Zookeeper	3.4.9

重要概念

- **KDC**
  全称：key distributed center
  作用：整个安全认证过程的票据生成管理服务，其中包含两个服务：AS 和 TGS。
- **AS**
  全称：authentication service
  作用：为 client 生成 TGT 的服务。
- **TGS**
  全称：ticket granting service
  作用：为 client 生成某个服务的 ticket。
- **AD**
  全称：account database
  作用：存储所有 client 的白名单，只有存在于白名单的 client 才能顺利申请到 TGT。
- **TGT**
  全称：ticket-granting ticket

作用：用于获取 ticket 的票据。

- **client**
  想访问某个 server 的客户端。
- **server**
  提供某种业务的服务。

其他概念

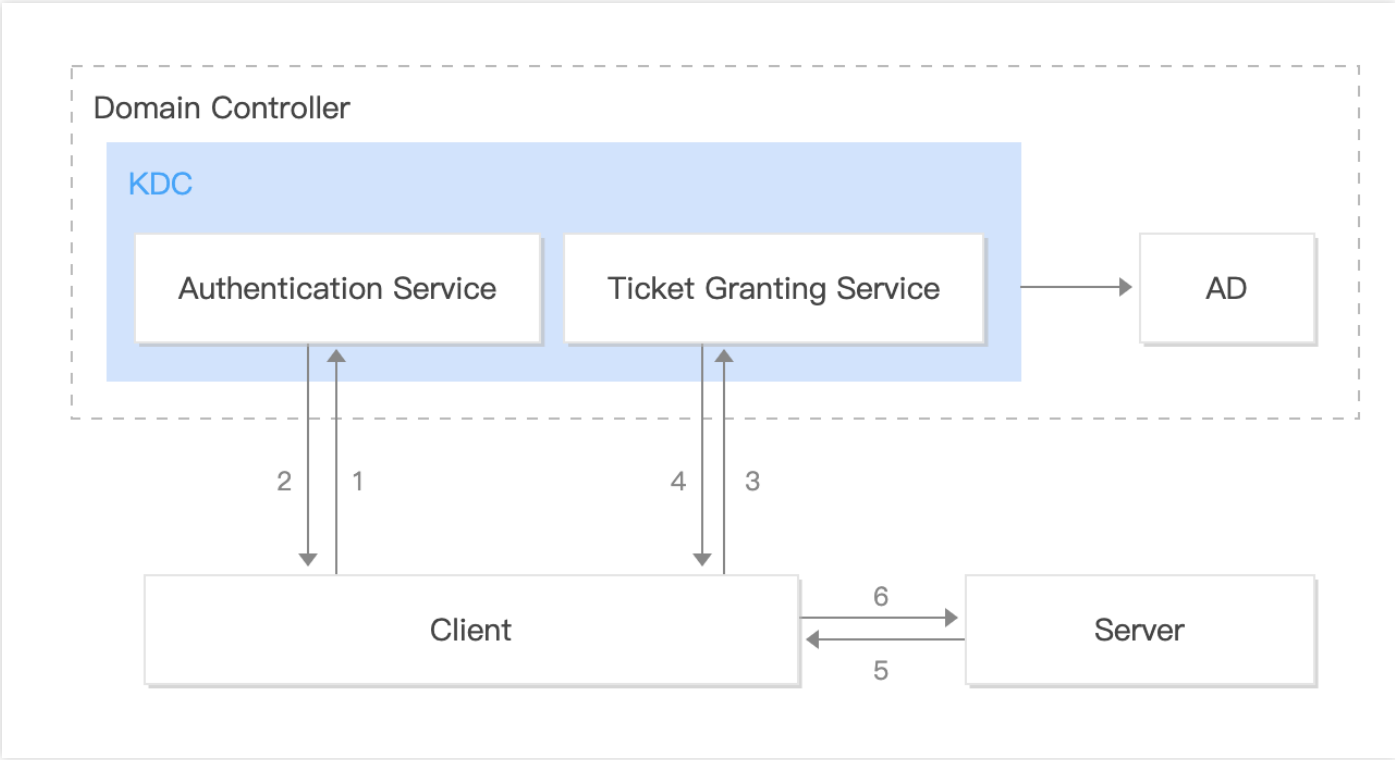
- **principal**
  认证的主体，即“用户名”。
- **realm**
  realm 有点像编程语言中的 namespace。在编程语言中，变量名只有在某个 namespace 里才有意义。同样的，一个 principal 只有在某个 realm 下才有意义。所以 realm 可以看成是 principal 的一个“容器”或者“空间”。
  相对应的，principal 的命名规则是 what_name_you_like@realm。
  在 kerberos，约定成俗用大写来命名 realm，例如：EXAMPLE.COM。
- **password**
  某个用户的密码，对应于 kerberos 中的 master_key。password 可以存在一个 keytab 文件中。所以 kerberos 中需要使用密码的场景都可以用一个 keytab 作为输入。
- **credential**
  credential 是“证明某个人确定是他自己/某一种行为的确可以发生”的凭据。在不同的使用场景下，credential 的具体含义也略有不同：
  - 对于某个 principal 个体而言，他的 credential 就是他的 password。
  - 在 kerberos 认证的环节中，credential 就意味着各种各样的 ticket。

认证流程

client 访问 server 的过程中，想确保 client 和 server 都是可靠的，必然要引入第三方公证平台，因此有了 AS 和 TGS 这两个服务，AS 与 TGS 通常位于同一个服务进程中，对于 mit 的 kerberos 实现，均由 kdc 提供。

认证流程分为以下三步：

1. client 向 kerberos 服务请求，希望获取访问 server 的权限。kerberos 首先判断 client 是否可信赖，通过在 AD 中存储黑名单和白名单来区分 client。成功后，AS 返回 TGT 给 client。
2. client 得到了 TGT 后，继续向 kerberos 请求，希望获取访问 server 的权限。kerberos 通过 client 消息中的 TGT，判断 client 拥有权限，给 client 访问 server 的权限 ticket。
3. client 得到 ticket 后，就可以访问 server 了，但这个 ticket 只是针对这个 server，访问其他 server 需要重新向 TGS 申请。


Domain Controller
KDC
Authentication Service
Ticket Granting Service
AD
2
1
4
3
Client
6
5
Server

Kerberos 使用说明

最近更新时间：2021-10-29 10:16:53

本文使用 mit 的 kerberos 来作为 kdc 服务。假设 kdc 服务已安装并启动，使用 kerberos，首先要创建域（realm），再添加相关角色的 principal（包括 server 和 client），然后生成 keytab 文件。

创建数据库

使用 `kdb5_util` 命令创建数据库，存放 principal 相关的信息。

```
kdb5_util -r EXAMPLE.COM create -s
Initializing database '/var/krb5/principal' for realm 'EXAMPLE.COM'
master key name 'K/M@EXAMPLE.COM'
You will be prompted for the database Master Password.
It is important that you NOT FORGET this password.
Enter KDC database master key: <Type the key>
Re-enter KDC database master key to verify: <Type it again>
```

添加 principal

```
kadmin.local
kadmin.local: add_principal -pw testpassword test/host@EXAMPLE.COM
WARNING: no policy specified fortest/host@EXAMPLE.COM; defaulting to no policy
Principal "test/host@EXAMPLE.COM" created.
```

生成密钥表文件

```
kadmin.local
kadmin.local: ktadd -k /var/krb5kdc/test.keytab test/host@EXAMPLE.COM
Entry for principal test/host@EXAMPLE.COM with kvno 2, encryption type des3-cbc-s
ha1 added to keytab WRFILE:/var/krb5kdc/test.keytab.
```

这里，我们创建了新的用户：test/host@EXAMPLE.COM ，并且将这个用户的密钥放置到 `/var/krb5kdc/test.keytab` 文件中。

启动 kdc

```
service krb5-kdc start
* Starting Kerberos KDC krb5kdc
```

kinit 验证

```
kinit -k -t /etc/krb5.keytab test-client/host@EXAMPLE.COM
```

kinit 对应的是向 kdc 获取 TGT 的步骤。它会向 `/etc/krb5.conf` 中指定的 kdc server 发送请求。如果 TGT 请求成功，使用 klist 即可看到。

```
klist
Ticket cache: FILE:/tmp/krb5cc_1000
Default principal: test-client/host@EXAMPLE.COM
Valid starting Expires Service principal
2019-01-15T17:50:25 2019-01-16T17:50:25 krbtgt/EXAMPLE.COM@EXAMPLE.COM
renew until 2019-01-16T00:00:25
```

在项目中使用

使用 kinit 验证成功后，可以把 keytab 文件复制到需要使用的 server 和 client 的服务器上，并配置相应的 principal 进行使用。

访问安全集群的 Hadoop

最近更新时间：2021-10-29 10:18:34

Hadoop 命令

未获取 ticket

当已启用 kerberos 时，执行 hadoop 命令时都需要提前获取 ticket。如果没有获取 ticket，则会出现如下错误信息：

```
hadoop fs -ls /
19/04/19 19:59:03 WARN ipc.Client: Exception encountered while connecting to the
server : javax.security.sasl.SaslException: GSS initiate failed [Caused by GSSExc
eption: No valid credentials provided (Mechanism level: Failed to find any Kerber
os tgt)]
ls: Failed on local exception: java.io.IOException: javax.security.sasl.SaslExcep
tion: GSS initiate failed [Caused by GSSException: No valid credentials provided
(Mechanism level: Failed to find any Kerberos tgt)]; Host Details : local host is
: "172.30.0.27/172.30.0.27"; destination host is: "172.30.0.27":4007;
```

获取 ticket

前提：hadoop@EMR 的 principal 已添加。

```
kinit -kt /var/krb5kdc/emr.keytab hadoop@EMR
```

执行命令即可正常访问。

```
hadoop fs -ls /
Found 8 items
-rw-r--r-- 3 hadoop supergroup 3809 2019-03-06 11:10 /README.md
drwxr-xr-x - hadoop supergroup 0 2019-01-14 21:43 /apps
drwxrwx--- - hadoop supergroup 0 2019-01-17 19:46 /emr
drwxr-xr-x - hadoop supergroup 0 2019-01-23 20:02 /hbase
drwxr-xr-x - hadoop supergroup 0 2019-03-07 11:10 /spark-history
drwx-wx-wx - hadoop supergroup 0 2019-03-06 20:23 /tmp
drwxr-xr-x - hadoop supergroup 0 2019-01-17 14:43 /user
drwxr-xr-x - hadoop supergroup 0 2019-01-17 19:43 /usr
```

Java 代码访问 HDFS

使用本地 ticket

> 注意：
> 需要提前执行 kinit 获取 ticket，ticket 过期后程序会访问异常。

```
public static void main(String[] args) throws IOException {
Configuration conf = new Configuration();
conf.addResource(new Path("/usr/local/service/hadoop/etc/hadoop/hdfs-site.xml"));
conf.addResource(new Path("/usr/local/service/hadoop/etc/hadoop/core-site.xml"));
UserGroupInformation.setConfiguration(conf);
UserGroupInformation.loginUserFromSubject(null);
FileSystem fileSystem = FileSystem.get(conf);
FileStatus[] fileStatus = fileSystem.listStatus(new Path("/"));
for (int i = 0; i < fileStatus.length; i++) {
System.out.println(fileStatus[i].getPath());
}
}
```

使用 keytab 文件

```
public static void main(String[] args) throws IOException {
Configuration conf = new Configuration();
conf.addResource(new Path("/usr/local/service/hadoop/etc/hadoop/hdfs-site.xml"));
conf.addResource(new Path("/usr/local/service/hadoop/etc/hadoop/core-site.xml"));
UserGroupInformation.setConfiguration(conf);
UserGroupInformation.loginUserFromKeytab("hadoop@EMR", "/var/krb5kdc/emr.keytab"
);
FileSystem fileSystem = FileSystem.get(conf);
FileStatus[] fileStatus = fileSystem.listStatus(new Path("/"));
for (int i = 0; i < fileStatus.length; i++) {
System.out.println(fileStatus[i].getPath());
}
}
```

Hadoop 接入 kerberos 示例

最近更新时间：2021-10-29 10:19:13

本文介绍 Hadoop 如何修改配置接入 kerberos。如果是通过腾讯云 EMR 购买的安全集群，系统会自动配置好，无需自行配置。

前提条件

- kdc 服务已搭建成功。
- Hadoop 相关的 principals 已创建完成。
- keytab 文件分发到了各台服务器上（假设 keytab 文件路径为 `/var/krb5kdc/emr.keytab` ）。

Hadoop 接入 kerberos

Hadoop 主要包含 HDFS 和 Yarn 服务，需要分别修改这两部分配置并重启服务进程。

HDFS 接入

修改 core-site.xml

```
hadoop.security.authentication: kerberos
hadoop.security.authorization: true
```

修改 hdfs-site.xml

```
dfs.namenode.kerberos.principal: hadoop/_HOST@EMR
dfs.namenode.keytab.file: /var/krb5kdc/emr.keytab
dfs.namenode.kerberos.internal.spnego.principal: HTTP/_HOST@EMR
dfs.secondary.namenode.kerberos.principal: hadoop/_HOST@EMR
dfs.secondary.namenode.keytab.file: /var/krb5kdc/emr.keytab
dfs.secondary.namenode.kerberos.internal.spnego.principal: HTTP/_HOST@EMR
dfs.journalnode.kerberos.principal: hadoop/_HOST@EMR
dfs.journalnode.keytab.file: /var/krb5kdc/emr.keytab
dfs.journalnode.kerberos.internal.spnego.principal: HTTP/_HOST@EMR
dfs.datanode.kerberos.principal: hadoop/_HOST@EMR
dfs.datanode.keytab.file: /var/krb5kdc/emr.keytab
dfs.datanode.data.dir.perm: 700
dfs.web.authentication.kerberos.keytab: /var/krb5kdc/emr.keytab
```

```
dfs.web.authentication.kerberos.principal: HTTP/_HOST@EMR
ignore.secure.ports.for.testing: true
```

注意：
ignore.secure.ports.for.testing 选项必须设置为 true，否则必须配置 sasl 模式，且 webhdfs 必须启用 HTTPS。

修改 httpfs-site.xml（如果启用 httpfs）

```
httpfs.authentication.type: kerberos
httpfs.hadoop.authentication.type: kerberos
httpfs.authentication.kerberos.principal: HTTP/_HOST@EMR
httpfs.hadoop.authentication.kerberos.principal: hadoop/_HOST@EMR
httpfs.authentication.kerberos.keytab: /var/krb5kdc/emr.keytab
httpfs.hadoop.authentication.kerberos.keytab: /var/krb5kdc/emr.keytab
```

Yarn 接入

修改 yarn-site.xml

```
yarn.resourcemanager.keytab: /var/krb5kdc/emr.keytab
yarn.resourcemanager.principal: hadoop/_HOST@EMR
yarn.nodemanager.keytab: /var/krb5kdc/emr.keytab
yarn.nodemanager.principal: hadoop/_HOST@EMR
```

修改 mapred-site.xml

```
mapreduce.jobhistory.keytab: /var/krb5kdc/emr.keytab
mapreduce.jobhistory.principal: hadoop/_HOST@EMR
```

Knox 开发指南
Knox 指引

最近更新时间：2021-12-17 15:37:02

目前 EMR-V1.3.1 和 EMR-V2.0.1 版本已经支持 Apache Knox，完成以下准备工作后，即可在公网直接访问 Yarn、HDFS 等服务的 Web UI。

准备工作

- 确认您已开通腾讯云，并且创建了一个 EMR 集群。
- EMR-V1.3.1 和 EMR-V2.0.1 默认在创建集群时，Knox 作为必选组件。如果您的集群是老集群，目前可通过 联系客服 安装 Knox。

开始访问 Knox

使用集群公网 IP 地址访问。建议修改集群拥有公网 IP 的 CVM 安全组规则，限定 TCP:30002 端口的访问 IP 端为您的 IP 端。

1. 通过集群详情查看公网 IP。
2. 在浏览器中访问相应服务的 URL。
   - HDFS UI：https://{集群公网ip}:30002/gateway/emr/hdfs
   - Yarn UI：https://{集群公网ip}:30002/gateway/emr/yarn
   - Hive UI：https://{集群公网ip}:30002/gateway/emr/hive
   - Hbase UI：https://{集群公网ip}:30002/gateway/emr/hbase/webui
   - HUE UI：https://{集群公网ip}:30002/gateway/emr/hue
   - Storm UI：https://{集群公网ip}:30002/gateway/emr/stormui
   - Ganglia UI：https://{集群公网ip}:30002/gateway/emr/ganglia/
   - Presto UI：https://{集群公网ip}:30002/gateway/emr/presto/
   - Oozie UI：https://{集群公网ip}:30002/gateway/emr/oozie/
3. 浏览器显示**您的链接不是私密链接**，是因为 Knox 服务使用了自签名证书，请再次确认访问的是自己集群的公网 IP，并且端口为30002。选择**高级>继续前往**。
4. 弹出的验证登录框，用户名为 root，默认密码为创建集群时输入的密码。这里建议您修改密码，可以在该页面中单击**重置原生UI密码**进行修改。

Knox 集成 tez 操作指南

最近更新时间：2022-05-16 12:52:25

早期 EMR 产品版本未提供 Tez-UI 集成必要的 Timelineserver 软件包和 Tomcat 软件包，需要根据本文进行操作。

> 注意：
> 由于 Tez 的 UI 启用需要运行 Timelineserver，Timelineserver 运行时有较高资源占用，需要评估对业务的影响，谨慎启用。

本文主要介绍 Knox 集成 tez 的具体操作步骤，主要有安装 tomcat 和 tez-ui、新建 role、配置 timelineserver、配置 tez和启动服务。其中，`172.**.**.9` 为主节点内网 IP，`159.**.**.70` 为主节点外网 IP，tez 版本为0.9.2 版本。

安装 Tomcat 和 tez-ui

```
cd /usr/local/service
wget https://jaydihu-package-1258469122.cos.ap-guangzhou.myqcloud.com/apache-tomcat-9.0.46.tar.gz
tar -zxvf apache-tomcat-9.0.46.tar.gz
mv /usr/local/service/apache-tomcat-9.0.46 /usr/local/service/tomcat
```

修改 tomcat 端口号：

```
vim /usr/local/service/tomcat/conf/server.xml
```

第一处：由8005修改为2020

```
      http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License.
-->
<!-- Note:  A "Server" is not itself a "Container", so you may not
     define subcomponents such as "Valves" at this level.
     Documentation at /docs/config/server.html
 -->
<Server port="2020" shutdown="SHUTDOWN">
  <Listener className="org.apache.catalina.startup.VersionLoggerListener" />
  <!-- Security listener. Documentation at /docs/config/listeners.html
  <Listener className="org.apache.catalina.security.SecurityListener" />
  -->
  <!--APR library loader. Documentation at /docs/apr.html -->
  <Listener className="org.apache.catalina.core.AprLifecycleListener" SSLEngine="on" />
  <!-- Prevent memory leaks due to use of particular java/javax APIs-->
  <Listener className="org.apache.catalina.core.JreMemoryLeakPreventionListener" />
  <Listener className="org.apache.catalina.mbeans.GlobalResourcesLifecycleListener" />
  <Listener className="org.apache.catalina.core.ThreadLocalLeakPreventionListener" />

  <!-- Global JNDI resources
       Documentation at /docs/jndi-resources-howto.html
  -->
  <GlobalNamingResources>
    <!-- Editable user database that can also be used by
         UserDatabaseRealm to authenticate users
```

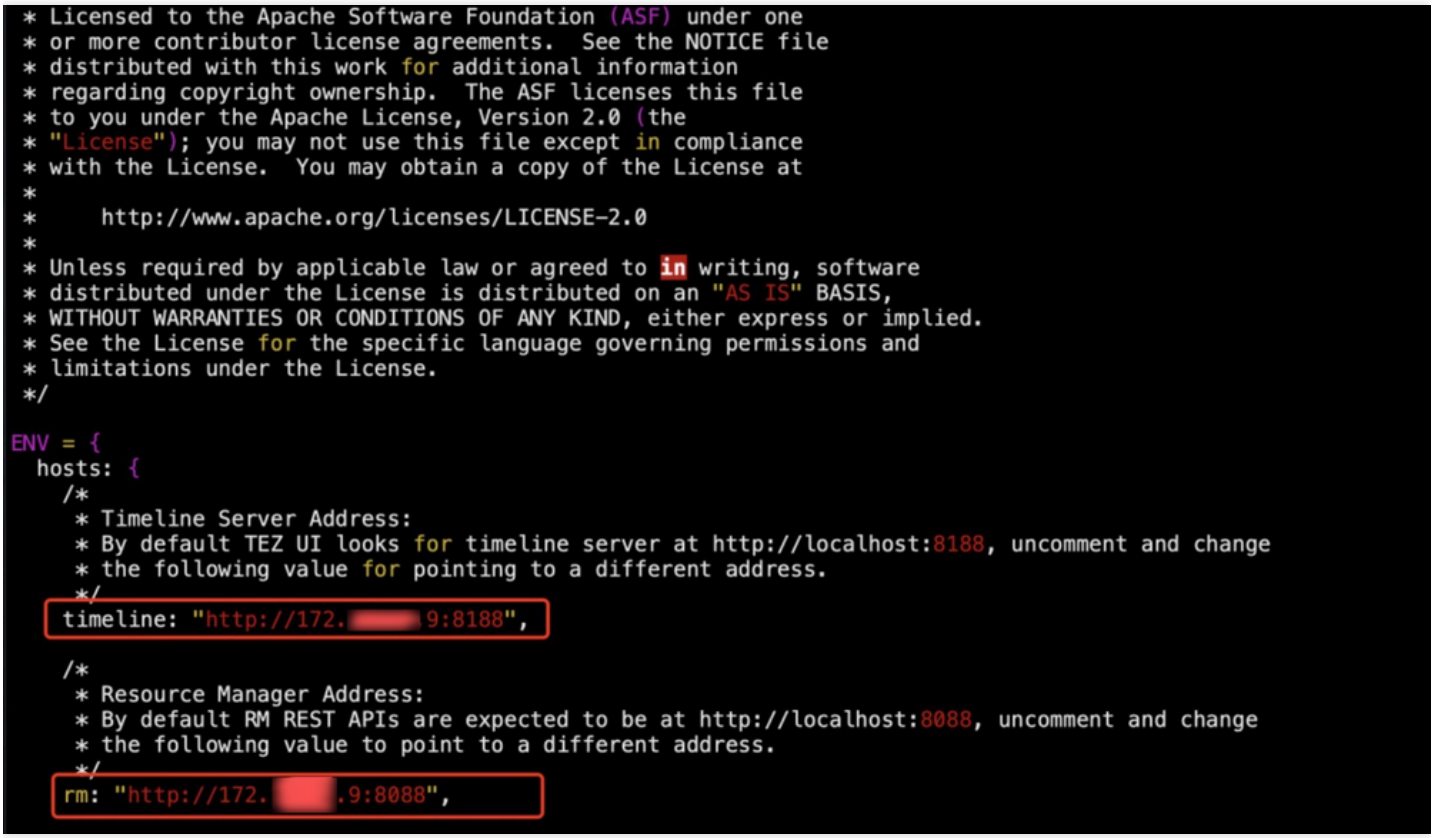
第二处：端口由8080改为2000

```
    <!-- A "Connector" represents an endpoint by which requests are received
         and responses are returned. Documentation at :
         Java HTTP Connector: /docs/config/http.html
         Java AJP  Connector: /docs/config/ajp.html
         APR (HTTP/AJP) Connector: /docs/apr.html
         Define a non-SSL/TLS HTTP/1.1 Connector on port 8080
    -->
    <Connector port="2000" protocol="HTTP/1.1"
               connectionTimeout="20000"
               redirectPort="8443" />
    <!-- A "Connector" using the shared thread pool-->
```

```
mkdir -p /usr/local/service/tomcat/webapps/tez-ui
cp /usr/local/service/tez/tez-ui-0.9.2.war /usr/local/service/tomcat/webapps/tez-ui/
cd /usr/local/service/tomcat/webapps/tez-ui
```

```
unzip tez-ui-0.9.2.war
vim ./config/configs.env
```

将 localhost 修改为当前服务器的内网 IP。



新建 role

```
vim /usr/local/service/knox/conf/topologies/emr.xml
```

修改 emr.xml 配置文件

第一处：添加内容如下

```
80          <!-- HA Provider -->
81          <provider>
82              <role>ha</role>
83              <name>HaProvider</name>
84              <enabled>true</enabled>
85              <param>
86                  <name>HDFSUI</name>
87                  <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
88              </param>
89              <param>
90                  <name>TEZUI</name>
91                  <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
92              </param>
93              <param>
94                  <name>APPLICATIONHISTORY</name>
95                  <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
96              </param>
97              <param>
98                  <name>WEBHDFS</name>
99                  <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
100             </param>
101             <param>
102                 <name>YARNUI</name>
103                 <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
104             </param>
105             <param>
106                 <name>GANGLIAUI</name>
107                 <value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
108             </param>
```

```
<param>
<name>TEZUI</name>
<value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
</param>
<param>
<name>APPLICATIONHISTORY</name>
<value>maxFailoverAttempts=3;failoverSleep=1000;enabled=true</value>
</param>
```

第二处：修改内容如下

```
<!--HA情况下，需要区分默认active和standby，对于active，url1是activeurl，url2是standbyurl，对于standby则相反-->

<service>
    <role>HDFSUI</role>
    <url>http://172.[redacted].9:4008</url>
    <version>2.7.0</version>
</service>

<service>
    <role>TEZUI</role>
    <url>http://172.[redacted]9:2000/tez-ui</url>
    <version>0.9.2</version>
</service>

<service>
    <role>APPLICATIONHISTORY</role>
    <url>http://172.[redacted].9:8188</url>
    <version>2.7.3</version>
</service>
```

```
<service>
<role>TEZUI</role>
<url>http://172.**.**.9:2000/tez-ui</url>
<version>0.9.2</version>
</service>
<service>
<role>APPLICATIONHISTORY</role>
<url>http://172.**.**.9:8188</url>
<version>2.7.3</version>
</service>
```

yarn 的 timelineserver 配置

在配置管理中修改 yarn-site.xml 配置文件，保存配置修改，重启配置发生变化组件。

参数	值
yarn.timeline-service.enabled	true
yarn.timeline-service.hostname	`172.**.**.9` （需替换为自己的 IP）
yarn.timeline-service.http-cross-origin.enabled	true
yarn.resourcemanager.system-metrics-publisher.enabled	true
yarn.timeline-service.address	`172.**.**.9:10201` （需替换为自己的 IP）
yarn.timeline-service.webapp.address	`172.**.**.9:8188` （需替换为自己的 IP）
yarn.timeline-service.webapp.https.address	`172.**.**.9:2191` （需替换为自己的 IP）
yarn.timeline-service.generic-application-history.enabled	true
yarn.timeline-service.handler-thread-count	24

tez 配置修改

在配置管理中的 tez-site.xml 配置文件中新增配置项，保存配置修改，重启配置发生变化组件。

参数	值
tez.tez-ui.history-url.base	`http://172.**.**.9:2000/tez-ui/` （需替换为自己的 IP）

参数	值
tez.history.logging.service.class	org.apache.tez.dag.history.logging.ats.ATSHistoryLoggingService

服务启动

1. 启动 timelineserver。

```
/usr/local/service/hadoop/sbin/yarn-daemon.sh start timelineserver
```

2. 启动 tomcat。

```
/usr/local/service/tomcat/bin/startup.sh
```

3. 重启 tez 服务。

```
su hadoop
rm -rf /usr/local/service/knox/data/deployments/*
/usr/local/service/knox/bin/ldap.sh stop
/usr/local/service/knox/bin/ldap.sh start
/usr/local/service/knox/bin/gateway.sh stop
/usr/local/service/knox/bin/gateway.sh start
```

4. tezui 访问地址。

> 说明：
> 账号密码与服务器登录账号密码相同。

```
https://{集群公网ip}:30002/gateway/emr/tez/
```

服务停止

如果运行一段时间后，发现 timelineserver 对业务影响较大，可参照如下操作停止相关服务。

1. 停止 tomcat。

```
/usr/local/service/tomcat/bin/shutdown.sh
```

2. 停止 timelineserver。

```
/usr/local/service/hadoop/sbin/yarn-daemon.sh stop timelineserver
```

Knox 集成 tez 0.8.5 版本说明

最近更新时间：2021-07-05 10:03:56

1. tez 0.8.5版本的 tez-ui 文件路径为 `/usr/local/service/tomcat/webapps/tez-ui/scripts/configs.js` 对应0.9.2版本中的 `configs.env` 。
2. yarn-site.xml 配置文件

   新建目录 `mkdir -p /usr/local/service/hadoop/filesystem/yarn/timeline`

参数	值
yarn.timeline-service.enabled	true
yarn.timeline-service.hostname	`172.**.**.9`
yarn.timeline-service.http-cross-origin.enabled	true
yarn.resourcemanager.system-metrics-publisher.enabled	true
yarn.timeline-service.address	`172.**.**.9:10201`
yarn.timeline-service.webapp.address	`172.**.**.9:8188`
yarn.timeline-service.webapp.https.address	`172.**.**.9:2191`
arn.timeline-service.generic-application-history.enabled	true
yarn.timeline-service.leveldb-timeline-store.path	`/usr/local/service/hadoop/filesystem/yarn/timeline`
yarn.timeline-service.handler-thread-count	24

3. 0.8.5版本需要修改 tez-site.xml 中对应的配置项参数为：

参数	值
tez.allow.disabled.timeline-domains	true
tez.history.logging.service.class	org.apache.tez.dag.history.logging.ats.ATSHistoryLoggingService

参数	值
tez.tez-ui.history-url.base	`http://172.**.**.9:2000/tez-ui/`

Knox 集成 tez 0.10.1 版本说明

最近更新时间：2022-06-15 15:12:30

1. 下载并编译 tez0.10.0 源码包

```
tar -zxvf apache-tez-0.10.1-src.tar.gz
chmod -R 777 apache-tez-0.10.1-src
cd apache-tez-0.10.1-src
mvn -X clean package -DskipTests=true -Dmaven.javadoc.skip=true
```

2. 解压编译的 war 包

3. yarn-site.xml 配置文件

新建目录 `mkdir -p /usr/local/service/hadoop/filesystem/yarn/timeline`

参数	值
yarn.timeline-service.enabled	true
yarn.timeline-service.hostname	`172.**.**.9`
yarn.timeline-service.http-cross-origin.enabled	true
yarn.resourcemanager.system-metrics-publisher.enabled	true
yarn.timeline-service.address	`172.**.**.9:10201`
yarn.timeline-service.webapp.address	`172.**.**.9:8188`
yarn.timeline-service.webapp.https.address	`172.**.**.9:2191`
yarn.timeline-service.generic-application-history.enabled	true
yarn.timeline-service.handler-thread-count	24

4. tez.xml 配置

参数	值
tez.tez-ui.history-url.base	`http://172.**.**.9:2000/tez-ui/`
tez.history.logging.service.class	org.apache.tez.dag.history.logging.ats.ATSHistoryLoggingService

Alluxio 开发指南
Alluxio 开发文档

最近更新时间：2021-10-29 10:28:23

背景

Alluxio 通过文件系统接口提供对数据的访问。Alluxio 中的文件提供一次写入语义：它们在被完整写下之后不可变，在完成之前不可读。Alluxio 提供了两种不同的文件系统 API：Alluxio API 和 Hadoop 兼容的 API。Alluxio API 提供了额外的功能，而 Hadoop 兼容的 API 为用户提供了无需修改现有代码使用 Hadoop API 的灵活性。

所有使用 Alluxio Java API 的资源都是通过 AlluxioURI 指定的路径实现。

获取文件系统客户端

要使用 Java 代码获取 Alluxio 文件系统客户端，请使用：

```
FileSystem fs = FileSystem.Factory.get();
```

创建一个文件

所有的元数据操作，以及打开一个文件读取或创建一个文件写入都通过 FileSystem 对象执行。由于 Alluxio 文件一旦写入就不可改变，创建文件的惯用方法是使用 `FileSystem#createFile(AlluxioURI)` ，它返回一个可用于写入文件的流对象。例如：

```
FileSystem fs = FileSystem.Factory.get();
AlluxioURI path = new AlluxioURI("/myFile");
// Create a file and get its output stream
FileOutStream out = fs.createFile(path);
// Write data
out.write(...);
// Close and complete file
out.close();
```

指定操作选项

对于所有的文件系统操作，可以指定一个额外的 options 字段，它允许用户可以指定操作的非默认设置。例如：

```
FileSystem fs = FileSystem.Factory.get();
AlluxioURI path = new AlluxioURI("/myFile");
// Generate options to set a custom blocksize of 128 MB
CreateFileOptions options = CreateFileOptions.defaults().setBlockSize(128 * Const
ants.MB);
FileOutStream out = fs.createFile(path, options);
```

IO 选项

Alluxio 使用两种不同的存储类型：Alluxio 管理存储和底层存储。Alluxio 管理存储是分配给 Alluxio worker 的内存 SSD 或 HDD。底层存储是由在最下层的存储系统（如 S3、Swift 或 HDFS）管理的资源。用户可以指定通过 ReadType 和 WriteType 与 Alluxio 管理的存储交互。ReadType 指定读取文件时的数据读取行为。WriteType 指定数据编写新文件时写入行为，例如，数据是否应该写入 Alluxio Storage。

下面是 ReadType 的预期行为表。读取总是偏好 Alluxio 存储优先于底层存储。

Read Type	Behavior
CACHE_PROMOTE	• 如果读取的数据在 Worker 上时，该数据被移动到 Worker 的最高层。 • 如果该数据不在本地 Worker 的 Alluxio 存储中，那么就将一个副本添加到本地 Alluxio Worker 中。 • 如果 `alluxio.user.file.cache.partially.read.block` 设置为 true，没有完全读取的数据块也会被**全部**存到 Alluxio 内。相反，一个数据块只有完全被读取时，才能被缓存。
CACHE	• 如果该数据不在本地 Worker 的 Alluxio 存储中，那么就将一个副本添加到本地 Alluxio Worker 中。 • 如果 `alluxio.user.file.cache.partially.read.block` 设置为 true，没有完全读取的数据块也会被**全部**存到 Alluxio 内。相反，一个数据块只有完全被读取时，才能被缓存。
NO_CACHE	仅读取数据，不在 Alluxio 中存储副本。

下面是 WriteType 的预期行为表。

Write Type	Behavior
CACHE_THROUGH	数据被同步地写入到 Alluxio 的 Worker 和底层存储系统。

Write Type	Behavior
MUST_CACHE	数据被同步地写入到 Alluxio 的 Worker。但不会被写入到底层存储系统。这是默认写类型。
THROUGH	数据被同步地写入到底层存储系统。但不会被写入到 Alluxio 的 Worker。
ASYNC_THROUGH	数据被同步地写入到 Alluxio 的 Worker，并异步地写入到底层存储系统。处于实验阶段。

位置策略

Alluxio 提供了位置策略来选择要存储文件块到哪一个 worker。使用 Alluxio 的 Java API，用户可以设置策略在 CreateFileOptions 中向 Alluxio 写入文件和在 OpenFileOptions 中读取文件。

用户可以轻松地覆盖默认的策略类配置文件中的属性 `alluxio.user.file.write.location.policy.class` 。内置的策略包括：

- LocalFirstPolicy (alluxio.client.file.policy.LocalFirstPolicy) ：首先返回本地节点，如果本地 worker 没有足够的块容量，它从活动 worker 列表中随机选择一名 worker。这是默认的策略。
- MostAvailableFirstPolicy (alluxio.client.file.policy.MostAvailableFirstPolicy)：返回具有最多可用字节的 worker。
- RoundRobinPolicy (alluxio.client.file.policy.RoundRobinPolicy)：以循环方式选择下一个 worker，跳过没有足够容量的 worker。
- SpecificHostPolicy (alluxio.client.file.policy.SpecificHostPolicy)：返回具有指定节点名的 worker。此策略不能设置为默认策略。

Alluxio 支持自定义策略，所以您也可以通过实现接口 `alluxio.client.file.policyFileWriteLocationPolicy` 制定适合自己的策略。

> 注意：
> 默认策略必须有一个空的构造函数。并使用 ASYNC_THROUGH 写入类型，所有块的文件必须写入同一个 worker。

Alluxio 允许客户在向本地 worker 写入数据块时选择一个层级偏好。目前这种策略偏好只适用于本地 worker 而不是远程 worker；远程 worker 会写到最高层。

默认情况下，数据被写入顶层。用户可以通过 `alluxio.user.file.write.tier.default` 配置项修改默认设置，或通过 `FileSystem#createFile(AlluxioURI)API` 调用覆盖它。

对现有文件或目录的所有操作都要求用户指定 AlluxioURI。使用 AlluxioURI，用户可以使用 FileSystem 中的任何方法来访问资源。

AlluxioURI 可用于执行 Alluxio FileSystem 操作，例如修改文件元数据、ttl 或 pin 状态，或者获取输入流来读取文件。例如，要读取一个文件：

```
FileSystem fs = FileSystem.Factory.get();
AlluxioURI path = new AlluxioURI("/myFile");
// Open the file for reading
FileInStream in = fs.openFile(path);
// Read data
in.read(...);
// Close file relinquishing the lock
in.close();
```

Alluxio 常用命令

最近更新时间：2021-10-29 10:30:25

操作	语法	描述
cat	cat "path"	将 Alluxio 中的一个文件内容打印在控制台中。
checkConsistency	checkConsistency "path"	检查 Alluxio 与底层存储系统的元数据一致性。
checksum	checksum "path"	计算一个文件的 md5 校验码。
chgrp	chgrp "group" "path"	修改 Alluxio 中的文件或文件夹的所属组。
chmod	chmod "permission" "path"	修改 Alluxio 中文件或文件夹的访问权限。
chown	chown "owner" "path"	修改 Alluxio 中文件或文件夹的所有者。
copyFromLocal	copyFromLocal "source path""remote path"	"remote path" 将 "source path" 指定的本地文件系统中的文件拷贝到 Alluxio 中 "remote path" 指定的路径，如果 "remote path" 已经存在该命令会失败。
copyToLocal	copyToLocal "remote path" "local path"	将 "remote path" 指定的 Alluxio 中的文件复制到本地文件系统中。
count	count "path"	输出 "path" 中所有名称匹配一个给定前缀的文件及文件夹的总数。
cp	cp "src" "dst"	在 Alluxio 文件系统中复制一个文件或目录。
du	du "path"	输出一个指定的文件或文件夹的大小。
fileInfo	fileInfo "path"	输出指定的文件的数据块信息。
free	free "path"	将 Alluxio 中的文件或文件夹移除，如果该文件或文件夹存在于底层存储中，那么仍然可以在那访问。
getCapacityBytes	getCapacityBytes	获取 Alluxio 文件系统的容量。
getUsedBytes	getUsedBytes	获取 Alluxio 文件系统已使用的字节数。

操作	语法	描述
help	help "cmd"	打印给定命令的帮助信息，如果没有给定命令，打印所有支持的命令的帮助信息。
leader	leader	打印当前 Alluxio leader master 节点名。
load	load "path"	将底层文件系统的文件或者目录加载到 Alluxio 中。
loadMetadata	loadMetadata "path"	将底层文件系统的文件或者目录的元数据加载到 Alluxio 中。
location	location "path"	输出包含某个文件数据的节点。
ls	ls "path"	列出给定路径下的所有直接文件和目录的信息，例如大小。
masterInfo	masterInfo	打印 Alluxio master 容错相关的信息，例如，leader 的地址、所有 master 的地址列表以及配置的 Zookeeper 地址。
mkdir	mkdir "path1" ... "pathn"	在给定路径下创建文件夹，以及需要的父文件夹，多个路径用空格或者 tab 分隔，如果其中的任何一个路径已经存在，该命令失败。
mount	mount "path" "uri"	将底层文件系统的 "uri" 路径挂载到 Alluxio 命名空间中的 "path" 路径下，"path" 路径事先不能存在并由该命令生成。没有任何数据或者元数据从底层文件系统加载。当挂载完成后，对该挂载路径下的操作会同时作用于底层文件系统的挂载点。
mv	mv "source" "destination"	将 "source" 指定的文件或文件夹移动到 "destination" 指定的新路径，如果 "destination" 已经存在该命令失败。
persist	persist "path1" ... "pathn"	将仅存在于 Alluxio 中的文件或文件夹持久化到底层文件系统中。
pin	pin "path"	将给定文件锁定到内容中以防止剔除。如果是目录，递归作用于其子文件以及里面新创建的文件。
report	report "path"	向 master 报告一个文件已经丢失。
rm	rm "path"	删除一个文件，如果输入路径是一个目录该命令失败。
setTtl	setTtl "path" "time"	设置一个文件的 TTL 时间，单位毫秒。
stat	stat "path"	显示文件和目录指定路径的信息。
tail	tail "path"	将指定文件的最后1KB内容输出到控制台。
test	test "path"	测试路径的属性，如果属性正确，返回0，否则返回1。

操作	语法	描述
touch	touch "path"	在指定路径创建一个空文件。
unmount	unmount "path"	卸载挂载在 Alluxio 中 "path" 指定路径上的底层文件路径，Alluxio 中该挂载点的所有对象都会被删除，但底层文件系统会将其保留。
unpin	unpin "path"	将一个文件解除锁定从而可以对其剔除，如果是目录则递归作用。
unsetTtl	unsetTtl "path"	删除文件的 ttl 值。

cat

背景

cat 命令将 Alluxio 中的一个文件内容全部打印在控制台中，这在用户确认一个文件的内容是否和预想的一致时非常有用。如果您想将文件拷贝到本地文件系统中，使用 copyToLocal 命令。

操作示例

例如，当测试一个新的计算任务时，cat 命令可以用来快速确认其输出结果。

```
$ ./bin/alluxio fs cat /output/part-00000
```

checkConsistency

背景

checkConsistency 命令会对比一给定路径下 Alluxio 以及底层存储系统的元数据，如果该路径是一个目录，那么其所有子内容都会被对比。该命令返回包含所有不一致的文件和目录的列表，系统管理员决定是否对这些不一致数据进行调整。为了防止 Alluxio 与底层存储系统的元数据不一致，应将您的系统设置为通过 Alluxio 来修改文件和目录，而不是直接访问底层存储系统进行修改。

如果使用了 -r 选项，那么 checkConsistency 命令会去修复不一致的文件或目录，对于只存在底层存储的文件或者文件夹会从 Alluxio 中删除，对于在底层文件系统中，但是，文件内容发生变化的文件，该文件的元数据会重新 load 到 Alluxio。

> 注意：
> 该命令需要请求将要被检查的目录子树的读锁，这意味着在该命令完成之前无法对该目录子树的文件或者目录进行写操作或者更新操作。

操作示例

例如，checkConsistency 命令可以用来周期性地检查命名空间的完整性。

列出不一致的文件或者目录：

```
$ ./bin/alluxio fs checkConsistency /
```

修复不一致的文件或者目录：

```
$ ./bin/alluxio fs checkConsistency -r /
```

checksum

背景

checksum 命令输出某个 Alluxio 文件的 md5 值。

操作示例

例如，checksum 可以用来验证 Alluxio 中的文件内容与存储在底层文件系统或者本地文件系统中的文件内容是否匹配。

```
$ ./bin/alluxio fs checksum /LICENSE
```

chgrp

背景

chgrp 命令可以改变 Alluxio 中的文件或文件夹的所属组，Alluxio 支持 POSIX 标准的文件权限，组在 POSIX 文件权限模型中是一个授权实体，文件所有者或者超级用户可以执行这条命令从而改变一个文件或文件夹的所属组。

加上 -R 选项可以递归的改变文件夹中子文件和子文件夹的所属组。

操作示例

使用 chgrp 命令能够快速修改一个文件的所属组。

```
$ ./bin/alluxio fs chgrp alluxio-group-new /input/file1
```

chmod

背景

chmod 命令修改 Alluxio 中文件或文件夹的访问权限，目前可支持八进制模式：三位八进制的数字分别对应于文件所有者、所属组以及其他用户的权限。以下是数字与权限的对应表：

Number	Permission	rwx
7	read, write and execute	rwx
6	read and write	rw-
5	read and execute	r-x
4	read only	r--
3	write and execute	-wx
2	write only	-w-
1	execute only	--x
0	none	---

加上 -R 选项可以递归的改变文件夹中子文件和子文件夹的权限。

操作示例

使用 chmod 命令可以快速修改一个文件的权限。

```
$ ./bin/alluxio fs chmod 755 /input/file1
```

chown

背景

chown 命令用于修改 Alluxio 中文件或文件夹的所有者，出于安全方面的考虑，只有超级用户能够更改一个文件的所有者。

加上 -R 选项可以递归的改变文件夹中子文件和子文件夹的所有者。

使用示例

使用 chown 命令可以快速修改一个文件的所有者。

```
$ ./bin/alluxio fs chown alluxio-user /input/file1
```

copyFromLocal

背景

copyFromLocal 命令将本地文件系统中的文件拷贝到 Alluxio 中，如果您运行该命令的机器上有 Alluxio worker，那么数据便会存放在这个 worker 上，否则，数据将会随机地复制到一个运行 Alluxio worker 的远程节点上。如果该命令指定的目标是一个文件夹，那么这个文件夹及其所有内容都会被递归复制到 Alluxio 中。

操作示例

使用 copyFromLocal 命令可以快速将数据复制到 alluxio 系统中以便后续处理。

```
$ ./bin/alluxio fs copyFromLocal /local/data /input
```

copyToLocal

背景

copyToLocal 命令将 Alluxio 中的文件复制到本地文件系统中，如果该命令指定的目标是一个文件夹，那么该文件夹及其所有内容都会被递归地复制。

操作示例

使用 copyToLocal 命令可以快速将输出数据下载下来从而进行后续研究或调试。

```
$ ./bin/alluxio fs copyToLocal /output/part-00000 part-00000
```

count

背景

count 命令输出 Alluxio 中所有名称匹配一个给定前缀的文件及文件夹的总数，以及它们总的大小，该命令对文件夹中的内容递归处理。当用户对文件有预定义命名习惯时，count 命令很有用。

操作示例

若文件是以它们的创建日期命名，使用 count 命令可以获取任何日期、月份以及年份的所有文件的数目以及它们的总大小。

```
$ ./bin/alluxio fs count /data/2014
```

cp

背景

cp 命令拷贝 Alluxio 文件系统中的一个文件或者目录，也可以在本地文件系统和 Alluxio 文件系统之间相互拷贝。filescheme 表示本地文件系统，alluxioscheme 或不写 scheme 表示 Alluxio 文件系统。如果使用了 -R 选项，并且源路径是一个目录，cp 将源路径下的整个子树拷贝到目标路径。

操作示例

例如，cp 可以在底层文件系统之间拷贝文件。

```
$ ./bin/alluxio fs cp /hdfs/file1 /s3/
```

du

背景

du 命令输出一个文件的大小，如果指定的目标为文件夹，该命令输出该文件夹下所有子文件及子文件夹中内容的大小总和。

操作示例

如果 Alluxio 空间被过分使用，使用 du 命令可以检测到哪些文件夹占用了大部分空间。

```
$ ./bin/alluxio fs du /\\*
```

fileInfo

背景

fileInfo 命令从1.5开始不再支持，请使用 stat 命令。

fileInfo 命令将一个文件的主要信息输出到控制台，这主要是为了让用户调试他们的系统。一般来说，在 Web UI 上查看文件信息要容易理解得多。

操作示例

使用 fileInfo 命令能够获取到一个文件的数据块的位置，这在获取计算任务中的数据局部性时非常有用。

```
$ ./bin/alluxio fs fileInfo /data/2015/logs-1.txt
```

free

背景

free 命令请求 Alluxio master 将一个文件的所有数据块从 Alluxio worker 中剔除，如果命令参数为一个文件夹，那么会递归作用于其子文件和子文件夹。该请求不保证会立即产生效果，因为该文件的数据块可能正在被读取。free 命令在被 master 接收后会立即返回。注意该命令不会删除底层文件系统中的任何数据，而只会影响存储在 Alluxio 中的数据。另外，该操作也不会影响元数据，这意味着如果运行 ls 命令，该文件仍然会被显示。

操作示例

使用 free 命令可以手动管理 Alluxio 的数据缓存。

```
$ ./bin/alluxio fs free /unused/data
```

getCapacityBytes

背景

getCapacityBytes 命令返回 Alluxio 被配置的最大字节数容量。

操作示例

使用 getCapacityBytes 命令能够确认您的系统是否正确启动。

```
$ ./bin/alluxio fs getCapacityBytes
```

getUsedBytes

背景

getUsedBytes 命令返回 Alluxio 中以及使用的空间字节数。

操作示例

使用 getUsedBytes 命令能够监控集群健康状态。

```
$ ./bin/alluxio fs getUsedBytes
```

leader

背景

leader 命令打印当前 Alluxio 的 leader master 节点名。

操作示例

```
$ ./bin/alluxio fs leader
```

load

背景

load 命令将底层文件系统中的数据载入到 Alluxio 中。如果运行该命令的机器上正在运行一个 Alluxio worker，那么数据将移动到该 worker 上，否则，数据会被随机移动到一个 worker 上。如果该文件已经存在在 Alluxio 中，该命令不进行任何操作。如果该命令的目标是一个文件夹，那么其子文件和子文件夹会被递归载入。

操作示例

使用 load 命令能够获取用于数据分析作用的数据。

```
$ ./bin/alluxio fs load /data/today
```

loadMetadata

背景

loadMetadata 命令查询本地文件系统中匹配给定路径名的所有文件和文件夹，并在 Alluxio 中创建这些文件的镜像。该命令只创建元数据，例如文件名及文件大小，而不会传输数据。

操作示例

当其他系统将数据输出到底层文件系统中（不经过 Alluxio），而在 Alluxio 上运行的某个应用又需要使用这些输出数据时，就可以使用 loadMetadata 命令。

```
$ ./bin/alluxio fs loadMetadata /hdfs/data/2015/logs-1.txt
```

location

背景

location 命令返回包含一个给定文件包含的数据块的所有 Alluxio worker 的地址。

操作示例

当使用某个计算框架进行作业时，使用 location 命令可以调试数据局部性。

```
$ ./bin/alluxio fs location /data/2015/logs-1.txt
```

ls

背景

ls 命令列出一个文件夹下的所有子文件和子文件夹及文件大小、上次修改时间以及文件的内存状态。对一个文件使用 ls 命令仅仅会显示该文件的信息。 ls 命令也将任意文件或者目录下的子目录的元数据从底层存储系统加载到 Alluxio 命名空间，如果 Alluxio 还没有这部分元数据的话。 ls 命令查询底层文件系统中匹配给定路径的文件或者目录，然后会在 Alluxio 中创建一个该文件的镜像文件。只有元数据，例如文件名和大小，会以这种方式加载而不发生数据传输。

选项：

-d 选项将目录作为普通文件列出。例如， `ls -d /` 显示根目录的属性。
-f 选项强制加载目录中的子目录的元数据。默认方式下，只有当目录首次被列出时，才会加载元数据。
-h 选项以可读方式显示文件大小。
-p 选项列出所有固定的文件。
-R 选项可以递归的列出输入路径下的所有子文件和子文件夹，并列出从输入路径开始的所有子树。

操作示例

使用 ls 命令可以浏览文件系统。

```
$ ./bin/alluxio fs mount /cos/data cosn://data-bucket/
```

验证：

```
$ ./bin/alluxio fs ls /s3/data/
```

masterInfo

背景

masterInfo 命令打印与 Alluxio master 容错相关的信息，例如 leader 的地址、所有 master 的地址列表以及配置的 Zookeeper 地址。如果 Alluxio 运行在单 master 模式下，masterInfo 命令会打印出该 master 的地址；如果 Alluxio 运行在多 master 容错模式下，masterInfo 命令会打印出当前的 leader 地址、所有 master 的地址列表以及 Zookeeper 的地址。

操作示例

使用 masterInfo 命令可以打印与 Alluxio master 容错相关的信息。

```
$ ./bin/alluxio fs masterInfo
```

mkdir

背景

mkdir 命令在 Alluxio 中创建一个新的文件夹。该命令可以递归创建不存在的父目录。注意在该文件夹中的某个文件被持久化到底层文件系统之前，该文件夹不会在底层文件系统中被创建。对一个无效的或者已存在的路径使用 mkdir 命令会失败。

操作示例

管理员使用 mkdir 命令可以创建一个基本文件夹结构。

```
$ ./bin/alluxio fs mkdir /users
$ ./bin/alluxio fs mkdir /users/Alice
$ ./bin/alluxio fs mkdir /users/Bob
```

mount

背景

mount 命令将一个底层存储中的路径链接到 Alluxio 路径，并且在 Alluxio 中该路径下创建的文件和文件夹会在对应的底层文件系统路径进行备份。访问统一命名空间获取更多相关信息。

选项：
--readonly 选项在 Alluxio 中设置挂载点为只读。
--option `<key>=<val>` 选项传递一个属性到这个挂载点（如 S3 credential）。

操作示例

使用 mount 命令可以让其他存储系统中的数据在 Alluxio 中也能获取。

```
$ ./bin/alluxio fs mount /mnt/hdfs hdfs://host1:9000/data/
```

mv

背景

mv 命令将 Alluxio 中的文件或文件夹移动到其他路径。目标路径一定不能事先存在或者是一个目录。如果是一个目录，那么该文件或文件夹会成为该目录的子文件或子文件夹。mv 命令仅仅对元数据进行操作，不会影响该文件的数据块。mv 命令不能在不同底层存储系统的挂载点之间操作。

操作示例

使用 mv 命令可以将过时数据移动到非工作目录。

```
$ ./bin/alluxio fs mv /data/2014 /data/archives/2014
```

persist

背景

persist 命令将 Alluxio 中的数据持久化到底层文件系统中。该命令是对数据的操作，因而其执行时间取决于该文件的大小。在持久化结束后，该文件即在底层文件系统中有了备份，因而该文件在 Alluxio 中的数据块被剔除甚至丢失的情况下，仍能够访问。

操作示例

在从一系列临时文件中过滤出包含有用数据的文件后，便可以使用 persist 命令对其进行持久化。

```
$ ./bin/alluxio fs persist /tmp/experimental-logs-2.txt
```

pin

背景

pin 命令对 Alluxio 中的文件或文件夹进行标记。该命令只针对元数据进行操作，不会导致任何数据被加载到 Alluxio 中。如果一个文件在 Alluxio 中被标记了，该文件的任何数据块都不会从 Alluxio worker 中被剔除。如果存在过多的

被锁定的文件，Alluxio worker 将会剩余少量存储空间，从而导致无法对其他文件进行缓存。

操作示例

如果管理员对作业运行流程十分清楚，那么可以使用 pin 命令手动提高性能。

```
$ ./bin/alluxio fs pin /data/today
```

report

背景

report 命令向 Alluxio master 标记一个文件为丢失状态。该命令应当只对使用 Lineage API 创建的文件使用。将一个文件标记为丢失状态将导致 master 调度重计算作业从而重新生成该文件。

操作示例

使用 report 命令可以强制重新计算生成一个文件。

```
$ ./bin/alluxio fs report /tmp/lineage-file
```

rm

背景

rm 命令将一个文件从 Alluxio 以及底层文件系统中删除。该命令返回后该文件便立即不可获取，但实际的数据要过一段时间才被真正删除。

加上 -R 选项可以递归的删除文件夹中所有内容后再删除文件夹自身。加上 -U 选项将会在尝试删除持久化目录之前不会检查将要删除的 UFS 内容是否与 Alluxio 一致。

操作示例

使用 rm 命令可以删除掉不再需要的临时文件。

```
$ ./bin/alluxio fs rm /tmp/unused-file
```

setTtl

背景

setTtl 命令设置一个文件或者文件夹的 ttl 时间，单位为毫秒。如果当前时间大于该文件的创建时间与 ttl 时间之和时，行动参数将指示要执行的操作。delete 操作（默认）将同时删除 Alluxio 和底层文件系统中的文件，而 free 操作将仅仅删除 Alluxio 中的文件。

操作示例

管理员在知道某些文件经过一段时间后便没用时，可使用带有 delete 操作的 setTtl 命令来清理文件；如果仅希望为 Alluxio 释放更多的空间，可使用带有 free 操作的 setTtl 命令来清理 Alluxio 中的文件内容。

```
$ ./bin/alluxio fs setTtl -action free /data/good-for-one-day 86400000
```

stat

背景

stat 命令将一个文件或者文件夹的主要信息输出到控制台，这主要是为了让用户调试他们的系统。一般来说，在 Web UI 上查看文件信息要容易理解得多。

可以指定 -f 来按指定格式显示信息：

- “%N”：文件名。
- “%z”：文件大小（bytes）。
- “%u”：文件拥有者。
- “%g”：拥有者所在组名。
- “%y” or “%Y”：编辑时间，`%y shows 'yyyy-MM-dd HH:mm:ss' (the UTC date), %Y` 为自从 January 1, 1970 UTC 以来的毫秒数。
- “%b”：为文件分配的数据块数。

操作示例

例如，使用 stat 命令能够获取到一个文件的数据块的位置，这在获取计算任务中的数据局部性时非常有用。

```
$ ./bin/alluxio fs stat /data/2015/logs-1.txt
$ ./bin/alluxio fs stat /data/2015
$ ./bin/alluxio fs stat -f %z /data/2015/logs-1.txt
```

tail

背景

tail 命令将一个文件的最后1kb内容输出到控制台。

操作示例

使用 tail 命令可以确认一个作业的输出是否符合格式或者包含期望的值。

```
$ ./bin/alluxio fs tail /output/part-00000
```

test

背景

test 命令测试路径的属性，如果属性为真，返回0，否则返回1。

选项：
-d 选项测试路径是否是目录。
-e 选项测试路径是否存在。
-f 选项测试路径是否是文件。
-s 选项测试路径是否为空。
-z 选项测试文件长度是否为0。

操作示例

```
$ ./bin/alluxio fs test -d /someDir
```

touch

背景

touch 命令创建一个空文件。由该命令创建的文件不能被覆写，大多数情况是用作标记。

操作示例

使用 touch 命令可以创建一个空文件用于标记一个文件夹的分析任务完成了。

```
$ ./bin/alluxio fs touch /data/yesterday/_DONE_
```

unmount

背景

unmount 将一个 Alluxio 路径和一个底层文件系统中的目录的链接断开。该挂载点的所有元数据和文件数据都会被删除，但底层文件系统会将其保留。访问 Unified Namespace 获取更多信息。

操作示例

当不再需要一个底层存储系统中的数据时，使用 unmont 命令可以移除该底层存储系统。

```
$ ./bin/alluxio fs unmount /s3/data
```

unpin

背景

unpin 命令将 Alluxio 中的文件或文件夹解除标记。该命令仅作用于元数据，不会剔除或者删除任何数据块。一旦文件被解除锁定，Alluxio worker 可以剔除该文件的数据块。

操作示例

当管理员知道数据访问模式发生改变时，可以使用 unpin 命令。

```
$ ./bin/alluxio fs unpin /data/yesterday/join-table
```

unsetTtl

背景

unsetTtl 命令删除 Alluxio 中一个文件的 TTL。该命令仅作用于元数据，不会剔除或者删除 Alluxio 中的数据块。该文件的 TTL 值可以由 setTtl 命令重新设定。

操作示例

在一些特殊情况下，当一个原本自动管理的文件需要手动管理时，可使用 unsetTtl 命令。

```
$ ./bin/alluxio fs unsetTtl /data/yesterday/data-not-yet-analyzed
```

挂载文件系统到 Alluxio 统一文件系统

最近更新时间：2022-08-08 11:04:32

背景

Alluxio 提供了统一命名空间机制，允许将其他文件系统挂载到 Alluxio 的文件系统中。该机制允许上层应用使用统一的命名空间，访问分散在不同系统中的数据。

挂载 COS

示例：将 COS 中某个 bucket 挂载到 alluxio 目录中。

```
bin/alluxio fs mount --option fs.cos.access.key=<COS_SECRET_ID> \
--option fs.cos.secret.key=<COS_SECRET_KEY> \
--option fs.cos.region=<COS_REGION> \
--option fs.cos.app.id=<COS_APP_ID> \
/cos cos://<COS_BUCKET>/
```

其中，--options 中配置 COS 的配置。

配置名称	解释
fs.cos.access.key	cos scecret id
fs.cos.secret.key	cos secret key
fs.cos.region	cos region 名称，例如 `ap-beijing`
fs.cos.app.id	用户 AppID
COS_BUCKET	COS BUCKET 名称。**只要名称，不要带 AppID 后缀**

该命令，将 COS 的目录（通过 `cos://bucket/xxx` 指定）挂载到 alluxio 的 `/cos` 目录下。

挂载 HDFS

示例：将 HDFS 某目录挂载到 alluxio 目录中。

```
`bin/alluxio fs mount /hdfs hdfs://data`
```

该命令将 HDFS 的 `/data` 目录挂载到 alluxio 的 `/hdfs` 子目录下。
挂载成功后，通过 `alluxio fs ls` 命令，查看挂载内容。

挂载 CHDFS

示例：将 CHDFS 通过 mount 挂载到 Alluxio

> 说明：
> 只有 EMR2.5.0 + alluxio2.3.0+ 以上才支持。

```
alluxio fs mount \
--option alluxio.underfs.hdfs.configuration=/usr/local/service/hadoop/etc/hadoop/core-site.xml \
/chdfs ofs://f4modr7kmvw-wMqw.chdfs.ap-chongqing.myqcloud.com
```

在腾讯云中使用 Alluxio 文档

最近更新时间：2022-08-19 11:30:47

概述

在腾讯云 EMR 上提供了开箱可用的 Alluxio 服务，以帮助腾讯云客户可以快速实现分布式内存级缓存加速、简化数据管理等。同时还可以通过腾讯云 EMR 控制台或 API 接口，使用配置下发功能，快速配置多层级缓存和元数据管理等，获取一站式监控告警等功能。

准备

腾讯云 EMR 的 Hadoop 标准版本2.1.0版本及以上。

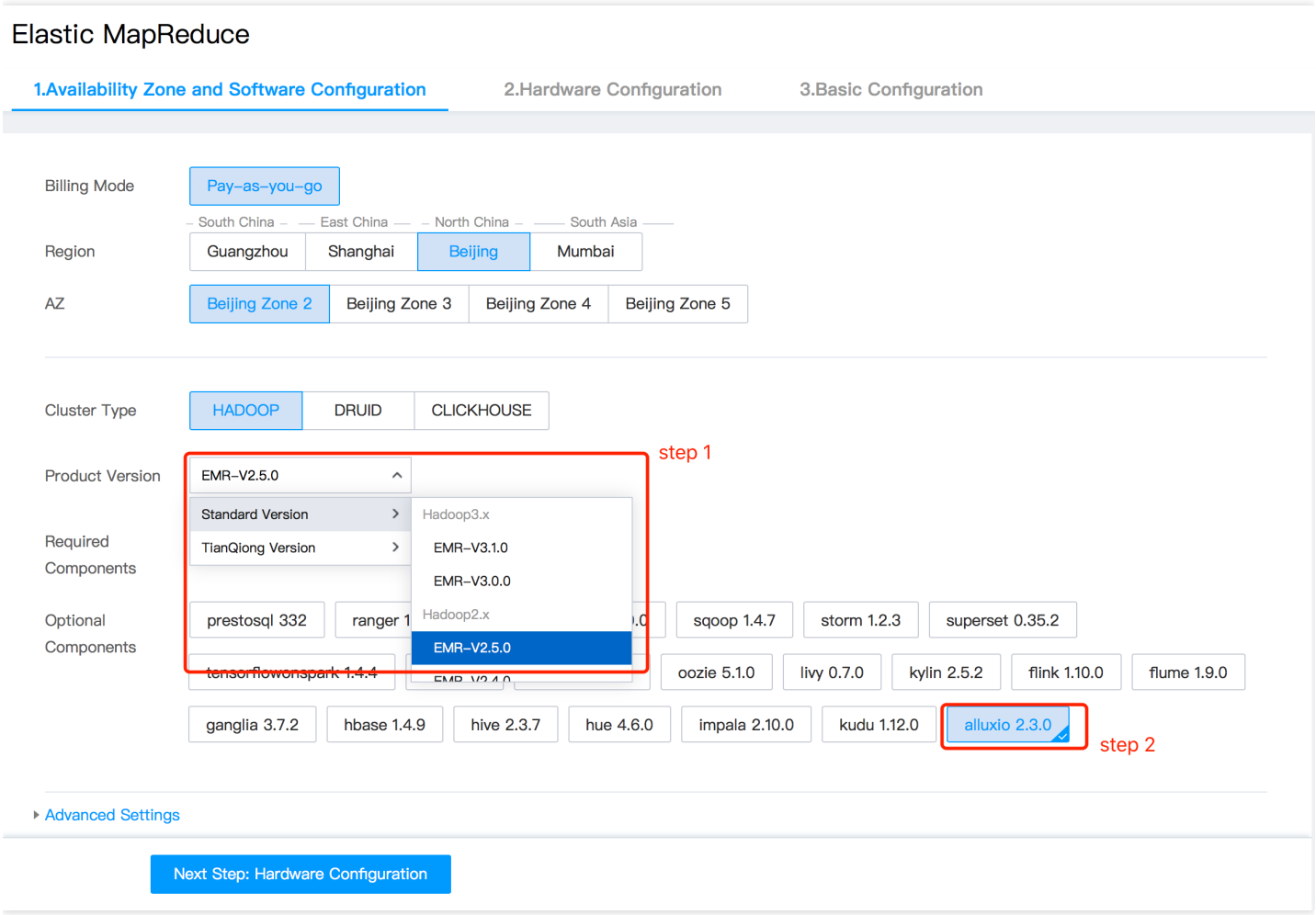
有关 EMR 中版本中支持具体的 Alluxio 的版本支持可参考 组件版本。

创建基于 Alluxio 的 EMR 集群

本节主要说明如何在腾讯云 EMR 上创建开箱即用的 Alluxio 集群。EMR 创建集群提供了购买页创建和 API 创建两种方式。

购买页创建集群

您需要登录腾讯云 EMR 购买页，在购买页选择支持的 Alluxio 发布版本，并且在可选组件列表中勾选 Alluxio 组件。



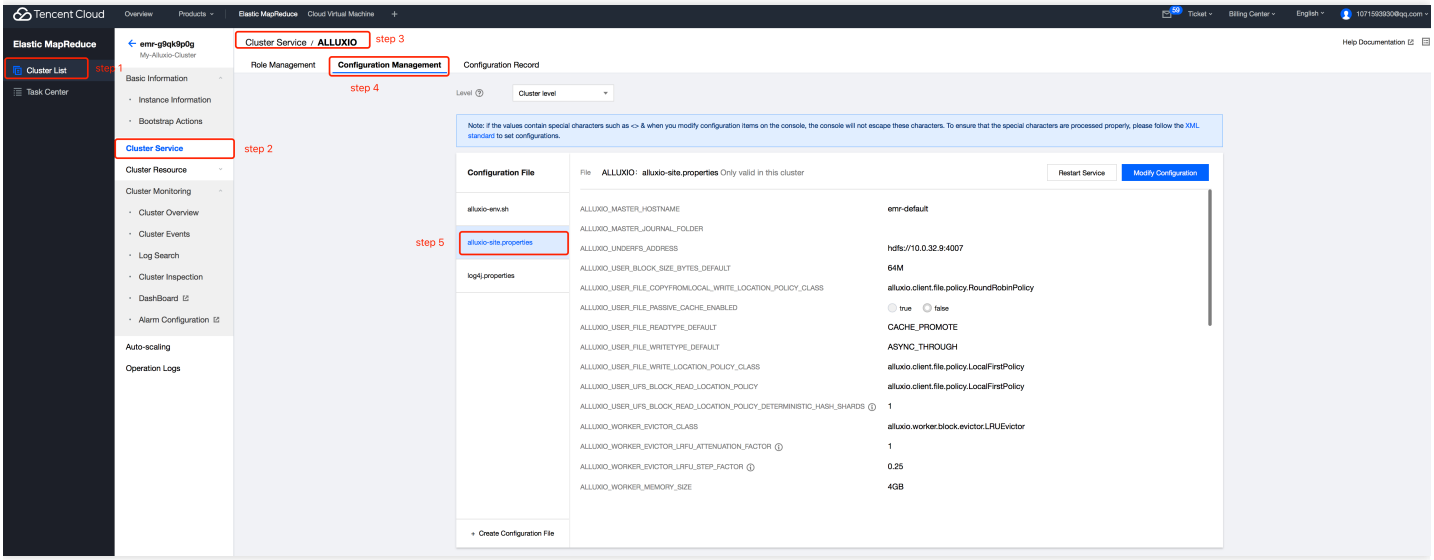
其他的选项可根据业务具体业务场景，进行个性化配置，创建过程中的具体选项可参考 创建 EMR 集群。

API 创建集群

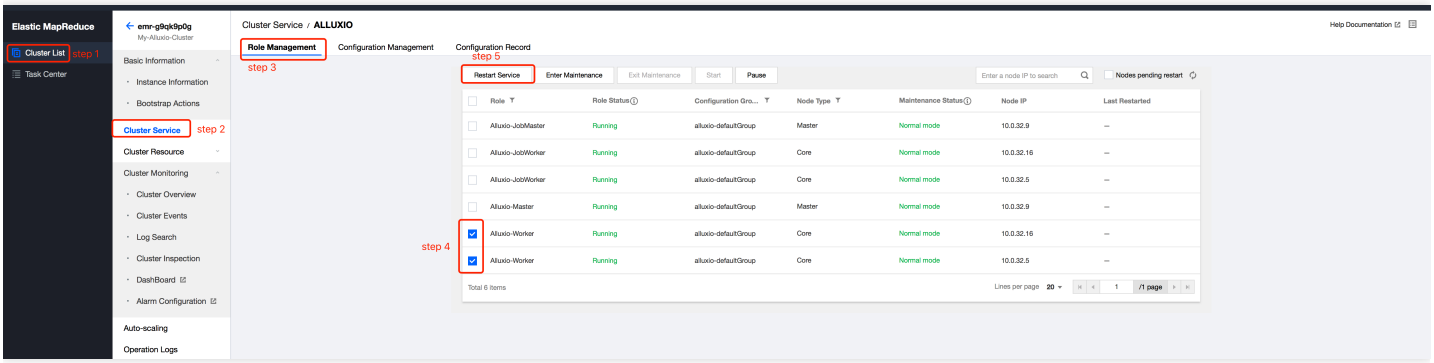
腾讯云 EMR 还提供了 API 方式构建基于 Alluxio 的大数据集群。具体可参考 查询硬件节点信息。

基础配置

创建一个带 Alluxio 组件的腾讯云 EMR 集群，默认会把 HDFS 挂载到 Alluxio 上，并使用内存作为单层 level0 存储。如果有需要更改更符合业务特性的多级存储，或者其他对应优化项，可以使用配置下发功能来完成相关配置。



在配置下发后，部分配置需要重启 Alluxio 服务才能生效。



更多配置下发和重启策略细节，可参考 配置下发 和 重启组件。

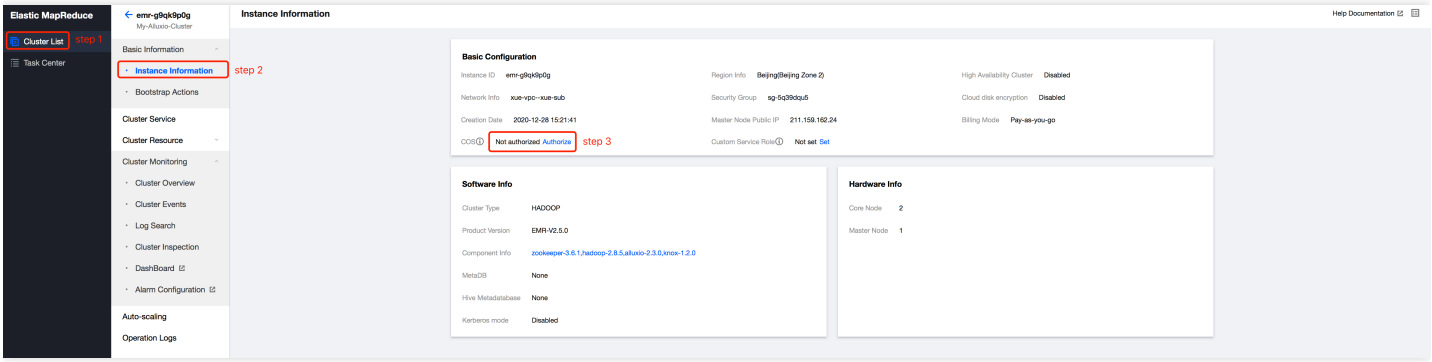
基于 Alluxio 加速计算存储分离

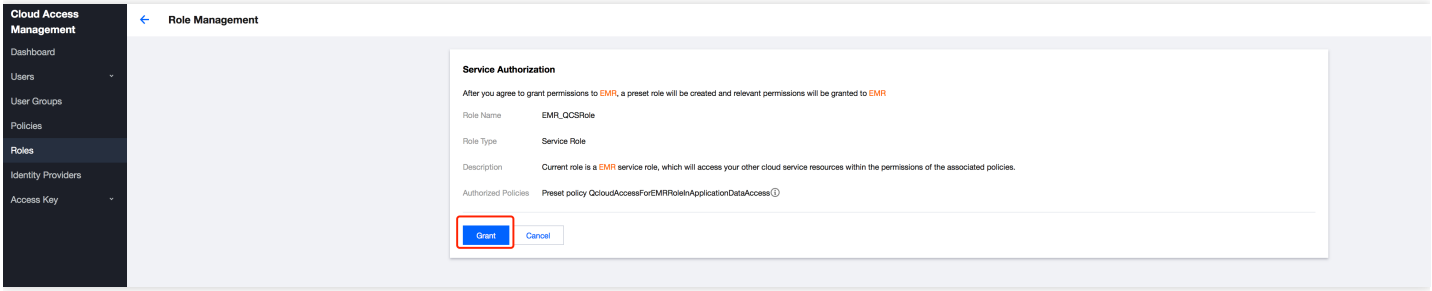
腾讯云 EMR 基于腾讯云对象存储（COS）提供了计算存储分离能力，默认直接访问对象存储中的数据时，应用程序没有节点级数据本地性或跨应用程序缓存。使用 Alluxio 加速将缓解这些问题。

在腾讯云 EMR 集群上默认已部署使用 COS 作为 UFS 的依赖 jar 包，只需授予访问 COS 的权限，并把 COS mount 到 Alluxio 上即可使用。

授权

若当前集群未开启对象存储，可在【访问管理控制台-角色】中进行授权，授权后 EMR 中节点可以通过临时密钥访问 COS 中数据。





Mount

登录到 EMR 任意一台机器，挂载 COS 到 Alluxio。

```
bin/alluxio fs mount <alluxio-path> <source-path>
//TODO,
```

更多在腾讯云 EMR 中使用 Alluxio 开发细节，可查阅 Alluxio 开发文档。

Alluxio 支持 COS 透明 URI

最近更新时间：2022-08-08 11:04:32

Alluxio 用户通常具有通过现有应用程序访问其底层存储系统（Under-FileSystem），将 Alluxio 添加到现有的生态系统中需求，但现有应用程序必须更改是需要在应用程序使用 Alluxio 的 URI。透明 URI 功能允许用户访问现有存储系统，且无需在应用程序级别更改 URI。

支持版本与配置 URI

1. 服务组件支持版本：Alluxio2.8.0版本。
2. 产品版本：Hadoop3.x 标准版本 EMR-V3.4.0 版本。
3. 配置支持透明 URI。使用 Alluxio 透明 URI，需要配置新的 Hadoop 兼容文件系统客户端实现。只要将客户端配置为接收外部 URI，此新的 ShimFileSystem 就会替换现有的 FileSystem。Hadoop 兼容的计算框架--Hadoop FileSystem 接口定义了从 FileSystem 方案到 FileSystem 实现的映射。为了配置 ShimFileSystem，请确保 core-site.xml 中配置了以下配置项：

配置项	配置项值
fs.cosn.impl	alluxio.hadoop.ShimFileSystem

4. Alluxio 在兼容透明 URL Schema 时需要对其进行转换兼容，请确保 alluxio-site.properties 中配置了以下配置项：

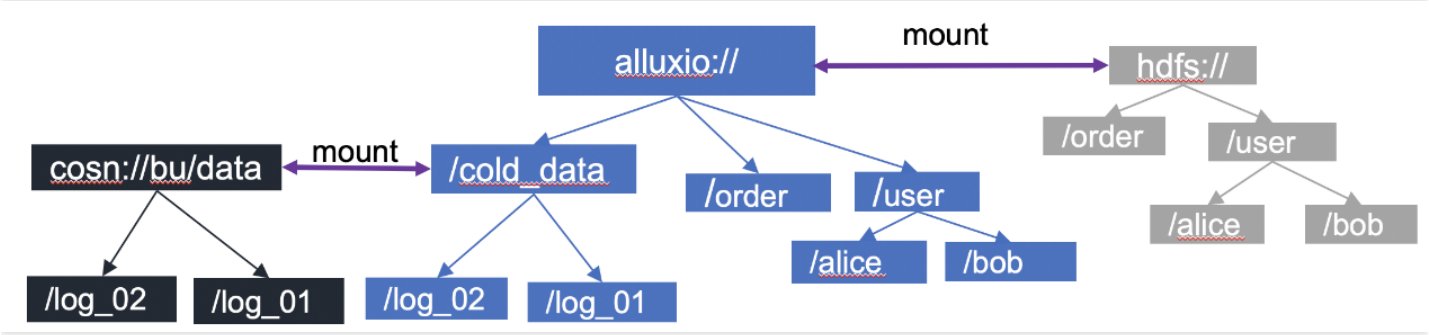
配置项	配置项值
alluxio.master.uri.translator.impl	alluxio.master.file.uritranslator.AutoMountUriTranslator
alluxio.user.shimfs.bypass.ufs.impl.list	fs.cosn.impl:org.apache.hadoop.fs.cosnative.NativeCosFileSystem

说明：

- 对 alluxio-site.properties 配置进行变更后需重启 Alluxio 服务。
- 一旦配置了 ShimFileSystem，master 将需要将外部存储系统本地的 URI 路由到 Alluxio 名称空间。这要求 cosn 已 mount 在 Alluxio 名称空间中。
- 关闭透明 URI 功能：只需回滚 core-site.xml中fs.cosn.impl 配置项。

mount

mount 命令可以说是 Alluxio 最有特色的命令之一。它类似于 Linux 里的 mount 命令---Linux 用户可以通过 Linux mount 把硬盘，SSD 等存储设备加载到这台 Linux 系统的本地文件系统中。而在 Alluxio 系统当中，mount 的概念进一步被扩展到了分布式系统一层：用户可以通过 Alluxio mount 把一个或多个其他的存储系统/云存储服务（例如 HDFS、COS 等），挂载到 Alluxio 这个分布式文件系统当中去。从而运行在 Alluxio 上的分布式应用，例如 Spark、Presto 或者 MapReduce 等，不需要去适配甚至了解具体的数据访问协议和路径，而只需要知道数据对应在 Alluxio 文件系统的路径就已足够，从而极大的方便了应用的开发和维护。



EMR-Alluxio 默认使用 hdfs 作为根目录挂载点

在 EMR-Alluxio2.5.1+后，Alluxio 的 UFS 开始支持 COSN 协议，COS UFS 存在读写性能较差以及不稳定的问题，为了解决此类问题，社区贡献了 COSN UFS 底层文件系统。COS 和 COSN UFS 都是用于访问腾讯云对象存储，COSN 相对于 COS 做了深度优化，其读写性能较COS 成倍提升，同时带来了更好的稳定性，所以强烈推荐使用 COSN。COS UFS 将于 EMR-Alluxio2.6.0 版本后停止维护。

Mount COSN 示例：

```
alluxio fs mount --option fs.cosn.userinfo.secretId=xx \
--option fs.cosn.userinfo.secretKey=xx \
--option fs.cosn.bucket.region=ap-xx \
--option fs.cosn.impl=org.apache.hadoop.fs.cosnative.NativeCosFileSystem \
--option fs.AbstractFileSystem.cosn.impl=org.apache.hadoop.fs.CosN \
--option fs.cosn.userinfo.appid=xx \
/cosn cosn://COS_BUCKET/path
```

其中，--options 中配置 COS 的配置。

配置项名称	解释
fs.cosn.userinfo.secretId	cos scecret id
fs.cosn.userinfo.secretKey	cos secret key

配置项名称	解释
fs.cosn.impl	固定值：`org.apache.hadoop.fs.CosFileSystem`
fs.AbstractFileSystem.cosn.impl	固定值：`org.apache.hadoop.fs.CosN`
fs.cosn.bucket.region cos region	名称，例如 ap-beijing
fs.cosn.userinfo.appid	用户主账号 AppID
COS_BUCKET COS BUCKET	名称。只要名称，不要带 AppID 后缀

Alluxio 支持鉴权

最近更新时间：2022-08-08 11:04:32

Alluxio 用户在对已有的统一命名空间访问 COS、HDFS、CHDFS 上的数据，或者用户使用透明 URL 来访问 Alluxio 中的缓存数据时，会出现无鉴权情况，也就是说任何用户只要拿着对应 URI 就能获取到数据。云上 Alluxio 对此类场景结合 Ranger 和 CosRanger 完善了鉴权场景。

> 说明：
> 为了设配鉴权特性，请确保集群有以下组件集成：
>
> - 如果 Alluxio 中只挂载了 HDFS，那么需要集成 Ranger 组件。
> - 如果 Alluxio 中挂载了 COS、CHDFS，那么需要集成 CosRanger 组件。

支持版本

- 服务组件支持版本：Alluxio2.8.0版本。
- 产品版本：Hadoop3.x 标准版本 EMR-V3.4.0版本。

配置鉴权

前置配置

```
#新增hive组件ranger-hive-security.xml配置项
ranger.plugin.hive.urlauth.filesystem.schemes==hdfs:,file:,wasb:,adl:,alluxio:
#新增presto组件hive.properties配置项
hive.hdfs.authentication.type=NONE
hive.metastore.authentication.type=NONE
hive.hdfs.impersonation.enabled=true
hive.metastore.thrift.impersonation.enabled=true
```

> 说明：
> 以上前置配置需客户根据集群现有组件来配置。

HDFS 鉴权

软链接 ranger 相关配置文件：

```
[hadoop@172 conf]$ pwd
/usr/local/service/alluxio/conf
[hadoop@172 conf]$ ln -s /usr/local/service/hadoop/etc/hadoop/ranger-hdfs-audit.x
ml
ranger-hdfs-audit.xml
[hadoop@172 conf]$ ln -s /usr/local/service/hadoop/etc/hadoop/ranger-hdfs-securit
y.xml ranger-hdfs-security.xml
```

alluxio-site.properties 配置

建议使用 EMR 控制台进行集群维度配置下发。

```
# 鉴权开关(默认false)
alluxio.security.authorization.plugins.enabled=true
alluxio.security.authorization.plugin.name=ranger
alluxio.security.authorization.plugin.paths=/usr/local/service/alluxio/conf
alluxio.underfs.security.authorization.plugin.name=ranger
alluxio.underfs.security.authorization.plugin.paths=/usr/local/service/alluxio/co
nf
alluxio.master.security.impersonation.hadoop.users=*
alluxio.security.login.impersonation.username=_HDFS_USER_
```

> 说明：
> 下发完成后需重启 Alluxio 服务。

COS 及 CHDFS 鉴权

```
#新增core-site.xml配置项
fs.ofs.ranger.enable.flag=true
```

alluxio-site.properties 配置

建议使用 EMR 控制台进行集群维度配置下发。

```
# 鉴权开关(默认false)
# 鉴权开关(默认false)
alluxio.security.authorization.plugins.enabled=true
alluxio.security.authorization.plugin.name=ranger
alluxio.security.authorization.plugin.paths=/usr/local/service/alluxio/conf
alluxio.underfs.security.authorization.plugin.name=ranger
alluxio.underfs.security.authorization.plugin.paths=/usr/local/service/alluxio/co
nf
```

```
alluxio.cos.qcloud.object.storage.ranger.service.config.dir=/usr/local/service/cosranger/conf
alluxio.master.security.impersonation.hadoop.users=*
alluxio.security.login.impersonation.username=_HDFS_USER_
# 重试次数默认为5次
alluxio.cos.qcloud.object.storage.permission.check.max.retry=5
```

说明：
下发完成后需重启 Alluxio 服务。

Kylin 开发指南

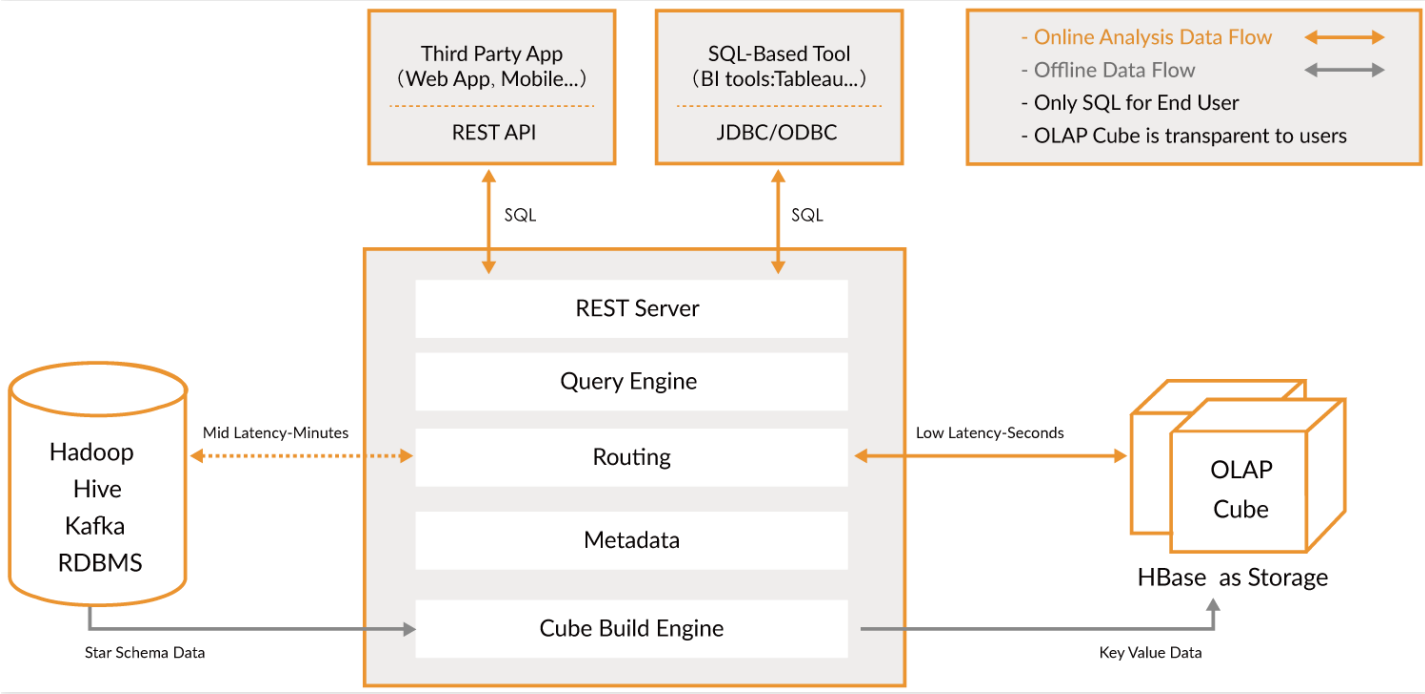
Kylin 简介

最近更新时间：2021-10-29 17:28:55

Apache Kylin™是一个开源的、分布式的分析型数据仓库，提供 Hadoop/Spark 之上的 SQL 查询接口及多维分析（OLAP）能力以支持超大规模数据，最初由 eBay 开发并贡献至开源社区。它能在亚秒内查询巨大的表。

Kylin 框架介绍

Kylin 能提供低延迟（sub-second latency）的秘诀就是预计算，即针对一个星型拓扑结构的数据立方体，预计算多个维度组合的度量，然后将结果保存在 hbase 中，对外提供 JDBC、ODBC、Rest API 的查询接口，即可实现实时查询。



Kylin 核心概念

- **表（table）**：表定义在 hive 中，是数据立方体（Data cube）的数据源，在 build cube 之前，必须同步在 kylin 中。
- **模型（model）**：模型描述了一个星型模式的数据结构，它定义了一个事实表（Fact Table）和多个查找表（Lookup Table）的连接和过滤关系。
- **Cube 描述**：描述一个 Cube 实例的定义和配置选项，包括使用了哪个数据模型、包含哪些维度和度量、如何将数据进行分区、如何处理自动合并等。

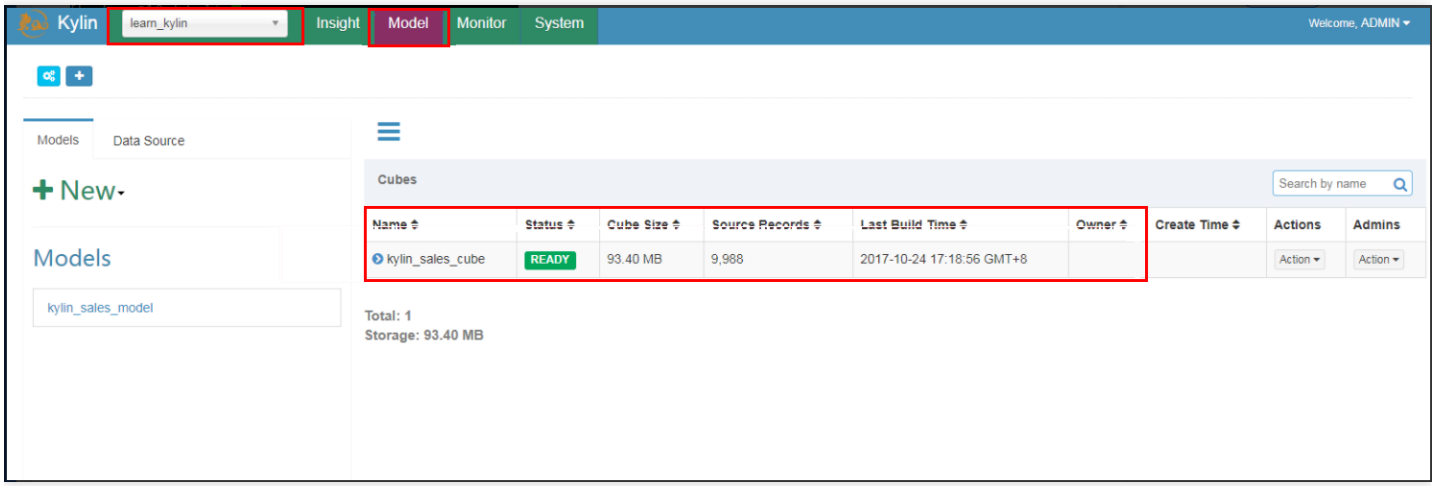
  - **Cube 实例**：通过 Cube 描述 Build 得到，包含一个或者多个 Cube Segment。
- **分区(Partition)**：用户可以在 Cube 描述中使用一个 DATA/STRING 的列作为分区的列，从而将一个 Cube 按照日期分割成多个 segment。
- **立方体段（cube segment）**：它是立方体构建（build）后的数据载体，一个 segment 映射 hbase 中的一张表，立方体实例构建（build）后，会产生一个新的 segment，一旦某个已经构建的立方体的原始数据发生变化，只需刷新（fresh）变化的时间段所关联的 segment 即可。
- **聚合组**：每一个聚合组是一个维度的子集，在内部通过组合构建 cuboid。
- **作业（job）**：对立方体实例发出构建（build）请求后，会产生一个作业。该作业记录了立方体实例 build 时的每一步任务信息。作业的状态信息反映构建立方体实例的结果信息。例如，作业执行的状态信息为 RUNNING 时，表明立方体实例正在被构建；作业状态信息为 FINISHED，表明立方体实例构建成功；作业状态信息为 ERROR，表明立方体实例构建失败。
- **DIMENSION & MEASURE 种类**
  - Mandotary：强制维度，所有 cuboid 必须包含的维度。
  - Hierarchy：层次关系维度，维度之间具有层次关系性，只需要保留一定层次关系的 cuboid 即可。
  - Derived：衍生维度，在 lookup 表中，有一些维度可以通过它的主键衍生得到，所以这些维度将不参加 cuboid 的构建。
  - Count Distinct(HyperLogLog)：直接进行 count distinct 是很难去计算的，一个近似的算法 HyperLogLog 可以保持错误率在一个很低的范围内。
  - Count Distinct(Precise)：将基于 RoaringBitMap 进行计算，目前只支持 int 和 BigInt。
- **Cube Action 种类**
  - BUILD：给定一个分区列指定的时间间隔，对 Cube 进行 Build，创建一个新的 cube Segment。
  - REFRESH：这个操作，将在一些分期周期内对 cube Segment 进行重新 build。
  - MERGE：这个操作将合并多个 cube segments。这个操作可以在构建 cube 时，设置为自动完成。
  - PURGE：清理一个 Cube 实例下的 segment，但是不会删除 HBase 表中的 Tables。
- **Job 状态**
  - NEW：表示一个 job 已经被创建。
  - PENDING：表示一个 job 已经被 job Scheduler 提交，等待执行资源。
  - RUNNING：表示一个 job 正在运行。
  - FINISHED：表示一个 job 成功完成。
  - ERROR：表示一个 job 因为错误退出。
  - DISCARDED：表示一个 job 被用户取消。
- **Job 执行**
  - RESUME：这个操作将从失败的 Job 的最后一个成功点继续执行该 Job。
  - DISCARD：无论工作的状态，用户可以结束它和释放资源。

Cube 快速入门示例

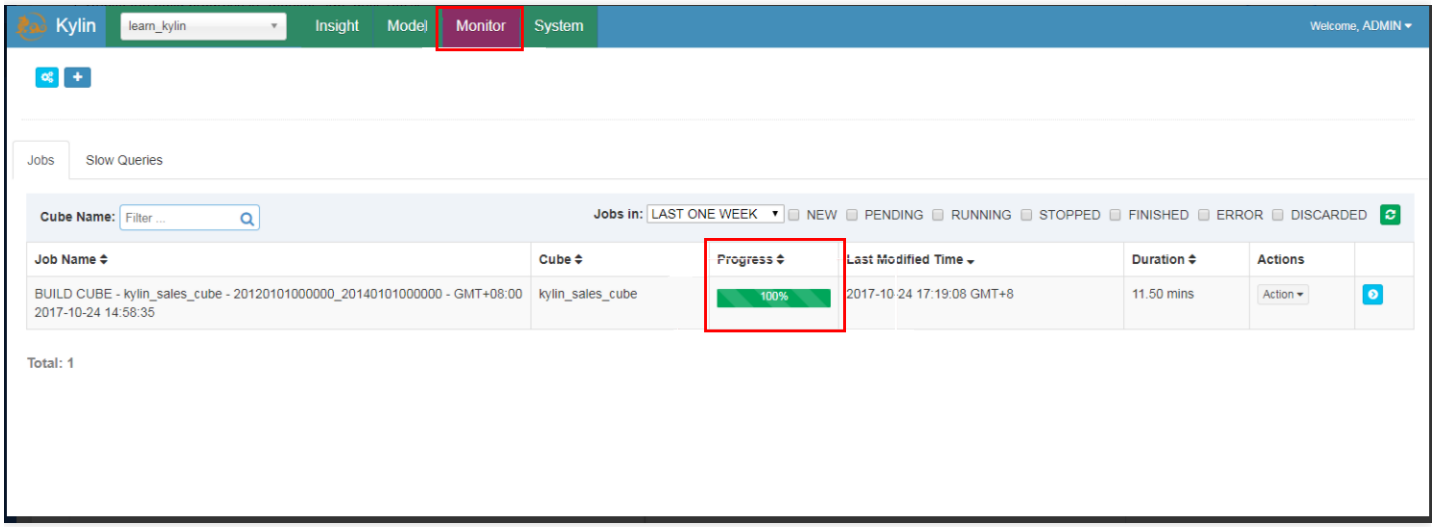
运行脚本，重启 Kylin 服务器刷新缓存。

```
/usr/local/service/kylin/bin/sample.sh
```

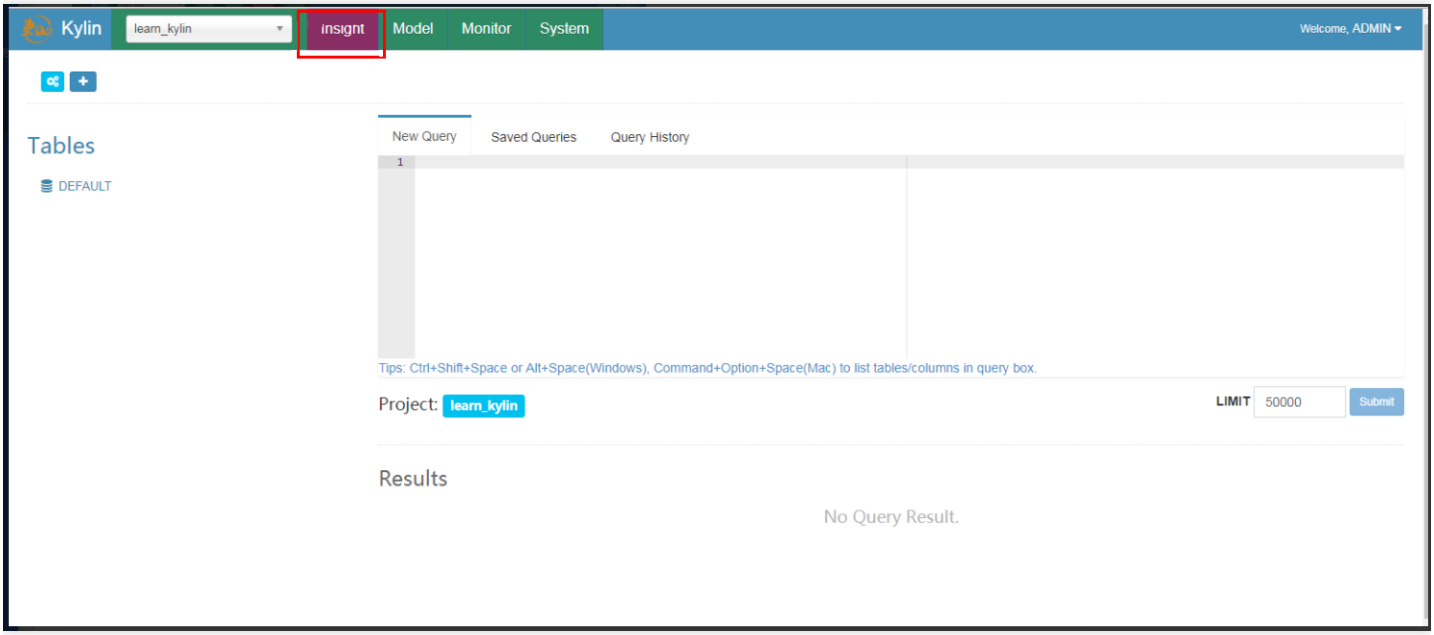
使用默认用户名和密码 ADMIN/KYLIN 登录 Kylin 网站，在左上角项目下拉框中选择 `learn_kylin` 工程，然后选择名为 `kylin_sales_cube` 的样例 Cube，选择**Actions>Build**，选择一个在2014-01-01之后的日期（覆盖所有的10000样例记录）。



单击**Monitor**，查看 build 进度直至 100%。



单击**Insight**，执行 SQLs，例如：




```
select part_dt, sum(price) as total_sold, count(distinct seller_id) as sellers fr
om kylin_sales group by part_dt order by part_dt
```

用 Spark 构建 Cube

1. 在 `kylin.properties` 中设置 `kylin.env.hadoop-conf-dir` 属性。

   ```
   kylin.env.hadoop-conf-dir=/usr/local/service/hadoop/etc/hadoop
   ```

2. 检查 Spark 配置

   Kylin 在 `$KYLIN_HOME/spark` 中嵌入一个 Spark binary (v2.1.2)，所有使用 `kylin.engine.spark-conf.` 作为前缀的 Spark 配置属性都能在 `$KYLIN_HOME/conf/kylin.properties` 中进行管理。这些属性当运行提交 Spark job 时会被提取并应用。例如，如果您配置 `kylin.engine.spark-conf.spark.executor.memory=4G`，Kylin 将会在执行 `spark-submit` 操作时使用 `-conf spark.executor.memory=4G` 作为参数。

   运行 Spark cubing 前，建议查看一下这些配置并根据您集群的情况进行自定义。下面是建议配置，开启了 Spark 动态资源分配：

```
kylin.engine.spark-conf.spark.master=yarn
kylin.engine.spark-conf.spark.submit.deployMode=cluster
kylin.engine.spark-conf.spark.dynamicAllocation.enabled=true
kylin.engine.spark-conf.spark.dynamicAllocation.minExecutors=1
kylin.engine.spark-conf.spark.dynamicAllocation.maxExecutors=1000
kylin.engine.spark-conf.spark.dynamicAllocation.executorIdleTimeout=300
kylin.engine.spark-conf.spark.yarn.queue=default
kylin.engine.spark-conf.spark.driver.memory=2G
kylin.engine.spark-conf.spark.executor.memory=4G
kylin.engine.spark-conf.spark.yarn.executor.memoryOverhead=1024
kylin.engine.spark-conf.spark.executor.cores=1
kylin.engine.spark-conf.spark.network.timeout=600
kylin.engine.spark-conf.spark.shuffle.service.enabled=true
#kylin.engine.spark-conf.spark.executor.instances=1
kylin.engine.spark-conf.spark.eventLog.enabled=true
kylin.engine.spark-conf.spark.hadoop.dfs.replication=2
kylin.engine.spark-conf.spark.hadoop.mapreduce.output.fileoutputformat.compress
=true
kylin.engine.spark-conf.spark.hadoop.mapreduce.output.fileoutputformat.compres
s.codec=org.apache.hadoop.io.compress.DefaultCodec
kylin.engine.spark-conf.spark.io.compression.codec=org.apache.spark.io.SnappyCo
mpressionCodec
kylin.engine.spark-conf.spark.eventLog.dir=hdfs\:///kylin/spark-history
kylin.engine.spark-conf.spark.history.fs.logDirectory=hdfs\:///kylin/spark-hist
ory
## uncomment for HDP
#kylin.engine.spark-conf.spark.driver.extraJavaOptions=-Dhdp.version=current
#kylin.engine.spark-conf.spark.yarn.am.extraJavaOptions=-Dhdp.version=current
#kylin.engine.spark-conf.spark.executor.extraJavaOptions=-Dhdp.version=current
```

为了在 Hortonworks 平台上运行，需要将 `hdp.version` 指定为 Yarn 容器的 Java 选项，因此需取消 `kylin.properties` 中最后三行的注释。

除此之外，为了避免重复上传 Spark jar 包到 Yarn，您可以手动上传一次，然后配置 jar 包的 HDFS 路径。**HDFS 路径必须是全路径名。**

```
jar cv0f spark-libs.jar -C $KYLIN_HOME/spark/jars/ .
hadoop fs -mkdir -p /kylin/spark/
hadoop fs -put spark-libs.jar /kylin/spark/
```

然后，要在 `kylin.properties` 中进行如下配置：

```
kylin.engine.spark-conf.spark.yarn.archive=hdfs://sandbox.hortonworks.com:8020/kylin/spark/spark-libs.jar
```

所有 `kylin.engine.spark-conf.*` 参数都可以在 Cube 或 Project 级别进行重写，这为用户提供了灵活性。

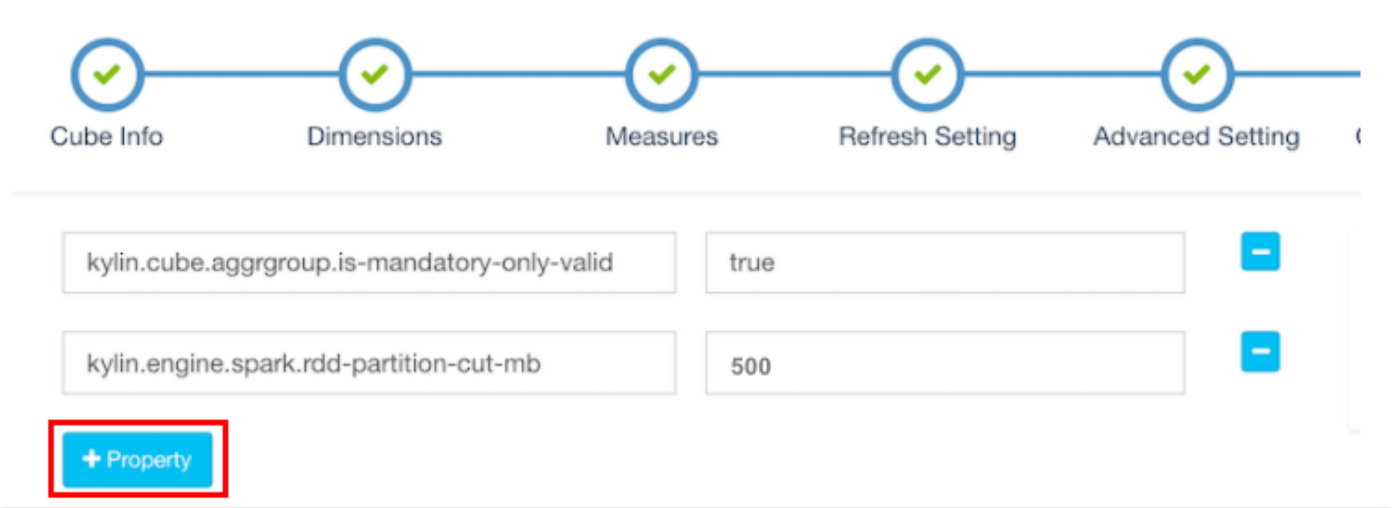
3. 创建和修改样例 cube

运行 `sample.sh` 创建样例 cube，然后启动 Kylin 服务器：

```
/usr/local/service/kylin/bin/sample.sh
/usr/local/service/kylin/bin/kylin.sh start
```

Kylin 启动后，访问 Kylin 网站，在“Advanced Setting”页，编辑名为 `kylin_sales` 的 cube，将**Cube Engine**由**MapReduce**修改为**Spark(Beta)**：

单击**Next**进入“Configuration Overwrites”页面，单击**+ Property**添加属性 `kylin.engine.spark.rdd-partition-cut-mb` 其值为500。

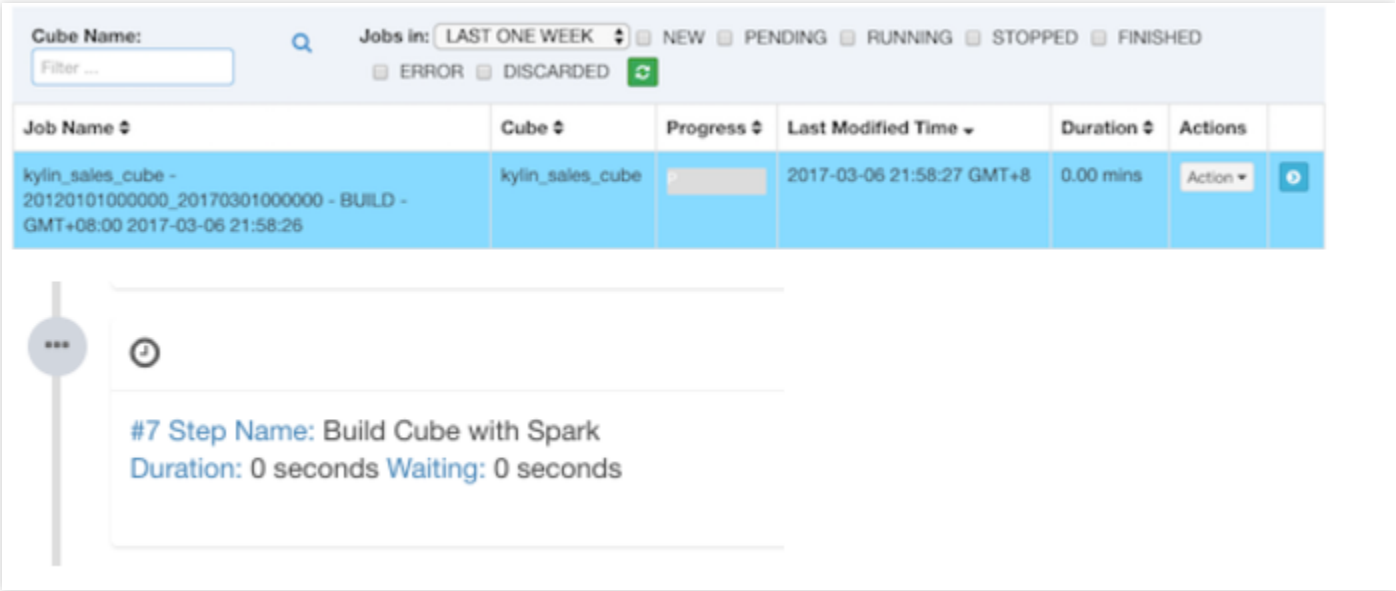


样例 cube 有两个耗尽内存的度量：COUNT DISTINCT 和 TOPN(100)。当源数据较小，它们预估的大小会比真实的大很多，导致了更多的 RDD partitions 被切分，使得 build 的速度降低。500是一个较为合理的数字。单击**Next**和**Save**保存 cube。

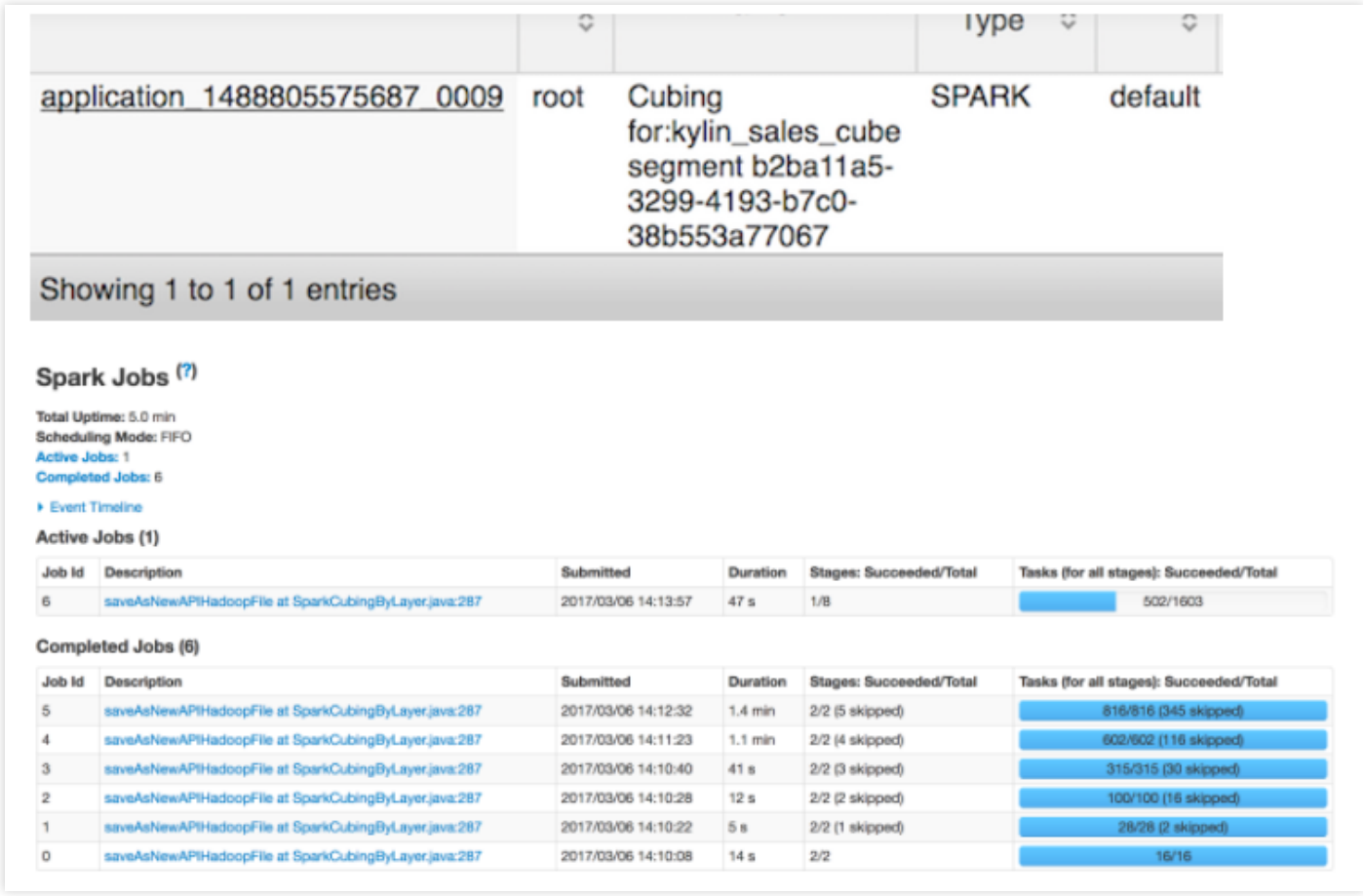
> 说明：
> 对于没有 COUNT DISTINCT 和 TOPN 的 cube，请保留默认配置。

4. 用 Spark 构建 Cube

单击**Build**，选择当前日期为 end date。Kylin 会在“Monitor”页生成一个构建 job，第7步是 Spark cubing。Job engine 开始按照顺序执行每一步。



当 Kylin 执行这一步时，您可以监视 Yarn 资源管理器中的状态，单击“Application Master”链接将会打开 Spark 的 UI 网页，它会显示每一个 stage 的进度以及详细的信息。



所有步骤成功执行后，Cube 的状态变为“Ready”，您即可正常进行查询。

Livy 开发指南

Livy 简介

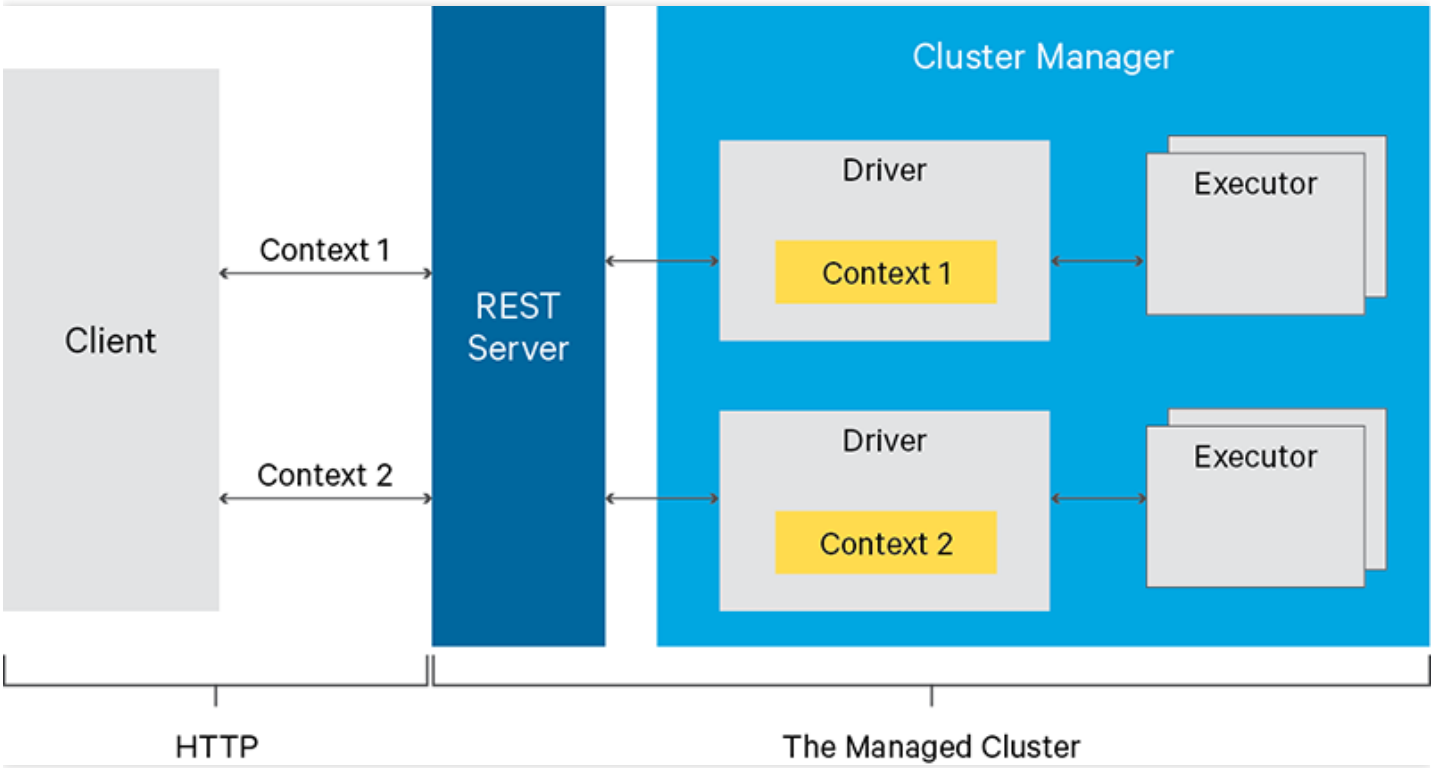
最近更新时间：2022-05-16 12:53:36

Apache Livy 是一个可以通过 REST 接口与 Spark 集群进行交互的服务，它可以提交 Spark 作业或者 Spark 代码片段，同步或者异步的进行结果检索以及 Spark Context 上下文管理，Apache Livy 简化 Spark 和应用程序服务器之间的交互，从而使 Spark 能够用于交互式 Web/移动应用程序。

Livy 特性

Livy 还支持如下功能：

- 由多个客户端长时间运行可用于多个 Spark 作业的 Spark 上下文。
- 跨多个作业和客户端共享缓存的 RDD 或数据帧。
- 可以同时管理多个 Spark 上下文，并且 Spark 上下文运行在群集（YARN/Mesos）而不是 Livy 服务器，以实现良好的容错性和并发性。
- 作业可以作为预编译的 jar，代码片段或通过 java/scala 客户端 API 提交。
- 通过安全的认证通信确保安全。



使用 Livy

1. 访问 `http://IP:8998/ui` 可以进入 Livy 的 UI 页面（**IP 为外网 IP，请自行为安装有 Livy 的机器申请外网 IP，并编辑设置安全组策略来开通对应的端口以进行访问**）。

2. 创建一个交互式会话。

```
curl -X POST --data '{"kind":"spark"}' -H "Content-Type:application/json" IP:8998/sessions
```

3. 查看 Livy 上存活的 sessions。

```
curl IP:8998/sessions
```

4. 执行代码片段，简单的加法操作（这里相当于指定的 session 0，如果有多个 session，也可以指定其他 session）。

```
curl -X POST IP:8998/sessions/0/statements -H "Content-Type:application/json" -d '{"code":"1+1"}'
```

5. 计算圆周率（执行 jar 包）。
   步骤1：上传 jar 包到 hdfs，如上传至/usr/local/spark-examples_2.11-2.4.3.jar。
   步骤2：执行命令：

```
curl -H "Content-Type: application/json" -X POST -d
'{ "file":"/usr/local/spark-examples_2.11-2.4.3.jar",
"className":"org.apache.spark.examples.SparkPi" }' IP:8998/batches
```

6. 查询代码片段执行是否成功，也可以直接在 UI 页面 `http://IP:8998/ui/session/0` 查看。

```
curl IP:8998/sessions/0/statements/0
```

7. 删除 session。

```
curl -X DELETE IP:8998/sessions/0
```

注意事项

开放的配置文件

目前开放的配置文件包括 `livy.conf` 和 `livy-env.sh`，这两个配置文件均可以通过配置下发的方式来修改配置。具体开放了哪几个配置文件，请以 EMR 控制台实际为准。

Livy 端口修改方法

目前的默认端口是8998，用户可以自行修改。修改配置文件 `livy.conf` 的 `livy.server.port` 属性即可。

若集群装有 Hue，由于 Hue 与 Livy 之间涉及联通性，因此 Hue 对应的配置端口也需要修改。其对应的配置文件为 `pseudo-distributed.ini`，路径为 `/usr/local/service/hue/desktop/conf`，对应的配置项为 `livy_server_port=8998`。修改后要重启对应的服务。

如非必要，建议不要修改 Livy 的端口，如涉及安全要求，可以通过安全组的方式来进行控制。修改后可能会导致一些其他的潜在问题。

Livy 部署方式

目前 Livy 默认会在所有 master 节点上部署，用户也可以通过扩容 router 的方式在 router 上进行部署。

Kyuubi 开发指南

Kyuubi 简介

最近更新时间：2022-05-16 12:52:25

Apache Kyuubi （Incubating）是一个 Thrift JDBC/ODBC 服务，目前对接了 Apache Spark 计算框架（正在对接 Apache Flink计算框架以及Trino），支持多租户和分布式等特性，可以满足企业内诸如 ETL、BI 报表等多种大数据场景的应用。

使用场景

- 替换 HiveServer2，轻松获得 10~100 倍性能提升。
  - Kyuubi 高度兼容 HiveServer2 接口及行为，支持无缝迁移。

  - Kyuubi 分层架构，消除客户端兼容性问题，支持无感升级。
  - Kyuubi 支持 Spark SQL 全链路优化及再增强，性能卓著。
  - 高可用、多租户、细粒度权限认证各种企业级特性都有。
- 构建 Serverless Spark 平台。
  - Serverless Spark 目标绝对不是让用户调用 Spark 的 API、继续写 Spark 作业。
  - 通过 Kyuubi 预置的 Engine 模块，用户无需理解 Spark 逻辑，入门门槛极低。
  - 用户只需通过 JDBC 及 SQL 操作数据专注自身业务开发即可，资源弹性伸缩，0运维。
  - 支持资源管理器（Kubernetes, YARN 等），Engine 生命周期，Spark 动态资源分配3级不同粒度全方位的资源弹性策略。
  - 支持 YARN/Kubernetes 多种资源管理器同时调度，保障历史作业安全迁移上云。
  - Spark 自适应查询引擎（AQE）及 Kyuubi AQE plus，提供澎湃动力。
- 构建统一数据湖探索分析管理平台（kyuubi-1.5以上版本）。
  - 支持 Spark 所有官方数据源及第三方数据源。
  - 支持 Spark DSv2 元数据管理，直观进行数据湖构建及管理。
  - 支持 Apache Iceberg/Hudi, DeltaLake 等所有主流数据湖框架。
  - 一个接口一个引擎一份数据，提供统一的分析查询、数据摄取、数据湖管理平台。
  - 批流一体，支持流式作业（Upcoming）。

Kyuubi 最佳实践

最近更新时间：2022-06-15 15:04:18

Beeline 连接 kyuubi

登录 EMR 集群的 Master 节点，切换到 Hadoop 用户并且进入 kyuubi 目录：

```
[root@172 ~]# su hadoop
[hadoop@172 root]$ cd /usr/local/service/kyuubi
```

连接 kyuubi：

```
[hadoop@10kyuubi]$ bin/beeline -u "jdbc:hive2://${zkserverip1}:${zkport},${zkserv
erip2}:${zkport},${zkserverip3}:${zkport}/default;serviceDiscoveryMode=zooKeeper;
zooKeeperNamespace=kyuubi" -n hadoop
```

或者

```
[hadoop@10kyuubi]$ bin/beeline -u "jdbc:hive2://${kyuubiserverip}:${kyuubiserverp
ort}" -n hadoop
```

`${zkserverip}:${zkport}` 见 `kyuubi-defaults.conf` 的 `kyuubi.ha.zookeeper.quorum` 配置。

`${kyuubiserverport}` 见 `kyuubi-defaults.conf` 的 `kyuubi.frontend.bind.port` 配置。

新建数据库并查看

在新建的数据库中新建一个表，并进行查看：

```
0: jdbc:hive2://ip:port> create database sparksql;
+---------+
| Result |
+---------+
+---------+
No rows selected (0.326 seconds)
```

向表中插入两行数据并查看：

```
0: jdbc:hive2://ip:port> use sparksql;
+---------+
| Result |
+---------+
```

```
+---------+
No rows selected (0.077 seconds)
0: jdbc:hive2://ip:port> create table sparksql_test(a int,b string);
+---------+
| Result |
+---------+
+---------+
No rows selected (0.402 seconds)
0: jdbc:hive2://ip:port> show tables;
+-----------+---------------+-------------+
| database | tableName | isTemporary |
+-----------+---------------+-------------+
| sparksql | sparksql_test | false |
+-----------+---------------+-------------+
1 row selected (0.108 seconds)
```

Hue 连接 kyuubi

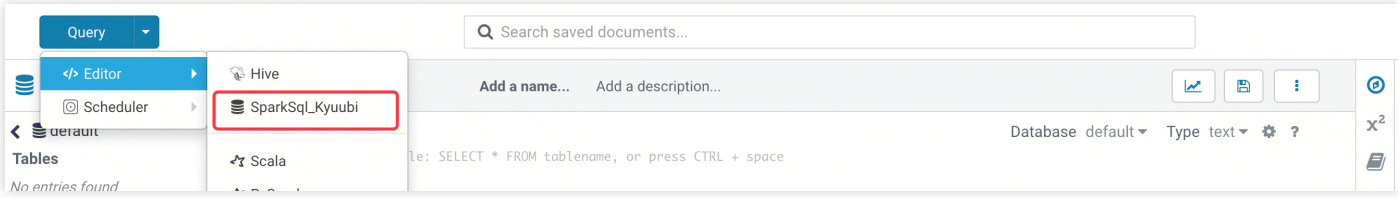
前提条件

针对在现有集群里后安装 kyuubi 场景，若想在 hue 上使用 kyuubi，需要做如下操作：

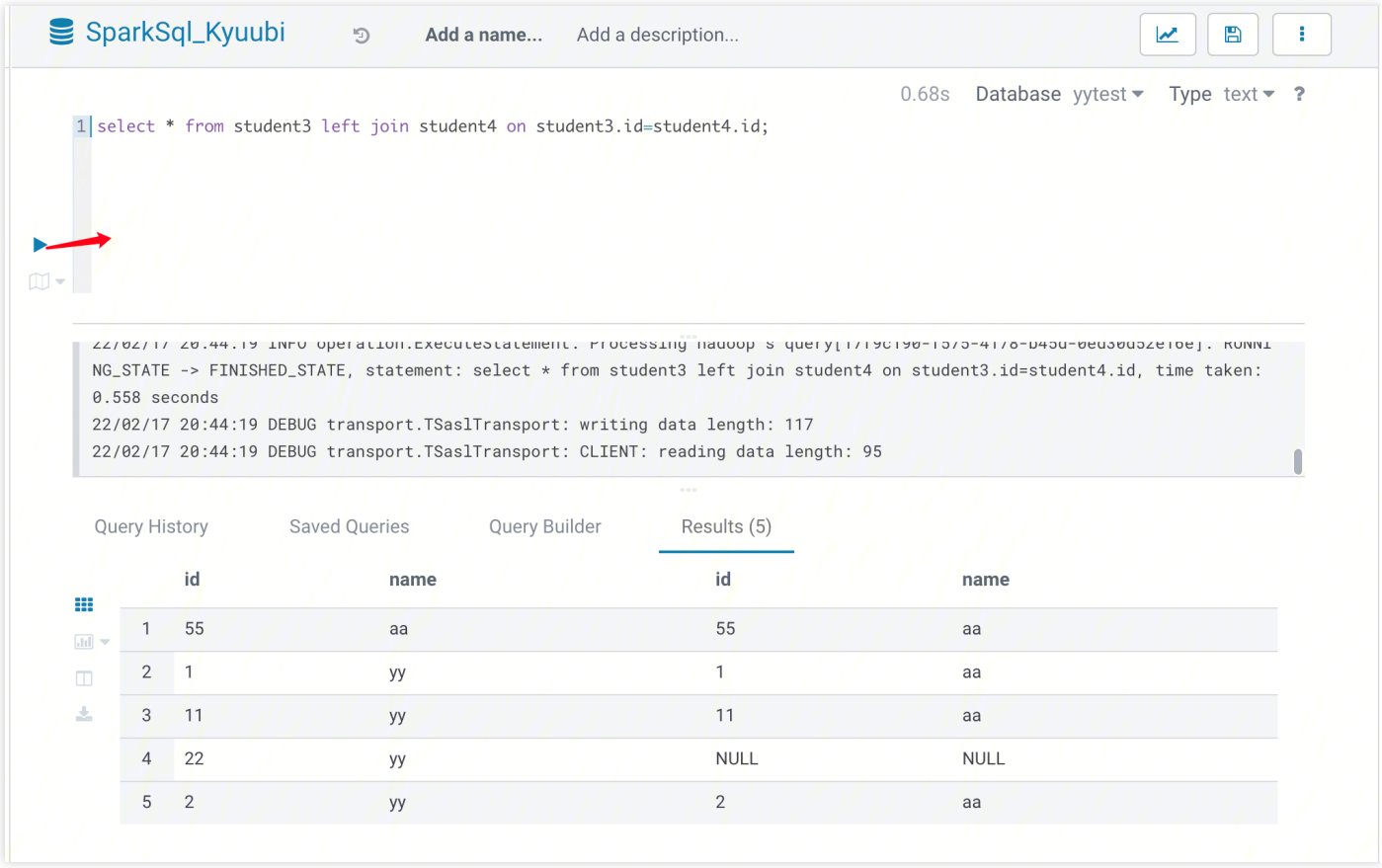
1. 进入 HDFS 的配置管理，core-site.xml 新增配置项 hadoop.proxyuser.hue.groups 值设为“ * ”和 hadoop.proxyuser.hue.hosts值设为“ * ” 。
2. 重启 kyuubi 服务和 hue 服务。
3. 访问 Hue 控 制台，详情请参见 登录 Hue 控制台。

Kyuubi 查询

1. 在 Hue 控制台上方，选择 Query > Editor > SparkSql_Kyuubi。



2. 在语句输入框中输入要执行语句，然后单击**执行**，执行语句。



	id	name	id	name
1	55	aa	55	aa
2	1	yy	1	aa
3	11	yy	11	aa
4	22	yy	NULL	NULL
5	2	yy	2	aa

通过 Java 连接 Kyuubi

KyuubiServer 中集成了 Thrift 服务。Thrift 是 Facebook 开发的一个软件框架，它用来进行可扩展且跨语言的服务的开发。Kyuubi 就是基于 Thrift 的，所以能让不同的语言如 Java、Python 来调用 Kyuubi 的接口。对于 Java，Kyuubi 复用了hive jdbc 驱动，用户可以使用 Java 代码来连接 Kyuubi 并进行一系列操作。本节将演示如何使用 Java 代码来连接 Kyuubi。

1. 开发准备。
   - 确认您已经开通了腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Kyuubi 及 Spark 组件。
   - Kyuubi 等相关软件安装在路径 EMR 云服务器的/usr/local/service/路径下。

2. 使用 Maven 来创建您的工程。
   推荐使用 Maven 来管理您的工程。Maven 是一个项目管理工具，能够帮助您方便的管理项目的依赖信息，即它可以通过 pom.xml 文件的配置获取 jar 包，而不用去手动添加。首先在本地下载并安装 Maven，配置好 Maven 的环

境变量，如果您使用 IDE，请在 IDE 中设置好 Maven 相关配置。
在本地 shell 下进入要新建工程的目录，例如：D://mavenWorkplace 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID -Dar
chetypeArtifactId=maven-archetype-quickstart
```

其中 $yourgroupID 即为您的包名；$yourartifactID 为您的项目名称；maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。创建成功后，在 D://mavenWorkplace 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
---pom.xml          核心配置，项目根下
---src
---main
---java          Java 源码目录
---resources     Java 配置文件目录
---test
---java          测试源码目录
---resources     测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。首先在 pom.xml 中添加 Maven 依赖：

```
<dependencies>
<dependency>
<groupId>org.apache.kyuubi</groupId>
<artifactId>kyuubi-hive-jdbc-shaded</artifactId>
<version>1.4.1-incubating</version>
</dependency>
<dependency>
<groupId>org.apache.hadoop</groupId>
<artifactId>hadoop-common</artifactId>
<!-- keep consistent with the build hadoop version -->
<version>2.8.5</version>
</dependency>
</dependencies>
```

继续在 pom.xml 中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
```

```
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

在 src>main>Java 下右键新建一个 Java Class，输入您的 Class 名，这里使用 KyuubiJDBCTest.java，在 Class 添加样例代码：

```
import java.sql.*;
public class KyuubiJDBCTest {
private static String driverName =
"org.apache.kyuubi.jdbc.KyuubiHiveDriver";
public static void main(String[] args)
throws SQLException {
try {
Class.forName(driverName);
} catch (ClassNotFoundException e) {
e.printStackTrace();
System.exit(1);
}
Connection con = DriverManager.getConnection(
"jdbc:hive2://$kyuubiserverhost:$kyuubiserverport/default", "hadoop", "");
Statement stmt = con.createStatement();
String tableName = "KyuubiTestByJava";
```

```
stmt.execute("drop table if exists " + tableName);
stmt.execute("create table " + tableName +
" (key int, value string)");
System.out.println("Create table success!");
// show tables
String sql = "show tables '" + tableName + "'";
System.out.println("Running: " + sql);
ResultSet res = stmt.executeQuery(sql);
if (res.next()) {
System.out.println(res.getString(1));
}
// describe table
sql = "describe " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(res.getString(1) + "\t" + res.getString(2));
}
sql = "insert into " + tableName + " values (42,\"hello\"),(48,\"world\")";
stmt.execute(sql);
sql = "select * from " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(String.valueOf(res.getInt(1)) + "\t"
+ res.getString(2));
}
sql = "select count(1) from " + tableName;
System.out.println("Running: " + sql);
res = stmt.executeQuery(sql);
while (res.next()) {
System.out.println(res.getString(1));
}
}
}
```

注意：
将程序中的参数 $kyuubiserverhost 和 $kyuubiserverport 分别修改为您查到的 KyuubiServer 的 ip 和端口号的值。

如果您的 Maven 配置正确并且成功的导入了依赖包，那么整个工程即可直接编译。在本地 shell 下进入工程目录，执行下面的命令对整个工程进行打包。

```
mvn package
```

3. 上传并运行程序。

首先需要把压缩好的 jar 包上传到 EMR 集群中，使用 scp 或者 sftp 工具来进行上传。在本地 shell 下运行：

```
scp $localfile root@公网IP地址:/usr/local/service/kyuubi
```

一定要上传具有依赖的 jar 包。登录 EMR 集群切换到 Hadoop 用户并且进入目录 /usr/local/service/kyuubi 。接下来可以执行程序：

```
[hadoop@172 kyuubi]$ yarn jar $package.jar KyuubiJDBCTest
```

其中 $package.jar 为您的 jar 包的路径 + 名字，KyuubiJDBCTest 为之前的 Java Class 的名字。运行结果如下：

```
Create table success!
Running: show tables 'KyuubiTestByJava'
default
Running: describe KyuubiTestByJava
key int
value string
Running: select * from KyuubiTestByJava
42 hello
48 world
Running: select count(1) from KyuubiTestByJava
2
```

Zeppelin 开发指南

Zeppelin 简介

最近更新时间：2023-05-08 15:22:21

Apache Zeppelin 是一款基于 Web 的 Notebook 产品，能够交互式数据分析。使用 Zeppelin，您可以使用丰富的预构建语言后端（或解释器）制作交互式的协作文档，例如 Scala(Apache Spark)、Python(Apache Spark)、SparkSQL、 Hive、Shell 等。

> 说明：
> EMR-V3.3.0及以上、EMR-V2.6.0及以上，已默认配置了 flink、hbase、kylin、livy、spark 的 Interpreter，其他版本和组件可参考 官方文档 根据 Zepplin 版本进行配置。

前提条件

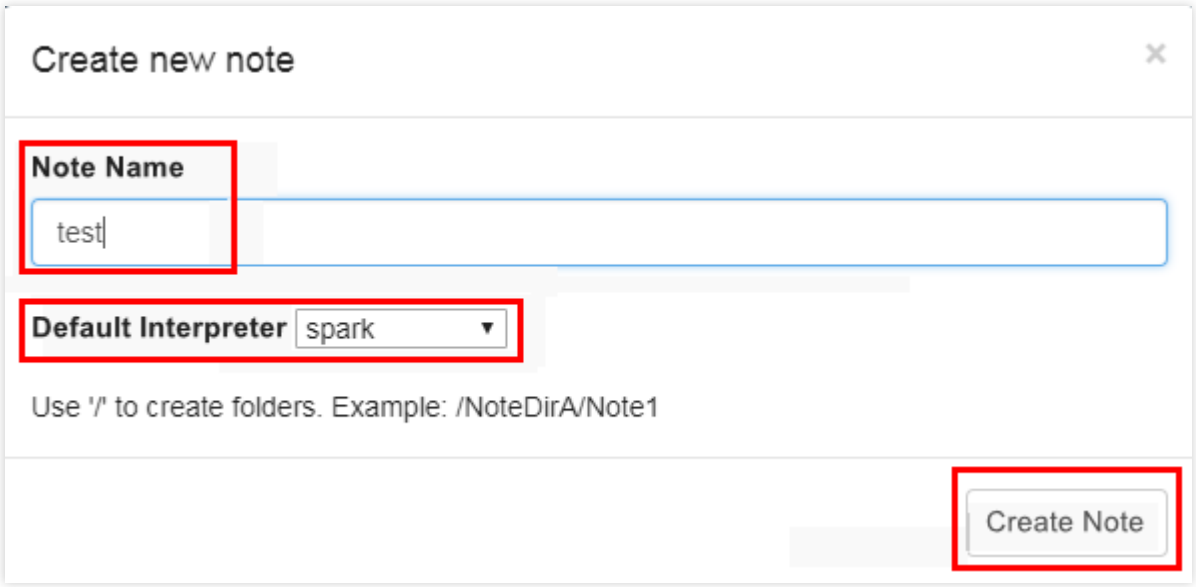
- 已创建集群，并选择 Zeppelin 服务，详情参见 创建 EMR 集群。
- 在集群的 EMR 安全组中，开启22、30001和18000端口（新建集群默认开启22和30001）及必要的内网通信网段，新安全组以 emr-xxxxxxxx_yyyyMMdd 命名，请勿手动修改安全组名称。
- 按需添加所需服务，如，Spark、Flink、HBase、Kylin。

登录 Zeppelin

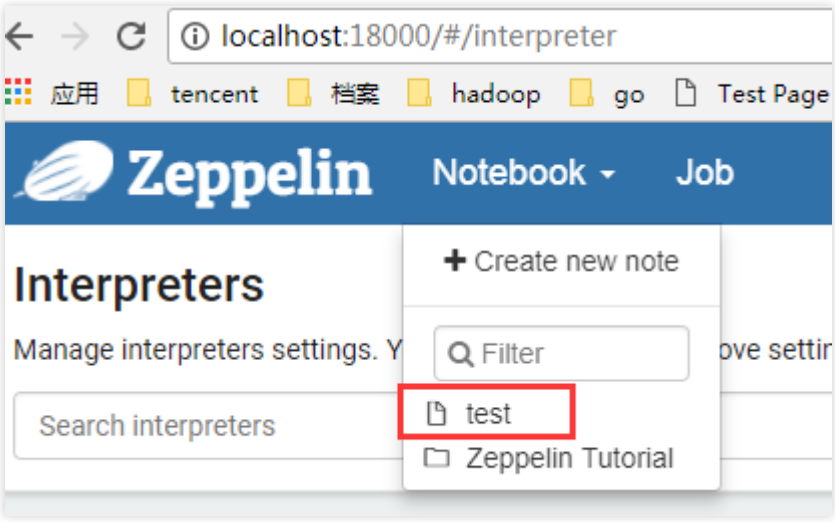
1. 创建集群，选择 Zeppelin 服务，详情参见 创建 EMR 集群。
2. 在 EMR 控制台 左侧的导航栏，选择集群服务。
3. 单击 Zeppelin 所在的卡片，单击 **Web UI 地址**，访问 Web UI 页面。
4. 在 EMR-V2.5.0 及以前版本、EMR-V3.2.1 及以前版本，设置了默认登录权限，用户名密码为 admin:admin。如需更改密码，可修改配置文件/usr/local/service/zeppelin-0.8.2/conf/shiro.ini 中的 users 和 roles 选项。更多配置说明，可参见 文档。
5. 在 EMR-V2.6.0 及以后版本、 EMR-V3.3.0 及以后版本，Zeppelin 登录已集成 Openldap 账户，只能用 Openldap 账户密码登录，新建集群后 Openldap 默认账户是 root 和 hadoop，默认密码是集群密码，且只有 root 账户拥有 zeppelin 管理员权限，有权访问解析器配置页面。

使用 spark 功能完成 wordcount

1. 单击页面左侧 **Create new note**，在弹出页面中创建 notebook。

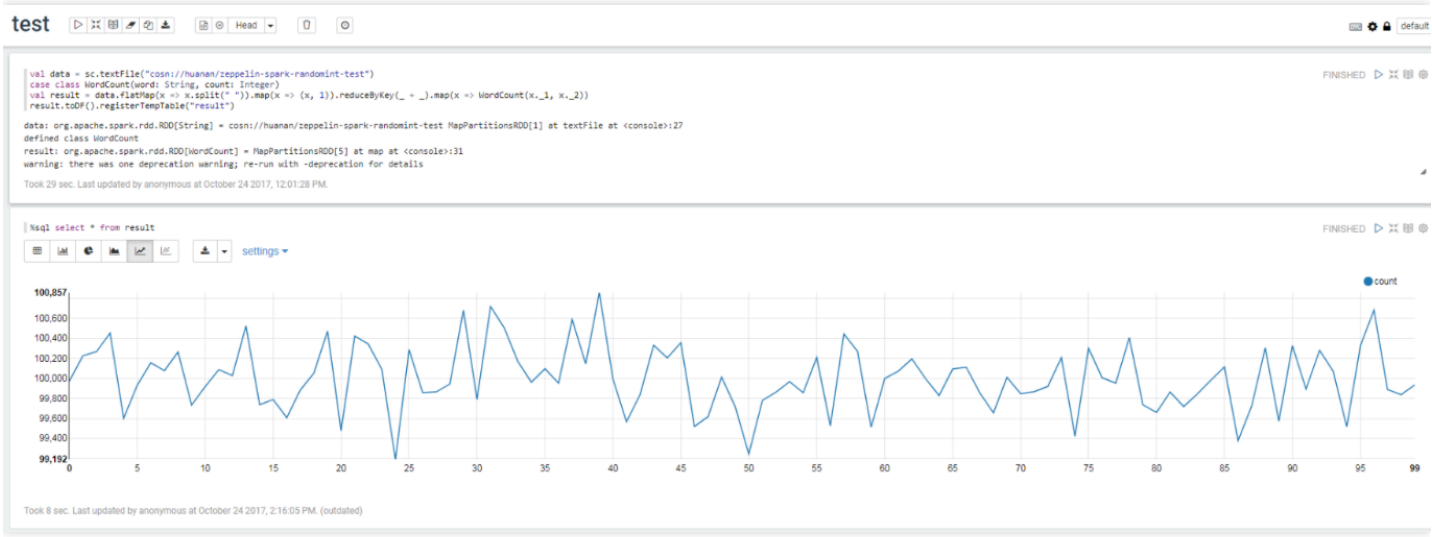


2. 2.EMR-V3.3.0 及以上、EMR-V2.6.0 及以上，已默认配置 Spark 对接 EMR 的集群（Spark On Yarn）。
   - 如果您的版本是 EMR-V3.1.0、EMR-V2.5.0、EMR-V2.3.0，请参考 文档 进行 Spark 解释器配置。
   - 如果您的版本是 EMR-V3.2.1，请参考 文档 进行 Spark 解释器配置。
3. 进入自己的 notebook。



4. 编写 wordcount 程序，并运行如下命令：

```
val data = sc.textFile("cosn://huanan/zeppelin-spark-randomint-test")
case class WordCount(word: String, count: Integer)
val result = data.flatMap(x => x.split(" ")).map(x => (x, 1)).reduceByKey(_ + _
).map(x => WordCount(x._1, x._2))
result.toDF().registerTempTable("result")
%sql select * from result
```


test
Head
default
val data = sc.textFile("cosn://huanan/zeppelin-spark-randomint-test")
case class WordCount(word: String, count: Integer)
val result = data.flatMap(x => x.split(" ")).map(x => (x, 1)).reduceByKey(_ + _).map(x => WordCount(x._1, x._2))
result.toDF().registerTempTable("result")
FINISHED
data: org.apache.spark.rdd.RDD[String] = cosn://huanan/zeppelin-spark-randomint-test MapPartitionsRDD[1] at textFile at <console>:27
defined class WordCount
result: org.apache.spark.rdd.RDD[WordCount] = MapPartitionsRDD[5] at map at <console>:31
warning: there was one deprecation warning; re-run with -deprecation for details
Took 29 sec. Last updated by anonymous at October 24 2017, 12:01:28 PM.
%sql select * from result
FINISHED
settings
count
100,857
100,600
100,400
100,200
100,000
99,800
99,600
99,400
99,192
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
99
Took 8 sec. Last updated by anonymous at October 24 2017, 2:16:05 PM. (outdated)

Zeppelin 解析器配置

最近更新时间：2023-05-08 15:22:21

本文以 Zepplin 0.91 以上版本为例，主要介绍常见 Zeppelin 解析的配置以及验证方法。

Spark 解析器

配置

```
SPARK_HOME: /usr/local/service/spark
spark.master: yarn
spark.submit.deployMode: cluster
spark.app.name: zeppelin-spark
```

验证

1. 先把 wordcount.txt 文件上传到 emr hdfs 的/tmp 路径下。
2. 通过 core-site.xml 找到 fs.defaultFS 配置项值 hdfs://HDFS45983。
3. 在 notebook 中执行 spark 相关代码。

```
%spark
val data = sc.textFile("hdfs://HDFS45983/tmp/wordcount.txt")
case class WordCount(word: String, count: Integer)
val result = data.flatMap(x => x.split(" ")).map(x => (x, 1)).reduceByKey(_ + _
).map(x => WordCount(x._1, x._2))
result.toDF().registerTempTable("result")
```

```
%sql
select * from result
```

Flink 解析器

配置

```
FLINK_HOME: /usr/local/service/flink

flink.execution.mode: yarn
```

验证

```
%flink
val data = benv.fromElements("hello world", "hello flink", "hello hadoop")
data.flatMap(line => line.split("\\s"))
.map(w => (w, 1))
.groupBy(0)
.sum(1)
.print()
```

HBase 解析器

配置

```
hbase.home: /usr/local/service/hbase

hbase.ruby.sources: lib/ruby

zeppelin.hbase.test.mode: false
```

注意：
此解析器依赖的 jar 包已集成到集群/usr/local/service/zeppelin/local-repo 路径下，因此不用配置 dependencies，如需已定义 jar 包才需配置dependencies。

验证

```
%hbase
help 'get'

%hbase
list
```

Livy 解析器

配置

```
zeppelin.livy.url: http://ip:8998
```

验证

```
%livy.spark
sc.version

%livy.pyspark
print "1"

%livy.sparkr
hello <- function( name ) {
sprintf( "Hello, %s", name );
}
hello("livy")
```

Kylin 解析器

配置

1. 在 Kylin 控制台页面中新建一个 default 的 Project。
2. 配置 zeppelin 的 kylin interpreter。

```
kylin.api.url: http://ip:16500/kylin/api/query


kylin.api.user: ADMIN

kylin.api.password: KYLIN


kylin.query.project: default
```

验证

```
%kylin(default)

select count(*) from table1
```

JDBC 解析器

1. MySQL 解析器配置

```
default.url: jdbc:mysql://ip:3306

default.user: xxx

default.password: xxx

default.driver: com.mysql.jdbc.Driver
```

> 注意：
> 此解析器依赖的 jar 包已集成到集群/usr/local/service/zeppelin/local-repo 路径下，因此不用配置 dependencies，如需已定义 jar 包才需配置 dependencies。

验证

```
%mysql
show databases
```

2. Hive 解析器配置

```
default.url: jdbc:hive2://ip:7001

default.user: hadoop

default.password:

default.driver: org.apache.hive.jdbc.HiveDriver
```

注意：
此解析器依赖的 jar 包已集成到集群/usr/local/service/zeppelin/local-repo 路径下，因此不用配置 dependencies，如需已定义 jar 包才需配置 dependencies。

验证

```
%hive
show databases

%hive
use default;
show tables;
```

3. Presto 解析器配置

```
default.url: jdbc:presto://ip:9000?user=hadoop

default.user: hadoop

default.password:

default.driver: io.prestosql.jdbc.PrestoDriver
```

注意：
此解析器依赖的 jar 包已集成到集群/usr/local/service/zeppelin/local-repo 路径下，因此不用配置 dependencies，如需已定义 jar 包才需配置 dependencies。

验证

```
%presto
show catalogs;

%presto
show schemas from hive;

%presto
show tables from hive.default;
```

更多版本及解析器配置请参考 Zeppelin 官网文档。

Hudi 开发指南
Hudi 简介

最近更新时间：2023-03-15 10:12:39

Apache Hudi 在 HDFS 的数据集上提供了插入更新和增量拉取的流原语。

一般来说，我们会将大量数据存储到 HDFS，新数据增量写入，而旧数据鲜有改动，特别是在经过数据清洗，放入数据仓库的场景。而且在数据仓库如 hive 中，对于 update 的支持非常有限，计算昂贵。另一方面，若是有仅对某段时间内新增数据进行分析的场景，则 hive、presto、hbase 等也未提供原生方式，而是需要根据时间戳进行过滤分析。

在此需求下，Hudi 可以提供这两种需求的实现。第一个是对 record 级别的更新，另一个是仅对增量数据的查询。且 Hudi 提供了对 Hive、presto、Spark 的支持，可以直接使用这些组件对 Hudi 管理的数据进行查询。

Hudi 是一个通用的大数据存储系统，主要特性：

- 摄取和查询引擎之间的快照隔离，包括 Apache Hive、Presto 和 Apache Spark。
- 支持回滚和存储点，可以恢复数据集。
- 自动管理文件大小和布局，以优化查询性能准实时摄取，为查询提供最新数据。
- 实时数据和列数据的异步压缩。

时间轴

在它的核心，Hudi 维护一条包含在不同的**即时**时间所有对数据集操作的**时间轴**，从而提供了从不同时间点出发得到不同的视图下的数据集。

Hudi 即时包含以下组件：

- 操作类型：对数据集执行的操作类型。
- 即时时间：即时时间通常是一个时间戳（例如20190117010349），该时间戳按操作开始时间的顺序单调增加。
- 状态：即时的状态。

文件组织

Hudi 将 DFS 上的数据集组织到 `基本路径` 下的目录结构中。数据集分为多个分区，这些分区是包含该分区的数据文件的文件夹，这与 Hive 表非常相似。

每个分区被相对于基本路径的特定 `分区路径` 区分开来。在每个分区内，文件被组织为 `文件组` ，由 `文件id` 唯一标识。每个文件组包含多个 `文件切片` ，其中每个切片包含在某个提交/压缩即时时间生成的基本列文

件 `*.parquet` 以及一组日志文件 `*.log*` ，该文件包含自生成基本文件以来对基本文件的插入/更新。

Hudi 采用 MVCC 设计，其中压缩操作将日志和基本文件合并以产生新的文件片，而清理操作则将未使用的/较旧的文件片删除以回收 DFS 上的空间。Hudi 通过索引机制将给定的 hoodie 键（记录键+分区路径）映射到文件组，从而提供了高效的 Upsert。

一旦将记录的第一个版本写入文件，记录键和 `文件组` / `文件id` 之间的映射就永远不会改变。简而言之，映射的文件组包含一组记录的所有版本。

存储类型

Hudi 支持以下存储类型：

- 写时复制：仅使用列文件格式（例如 parquet）存储数据。通过在写入过程中执行同步合并以更新版本并重写文件。
- 读时合并：使用列式（例如 parquet）+ 基于行（例如 avro）的文件格式组合来存储数据。 更新记录到增量文件中，然后进行同步或异步压缩以生成列文件的新版本。

下表总结了这两种存储类型之间的权衡：

权衡	写时复制	读时合并
数据延迟	更高	更低
更新代价（I/O）	更高（重写整个 parquet 文件）	更低（追加到增量日志）
Parquet 文件大小	更小（高更新代价（I/o））	更大（低更新代价）
写放大	更高	更低（取决于压缩策略）

Hudi 对 EMR 底层存储支持

- HDFS
- COS

安装 Hudi

进入 EMR 购买页，选择**产品版本**为 EMR-V2.2.0，选择**可选组件**为 **hudi 0.5.1**。hudi 组件默认安装在 master 和 router 节点上。

注意：
hudi 组件依赖 hive 和 spark 组件， 如果选择安装 hudi 组件，EMR 将自动安装 hive 和 spark 组件。

使用示例

以下示例适用于 hudi 0.11.0及以后版本，其它版本示例可参考 hudi 官网示例：

1. 登录 master 节点，切换为 hadoop 用户。
2. 加载 spark 配置。

```
cd /usr/local/service/hudi
ln -s /usr/local/service/spark/conf/spark-defaults.conf /usr/local/service/hudi
/demo/config/spark-defaults.conf
```

上传配置到 hdfs：

```
hdfs dfs -mkdir -p /hudi/config
hdfs dfs -copyFromLocal demo/config/* /hudi/config/
```

3. 修改 kafka 数据源。

```
/usr/local/service/hudi/demo/config/kafka-source.properties
bootstrap.servers=kafka_ip:kafka_port
```

上传第一批次数据：

```
cat demo/data/batch_1.json | kafkacat -b [kafka_ip] -t stock_ticks -P
```

4. 摄取第一批数据。

```
spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreame
r --master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type
COPY_ON_WRITE --source-class org.apache.hudi.utilities.sources.JsonKafkaSource
--source-ordering-field ts --target-base-path /usr/hive/warehouse/stock_ticks_c
ow --target-table stock_ticks_cow --props /hudi/config/kafka-source.properties
--schemaprovider-class org.apache.hudi.utilities.schema.FilebasedSchemaProvider
spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreame
r --master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type
```

```
MERGE_ON_READ --source-class org.apache.hudi.utilities.sources.JsonKafkaSource
--source-ordering-field ts --target-base-path /usr/hive/warehouse/stock_ticks_m
or --target-table stock_ticks_mor --props /hudi/config/kafka-source.properties
--schemaprovider-class org.apache.hudi.utilities.schema.FilebasedSchemaProvider
--disable-compaction
```

5. 查看 hdfs 数据。

```
hdfs dfs -ls /usr/hive/warehouse/
```

6. 同步 hive 元数据。

```
bin/run_sync_tool.sh --jdbc-url jdbc:hive2://[hiveserver2_ip:hiveserver2_port]
--user hadoop --pass [password] --partitioned-by dt --base-path /usr/hive/wareh
ouse/stock_ticks_cow --database default --table stock_ticks_cow
bin/run_sync_tool.sh --jdbc-url jdbc:hive2://[hiveserver2_ip:hiveserver2_port]
--user hadoop --pass [password]--partitioned-by dt --base-path /usr/hive/wareho
use/stock_ticks_mor --database default --table stock_ticks_mor --skip-ro-suffix
```

7. 使用计算引擎查询数据。

- hive 引擎

```
beeline -u jdbc:hive2://[hiveserver2_ip:hiveserver2_port] -n hadoop --hiveconf
hive.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat --hiveconf hive.
stats.autogather=false
```

或者 spark 引擎

```
spark-sql --master yarn --conf spark.sql.hive.convertMetastoreParquet=false
```

hive/spark 引擎执行如下 sql 语句：

```
select symbol, max(ts) from stock_ticks_cow group by symbol HAVING symbol = 'GOO
G';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_co
w where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor group by symbol HAVING symbol = 'GOO
G';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_mo
r where symbol = 'GOOG';
```

```
select symbol, max(ts) from stock_ticks_mor_rt group by symbol HAVING symbol = 'G
OOG';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_mo
r_rt where symbol = 'GOOG';
```

- 进入 presto 引擎

```
/usr/local/service/presto-client/presto --server localhost:9000 --catalog hive
--schema default --user Hadoop
```

presto 查询有下划线的字段需要用双引号，例如 `"_hoodie_commit_time"`，执行如下 sql 语句：

```
select symbol, max(ts) from stock_ticks_cow group by symbol HAVING symbol = 'GOO
G';
select "_hoodie_commit_time", symbol, ts, volume, open, close from stock_ticks_co
w where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor group by symbol HAVING symbol = 'GOO
G';
select "_hoodie_commit_time", symbol, ts, volume, open, close from stock_ticks_mo
r where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor_rt group by symbol HAVING symbol = 'G
OOG';
```

8. 上传第二批数据。

```
cat demo/data/batch_2.json | kafkacat -b 10.0.1.70 -t stock_ticks -P
```

9. 摄取第二批增量数据。

```
spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreame
r --master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type
COPY_ON_WRITE --source-class org.apache.hudi.utilities.sources.JsonKafkaSource
--source-ordering-field ts --target-base-path /usr/hive/warehouse/stock_ticks_c
ow --target-table stock_ticks_cow --props /hudi/config/kafka-source.properties
--schemaprovider-class org.apache.hudi.utilities.schema.FilebasedSchemaProvider
spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreame
r --master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type
MERGE_ON_READ --source-class org.apache.hudi.utilities.sources.JsonKafkaSource
--source-ordering-field ts --target-base-path /usr/hive/warehouse/stock_ticks_m
or --target-table stock_ticks_mor --props /hudi/config/kafka-source.properties
--schemaprovider-class org.apache.hudi.utilities.schema.FilebasedSchemaProvider
--disable-compaction
```

0. 查询增量数据，查询方法同步骤7。

1. 使用 hudi-cli 工具。

```
cli/bin/hudi-cli.sh
connect --path /usr/hive/warehouse/stock_ticks_mor
compactions show all
compaction schedule
合并执行计划
compaction run --compactionInstant [requestID] --parallelism 2 --sparkMemory 1G
--schemaFilePath /hudi/config/schema.avsc --retry 1
```

2. 使用 `tez/spark` 引擎查询。

```
beeline -u jdbc:hive2://[hiveserver2_ip:hiveserver2_port] -n hadoop --hiveconf
hive.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat --hiveconf hive.
stats.autogather=false
set hive.execution.engine=tez;
set hive.execution.engine=spark;
```

然后执行 sql 查询，可参考步骤7。

与对象存储结合使用

与 hdfs 类似，需要在存储路径前加上 `cosn://[bucket]` 。参考如下操作：

```
bin/kafka-server-start.sh config/server.properties &
cat demo/data/batch_1.json | kafkacat -b kafkaip -t stock_ticks -P
cat demo/data/batch_2.json | kafkacat -b kafkaip -t stock_ticks -P
kafkacat -b kafkaip -L
hdfs dfs -mkdir -p cosn://[bucket]/hudi/config
hdfs dfs -copyFromLocal demo/config/* cosn://[bucket]/hudi/config/

spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreamer
--master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type COPY
_ON_WRITE --source-class org.apache.hudi.utilities.sources.JsonKafkaSource --sour
ce-ordering-field ts --target-base-path cosn://[bucket]/usr/hive/warehouse/stock_
ticks_cow --target-table stock_ticks_cow --props cosn://[bucket]/hudi/config/kafk
a-source.properties --schemaprovider-class org.apache.hudi.utilities.schema.Fileb
asedSchemaProvider
```

```
spark-submit --class org.apache.hudi.utilities.deltastreamer.HoodieDeltaStreamer
--master yarn ./hudi-utilities-bundle_2.11-0.5.1-incubating.jar --table-type MERG
E_ON_READ --source-class org.apache.hudi.utilities.sources.JsonKafkaSource --sour
ce-ordering-field ts --target-base-path cosn://[bucket]/usr/hive/warehouse/stock_
ticks_mor --target-table stock_ticks_mor --props cosn://[bucket]/hudi/config/kafk
a-source.properties --schemaprovider-class org.apache.hudi.utilities.schema.Fileb
asedSchemaProvider --disable-compaction

bin/run_sync_tool.sh --jdbc-url jdbc:hive2://[hiveserver2_ip:hiveserver2_port] --
user hadoop --pass isd@cloud --partitioned-by dt --base-path cosn://[bucket]/usr/
hive/warehouse/stock_ticks_cow --database default --table stock_ticks_cow

bin/run_sync_tool.sh --jdbc-url jdbc:hive2://[hiveserver2_ip:hiveserver2_port] --
user hadoop --pass hive --partitioned-by dt --base-path cosn://[bucket]/usr/hive/
warehouse/stock_ticks_mor --database default --table stock_ticks_mor --skip-ro-su
ffix


beeline -u jdbc:hive2://[hiveserver2_ip:hiveserver2_port] -n hadoop --hiveconf hi
ve.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat --hiveconf hive.stat
s.autogather=false

spark-sql --master yarn --conf spark.sql.hive.convertMetastoreParquet=false

hivesqls:
select symbol, max(ts) from stock_ticks_cow group by symbol HAVING symbol = 'GOO
G';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_co
w where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor group by symbol HAVING symbol = 'GOO
G';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_mo
r where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor_rt group by symbol HAVING symbol = 'G
OOG';
select `_hoodie_commit_time`, symbol, ts, volume, open, close from stock_ticks_mo
r_rt where symbol = 'GOOG';

prestosqls:
/usr/local/service/presto-client/presto --server localhost:9000 --catalog hive --
schema default --user Hadoop
select symbol, max(ts) from stock_ticks_cow group by symbol HAVING symbol = 'GOO
G';
select "_hoodie_commit_time", symbol, ts, volume, open, close from stock_ticks_co
w where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor group by symbol HAVING symbol = 'GOO
G';
```

```
select "_hoodie_commit_time", symbol, ts, volume, open, close from stock_ticks_mo
r where symbol = 'GOOG';
select symbol, max(ts) from stock_ticks_mor_rt group by symbol HAVING symbol = 'G
OOG';
select "_hoodie_commit_time", symbol, ts, volume, open, close from stock_ticks_mo
r_rt where symbol = 'GOOG';

cli/bin/hudi-cli.sh
connect --path cosn://[bucket]/usr/hive/warehouse/stock_ticks_mor
compactions show all
compaction schedule
compaction run --compactionInstant [requestid] --parallelism 2 --sparkMemory 1G -
-schemaFilePath cosn://[bucket]/hudi/config/schema.avsc --retry 1
```

Superset 开发指南

Superset 简介

最近更新时间：2022-05-16 12:52:26

Apache Superset 是一个数据浏览和可视化 Web 应用程序。EMR 上的 Superset，原装了对 Mysql、Hive、Presto、Impala、Kylin、Druid、Clickhouse 的支持。

Superset 特性

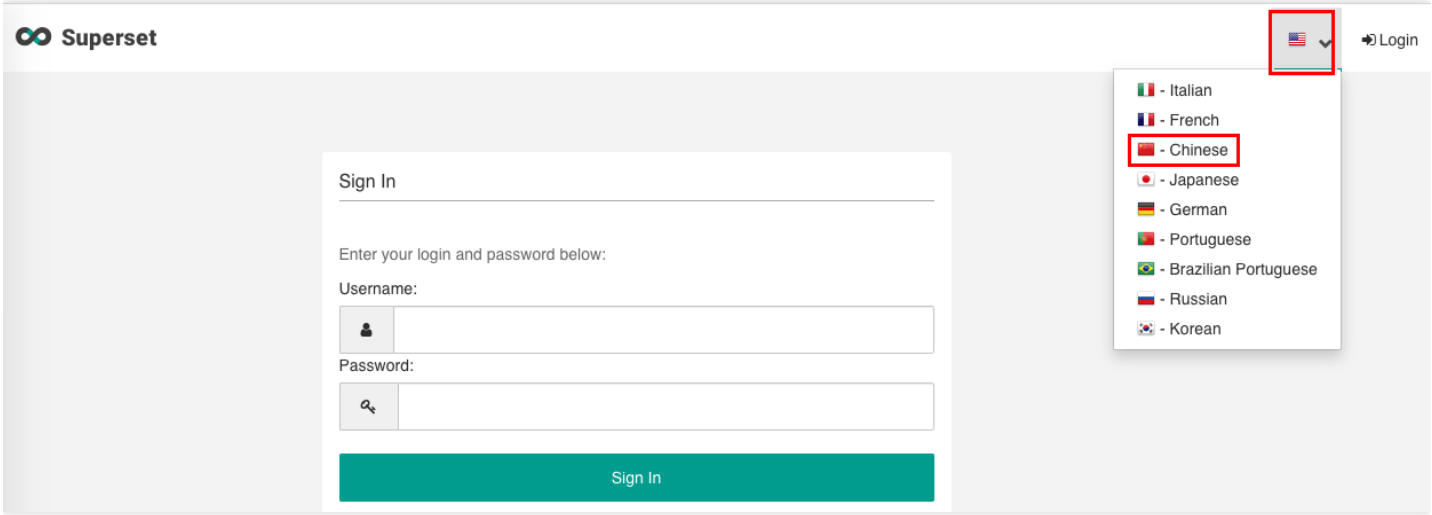
- 支持几乎所有主流的数据库，包括 MySQL、PostgresSQL、Oracle、SQL Server、SQLite、SparkSQL 等，并深度支持 Druid。
- 丰富的可视化展示，支持自定义创建 dashboard。
- 数据的展示完全可控，可自定义展示字段、聚合数据、数据源等。

前提条件

1. 已创建弹性-MapReduce（简称EMR）的 Hadoop 或 Druid 集群，并选择了 Superset 服务，详情请参见 创建 EMR 集群。
2. Superset 默认安装在集群的 master 节点上，打开 master 节点的安全组策略，确保您的网络可以访问 master 节点的18088端口。

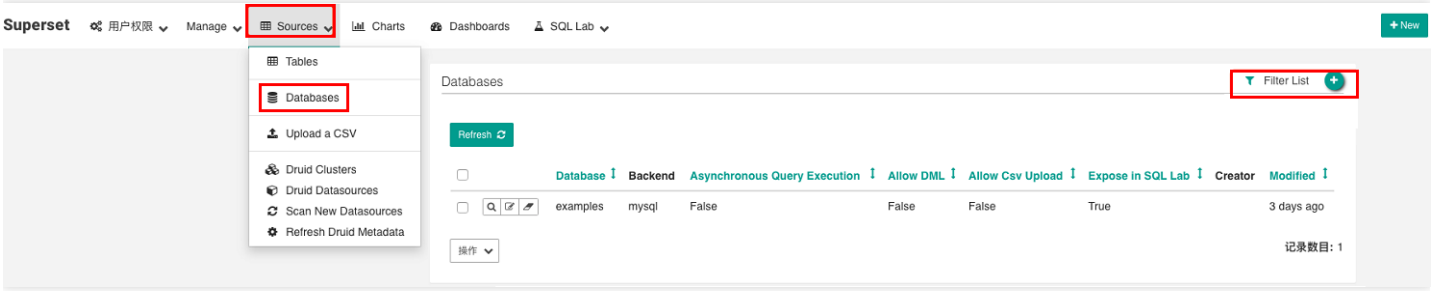
登录

在浏览器地址栏中输入 `http://${master_ip}:18088` （或者通过 **EMR 控制台** > **集群服务**），打开 Supserset 登录界面，默认用户名为 admin，密码为您创建集群时的密码。

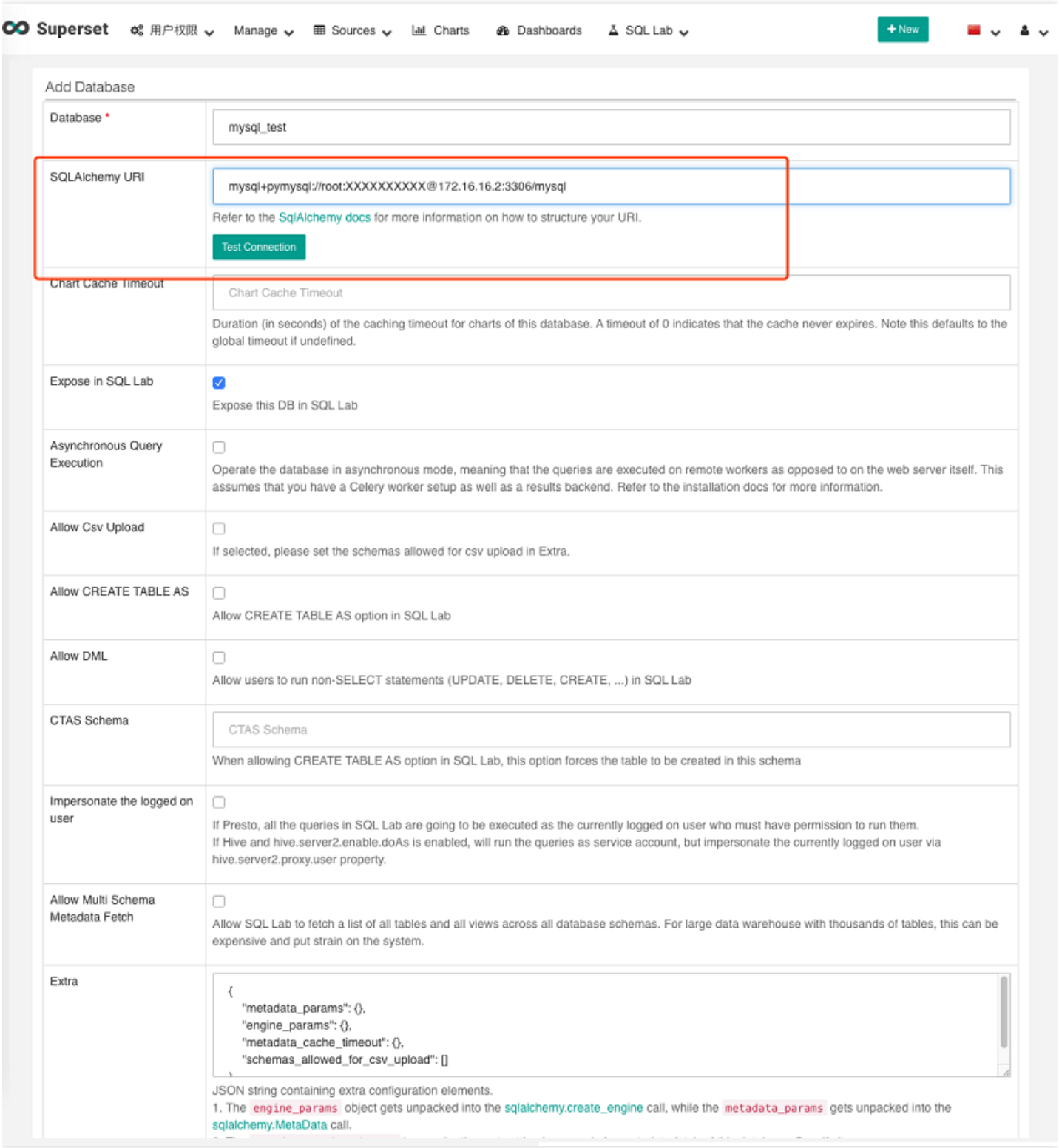


添加 DataBase

进入 **Sources > Databases** 界面，单击 **Filter List**。



进入如下页面，在 SQLAlchemy URI 中加入您需要添加的组件的 URI。



各个数据库的链接 SQLAlchemy URI 如下：

名称	SQLAlchemy URI	备注

名称	SQLAlchemy URI	备注
Mysql	`mysql+pymysql://<mysqlname>:<password>@<mysql_ip>:<mysql_port>/<your_database>`	• mysqlname：连接 mysql 使用的用户名 • password：mysql 密码 • your_database：需要连接的 mysql 数据库

| Hive | `hive://hadoop@<master_ip>:7001/default?auth=NONE` | Master_ip：EMR 集群的 master_ip |

| presto | `presto://hive@<master_ip>:9000/hive/<hive_db_name>` |

- Master_ip：EMR 集群的 master_ip
- hive_db_name：hive 中的数据库名称，不填默认为 default |

| impala | `impala://<core_ip>:27000` | core_ip：EMR 集群中的 core ip |

| kylin | `kylin://<kylin_user>:<password>@<master_ip>:16500/<kylin_project>` |

- kylin_user：kylin 的用户名
- password：kylin 的密码
- master_ip：EMR 集群的 master_ip
- kylin_project：kylin 的项目 |

|Clickhouse | `clickhouse://<user_name>:<password>@<clickhouse-server-endpoint>:8123/<database_name>` |

`clickhouse://default:password@localhost:8123/default`

- user_name：用户名

password：密码

- clickhouse-server-endpoint：ch 服务的服务 endpoint
- database_name：需要访问的 DB 名字 |

自行添加新 Database

Superset 支持 Database。如果您需要安装其他的数据库，可通过如下操作进行：

1. 登录 EMR 集群 master 所在机器。
2. 执行命令 `source /usr/local/service/superset/bin/activate` 。
3. pip3 install 对应的 Python 库。
4. 重启 Superset。

Impala 开发指南
Impala 简介

最近更新时间：2023-02-23 14:17:19

Apache Impala 项目为存储在 Apache Hadoop 文件格式的数据提供高性能、低延迟的 SQL 查询。它对查询进行快速响应，同时支持对分析查询进行交互式的数据探索和查询调整，而不是传统上那种与 SQL-on-Hadoop 技术相关联的长时间批量作业。

Impala 不同于 hive，hive 底层执行使用的是 MapReduce 引擎，仍然是一个批处理过程。而 impala 的中间结果不写入磁盘，即时通过网络以流的形式传递，极大降低了节点的 IO 开销。

Impala 与 Apache Hive 数据库集成，在两个组件之间共享数据库和表。通过与 Hive 的高度集成，以及与 HiveQL 语法的兼容性，您可以使用 Impala 或 Hive 创建表、发起查询、加载数据等。

前提条件

- 确认已开通腾讯云，并且创建了一个 EMR 集群。在创建 EMR 集群时，需要在软件配置界面选择 Impala 组件。
- Impala 安装在路径 EMR 云服务器的 `/data/` 路径下（ `/data/Impala` ）。

准备数据

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例，可选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并进入 Impala 文件夹。

```
[root@10 ~]# su hadoop
[hadoop@10 root]$ cd /data/Impala/
```

新建一个 bash 脚本文件 gen_data.sh，在其中添加以下代码：

```
#!/bin/bash
MAXROW=1000000 #指定生成数据行数
for((i = 0; i < $MAXROW; i++))
do
echo $RANDOM, \"$RANDOM\"
```

```
done
```

然后执行如下命令：

```
[hadoop@10 ~]$ ./gen_data.sh > impala_test.data
```

这个脚本文件会生成1000000个随机数对，并且保存到文件 `impala_test.data` 中。然后把生成的测试数据上传到 HDFS 中，执行如下命令：

```
[hadoop@10 ~]$ hdfspath="/impala_test_dir"
[hadoop@10 ~]$ hdfs dfs -mkdir $hdfspath
[hadoop@10 ~]$ hdfs dfs -put ./impala_test.data $hdfspath
```

其中 $hdfspath 为 HDFS 中您存放文件的路径。最后可用如下命令，验证数据是否正常放到 hdfs 上。

```
[hadoop@10 ~]$ hdfs dfs -ls $hdfspath
```

Impala 基础操作

由于不同 Impala 版本社区组件接口协议及路径默认值变化，不同版本 impala-shell 路径如下表所示。

Imapala 版本	impala-shell 路径	impala-shell 默认链接端口
4.1.0/4.0.0	/data/Impala/shell	27009
3.4.0	/data/Impala/shell	27001
2.10.0	/data/Impala/bin	27001

以下操作以 Impala3.4.0版本为示例：

连接 Impala

登录 EMR 集群的 Master 节点，切换到 Hadoop 用户并且进入 Impala 目录，并连接 Impala：

```
[root@10 Impala]# cd /data/Impala/shell;./impala-shell -i $core_ip:27001
```

其中 core_ip 为 EMR 集群的 core 节点 IP，也可以用 task 节点的 IP，正常登录后显示如下：

```
Connected to $core_ip:27001
Server version: impalad version 3.4.1-RELEASE RELEASE (build Could not obtain git
```

```
hash)
*********************************************************************************
**
Welcome to the Impala shell.
(Impala Shell 3.4.1-RELEASE (ebled66) built on Tue Nov 20 17:28:10 CST 2021)

The SET command shows the current value of all shell and query options.
*********************************************************************************
**
[$core_ip:27001] >
```

也可以登录 core 节点或者 task 节点后，直接连接，执行语句如下：

```
cd /data/Impala/shell;./impala-shell -i localhost:27001
```

创建 Impala 库

在 Impala 下执行以下语句，查看数据库：

```
[10.1.0.215:27001] > show databases;
Query: show databases
+------------------+-----------------------------------------------+
| name | comment |
+------------------+-----------------------------------------------+
| _impala_builtins | System database for Impala builtin functions |
| default | Default Hive database |
+------------------+-----------------------------------------------+
Fetched 2 row(s) in 0.09s
```

使用 `create` 指令创建一个数据库：

```
[localhost:27001] > create database experiments;
Query: create database experiments
Fetched 0 row(s) in 0.41s
```

使用 `use` 指令转到刚创建的 test 数据库下：

```
[localhost:27001] > use experiments;
Query: use experiments
```

查看当前所在库，执行如下语句：

```
select current_database();
```

创建 Impala 表

使用 `create` 指令在 experiments 数据库下创建一个新的名为 impala_test 的内部表：

```
[localhost:27001] > create table t1 (a int, b string) ROW FORMAT DELIMITED FIELDS
TERMINATED BY ',';
Query: create table t1 (a int, b string)
Fetched 0 row(s) in 0.13s
```

查看所有表：

```
[localhost:27001] > show tables;
Query: show tables
+------+
| name |
+------+
| t1 |
+------+
Fetched 1 row(s) in 0.01s
```

查看表结构：

```
[localhost:27001] > desc t1;
Query: describe t1
+------+--------+---------+
| name | type | comment |
+------+--------+---------+
| a | int | |
| b | string | |
+------+--------+---------+
Fetched 2 row(s) in 0.01s
```

将数据导入表中

对于存放在 HDFS 中的数据，使用如下指令来将其导入表中：

```
LOAD DATA INPATH '$hdfspath/impala_test.data' INTO TABLE t1;
```

其中 $hdfspath 为 HDFS 中您存放文件的路径。导入完成后，HDFS 上导入路径上的源数据文件将会被删除。存放到 Impala 内部表的存放路径 `/usr/hive/warehouse/experiments.db/t1` 下。也可以建立外部表，语句如下：

> 注意：

这里只有一条指令，如果不输入分号";"，可以把一条指令放在多行输入。

```
CREATE EXTERNAL TABLE t2
(
a INT,
b string
)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
LOCATION '/impala_test_dir';
```

执行查询

```
[localhost:27001] > select count(*) from experiments.t1;
Query: select count(*) from experiments.t1
Query submitted at: 2019-03-01 11:20:20 (Coordinator: http://10.1.0.215:20004)
Query progress can be monitored at: http://10.1.0.215:20004/query_plan?query_id=f1441478dba3a1c5:fa7a8eef00000000
+----------+
| count(*) |
+----------+
| 1000000 |
+----------+
Fetched 1 row(s) in 0.63s
```

最后输出结果为1000000。

删除表

```
[localhost:27001] > drop table experiments.t1;
Query: drop table experiments.t1
```

更多 Impala 的操作，详见 官方文档。

通过 JDBC 连接 Impala

Impala 也可以通过 Java 代码来连接，步骤类似于 通过 Java 连接 Hive。

唯一区别的是，`$hs2host` 和 `$hsport`，其中 `$hs2host` 是 EMR 集群中任意 core 节点或者 task 节点的 IP。而 hsport 可以在对应节点的 Impala 目录下，配置文件 `conf/impalad.flgs` 中查看。

```
[root@10 ~]# su hadoop
[hadoop@10 root]$ cd /data/Impala/
[hadoop@10 Impala]$ grep hs2_port conf/impalad.flgs
```

如何映射 Hbase 表

Impala 会使用 hive 的元数据信息，所有在 Hive 中的表，都可以在 Impala 中读到。可通过 在 hive 中映射 Hbase 表 达到在 Impala 中映射 Hbase 表。

Impala 运维手册

最近更新时间：2021-06-30 17:42:25

数据变大，Impala 启动失败

背景

当 Impala 中的元数据信息太多（例如几百个库、几万个表），Impala 在启动的时候，需要把这些元数据信息广播到所有节点上，默认这个动作的超时时间是10s。当元数据较大且较易触发时，可通过在 Impala 的启动配置文件 `/data/Impala/conf/impalad.flgs` 中设置 `-statestore_subscriber_timeout_seconds=100` 。

问题排查确定

通常出现这个问题，会在 Impala 的日志 `/data/emr/impala/logs` 中出现如下内容：

```
Connection with state-store lost
Trying to re-register with state-store
```

配置低导致，Impala 查询慢

虽然 Impala 不是内存数据库，但在处理大型表、大型数据时，还是应该为 Impala 分配更多的物理内存，一般建议是使用128G或者更多的内存，并分配80%给到 Impala 进程。

SELECT 语句失败

可能原因如下：

1. 由于某个特定节点的性能，容量或网络问题而导致超时。查看 impala log，确定是什么节点，检查该节点机器网络是否有问题。
2. join 查询使用过多的内存，导致查询自动取消。检查 join 语句是否合理，或者加大机器内存。
3. 受节点上如何生成本机代码以处理查询中特定 WHERE 子句的影响。例如，可以生成特定节点的处理器不支持的机器指令。如果日志中的错误消息表明原因是非法指令，请考虑暂时关闭本机代码生成，然后再次尝试查询。
4. 格式错误的输入数据，例如具有超长行的文本数据文件，或者与 CREATE TABLE 语句的 FIELDS TERMINATED BY 子句中指定的字符不匹配的分隔符。检查是否有超长的数据。并检查建表语句，是否制定了正确的分隔符。

设置查询的使用内存限制

```
[localhost:27001] > set mem_limit=3000000000;
MEM_LIMIT set to 3000000000
[localhost:27001] > select 5;
Query: select 5
+---+ |5 | +---+ |5 | +---+
[localhost:27001] > set mem_limit=3g;
MEM_LIMIT set to 3g
[localhost:27001] > select 5;
Query: select 5
+---+ |5 | +---+ |5 | +---+
[localhost:27001] > set mem_limit=3gb;
MEM_LIMIT set to 3gb
[localhost:27001] > select 5;
+---+
|5 | +---+ |5 | +---+
[localhost:27001] > set mem_limit=3m;
MEM_LIMIT set to 3m
[localhost:27001] > select 5;
+---+
|5 |
+---+
|5 |
+---+
[localhost:27001] > set mem_limit=3mb; MEM_LIMIT set to 3mb [localhost:21000] > s
elect 5;
+---+ |5 | +---+
```

分析 COS/CHDFS 上的数据

最近更新时间：2021-07-08 10:43:45

本节将基于腾讯云对象存储 COS 展示 Impala 更多使用方法，数据来源于直接插入数据、COS 数据。

开发准备

1. 由于任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建一个存储桶（Bucket）。
2. 确认您已开通腾讯云，且已创建一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Impala 组件，并且在基础配置页面开启对象存储的授权。
3. Impala 等相关软件安装在路径 EMR 云服务器的 `/usr/local/service/` 路径下。

操作步骤

登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 登录 Linux 实例，可选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。

在 EMR 命令行先使用以下指令切换到 Hadoop 用户，并连接到 impala：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]$ impala-shel.sh -i $host:27001
```

$host 为您的 impala 数据节点所在的内网 IP。

步骤一：创建表（record）

```
[$host:27001 ] > create table record(id int, name string) row format delimited fi
elds terminated by ',' stored as textfile location 'cosn://$bucketname/';
Query: create table record(id int, name string) row format delimited fields termi
nated by ',' stored as textfile location 'cosn://$bucketname/'
Fetched 0 row(s) in 3.07s
其中 $bucketname 为您的 COS 存储桶名加路径，如果使用 CHDFS 将 location 的值换成 ofs://$m
ountname/，$mountname 为您的 CHDFS 挂在地址加路径
查看表信息，确认 location 是 cos 路径
[$host:27001 ] > show create table record2;
Query: show create table record2
+------------------------------------------------------------------------------+
| result |
+------------------------------------------------------------------------------+
```

```
| CREATE TABLE default.record2 ( |
| id INT, |
| name STRING |
| ) |
| ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' |
| WITH SERDEPROPERTIES ('field.delim'=',', 'serialization.format'=',') |
| STORED AS TEXTFILE |
| LOCATION 'cosn://$bucketname' |
| TBLPROPERTIES ('numFiles'='19', 'totalSize'='1870') |
+----------------------------------------------------------------------------+
Fetched 1 row(s) in 5.90s
```

步骤二：向表中插入数据

```
[$host:27001] > insert into record values(1,"test");
Query: insert into record values(1,"test")
Query submitted at: 2020-08-03 11:29:16 (Coordinator: http://$host:27004)
Query progress can be monitored at: http:/$host:27004/query_plan?query_id=b246d31
94efb7a8f:bc60721600000000
Modified 1 row(s) in 0.64s
```

步骤三：使用 impala 查询表

```
[$host:27001] > select * from record;
Query: select * from record
Query submitted at: 2020-08-03 11:29:31 (Coordinator: http://172.30.1.136:27004)
Query progress can be monitored at: http://$host:27004/query_plan?query_id=8148da
96f8c0d369:4b26432a00000000
+----+---------+
| id | name |
+----+---------+
| 1 | test |
+----+---------+
Fetched 1 row(s) in 0.37s
```

ClickHouse 开发指南

ClickHouse 简介

最近更新时间：2023-07-14 15:18:32

腾讯云弹性 MapReduce（EMR）- ClickHouse 提供开源列式数据库 ClickHouse 的云上托管服务，提供了便捷的 ClickHouse 集群部署、配置修改、监控告警等功能，为企业及用户提供安全稳定的 OLAP 解决方案。ClickHouse 具备极佳的查询性能，非常适合数据在线查询分析场景。

ClickHouse 功能特征

架构

支持单节点、多节点以及多副本架构。根据业务需求，灵活选择。

运维

在控制台提供了开箱即用的监控、日志检索、参数调整等服务。

数据

弹性 MapReduce（EMR）- ClickHouse 提供了完整的数据导入导出支持，可方便地将数据从 COS、HDFS、KAFKA、MySQL 等其他数据源导入或导出到 ClickHouse 集群。

ClickHouse 优势

高性能

充分发挥多核并行优势（SIMD 高效指令集、向量化执行引擎）并借助分布式技术，加速计算提供实时分析能力。开源公开 benchmark 显示比传统方法快100 - 1000倍，提供50MB/s - 200MB/s的高吞吐实时导入能力。

低成本

借助于精心设计的列式存储、高效的数据压缩算法，提供高达10倍的压缩比，大幅提升单机数据存储和计算能力，大幅降低使用成本，是构建海量数据仓库的绝佳方案。

易用

- 提供完善 SQL 支持，上手十分简单。
- 提供 json、map、array 等灵活数据类型适配业务快速变化。
- 支持近似计算、概率数据结构等应对海量数据处理。

EMR-ClickHouse 表引擎介绍

表引擎的作用

表引擎（即表的类型）决定了：

- 数据的存储方式和位置，写到哪里以及从哪里读取数据。
- 支持哪些查询以及如何支持。
- 并发数据访问。
- 索引的使用（如果存在）。
- 是否可以执行多线程请求。
- 数据复制参数。

引擎类型

MergeTree 系列

适用于高负载任务的最通用和功能最强大的表引擎。这些引擎的共同特点是可以快速插入数据并进行后续的后台数据处理。MergeTree 系列引擎支持数据复制（使用 Replicated* 的引擎版本），分区和一些其他引擎不支持的功能。

该类型的引擎：

- ReplacingMergeTree：该引擎和 MergeTree 的不同之处在于它会删除具有相同主键的重复项。
- SummingMergeTree：该引擎继承自 MergeTree。区别在于，当合并 SummingMergeTree 表的数据片段时，ClickHouse 会把所有具有相同主键的行合并为一行，该行包含了被合并的行中具有数值数据类型的列的汇总值。如果主键的组合方式使得单个键值对应于大量的行，则可以显著的减少存储空间并加快数据查询的速度。
- AggregatingMergeTree：该引擎继承自 MergeTree，并改变了数据片段的合并逻辑。ClickHouse 会将相同主键的所有行（在一个数据片段内）替换为单个存储一系列聚合函数状态的行。
- CollapsingMergeTree：该引擎继承于 MergeTree，并在数据块合并算法中添加了折叠行的逻辑。
- VersionedCollapsingMergeTree：VersionedCollapsingMergeTree 是 collapsingmergetree 的升级，使用不同的 collapsing 算法，该算法允许使用多个线程以任何顺序插入数据。
- GraphiteMergeTree：应用于 Graphite data 的数据汇总，该引擎减少了存储容量，提高了 Graphite 查询的效率。

Log 系列

具有最小功能的轻量级引擎。当您需要快速写入许多小表（最多约100万行）并在以后整体读取它们时，该类型的引擎是最有效的。

该类型的引擎：

- TinyLog：最简单的表引擎，用于将数据存储在磁盘上。每列都存储在单独的压缩文件中。写入时，数据将附加到文件末尾。
- StripeLog：该引擎属于日志引擎系列。在需要写入许多小数据量（小于一百万行）的表的场景下使用这个引擎。

- Log：日志与 TinyLog 的不同之处在于，“标记”的小文件与列文件存在一起。这些标记写在每个数据块上，并且包含偏移量，这些偏移量指示从哪里开始读取文件以便跳过指定的行数。这使得可以在多个线程中读取表数据。对于并发数据访问，可以同时执行读取操作，而写入操作则阻塞读取和其它写入。Log 引擎不支持索引。同样，如果写入表失败，则该表将被破坏，并且从该表读取将返回错误。Log 引擎适用于临时数据，write-once 表以及测试或演示目的。

Intergation engines

用于与其他的数据存储与处理系统集成的引擎。

- Kafka：此引擎与 Apache Kafka 结合使用。
- MySQL：MySQL 引擎可以对存储在远程 MySQL 服务器上的数据执行 SELECT 查询。
- ODBC：此引擎允许 ClickHouse 通过 ODBC 接口访问外部数据库。
- JDBC：此引擎允许 ClickHouse 通过 JDBC 接口访问外部数据库。
- HDFS：此引擎允许 ClickHouse 访问 HDFS 上的数据。

用于其他特定功能的引擎

- Distributed：分布式引擎本身不存储数据，但可以在多个服务器上进行分布式查询。读是自动并行的。读取时，远程服务器表的索引（如果有的话）会被使用。
- MaterializedView：物化视图的使用（更多信息请参阅 CREATE TABLE）。它需要使用一个不同的引擎来存储数据，这个引擎要在创建物化视图时指定。当从表中读取时，就会使用该引擎。
- Dictionary：Dictionary 引擎将字典数据展示为一个 ClickHouse 的表。
- Merge：Merge 引擎 (不要和 MergeTree 引擎混淆) 本身不存储数据，但可用于同时从任意多个其他的表中读取数据。
- File：数据源是以 Clickhouse 支持的一种输入格式（TabSeparated，Native 等）存储数据的文件。
- Null：当写入 Null 类型的表时，将忽略数据。从 Null 类型的表中读取时，返回空。但是，可以在 Null 类型的表上创建物化视图。写入表的数据将转发到视图中。
- Set：始终存在于 RAM 中的数据集。
- Join：加载好的 JOIN 表数据会常驻内存中。
- URL：用于管理远程 HTTP/HTTPS 服务器上的数据。该引擎类似 File 引擎。
- View：用于构建视图（有关更多信息，请参阅 CREATE VIEW 查询）。它不存储数据，仅存储指定的 SELECT 查询。 从表中读取时，会运行此查询（并从查询中删除所有不必要的列）。
- Memory：Memory 引擎以未压缩的形式将数据存储在 RAM 中。
- Buffer：缓冲数据写入 RAM 中，周期性地将数据刷新到另一个表。在读取操作时，同时从缓冲区和另一个表读取数据。

虚拟列

虚拟列是表引擎组成的一部分，它在对应的表引擎的源代码中定义。

不能在 CREATE TABLE 中指定虚拟列，并且虚拟列不会包含在 SHOW CREATE TABLE 和 DESCRIBE TABLE 的查询结果中。虚拟列是只读的，所以不能向虚拟列中写入数据。

如果想要查询虚拟列中的数据，您必须在 SELECT 查询中包含虚拟列的名字。 `SELECT *` 不会返回虚拟列的内容。

若您创建的表中有一列与虚拟列的名字相同，那么虚拟列将不能再被访问（不建议这样操作）。为了避免这种列名的冲突，虚拟列的名字一般都以下划线开头。

ClickHouse 使用
ClickHouse SQL 语法

最近更新时间：2021-10-29 09:57:08

数据类型

ClickHouse 支持整数、浮点数、字符型、日期、枚举值、数组等多种数据类型。

类型列表

类别	名称	类型标识	数据范围或描述
整数	单字节整数	Int8	[-128，127]
	双字节整数	Int16	[-32768，32767]
	四字节整数	Int32	[-2147483648，2147483647]
	八字节整数	Int64	[-9223372036854775808，9223372036854775807]
	无符号单字节整数	UInt8	[0，255]
	无符号双字节整数	UInt16	[0，65535]
	无符号四字节整数	UInt32	[0，4294967295]
	无符号八字节整数	UInt64	[0，18446744073709551615]
浮点数	单精度浮点数	Float32	浮点数有效数字6 - 7位
	双精度浮点数	Float64	浮点数有效数字15 - 16位
	自定义浮点	Decimal32(S)	浮点数有效数字 S，S 取值范围[1，9]
		Decimal64(S)	浮点数有效数字 S，S 取值范围[10，18]
		Decimal128(S)	浮点数有效数字 S，S 取值范围[19，38]
字符型	任意长度字符	String	不限定字符串长度

	固定长度字符	FixedString(N)	固定长度的字符串
	唯一标识 UUID 类型	UUID	通过内置函数 generateUUIDv4 生成唯一的标志符
时间类型	日期类型	Date	存储年月日时间，格式 yyyy-MM-dd
	时间戳类型 （秒级）	DateTime(timezone)	Unix 时间戳，精确到秒
	时间戳类型 （自定义）	DateTime(precision, timezone)	可以指定时间精度
枚举类型	单字节枚举	Enum8	提供[-128，127]共256个值
	双字节枚举	Enum16	提供[-32768，32767]共65536个值
数组类型	数组类型	Array(T)	表示由 T 类型组成的数组类型，不推荐使用嵌套数组

- 可以使用 UInt8 来存储布尔类型，将取值限制为0或1。
- 其他数据类型官方文档。

使用举例

枚举类型应用

存储某站点用户的性别信息。

```
CREATE TABLE user (uid Int16, name String, gender Enum('male'=1, 'female'=2)) ENG
INE=Memory;

INSERT INTO user VALUES (1, 'Gary', 'male'), (2,
'Jaco', 'female');

# 查询数据
SELECT * FROM user;

┌─uid─┬─name─┬─gender─┐
│ 1 │ Gary │ male │
│ 2 │ Jaco │ female │
└─────┴──────┴────────┘

# 使用CAST函数查询枚举整数值
SELECT uid, name, CAST(gender, 'Int8') FROM user;

┌─uid─┬─name─┬─CAST(gender, 'Int8')─┐
```

```
| 1 | Gary | 1 |
| 2 | Jaco | 2 |
└──────┴──────┴──────────────────────────┘
```

数组类型应用

某站点记录每天登录用户的 ID，用来分析活跃用户。

```
CREATE TABLE userloginlog (logindate Date, uids Array(String)) ENGINE=TinyLog;

INSERT INTO userloginlog VALUES ('2020-01-02', ['Gary', 'Jaco']), ('2020-02-03',
['Jaco', 'Sammie']);

# 查询结果
SELECT * FROM userloginlog;

┌─logindate─┬─uids─────────────┐
| 2020-01-02 | ['Gary','Jaco'] |
| 2020-02-03 | ['Jaco','Sammie'] |
└──────────┴──────────────────┘
```

创建数据库或表

ClickHouse 使用 CREATE 语句来完成数据库或表的创建。

```
CREATE DATABASE [IF NOT EXISTS] db_name [ON CLUSTER cluster] [ENGINE = engine
(...)]

CREATE TABLE [IF NOT EXISTS] [db.]table_name [ON CLUSTER cluster]
(
name1 [type1] [DEFAULT|MATERIALIZED|ALIAS expr1] [compression_codec] [TTL expr1],
name2 [type2] [DEFAULT|MATERIALIZED|ALIAS expr2] [compression_codec] [TTL expr2],
...
) ENGINE = engine
```

数据库和表都支持本地和分布式两种，分布式方式的创建有以下两种方法：

- 在每台 clickhouse-server 所在机器上都执行创建语句。
- 使用 ON CLUSTER 子句，配合 ZooKeeper 服务完成创建动作。

当使用 clickhouse-client 进行查询时，若在 A 机上查询 B 机的本地表则会报错“Table xxx doesn't exist..”。若希望集群内的所有机器都能查询某张表，推荐使用分布式表。

相关官方文档 CREATE Queries。

查询

ClickHouse 使用 SELECT 语句来完成数据查询。

```
SELECT [DISTINCT] expr_list
[FROM [db.]table | (subquery) | table_function] [FINAL]
[SAMPLE sample_coeff]
[GLOBAL] [ANY|ALL] [INNER|LEFT|RIGHT|FULL|CROSS] [OUTER] JOIN (subquery)|table US
ING columns_list
[PREWHERE expr]
[WHERE expr]
[GROUP BY expr_list] [WITH TOTALS]
[HAVING expr]
[ORDER BY expr_list]
[LIMIT [offset_value, ]n BY columns]
[LIMIT [n, ]m]
[UNION ALL ...]
[INTO OUTFILE filename]
[FORMAT format]
```

相关官方文档 SELECT Queries Syntax。

批量写入

ClickHouse 使用 INSERT INTO 语句来完成数据写入。

```
INSERT INTO [db.]table [(c1, c2, c3)] VALUES (v11, v12, v13), (v21, v22, v23),
...
INSERT INTO [db.]table [(c1, c2, c3)] SELECT ...
```

相关官方文档 INSERT。

删除数据

ClickHouse 使用 DROP 或 TRUNCATE 语句来完成数据删除。

> 说明：

DROP 删除元数据和数据，TRUNCATE 只删除数据。

```
DROP DATABASE [IF EXISTS] db [ON CLUSTER cluster]
DROP [TEMPORARY] TABLE [IF EXISTS] [db.]name [ON CLUSTER cluster]
TRUNCATE TABLE [IF EXISTS] [db.]name [ON CLUSTER cluster]
```

修改表结构

ClickHouse 使用 ALTER 语句来完成表结构修改。

```
# 对表的列操作
ALTER TABLE [db].name [ON CLUSTER cluster] ADD COLUMN [IF NOT EXISTS] name [type]
[default_expr] [codec] [AFTER name_after]
ALTER TABLE [db].name [ON CLUSTER cluster] DROP COLUMN [IF EXISTS] name
ALTER TABLE [db].name [ON CLUSTER cluster] CLEAR COLUMN [IF EXISTS] name IN PARTI
TION partition_name
ALTER TABLE [db].name [ON CLUSTER cluster] COMMENT COLUMN [IF EXISTS] name 'comme
nt'
ALTER TABLE [db].name [ON CLUSTER cluster] MODIFY COLUMN [IF EXISTS] name [type]
[default_expr] [TTL]
# 对表的分区操作
ALTER TABLE table_name DETACH PARTITION partition_expr
ALTER TABLE table_name DROP PARTITION partition_expr
ALTER TABLE table_name CLEAR INDEX index_name IN PARTITION partition_expr
# 对表的属性操作
ALTER TABLE table-name MODIFY TTL ttl-expression
```

相关官方文档 ALTER。

查看信息

- SHOW 语句

  展现数据库、处理列表、表、字典等信息。

  ```
  SHOW DATABASES [INTO OUTFILE filename] [FORMAT format]
  SHOW PROCESSLIST [INTO OUTFILE filename] [FORMAT format]
  SHOW [TEMPORARY] TABLES [{FROM | IN} <db>] [LIKE '<pattern>' | WHERE expr] [LIM
  IT <N>] [INTO OUTFILE <filename>] [FORMAT <format>]
  SHOW DICTIONARIES [FROM <db>] [LIKE '<pattern>'] [LIMIT <N>] [INTO OUTFILE <fil
  ename>] [FORMAT <format>]
  ```

相关官方文档 SHOW Queries。

- DESCRIBE 语句

  查看表的元数据信息。

```
DESC|DESCRIBE TABLE [db.]table [INTO OUTFILE filename] [FORMAT format]
```

函数

ClickHouse 函数有两种类型：常规函数和聚合函数，区别是常规函数可以通过一行数据产生结果，聚合函数则需要一组数据来产生结果。

常规函数

算数函数

数据表中各字段参与数学计算函数。

函数名称	用途	使用场景
plus(a, b), a + b	计算两个字段的和	plus(table.field1, table.field2)
minus(a, b), a - b	计算两个字段的差	minus(table.field1, table.field2)
multiply(a, b), a * b	计算两个字段的积	multiply(table.field1, table.field2)
divide(a, b), a / b	计算两个字段的商	divide(table.field1, table.field2)
modulo(a, b), a % b	计算两个字段的余数	modulo(table.field1, table.field2)
abs(a)	取绝对值	abs(table.field1)
negate(a)	取相反数	negate(table.field1)

比较函数

函数名称	用途	使用场景
=, ==	判断是否相等	table.field1 = value
!=, <>	判断是否不相等	table.field1 != value
>	判断是否大于	table.field1 > value
>=	判断是否大于等于	table.field1 >= value

函数名称	用途	使用场景
<	判断是否小于	table.field1 < value
<=	判断是否小于等于	table.field1 <= value

逻辑运算函数

函数名称	用途	使用场景
AND	两个条件都必须满足	-
OR	两个条件满足其中之一	-
NOT	取条件判断的相反	-

类型转换函数

转换函数可能会溢出，溢出后的数字与C语言中数据类型保持一致。

函数名称	用途	使用场景
toInt(8|16|32|64)	将字符型转化为整数型	toInt8('128') 结果为-127
toUInt(8|16|32|64)	将字符型转化为无符号整数型	toUInt8('128') 结果为128
toInt(8|16|32|64)OrZero	将整数字符型转化为整数型，异常时返回0	toInt8OrZero('a') 结果为0
toUInt(8|16|32|64)OrZero	将整数字符型转化为整数型，异常时返回0	toUInt8OrZero('a') 结果为0
toInt(8|16|32|64)OrNull	将整数字符型转化为整数型，异常时返回NULL	toInt8OrNull('a') 结果为 NULL
toUInt(8|16|32|64)OrNull	将整数字符型转化为整数型，异常时返回NULL	toUInt8OrNull('a') 结果为NULL

浮点数类型或日期类型也有上述类似的函数。

相关官方文档 Type Conversion Functions。

日期函数

相关官方文档 Functions for working with dates and times。

字符串函数

相关官方文档 Functions for working with strings。

UUID

相关官方文档 Functions for working with UUID。

JSON 处理函数

相关官方文档 Functions for working with JSON。

聚合函数

函数名称	用途	使用场景
count	统计行数或者非 NULL 值个数	count(expr)、COUNT(DISTINCT expr)、count()、count(*)
any(x)	返回第一个遇到的值，结果不确定	any(column)
anyHeavy(x)	基于 heavy hitters 算法，返回经常出现的值。通常结果不确定	anyHeavy(column)
anyLast(x)	返回最后一个遇到的值，结果不确定	anyLast(column)
groupBitAnd	按位与	groupBitAnd(expr)
groupBitOr	按位或	groupBitOr(expr)
groupBitXor	按位异或	groupBitXor(expr)
groupBitmap	求基数（cardinality）	groupBitmap(expr)
min(x)	求最小值	min(column)
max(x)	求最大值	max(x)
argMin(arg, val)	返回 val 最小值行的 arg 的值	argMin(c1, c2)
argMax(arg, val)	返回 val 最大值行的 arg 的值	argMax(c1, c2)
sum(x)	求和	sum(x)
sumWithOverflow(x)	求和，结果溢出则返回错误	sumWithOverflow(x)
sumMap(key, value)	用于数组类型，对相同 key 的 value 求和，返回两个数组的 tuple，第一个为排序后的 key，第二个为对应 key 的 value 之和	-
skewPop	求 偏度	skewPop(expr)

函数名称	用途	使用场景
skewSamp	求 样本偏度	skewSamp(expr)
kurtPop	求 峰度	kurtPop(expr)
kurtSamp	求 样本峰度	kurtSamp(expr)
timeSeriesGroupSum(uid, timestamp, value)	对 uid 分组的时间序列对应时间点求和，求和前缺失的时间点线性插值	-
timeSeriesGroupRateSum(uid, ts, val)	对 uid 分组的时间序列对应时间点的变化率求和	-
avg(x)	求平均值	-
uniq	计算不同值的近似个数	uniq(x[, ...])
uniqCombined	计算不同值的近似个数，相比uniq消耗的内存更少，精度更高，但是性能稍差	uniqCombined(HLL_precision)(x[, ...])、uniqCombined(x[, ...])
uniqCombined64	uniqCombined 的 64bit 版本，结果溢出的可能性降低	-
uniqHLL12	计算不同值的近似个数，不建议使用。请用 uniq、uniqCombined	-
uniqExact	计算不同值的精确个数	uniqExact(x[, ...])
groupArray(x), groupArray(max_size)(x)	返回 x 取值的数组，数组大小可由 max_size 指定	-
groupArrayInsertAt(value, position)	在数组的指定位置 position 插入值 value	-
groupArrayMovingSum	-	-
groupArrayMovingAvg	-	-
groupUniqArray(x), groupUniqArray(max_size)(x)	-	-
quantile	-	-
quantileDeterministic	-	-
quantileExact	-	-

函数名称	用途	使用场景
quantileExactWeighted	-	-
quantileTiming	-	-
quantileTimingWeighted	-	-
quantileTDigest	-	-
quantileTDigestWeighted	-	-
median	-	-
quantiles(level1, level2, ...)(x)	-	-
varSamp(x)	-	-
varPop(x)	-	-
stddevSamp(x)	-	-
stddevPop(x)	-	-
topK(N)(x)	-	-
topKWeighted	-	-
covarSamp(x, y)	-	-
covarPop(x, y)	-	-
corr(x, y)	-	-
categoricalInformationValue	-	-
simpleLinearRegression	-	-
stochasticLinearRegression	-	-
stochasticLogisticRegression	-	-
groupBitmapAnd	-	-
groupBitmapOr	-	-
groupBitmapXor	-	-

字典

一个字典是一个映射（key -> attributes），能够作为函数被用于查询，相比引用（reference）表 `JOIN` 的方式更简单和高效。

数据字典有两种，一个是内置字典，另一个是外置字典。

内置字典

ClickHouse 支持一种 内置字典 geobase，支持的函数可参考 Functions for working with Yandex.Metrica dictionaries。

外置字典

ClickHouse 可以从多个数据源添加 外置字典，支持的数据源可参考 Sources Of External Dictionaries。

客户端介绍

最近更新时间：2020-05-07 17:09:57

ClickHouse 本身提供两种客户端接口，分别基于 HTTP 和 TCP 协议。

基于 HTTP 协议

主要用来支持轻量级的简单操作，方便跨平台和编程语言。EMR 集群内的 clickhouse-server 进程会启动8123的 HTTP 服务，可以发送简单的 GET 请求检查服务是否正常。

```
$ curl http://127.0.0.1:8123
Ok.
```

还可以通过 query 参数发送请求，例如查询 testdb 中 account 表的数据。

```
$ wget -q -O- 'http://127.0.0.1:8123/?query=SELECT * from testdb.account'
1 GHua WuHan Hubei 1990
2 SLiu ShenZhen Guangzhou 1991
3 JPong Chengdu Sichuan 1992
```

其他用法可以参照官方文档 HTTP Interface。

基于 TCP 协议

主要在 clickhouse-client 端使用，在 EMR 集群内输入 clickhouse-client 命令，会输出版本信息、连接到的 clickhouse-server 地址、默认使用的数据库等。**可以通过 quit、exit 或 q 等退出使用。**

```
$ clickhouse-client
ClickHouse client version 19.16.12.49.
Connecting to localhost:9000 as user default.
Connected to ClickHouse server version 19.16.12 revision 54427.
```

clickhouse-client 使用的主要参数有以下几个：

- -C --config-file：指定客户端使用的配置文件。
- -h --host：指定 clickhouse-server 的 IP 地址。
- --port：指定 clickhouse-server 的端口地址。
- -u --user 用户名。
- --password 密码。

- -d --database：数据库名称。
- -V --version：查看客户端版本。
- -E --vertical：查询结果按照垂直格式显示。
- -q --query：非交互模式下传入的SQL语句。
- -t --time：非交互模式下显示执行时间。
- --log-level：客户端日志级别。
- --send_logs_level：指定服务端返回日志数据的级别。
- --server_logs_file：指定服务端日志保存路径。

其他参数可以参照官方文档 Command-line Client。

快速开始

最近更新时间：2020-05-07 17:09:57

1. 在弹性 MapReduce（EMR）控制台上新建 ClickHouse 集群，通过外网 IP 地址登录到某一台机器上。
2. 准备 csv 格式的数据，在 `/data` 目录下创建 `account.csv` 文件。

```
AccountId, Name, Address, Year
1, 'GHua', 'WuHan Hubei', 1990
2, 'SLiu', 'ShenZhen Guangzhou', 1991
3, 'JPong', 'Chengdu Sichuan', 1992
```

3. 使用 clickhouse-client 连接服务，创建数据库和表。

```
CREATE DATABASE testdb;
CREATE TABLE testdb.account (accountid UInt16, name String, address String, yea
r UInt64) ENGINE=MergeTree ORDER BY(accountid);
```

4. 导入数据到数据表。

```
cat /data/account.csv | clickhouse-client --database=testdb --query="INSERT INT
O account FORMAT CSVWithNames"
```

5. 查询导入的数据。

```
select * from testdb.account;
```

```
SELECT *
FROM testdb.account

┌─accountid─┬─name──┬─address──────────────┬─year─┐
│         1 │ GHua  │ WuHan Hubei          │ 1990 │
│         2 │ SLiu  │ ShenZhen Guangzhou   │ 1991 │
│         3 │ JPong │ Chengdu Sichuan      │ 1992 │
└───────────┴───────┴──────────────────────┴──────┘

3 rows in set. Elapsed: 0.001 sec.
```

ClickHouse 数据导入
COS 数据导入

最近更新时间：2020-09-11 14:48:03

ClickHouse 提供了两种方法将外部云对象存储的数据导入到表中：

- 表引擎：通过创建 Engine=S3 的外部表将数据导入。
- 表函数：通过使用内置函数将数据导入。

使用上述两种方法，EMR 目前提供的版本支持对象存储的访问属性设置如下表所示：

EMR 版本	ClickHouse 版本	对象存储访问属性	支持的表引擎
1.0.0	19.16.12.49	公共读写	S3
1.1.0	20.3.10.75	公共读写、认证读写（通过 secretId 和 secretKey）	S3
1.2.0+	20.7.2.30	公共读写、认证读写（通过 secretId 和 secretKey）	S3、COSN

表引擎方式

通过创建 Engine 为 S3 的外表和目的数据表，然后使用 INSERT INTO 语句批量插入数据。

```
CREATE TABLE testdb.costb (
column1 UInt32,
column2 String,
column3 String
) ENGINE=S3 ('http://${bucket-name}.cos.${region}.myqcloud.com/data1.csv', 'CSV'
);
或者
CREATE TABLE testdb.costb (
column1 UInt32,
column2 String,
column3 String
) ENGINE=S3('http://${bucket-name}.cos.${region}.myqcloud.com/data1.csv', 'secret
Id', 'secretKey', 'CSV');
或者
CREATE TABLE testdb.costb (
column1 UInt32,
column2 String,
column3 String
```

```
) ENGINE=COSN('http://${bucket-name}.cos.${region}.myqcloud.com/data1.csv', 'secr
etId', 'secretKey', 'CSV');


CREATE TABLE testdb.chtb (
column1 UInt32,
column2 String,
column3 String
) ENGINE=MergeTree() ORDER BY(column1);

INSERT INTO testdb.chtb SELECT * FROM testdb.costb;
```

EMR 1.2.0 以上的版本（ClickHouse 20.7.2.30+）可以将 S3 表引擎修改为 COSN 引擎，用法和效果与 S3 一样。

表函数方式

在创建数据表时使用 s3 内置函数直接将数据导入到表中。

```
CREATE TABLE testdb.chtb
ENGINE=MergeTree()
ORDER BY(column1)
AS SELECT * FROM s3(
'http://${bucket-name}.cos.${region}.myqcloud.com/data1.csv',
'CSV', 'column1 UInt32, column2 String, column3 String');
或者
CREATE TABLE testdb.chtb
ENGINE=MergeTree()
ORDER BY(column1)
AS SELECT * FROM s3(
'http://${bucket-name}.cos.${region}.myqcloud.com/data1.csv', 'secretId', 'secret
Key',
'CSV', 'column1 UInt32, column2 String, column3 String');
```

参考资料

- clickhouse sql
- clickhouse issue

HDFS 数据导入

最近更新时间：2021-07-01 14:29:45

概述

本文介绍了 HDFS 数据导入 ClickHouse 集群的两种方案。本文介绍两种可行的实践方案，一种适合数据量较少的场景。另一种适合大数据场景。**本文实战基于版本19.16.12.49。**

> 说明：
> 想获取更多关于 ClickHouse 技术交流，可 提交工单，我们将您拉入 ClickHouse 技术交流群。

详细步骤

外表导入方案

这种方式适合数据量较少的场景，导入步骤如下：

- 在 ClickHouse 中建立 HDFS Engine 外表，用于读取 HDFS 数据。
- 在 ClickHouse 中创建普通表（通常是 MergeTree 系列）存储 HDFS 中的数据。
- 从外表中 SELECT 数据 INSERT 到普通表，完成数据导入。

步骤1：创建 HDFS Engine 外表

```
CREATE TABLE source
(
`id` UInt32,
`name` String,
`comment` String
)
ENGINE = HDFS('hdfs://172.30.1.146:4007/clickhouse/globs/*.csv', 'CSV')
```

HDFS Engine 使用方法 `ENGINE = HDFS(URI, format)` ，可参考 Table Engine HDFS。

`URI` 为 HDFS 路径，如果包含通配符，则表是只读的。通配符的文件匹配在查询时执行，而不是在创建表时执行。也就是说，如果两次查询之间匹配的文件数目或者内容有变化，两次查询的结果才能够体现这种差异。支持的通配符如下：

- `*` 匹配除路径分隔符 `/` 外的任意数量的字符，包括空字符串。

- `?` 匹配一个字符。
- `{some_string,another_string,yet_another_one}` 匹配 `some_string` 、 `another_string` 或者 `yet_another_one` 。
- `{N..M}` 匹配 N 到 M 的数字，包括 N 和 M，例如 `{1..3}` 匹配1、2、3。

`format` 支持的格式，详见 Formats for Input and Output Data。

步骤2：创建普通表

```
CREATE TABLE dest
(
`id` UInt32,
`name` String,
`comment` String
)
ENGINE = MergeTree()
ORDER BY id
```

步骤3：导入数据

```
INSERT INTO dest SELECT *
FROM source
```

步骤4：查询数据

```
SELECT *
FROM dest
LIMIT 2
```

JDBC Driver 并行导入方案

ClickHouse 提供了 JDBC 的访问方式，并提供了官方的 Driver，此外还有第三方的 Driver，详情可参见 JDBC Driver。

ClickHouse 与 Hadoop/Spark 等大数据生态紧密结合，通过开发 Spark 或者 MapReduce 应用，利用大数据平台的并发处理能力，可以将 HDFS 上的大批量数据的快速导入 ClickHouse。Spark 还支持 Hive 等其他数据源，因此这种方式也可实现 Hive 等其他数据源的导入。

下面举例说明 Spark Python 并发导入数据：

步骤1：创建普通表

```
CREATE TABLE default.hdfs_loader_table
(
`id` UInt32,
`name` String,
`comment` String
)
ENGINE = MergeTree()
PARTITION BY id
ORDER BY id
```

步骤2：开发 Spark Python 应用

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-
from pyspark.sql import SparkSession
import sys
if __name__ == '__main__':
if len(sys.argv) != 2:
print("Usage: clickhouse-spark <path>", file=sys.stderr)
sys.exit(-1)
spark = SparkSession.builder \
.appName("clickhouse-spark") \
.enableHiveSupport() \
.getOrCreate()
url = "jdbc:clickhouse://172.30.1.15:8123/default"
driver = 'ru.yandex.clickhouse.ClickHouseDriver'
properties = {
'driver': driver,
"socket_timeout": "300000",
"rewriteBatchedStatements": "true",
"batchsize": "1000000",
"numPartitions": "4",
'user': 'default',
'password': 'test'
}
spark.read.csv(sys.argv[1], schema="""id INT, name String, comment String""").wri
te.jdbc(
url=url,
table='hdfs_loader_table',
mode='append',
properties=properties,
)
```

url 的格式为 `jdbc:clickhouse://host:port/database`，其中 port 为 clickhosue 的 HTTP 协议端口，默认为8123。

properties 中的部分参数含义如下：

- `socket_timeout` 为超时时间，单位 ms，详情可参考 这里。
- `rewriteBatchedStatements` 打开 JDBC Driver 的批量执行 SQL 功能。
- `batchsize` 批量写入数据条数，可以适当调高，有利于提高写入性能。
- `numPartitions` 数据写入并行度，也决定了 JDBC 并发连接数。`batchsize` 和 `numPartitions` 可参考 JDBC To Other Databases。

步骤3：提交 Spark 任务

```
#!/usr/bin/env bash
spark-submit \
--master yarn \
--jars ./clickhouse-jdbc-0.2.4.jar,./guava-19.0.jar \
clickhouse-spark.py hdfs:///clickhouse/globs
```

Spark Python 需要注意 `clickhouse-jdbc-0.2.4.jar` 依赖的 jar 版本，可以解压该 jar 文件，查看 pom.xml 中的配置，对比 Spark 环境的 jar 包是否版本匹配。版本不匹配时可能会出现错误 Could not initialize class ru.yandex.clickhouse.ClickHouseUtil。这时需要下载正确版本的 jar 包，通过 spark-submit 命令行参数 `--jars` 提交。

步骤4：查询数据

```
SELECT *
FROM hdfs_loader_table
LIMIT 2
```

补充阅读

下面介绍两种直接读写 HDFS 的方式，一般用作从 HDFS 导入数据到 ClickHouse。这两种方式的读写速度比较慢，且不支持如下功能，可参考 Table Engine HDFS：

- `ALTER`、`SELECT...SAMPLE` 操作
- 索引（Indexes）
- 复制（Replication）

Table Engine

1. 创建表

```
CREATE TABLE hdfs_engine_table(id UInt32, name String, comment String) ENGINE=H
DFS('hdfs://172.30.1.146:4007/clickhouse/hdfs_engine_table', 'CSV')
```

2. Insert 测试数据

```
INSERT INTO hdfs_engine_table VALUES(1, 'zhangsan', 'hello zhangsan'),(2, 'lis
i', 'hello lisi')
```

3. 查询

```
SELECT * FROM hdfs_engine_table
┌─id─┬─name─────┬─comment─────────┐
│ 1 │ zhangsan │ hello zhangsan │
│ 2 │ lisi │ hello lisi │
└────┴──────────┴─────────────────┘
```

4. 查看 HDFS 文件

```
hadoop fs -cat /clickhouse/hdfs_engine_table
1,"zhangsan","hello zhangsan"
2,"lisi","hello lisi"
```

Table Function

在使用上与 Table Engine 方式的区别仅是创建表语法稍有差异，示例如下：

```
CREATE TABLE hdfs_function_table AS hdfs('hdfs://172.30.1.146:4007/clickhouse/hdf
s_function_table', 'CSV', 'id UInt32, name String, comment String')
```

参考资料

- ClickHouse Documentation - Table Engine HDFS
- ClickHouse Documentation - Table Function hdfs
- 如何从 HDFS 导入数据到 ClickHouse？
- How to import my data from hdfs？
- ClickHouse Documentation - JDBC Driver
- Spark JDBC 写 clickhouse 操作总结
- 将数据通过 spark 从 hive 导入到 Clickhouse

Kafka 数据导入

最近更新时间：2020-11-23 17:35:54

概述

本文介绍如何将 Kafka 中的数据导入到 ClickHouse 集群的方案。

> **说明：**
>
> 想获取更多关于 ClickHouse 技术交流，可 提交工单，我们将您拉入 ClickHouse 技术交流群。

Kafka 是目前应用非常广泛的开源消息中间件，常用场景就是做数据总线收集各个服务的数据，下游各种数据服务订阅消费数据，生成各种报表或数据应用等。Clickhouse 自带了 Kafka Engine，使得 Clickhouse 和 Kafka 的集成变得非常容易。

将 Kafka 中数据导入到 ClickHouse 的标准流程是：

- 在 ClickHouse 中建立 Kafka Engine 外表，作为 Kafka 数据源的一个接口。
- 在 ClickHouse 中创建普通表（通常是 MergeTree 系列）存储 Kafka 中的数据。
- 在 ClickHouse 中创建 Materialized View，监听 Kafka 中的数据，并将数据写入到 ClickHouse 存储表中。

完成上述三个步骤，就可以将 Kafka 中的数据导入到 ClickHouse 集群中。

Kafka 数据导入到 ClickHouse

ClickHouse 提供了 Kafka Engine 作为访问 Kafka 集群的一个接口（数据流），具体步骤如下：

- **步骤1**：创建 Kafka Engine

```
CREATE TABLE source
(
`ts` DateTime,
`tag` String,
`message` String
)
ENGINE = Kafka()
SETTINGS kafka_broker_list = '172.19.0.47:9092',
kafka_topic_list = 'tag',
kafka_group_name = 'clickhouse',
kafka_format = 'JSONEachRow',
```

```
kafka_skip_broken_messages = 1,
kafka_num_consumers = 2
```

参数	必填	说明
kafka_broker_list	是	填写 Kafka 服务的 broker 列表，用逗号分隔
kafka_topic_list	是	填写 Kafka topic，多个 topic 用逗号分隔
kafka_group_name	是	填写消费者 group 名称
kafka_format	是	Kafka 数据格式，ClickHouse 支持的 Format 详见 Formats for Input and Output Data
kafka_skip_broken_messages	否	填写大于等于0的整数，表示忽略解析异常的 Kafka 数据的条数。如果出现了 N 条异常后，后台线程结束，Materialized View 会被重新安排后台线程去监听数据
kafka_num_consumers	否	单个 Kafka Engine 的消费者数量，通过增加该参数，可以提高消费数据吞吐，但总数不应超过对应 topic 的 partitions 总数
kafka_row_delimiter	否	消息分隔符
kafka_schema	否	对于 kafka_format 需要 schema 定义时，其 schema 由该参数确定
kafka_max_block_size	否	该参数控制 Kafka 数据写入目标表的 Block 大小，超过该数值后，就将数据刷盘

- **步骤2**：创建存储 Kafka 数据的目标表，该表就是最终存储 Kafka 数据

  本文采用 MergeTree 来存储 Kafka 数据：

```
CREATE TABLE target
(
`ts` DateTime,
`tag` String
)
ENGINE = MergeTree()
PARTITION BY toYYYYMM(ts)
ORDER BY tag
```

- **步骤3**：创建 Materialized View 抓取数据

  本文采用如下语句创建 MV：

```
CREATE MATERIALIZED VIEW source_mv TO target AS
SELECT
ts,
```

```
    tag
    FROM source
```

完成上述三个步骤，即可在表 target 中查询到来自 Kafka 的数据。

在上述数据导入流程中，Materialized View 起到了一个中间管道作用，将 Kafka Engine 代表的数据流，写入到目标表中。实际上，一个数据流可以关联多个 Materialized View，将 Kafka 中的数据同时导入到多个不同目的的表中。也可以通过 DETACH/ATTACH 来取消关联，或者重新关联到某个目标表。

MySQL 数据导入

最近更新时间：2021-07-01 10:54:07

概述

本文介绍两种将 MySQL 数据库中的数据导入到 ClickHouse 集群的方案。

- 利用 ClickHouse 支持 MySQL 外表的特性来实现。
- 使用 Altinity 提供的 `clickhouse-mysql-data-reader` 工具来实现数据导入。

本文示例中，将 MySQL 数据表 test.clickhouse_test 中的数据导入到 ClickHouse 集群中，该表的 Schema 如下：

```
MySQL [test]> desc clickhouse_test;
+-------+-------------+------+-----+-------------------+-----------------------------+
| Field | Type        | Null | Key | Default           | Extra                       |
+-------+-------------+------+-----+-------------------+-----------------------------+
| id    | int(20)     | YES  |     | NULL              |                             |
| name  | varchar(64) | YES  |     | NULL              |                             |
| ts    | timestamp   | NO   |     | CURRENT_TIMESTAMP | on update CURRENT_TIMESTAMP |
+-------+-------------+------+-----+-------------------+-----------------------------+
3 rows in set (0.00 sec)
```

基于 MySQL 表引擎来实现数据导入（简易方案）

ClickHouse 的 MySQL 表引擎可以对存储在远程 MySQL 服务器上的数据执行 SELECT 查询。基于这样能力，利用 `CREATE ... SELECT * FROM` 或者 `INSERT INTO ... SELECT * FROM` 语句即可完成数据导入。

具体步骤：

- 步骤1：在 ClickHouse 中创建 MySQL 表引擎。

```
CREATE TABLE clickhouse_test2
(
    `id` Nullable(Int32),
    `name` Nullable(String),
    `ts` DateTime
)
ENGINE = MySQL('172.19.0.31:3306', 'test', 'clickhouse_test', 'root', '[redacted]loud')
```

- 步骤2：建立 ClickHouse 表。

```
CREATE TABLE clickhouse_test3
(
    `id` Nullable(Int32),
    `name` Nullable(String),
    `ts` DateTIme
)
ENGINE = TinyLog()
```

- 步骤3：将步骤1中的外表中数据，导入到 ClickHouse 表中。

```
INSERT INTO clickhouse_test3 (id, name, ts) SELECT *
FROM clickhouse_test2
```

还可以将步骤2/3合并成一个步骤，即采用 `CREATE TABLE AS SELECT * FROM` 方式来达到同样效果。

ClickHouse 支持 MySQL 外表引擎，是否还有必要将数据导入到 ClickHouse 中？
是非常有必要的。MySQL 外表引擎，本身不存储数据，数据存储在 MySQL 中。在复制查询中，特别是有 JOIN 的情况下，访问外表是相当慢的，甚至不可能完成。该方案有明显缺陷，无法增量导入数据。

基于 Altinity 的工具实现数据导入（推荐方案）

Altinity 提供了一个工具 clickhouse-mysql-data-reader 来实现数据导入。该工具可以实现 MySQL 的存量数据导出和增量数据的导出。

按照官网推荐，使用 pypy 工具能够显著提升 clickhouse-mysql-data-reader 导入数据的性能。

工具准备

- 步骤1：下载 pypy3.6-7.2.0，解压到 pypy 目录下。
- 步骤2：安装 clickhouse-mysql。**如果是在腾讯云 ClickHouse 集群，完成下面安装操作后，工具已经集成，开箱即用，无需配置。**
  - 安装 pip：执行 `pypy/bin/pypy3 -m ensurepip` 。
  - 安装 mysql-replication,clickhouse-driver，执行 `pypy/bin/pip3 install mysql-replication` 和 `pypy/bin/pip3 install clickhouse-driver` 。
  - 安装 clickhouse-mysql 并初始化，执行 `pypy/bin/pip3 install clickhouse-mysql` ，执行 `pypy/bin/clickhouse-mysql --install` 。
  - 安装 clickhouse-client，执行 `yum install -y clickhouse-client` 。
  - 安装 mysql-community-devel，执行 `yum install -y mysql-community-devel` 。

- 步骤3：数据库权限准备，所需权限为 SUPER、REPLICATION CLIENT。

```
CREATE USER 'root'@'%' IDENTIFIED BY 'cloud';
GRANT SELECT, REPLICATION CLIENT, REPLICATION SLAVE, SUPER ON *.* TO 'root'@'%'
;
FLUSH PRIVILEGES;
```

数据导入

准备工作完成后，即可使用该工具完成数据从 MySQL 导入到 ClickHouse 集群中。具体步骤如下：

1. 使用 clickhouse-mysql-data-reader 生成建表 SQL。

```
pypy/bin/pypy3 pypy/bin/clickhouse-mysql \
               --src-host=172.19.0.31  \
               --src-user=root \
               --src-password=cloud \
               --create-table-sql-template \
               --with-create-database \
               --src-tables=test.clickhouse_test > create.sql
```

   然后修改 SQL 语句，选择合适的表引擎（在本示例中使用 TinyLog）。执行建表语句 `clickhouse-client -m < create.sql` 。

2. 导入存量数据

```
pypy/bin/pypy3 pypy/bin/clickhouse-mysql \
               --src-host=172.19.0.31 \
               --src-user=root \
               --src-password=cloud \
               --with-create-database \
               --src-tables=test.clickhouse_test \
               --migrate-table  \
               --dst-host=localhost
```

3. 导入增量数据

```
pypy/bin/pypy3 pypy/bin/clickhouse-mysql \
               --src-host=172.19.0.31 \
               --src-user=root \
               --src-password=cloud \
               --src-tables=test.clickhouse_test \
               --dst-host=127.0.0.1 \
               --mempool-max-flush-interval=60 \
               --mempool-max-events-num=10000 \
               --pump-data \
               --src-server-id=1 \
               --src-resume \
               --src-wait \
               --nice-pause=1
```

其中，参数含义如下：

参数	说明
src-host	MySQL 数据库 IP
src-user	MySQL 数据库用户名
src-password	MySQL 数据库密码
create-table-sql-template	生产 ClickHouse 的建表脚本
with-create-database	建表脚本中增加创建数据库语句
src-tables	源表（MySQL 表）
mempool-max-flush-interval	mempool flush 的时间周期
src-server-id	源 MySQL 是否为 master 节点
src-resume	断点续传
src-wait	等待数据
nice-pause	如果没有数据，睡眠的时间间隔

ClickHouse 运维监控

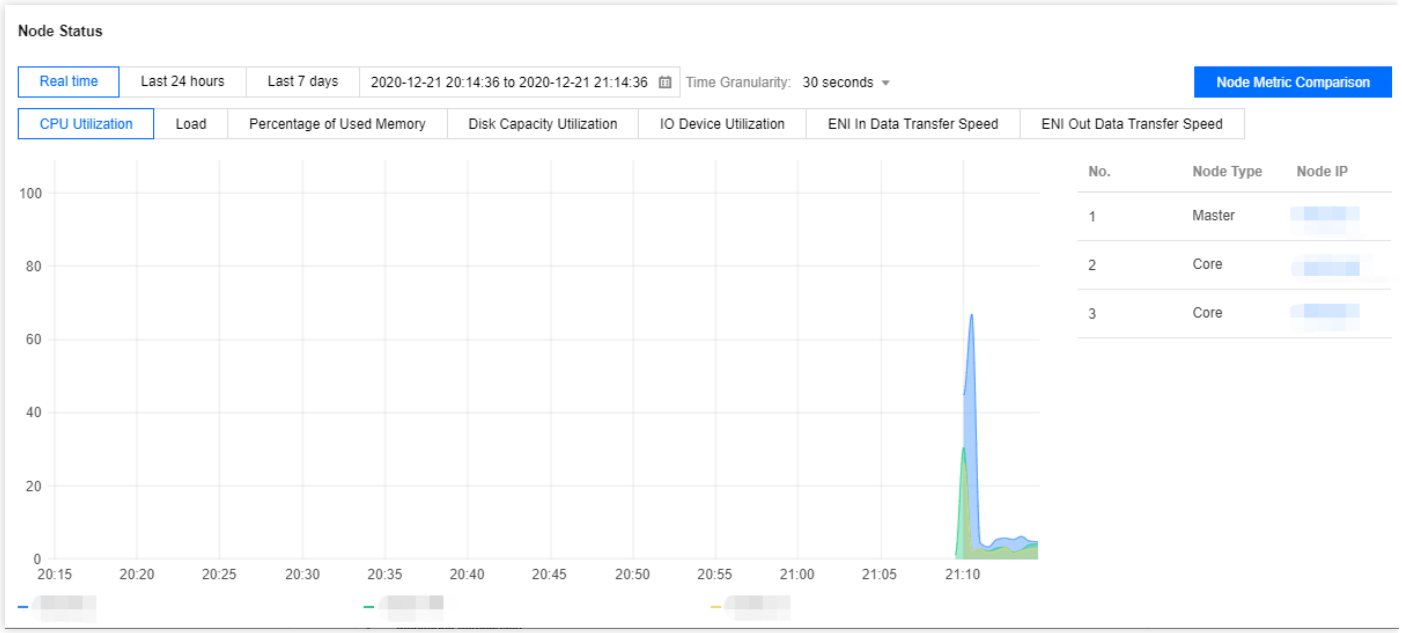
最近更新时间：2021-10-08 15:33:03

腾讯云弹性 MapReduce（EMR）为 ClickHouse 集群提供了完备的监控体系，分为**集群概览**、**服务监控**、**节点监控**三个维度的监控。

集群概览

集群概览页展示的是 ClickHouse 集群的概览信息，如集群的运行状态、节点数量、Zookeeper 状态等。同时也提供了集群聚合维度的服务指标和节点指标的聚合情况，可以直观地看到集群整体的运行情况。

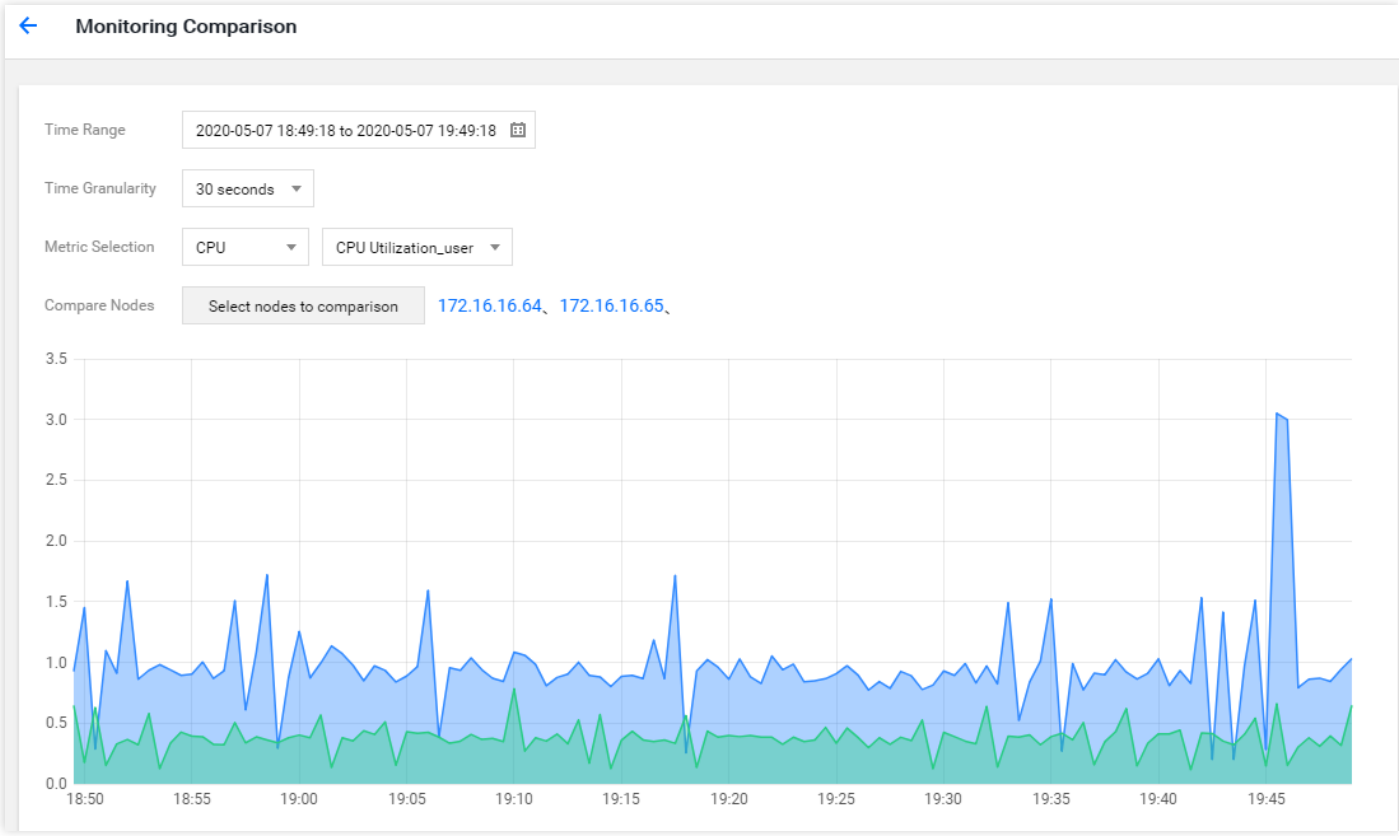
- 服务监控的四个聚合指标为：查询数量、活跃数据块的数量、操作队列大小、网络连接数。集群概览页上的聚合图展示的是 ClickHouse 集群所有节点上的对应指标的总和。
- 节点监控的四个聚合指标为：CPU 使用率、内存使用率、磁盘使用率、网络流量。集群概览页上的聚合图展示的是 ClickHouse 集群整体的节点资源使用情况。



- 部署状态栏，提供了对集群进程状态的实时监控，若进程出现缺失会及时在监控页展示出来。

- 节点状态栏，可以查看最近7天内，机器资源使用量最高的10个机器的使用情况，以便用户能及时定位到集群瓶颈是在哪些机器上。

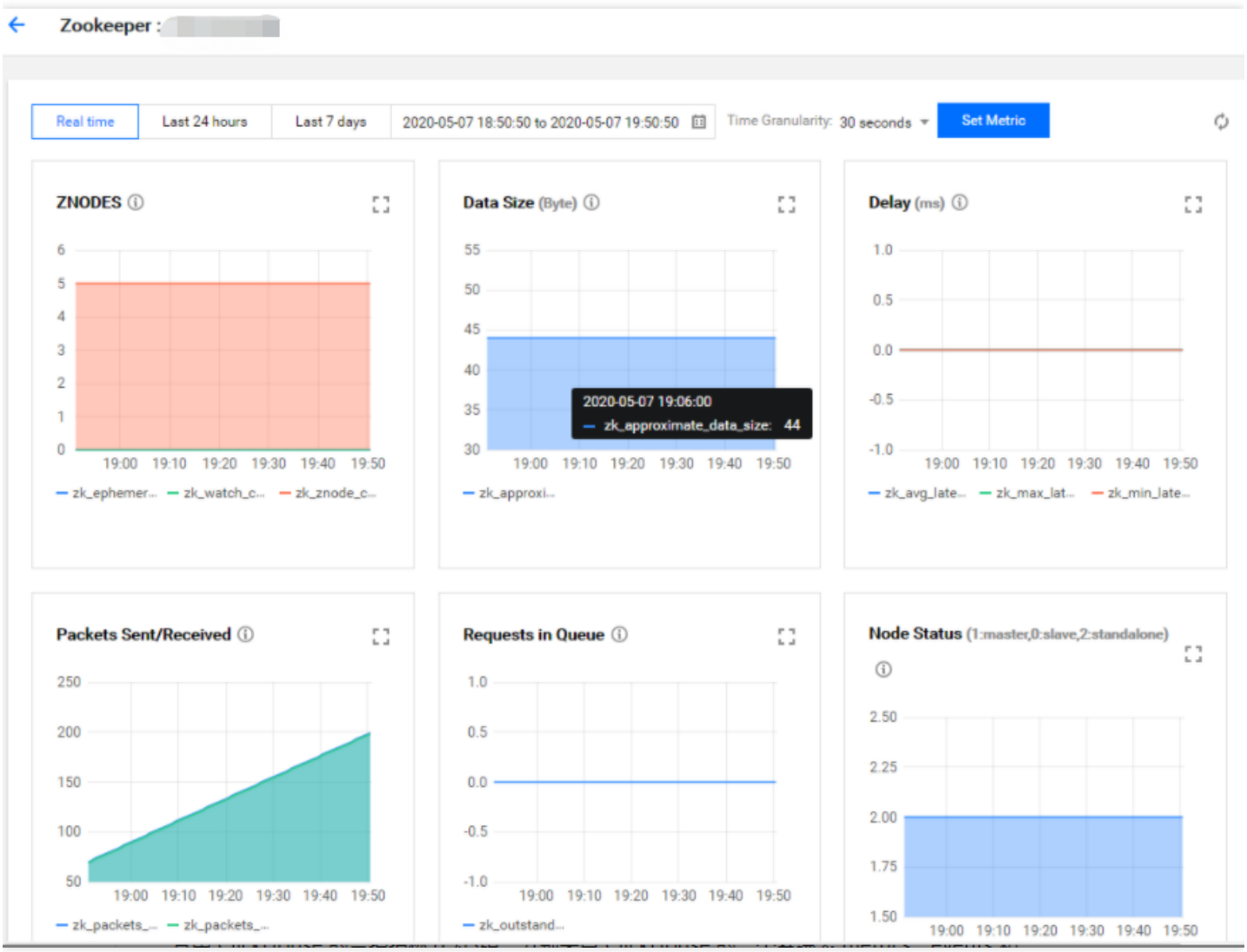
  同时，单击**节点指标对比**，还可以比较若干台节点之间某个时间段对资源的使用情况。



服务监控

ClickHouse 集群服务监控比较简单，服务只有 ClickHouse 和 Zookeeper（如果是 HA 集群），在服务监控页的角色列表中，可以看到角色的简单情况，Zookeeper 的角色只有一个 Zookeeper，ClickHouse 的角色则只有 ClickHouse-Server。从节点 IP 栏可以进到具体的节点监控中。

在具体的角色中，可以查看到具体的服务监控数据，最多支持30天的历史监控数据查看，时间粒度可以根据需要选择。同时，用户可以自定义想要展示的指标，单击**设置指标**可以看到该角色所有的监控指标，每个指标支持预览，如果合适可以勾选上默认展示出来，目前最多支持展示12个指标。



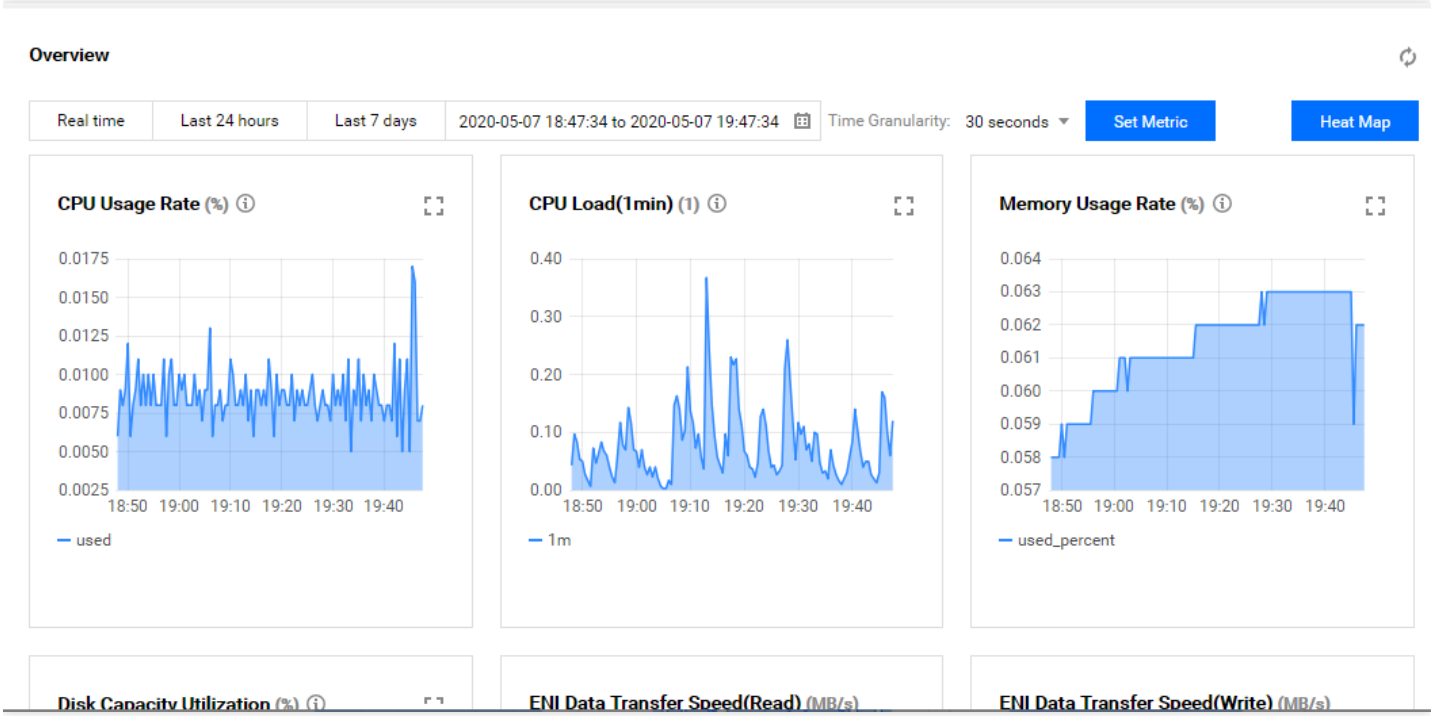
其中 ClickHouse 的监控指标分为3组，分别来自 ClickHouse 的三个系统表 metrics、events 和 asynchronous_metrics。

节点监控

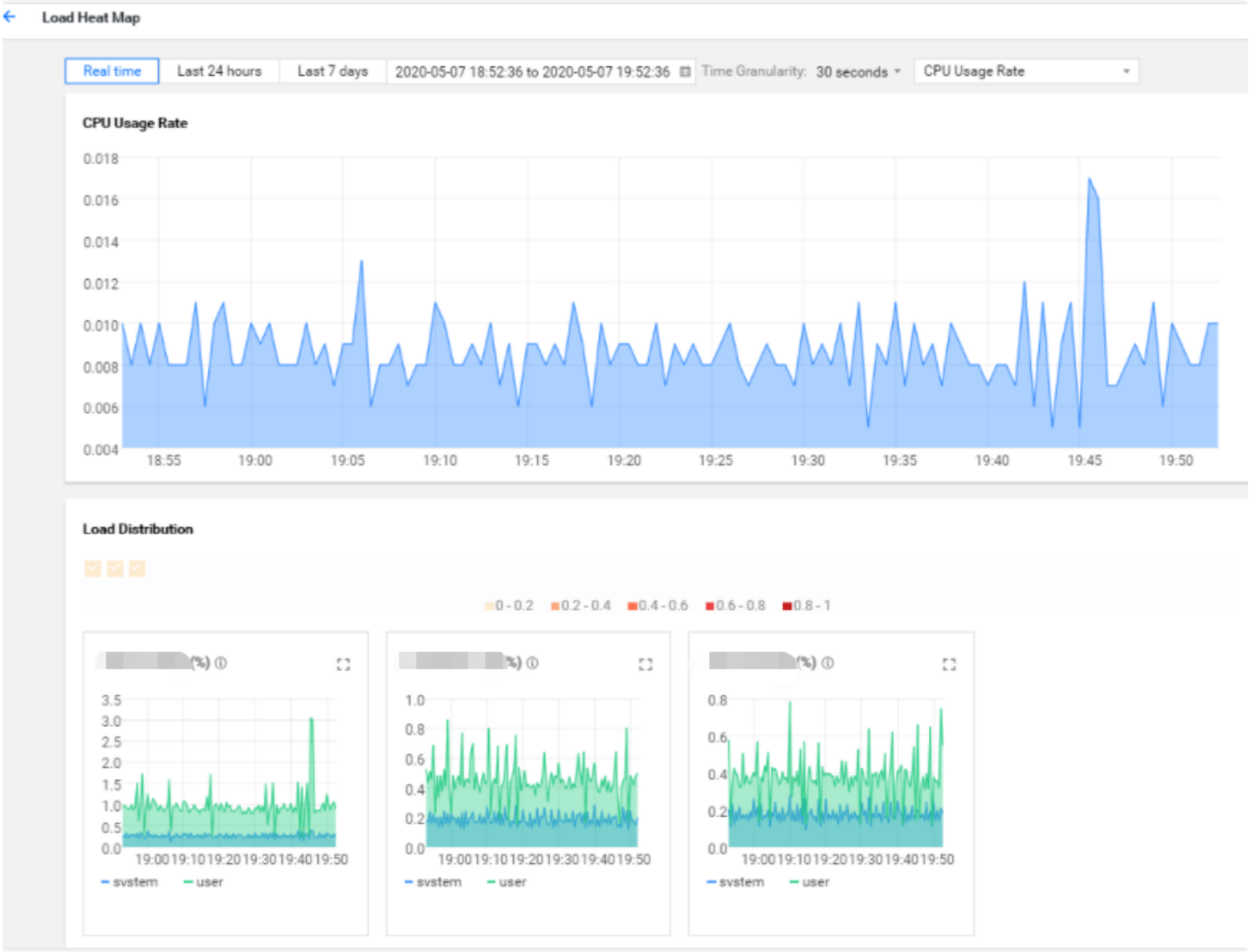
节点监控又分为节点监控概览页和节点监控详情页。

节点监控概览页

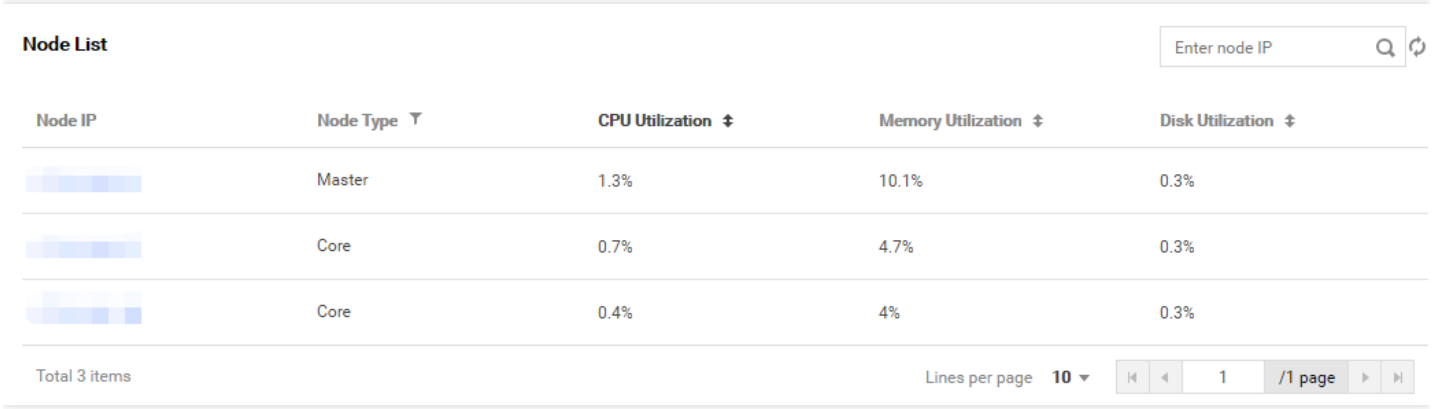
节点监控概览页展示的是集群聚合维度的节点监控指标，目前提供了 CPU、内存、磁盘、网络等四个维度共12个可选的聚合指标，反映的是集群整体的节点资源利用情况。与服务监控类似，用户可以在**设置指标**定义要展示的聚合指标。



同时，节点监控提供了热力图功能，可以查看针对某个节点指标、在某个时间段内每个机器的负载情况。以内存使用率为例，上面的曲线表示集群聚合维度的内存使用率情况。负载分布中，每个小方格表示一个节点，不同颜色表示该节点的内存使用率处于不同的类别，颜色越深表示内存使用率越高。负载分布默认倒序展示，同时默认展示 Top3 的节点内存使用率情况，方便对比机器之间的差异。



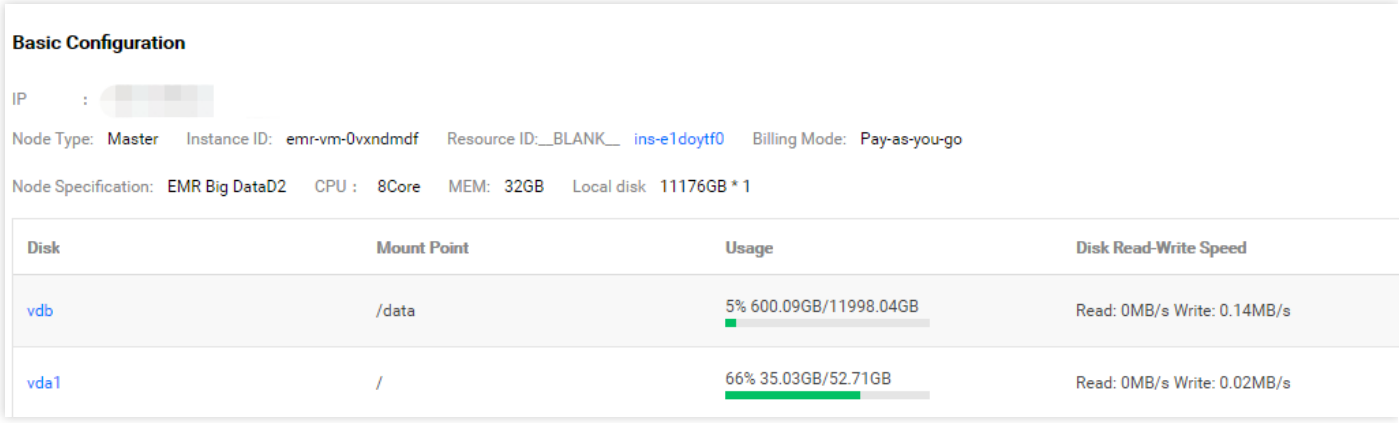
节点监控概览页还有节点列表，是集群所有节点的列表。可以根据节点类型进行筛选，根据 IP 进行搜索，也可以根据 CPU 使用率、内存使用率、磁盘利用率来排序展示。单击具体的节点 IP 可以进入到具体机器的节点监控详情页。



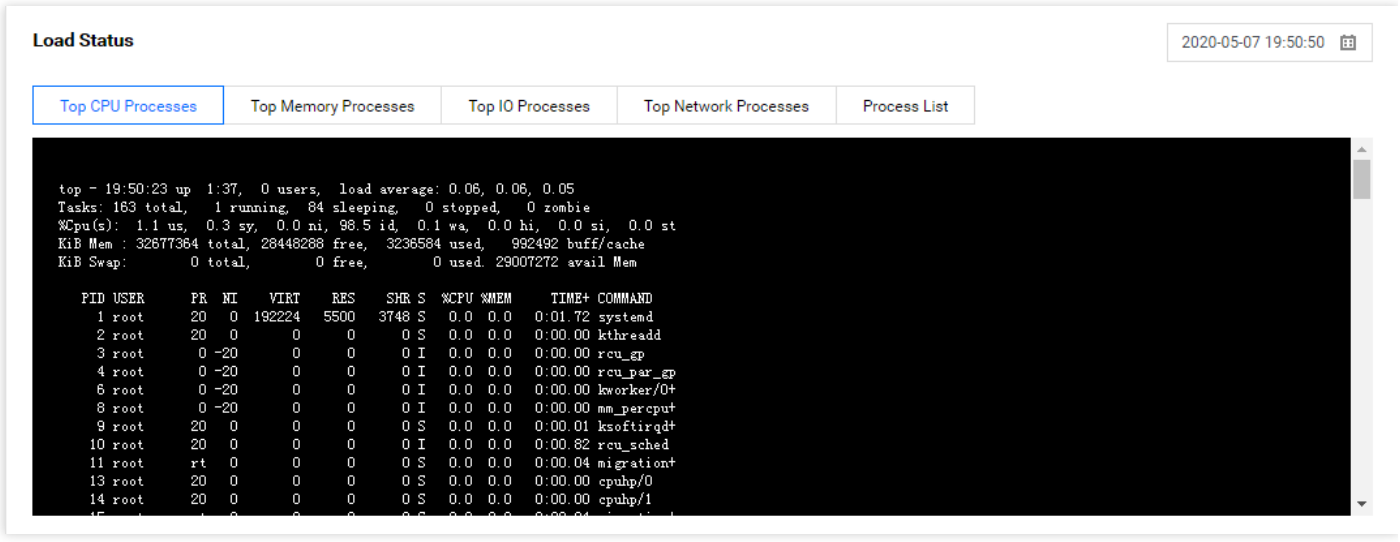
节点监控详情页

节点监控详情页分为基本配置、部署状态、负载状态、节点监控四个部分。

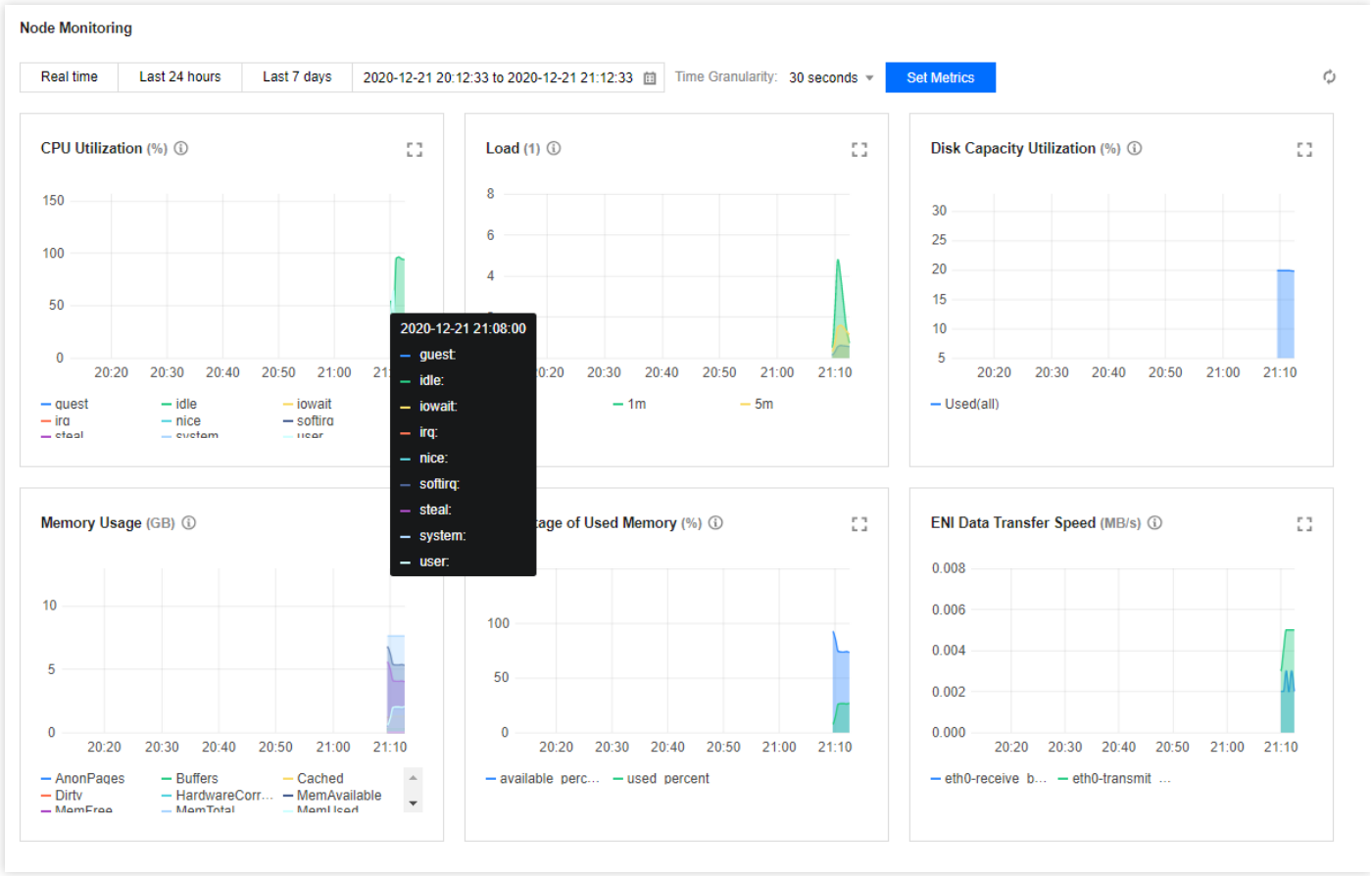
- 基本配置展示的是节点的基本硬件信息，同时还有一个磁盘列表，展示了该节点的磁盘信息。如果磁盘名和磁盘挂载点发生改变，将在30分钟内感知到。单击具体的磁盘名可以查看该磁盘的监控指标。



- 部署状态展示了该节点具体部署服务的进程的实时状态，方便用户监控机器进程的情况。
- 负载状态是某个时刻机器的快照信息，展示了某个时刻机器上进程占用 CPU、内存、IO、网络的情况，也能看到某个时间的节点的进程列表，方便用户查看每个时刻机器的快照信息。



- 节点监控展示了节点具体的监控指标的情况，节点的监控指标包含 CPU、内存、文件句柄、磁盘、网络、进程等多个方面，与服务监控相同，用户可以自定义要展示的指标。



总结

通过集群概览、服务监控、节点监控三个部分，构建了对 ClickHouse 集群完整的监控体系，对 ClickHOuse 集群的运维有很大的帮助。

配置说明

最近更新时间：2021-03-03 19:31:16

ClickHouse-Server 的配置文件在 `/etc/clickhouse-server` 路径下，配置文件主要有三个：config.xml、metrika.xml、users.xml，其中 config.xml 为 ClickHouse-Server 的主配置文件。

config.xml

在 config.xml 配置文件同级的 conf.d 和 config.d 文件夹下，用户可以新建 `*.xml` 文件来覆盖 config.xml 文件中的配置（**这里不建议用户手动更改，建议在控制台提供的配置下发功能中统一下发配置**）。

例如，在 config.xml 文件同级的目录下新建 config.d 目录：

1. 修改 clickhouse-server 监听的 TCP 端口，在 config.xml 文件中默认为9000端口。在 config.d 文件中新建 tcp_port.xml 文件，内容如下：

   ```
   <yandex>
   <tcp_port>9900</tcp_port>
   </yandex>
   ```

   重启 clickhouse-server，可以发现监听的 TCP 端口变成了9000。
2. 新增 metric_log 配置。新建 metric_log.xml 文件，内容如下：

   ```
   <yandex>
   <metric_log>
   <database>system</database>
   <table>metric_log</table>
   <flush_interval_milliseconds>7500</flush_interval_milliseconds>
   <collect_interval_milliseconds>1000</collect_interval_milliseconds>
   </metric_log>
   </yandex>
   ```

   重启 clickhouse-server，clickhouse-client 执行命令 `SELECT * FROM system.metric_log LIMIT 1 FORMAT Vertical;` ，自动创建 system.metric_log 表。

metrika.xml

ClickHouse 配置提供了“替换”的功能，可以使用**incl**属性将指定文件中的配置替换到 config.xml 配置文件中来，这样使得主配置文件不至于过于冗余，方便维护。在默认的 config.xml 配置中，我们可以看到 `<remote_servers>` 、 `<macros>` 、 `<zookeeper>` 这三个标签都是有 **incl** 属性的，可以将 metrika.xml

文件中对应的 `<clickhouse_remote_servers>` 、 `<macros>` 、 `<zookeeper-servers>` 配置加载到 config.xml 文件中。替换文件的路径默认为 `/etc/metrika.xml` ，可通过 `<include_from>` 配置来修改。

> ⚠ **注意：**
>
> 如果 incl 属性中替换的配置是不存在的，会在日志中记录下来。如果您不想让日志记录这些不存在的替换配置，使用 `optional="true"` 属性即可。

users.xml

在 config.xml 配置中，users_config 表示用户相关的配置文件的路径，默认设置为 users.xml。users.xml 中可以配置用户的密码、权限、profile、quota 等信息，针对不同的用户可以有不同的配置。

在 user.xml 统计目录下，可以新建 users.d 目录，在该目录下添加用户相关的配置，可以降低 users.xml 文件的冗余，易于维护。例如： `/etc/clickhouse-server/users.d/testUser.xml`

```
<yandex>
<users>
<testUser>
<profile>default</profile>
<networks>
<ip>::/0</ip>
</networks>
<password>12345</password>
<quota>default</quota>
</testUser>
</users>
</yandex>
```

具体的 Server 参数配置和 Settings 配置可参考官网 Server 参数配置 和 Settings 配置。

日志说明

最近更新时间：2021-07-01 10:47:26

ClickHouse 服务端日志说明

ClickHouse 服务端的日志配置默认在 `/etc/clickhouse-server` 目录的 config.xml 文件中。

```
<logger>
<level>trace</level>
<log>/data/clickhouse/clickhouse-server/logs/clickhouse-server.log</log>
<errorlog>/data/clickhouse/clickhouse-server/logs/clickhouse-server.err.log</erro
rlog>
<size>100M</size>
<count>10</count>
</logger>
```

- level 记录了服务端的日志级别，可以选择的级别有 trace、debug、information、warning、error。
- log 记录了文件路径。
- errlog 记录错误日志的文件路径。
- size 和 count 记录历史日志文件的容量限制和保留的数目。

ClickHouse 客户端日志说明

使用 clickhouse-client 命令行执行 sql 时，可以在交互模式内设置参数 send_logs_level 查看每次执行的日志。

```
172.30.1.15 :) set send_logs_level='trace';
SET send_logs_level = 'trace'
Ok.
0 rows in set. Elapsed: 0.001 sec.
```

在 clickhouse-client 启动时，可以指定 send_logs_level 和 log-level 参数。

```
clickhouse-client --send_logs_level=trace --log-level=trace
```

可以配合 `--server_logs_file` 使用，将日志保存到指定的文件中。

```
clickhouse-client --send_logs_level=trace --log-level=trace --server_logs_file='/
data/query.log'
```

数据备份

最近更新时间：2021-07-01 10:49:02

ClickHouse 数据目录的配置默认在 config.xml 中的 path 字段，默认目录为 `/data/clickhouse/clickhouse-server/data` 。

数据目录格式

某个 clickhouse-server 的数据目录结构如下所示。

```
test
|-- account
| |-- all_1_1_0
| | |-- checksums.txt
| | |-- columns.txt
| | |-- count.txt
| | |-- name.bin
| | |-- name.mrk2
| | |-- id.bin
| | |-- id.mrk2
| | |-- age.bin
| | |-- age.mrk2
| | |-- primary.idx
| |-- detached
|-- |-- format_version.txt
```

- 第一级目录显示数据库名称，例如 test。
- 第二级目录为表名称，上述例子中 test 数据库中存在 account 数据表。
- 第三级目录显示分区目录。
  - checksum.txt 存储数据校验信息。
  - columns.txt 保存数据列的列名和数据类型。
  - count.txt 存储该分区的数据总数。
  - `.bin` 和 `.mrk2` 文件保存数据信息。

数据备份方法

ClickHouse 官方提供了 clickhouse-backup 备份工具，该工具可以将 clickhouse-server 进程的数据目录备份到公有云的对象存储上。

步骤一：准备配置文件 config.yml

在文件中添加公有云对象存储的 url、secret_id、secret_key、clickhouse-server 端的地址和访问信息等。新建 `/etc/clickhouse-backup` 目录，将配置文件移动到该目录下。

```
general:
remote_storage: cos
disable_progress_bar: false
backups_to_keep_local: 0
backups_to_keep_remote: 0
clickhouse:
username: default
password: ""
host: 127.0.0.1
port: 9000
data_path: ""
skip_tables:
- system.*
timeout: 5m
cos:
url: "https://${bucket-name}.cos.${region}.myqcloud.com"
timeout: 100
secret_id: "xxxxxx"
secret_key: "yyyyy"
path: "/"
compression_format: gzip
compression_level: 1
debug: false
```

步骤二：使用 clickhouse-backup 工具创建备份目录

```
$ ./clickhouse-backup create testbak1
# 使用list查看是否创建成功
$ ./clickhouse-backup list local
- 'testbak1' (created at 26-03-2020 02:55:23)
```

步骤三：使用 clickhouse-backup 工具打包数据到对象存储上

执行命令后，查看 COS 桶相应的目录下是否存在压缩文件 `testbak1.tar.gz` 。

```
$ ./clickhouse-backup upload testbak1 2020/03/25 12:58:25 Upload backup 'testbak
1' 1.27 MiB / 1.27 MiB [============================] 100.00% 0s 2020/03/25 12:58:
26 Done.
```

参考资料

clickhouse-backup tool

系统表说明

最近更新时间：2022-04-25 12:30:06

本文来自 ClickHouse 官方文档。

- 系统表用于实现部分系统功能，并提供途径来获取与系统运行状态相关的信息。
- 系统表无法删除（但可以执行 DETACH）。
- 系统表中的数据或者元数据没有以文件的方式存储在磁盘上。server 启动时将创建所有系统表。
- 系统表是只读的。
- 系统表位于“system”数据库中。

system.asynchronous_metrics

system.asynchronous_metrics 表包含在后台定期计算的指标。例如，正在使用的 RAM 数量。

列信息：

列名	类型	描述
metric	String	指标名
value	Float64	指标值

示例：

```
SELECT * FROM system.asynchronous_metrics LIMIT 10
```

```
┌─metric──────────────────────────────────┬──────value─┐
│ jemalloc.background_thread.run_interval │          0 │
│ jemalloc.background_thread.num_runs     │          0 │
│ jemalloc.background_thread.num_threads  │          0 │
│ jemalloc.retained                       │  299642880 │
│ jemalloc.mapped                         │  964415488 │
│ jemalloc.resident                       │  942026752 │
│ jemalloc.metadata_thp                   │          0 │
│ jemalloc.metadata                       │    5601816 │
│ jemalloc.allocated                      │  936427032 │
│ ReplicasMaxQueueSize                    │          0 │
└─────────────────────────────────────────┴────────────┘
```

列名	类型	描述
cluster	String	集群名称
shard_num	UInt32	集群中分片号
shard_weight	UInt32	写数据时分片的权重
replica_num	UInt32	分片中的副本号
host_name	String	节点名
host_address	String	节点 IP 地址
port	UInt16	用于连接 server 的端口号
user	String	用于连接 server 的用户名
errors_count	UInt32	该节点无法到达副本的次数
estimated_recovery_time	UInt32	从当前到副本错误数为零且被认为恢复正常的时间，单位为秒

system.clusters

system.clusters 表包含配置文件中可用集群以及配置文件中 server 的信息。

列信息：

对于集群的每一个查询，errors_count 都会更新一次，但 estimated_recovery_time 会根据需求重新计算，所以可能会出现这样的情况，当 errors_count 非零、estimated_recovery_time 为零时，下次查询会把 errors_count 置为零，并且尝试使用副本，就好像没有错误一样。

system.columns

system.columns 表包含了所有表的列信息。您可以使用此表获取与 DESCRIBE TABLE 查询类似的信息，并且一次获取多个表的信息。

列信息：

列名	类型	描述
database	String	数据库名
table	String	表名
name	String	列名
type	String	列类型

列名	类型	描述
default_kind	String	默认值的表达式类型（DEFAULT、MATERIALIZED、ALIAS），或者为空字符串（如果未定义）
default_expression	String	默认值的表达式，或者为空字符串（如果未定义）
data_compressed_bytes	UInt64	压缩数据的大小，以字节为单位
data_uncompressed_bytes	UInt64	解压缩数据的大小，以字节为单位
marks_bytes	UInt64	marks 的大小，以字节为单位
comment	String	列的注释，或者为空字符串（如果未定义）
is_in_partition_key	UInt8	表示列是否在分区表达式中的标志
is_in_sorting_key	UInt8	表示列是否在排序键表达式中的标志
is_in_primary_key	UInt8	表示列是否在主键表达式中的标志
is_in_sampling_key	UInt8	表示列是否在采样键表达式中的标志

system.contributors

system.contributors 包含有关贡献者的信息，所有贡献者以随机顺序排列，该顺序在查询执行时是随机的。

system.databases

system.databases 该表仅包含一个名为“name”的 String 列-数据库名称。server 知道每个数据库在表中都有一个对应的条目。该系统表用于实现 SHOW DATABASES 查询。

system.detached_parts

system.detached_parts 包含有关 MergeTree 表的分离 part 的信息。reason 列表示为什么该 part 要分离。对于 user-detache 的部分，reason 列为空。有关其他列的描述，请参见 system.parts。如果 part 的名字非法，则某些列的值可能为 NULL，可以使用 `ALTER TABLE DROP DETACHED PART` 删除这些 parts。

system.dictionaries

system.dictionaries 包含有关外部词典的信息。

列信息：

列名	类型	描述
name	String	字典名

列名	类型	描述
type	String	字典类型，Flat、Hashed、Cache
origin	String	描述字典的配置文件的路径
attribute.names	Array(String)	字典提供的属性名称数组
attribute.types	Array(String)	字典提供的属性类型的对应数组
has_hierarchy	UInt8	字典是否分层
bytes_allocated	UInt64	字典使用的 RAM 数量
hit_rate	Float64	对于缓存字典，该值在缓存中的使用百分比
element_count	UInt64	词典中存储的项目数
load_factor	Float64	字典中填充的百分比
creation_time	DateTime	创建字典或上次成功重新加载字典的时间
last_exception	String	错误文本，记录的是因为无法创建字典导致的，创建或重新加载字典时发生的错误
source	String	描述字典数据源的文本

system.events

system.events 包含有关系统中发生的事件数的信息。例如，您可以在 system.events 表中找到自 ClickHouse server 启动以来已处理的 SELECT 查询数量。

列信息：

列名	类型	描述
event	String	事件名
value	UInt64	事件发生的次数
description	String	事件描述

示例：

```
SELECT * FROM system.events LIMIT 5
```

```
event                                 value   description
SelectQuery                           31663   Same as Query, but only for SELECT queries.
ReadBufferFromFileDescriptorRead          2   Number of reads (read/pread) from a file descriptor. Does not include sockets.
WriteBufferFromFileDescriptorWrite        1   Number of writes (write/pwrite) to a file descriptor. Does not include sockets.
ReadCompressedBytes                     360
CompressedReadBufferBlocks               10
```

system.functions

system.functions 包含有关普通函数和聚合函数的信息。

列信息：

列名	类型	描述
name	String	函数名
is_aggregate	UInt8	是否为聚合函数

system.graphite_retentions

system.graphite_retentions 包含与 `* GraphiteMergeTree` 引擎一起在表中使用的关于参数 graphip_rollup 的信息。

列信息：

列名	类型	描述
config_name	String	graphite_rollup 变量名
regexp	String	指标名称的模式
function	String	汇总函数的名称
age	UInt64	数据的最小使用期限（以秒为单位）
precision	UInt64	数据使用期限的精度（以秒为单位）
priority	UInt16	模式优先级
is_default	UInt8	模式是否为默认
Tables.database	Array(String)	使用 config_name 参数的数据库表的名称数组
Tables.table	Array(String)	使用 config_name 参数的表名数组

system.merges

system.merges 表包含有关 MergeTree 系列表中当前正在处理的 merge 和 part 突变的信息。

列信息：

列名	类型	描述
database	String	表所在的数据库的名称
table	String	表名
elapsed	Float64	自合并开始起经过的时间（以秒为单位）
progress	Float64	已完成工作（从0到1）的百分比
num_parts	UInt64	合并件数
result_part_name	String	合并后将形成的 part 的名称
is_mutation	UInt8	如果此过程是 part 突变，则为1
total_size_bytes_compressed	UInt64	合并块中压缩数据的总大小
total_size_marks	UInt64	合并 parts 中的 marks 总数
bytes_read_uncompressed	UInt64	读取的字节数，未压缩
rows_read	UInt64	读取的行数
bytes_written_uncompressed	UInt64	写入的字节数，未压缩
rows_written	UInt64	写入的行数

system.metrics

system.metrics 表包含可以立即计算或具有当前值的指标。例如，同时处理的查询数或当前副本延迟。该表始终是最新的。

列信息：

列名	类型	描述
metric	String	指标名
value	Int64	指标值
description	String	指标描述

所支持的指标可以在 ClickHouse 的源码文件中找到 dbms/src/Common/CurrentMetrics.cpp。

示例：

```
SELECT * FROM system.metrics LIMIT 10
```

```
┌─metric──────────────┬─value─┬─description──────────────────────────────────────┐
│ Query               │     1 │ Number of executing queries                      │
│ Merge               │     0 │ Number of executing background merges            │
│ PartMutation        │     0 │ Number of mutations (ALTER DELETE/UPDATE)        │
│ ReplicatedFetch     │     0 │ Number of data parts being fetched from replica  │
│ ReplicatedSend      │     0 │ Number of data parts being sent to replicas      │
└─────────────────────┴───────┴──────────────────────────────────────────────────┘
```

```
Row 1:
──────
event_date:                                 2020-03-24
event_time:                                 2020-03-24 11:06:10
milliseconds:                               437
ProfileEvent_Query:                         0
ProfileEvent_SelectQuery:                   0
ProfileEvent_InsertQuery:                   0
ProfileEvent_FileOpen:                      0
ProfileEvent_Seek:                          0
ProfileEvent_ReadBufferFromFileDescriptorRead:          1
ProfileEvent_ReadBufferFromFileDescriptorReadFailed:    0
ProfileEvent_ReadBufferFromFileDescriptorReadBytes:     0
ProfileEvent_WriteBufferFromFileDescriptorWrite:        1
ProfileEvent_WriteBufferFromFileDescriptorWriteFailed:  0
ProfileEvent_WriteBufferFromFileDescriptorWriteBytes:   60
ProfileEvent_ReadBufferAIORead:             0
ProfileEvent_ReadBufferAIOReadBytes:        0
ProfileEvent_WriteBufferAIOWrite:           0
ProfileEvent_WriteBufferAIOWriteBytes:      0
```

system.metric_log

system.metric_log 表包含来自表 system.metrics 和 system.events 的指标值的历史记录，定期刷新到磁盘。

要在 system.metric_log 上打开指标历史采集，请按以下内容创建 `/etc/clickhouse-server/config.d/metric_log.xml` ，并重启 clickhouse-server：

```
<yandex>
<metric_log>
<database>system</database>
<table>metric_log</table>
<flush_interval_milliseconds>7500</flush_interval_milliseconds>
<collect_interval_milliseconds>1000</collect_interval_milliseconds>
</metric_log>
</yandex>
```

示例：

```
SELECT * FROM system.metric_log LIMIT 1 FORMAT Vertical;
```

```
Row 1:
──────
event_date:                                  2020-03-24
event_time:                                  2020-03-24 11:06:10
milliseconds:                                437
ProfileEvent_Query:                          0
ProfileEvent_SelectQuery:                    0
ProfileEvent_InsertQuery:                    0
ProfileEvent_FileOpen:                       0
ProfileEvent_Seek:                           0
ProfileEvent_ReadBufferFromFileDescriptorRead:          1
ProfileEvent_ReadBufferFromFileDescriptorReadFailed:    0
ProfileEvent_ReadBufferFromFileDescriptorReadBytes:     0
ProfileEvent_WriteBufferFromFileDescriptorWrite:        1
ProfileEvent_WriteBufferFromFileDescriptorWriteFailed:  0
ProfileEvent_WriteBufferFromFileDescriptorWriteBytes:   60
ProfileEvent_ReadBufferAIORead:              0
ProfileEvent_ReadBufferAIOReadBytes:         0
ProfileEvent_WriteBufferAIOWrite:            0
ProfileEvent_WriteBufferAIOWriteBytes:       0
```

system.numbers

该表仅包含一个名为“number”的 UInt64 列，其中几乎包含所有从零开始的自然数。您可以使用该表进行测试，或者根据需要进行暴力搜索。**从该表读取的数据不会并行化。**

system.numbers_mt

与 system.numbers 表相同，但读取是并行的。可以按任何顺序返回数字（用于测试）。

system.one

该表只有一行一列，其中“dummy”列为 UInt8 类型，值为0。如果 SELECT 查询未指定 FROM 子句，则使用此表。这类似于在其他 DBMS 中找到的 DUAL 表。

system.parts

system.parts 包含 MergeTree 表的 parts 信息，每一行描述了一个数据分块。

列信息：

列名	类型	描述
partition	String	分区名称
name	String	数据分块名

列名	类型	描述
active	UInt8	表示数据部分是否处于活跃状态的标志。如果数据分块处于活跃状态，则会在表中使用它。否则，它将被删除。合并后，不活跃的数据分块仍然保留
marks	UInt64	marks 数量
rows	UInt64	行数
bytes_on_disk	UInt64	所有数据分块文件的总大小（以字节为单位）
data_compressed_bytes	UInt64	数据分块中压缩数据的总大小。不包括所有辅助文件（例如，带标记的文件）
data_uncompressed_bytes	UInt64	数据分块中未压缩数据的总大小。不包括所有辅助文件（例如，带标记的文件）
marks_bytes	UInt64	带标记的文件大小
modification_time	DateTime	包含数据分块的目录被修改的时间
remove_time	DateTime	数据分块变为非活跃状态的时间
refcount	UInt32	使用数据分块的位置数。大于2的值表示该数据分块用于查询或合并
min_date	Date	数据分块中日期键的最小值
max_date	Date	数据分块中日期键的最大值
min_time	DateTime	数据分块中日期和时间键的最小值
max_time	DateTime	数据分块中日期和时间键的最大值
partition_id	String	分区 ID
min_block_number	UInt64	合并后组成当前分块的最小数据分块数
max_block_number	UInt64	合并后组成当前分块的最大数据分块数
level	UInt32	merge tree 的深度。0表示当前分块是通过插入而不是通过合并其他分块来创建的
data_version	UInt64	用于确定应将哪些突变应用于数据分块（版本高于 data_version 的突变）
primary_key_bytes_in_memory	UInt64	主键值使用的内存量（以字节为单位）

列名	类型	描述
primary_key_bytes_in_memory_allocated	UInt64	主键值保留的内存量（以字节为单位）
is_frozen	UInt8	表示存在分区数据备份的标志。1，备份存在。0，表示备份不存在
database	String	数据库名
table	String	表名
engine	String	不带参数的表引擎名称
path	String	包含数据分块文件的文件夹的绝对路径
disk	String	存储数据分块的磁盘名称
hash_of_all_files	String	sipHash128 的压缩文件
hash_of_uncompressed_files	String	sipHash128 的未压缩文件（带有标记的文件，索引文件等）
uncompressed_hash_of_compressed_files	String	sipHash128 压缩文件中的数据，就好像它们是未压缩的一样
bytes	UInt64	bytes_on_disk 的别名
marks_size	UInt64	mark_bytes 的别名

system.part_log

仅当指定 part_log server 设置时，才会创建 system.part_log 表。该表包含有关 MergeTree 系列表中数据分块发生的事件的信息，例如添加或合并数据。

列信息：

列名	类型	描述
event_type	Enum	数据分块发生的事件的类型。可取值为 NEW_PART、MERGE_PARTS、DOWNLOAD_PART、REMOVE_PART、MUTATE_PART、MOVE_PART 其中之一。
event_date	Date	事件日期
event_time	DateTime	事件时间
duration_ms	UInt64	持续时间

列名	类型	描述
database	String	数据分块所在的数据库的名称
table	String	数据分块所在的表的名称
part_name	String	数据分块名
partition_id	String	数据分块插入到的分区的 ID
rows	UInt64	数据分块的行数
size_in_bytes	UInt64	数据分块的大小（以字节为单位）
merged_from	Array(String)	组成当前分块的分块名称数组（合并后）
bytes_uncompressed	UInt64	未压缩字节的大小
read_rows	UInt64	合并期间读取的行数
read_bytes	UInt64	合并期间读取的字节数
error	UInt16	发生错误的错误码
exception	String	发生错误的错误信息

system.part_log 表在第一次将数据插入到 MergeTree 表之后创建。

system.processes

该系统表用于实现 SHOW PROCESSLIST 查询。

列信息：

列名	类型	描述
user	String	进行查询的用户
address	String	发出请求的 IP 地址
elapsed	Float64	自请求执行开始以来的时间（以秒为单位）
rows_read	UInt64	从表中读取的行数。对于分布式处理，在请求者 server 上，这是所有远程 server 的总数
bytes_read	UInt64	从表中读取的未压缩字节数。对于分布式处理，在请求者 server 上，这是所有远程 server 的总数

列名	类型	描述
total_rows_approx	UInt64	应该读取的总行数的近似值。对于分布式处理，在请求者 server 上，这是所有远程 server 的总数
memory_usage	UInt64	请求使用的 RAM 量。它可能不包括某些类型的专用内存
query	String	查询文本。对于 INSERT，不包含要插入的数据
query_id	String	查询 ID（如果已定义）

system.text_log

system.text_log 表包含日志记录条目。可以使用 text_log.level 服务器设置来限制进入该表的日志记录级别。

列信息：

列名	类型	描述
event_date	Date	条目日期
event_time	DateTime	条目时间
microseconds	UInt32	条目时间
thread_name	String	进行日志记录的线程的名称
thread_id	UInt64	操作系统线程 ID
level	Enum8	日志级别。'Fatal' = 1，'Critical' = 2，'Error' = 3，'Warning' = 4，'Notice' = 5，'Information' = 6，'Debug' = 7，'Trace' = 8
query_id	String	查询的 ID
logger_name	LowCardinality(String)	记录器的名称（如 DDLWorker）
message	String	日志信息
revision	UInt32	ClickHouse 版本
source_file	LowCardinality(String)	完成日志记录的源文件
source_line	UInt64	完成记录的源代码行

system.query_log

system.query_log 表包含有关查询执行的信息。对于每个查询，您可以查看处理开始时间、处理持续时间、错误消息和其他信息。

- 该表不包含 INSERT 查询的输入数据。
- 只有指定了 query_log server 参数，ClickHouse 才会创建此表。此参数设置日志记录规则，例如日志记录间隔或将要登录查询的表的名称。
- 要启用查询日志记录，请将 log_queries 参数设置为1。

system.query_log 表注册两种查询：

- 客户端直接运行的初始查询。
- 由其他查询（用于分布式查询执行）启动的子查询。对于这些类型的查询，有关父查询的信息显示在 `initial_*` 列中。

每个查询根据查询的状态在 query_log 表中创建一两行：

- 如果查询执行成功，则将创建两个 type 为1和2的事件（请参见 type 列）。
- 如果查询处理期间发生错误，则会创建两个 type 为1和4的事件。
- 如果在启动查询之前发生错误，则会创建一个 type 为3的事件。

默认情况下，日志以7.5秒的间隔添加到表中。您可以在 query_log server 中设置此间隔（请参见 flush_interval_milliseconds 参数）。要将日志从内存缓冲区强行刷新到表中，请使用 SYSTEM FLUSH LOGS 查询。

手动删除表后，将即时自动创建它。**所有先前的日志将被删除。**

NOTE：日志的存储期限不受限制。日志不会自动从表格中删除，用户需自行删除过时的日志。您可以在 query_log server 中为 system.query_log 表指定一个任意分区键（请参阅 partition_by 参数）。

system.query_thread_log

system.query_thread_log 表包含有关每个查询执行线程的信息。

- 只有指定了 query_thread_log server 参数，ClickHouse 才会创建此表。此参数设置日志记录规则，例如日志记录间隔或将要登录查询的表的名称。
- 要启用查询日志记录，请将 log_query_threads 参数设置为1。

默认情况下，日志以7.5秒的间隔添加到表中。您可以在 query_log server 中设置此间隔（请参见 flush_interval_milliseconds 参数）。要将日志从内存缓冲区强行刷新到表中，请使用 SYSTEM FLUSH LOGS 查询。

手动删除表后，将即时自动创建它。**所有先前的日志将被删除。**

NOTE：日志的存储期限不受限制。日志不会自动从表格中删除，用户需自行删除过时的日志。您可以在 query_log server 中为 system.query_log 表指定一个任意分区键（请参阅 partition_by 参数）。

system.trace_log

system.trace_log 表包含由采样查询分析器收集的堆栈跟踪记录。设置 trace_log server 配置部分后，ClickHouse 会创建此表。另外，还应该设置 query_profiler_real_time_period_ns 和 query_profiler_cpu_time_period_ns。要分析日志，请使用 addressToLine、addressToSymbol 和 demangle 自我检查函数。

列信息：

列名	类型	描述
event_date	Date	采样日期
event_time	DateTIme	采样时间
revision	UInt32	ClickHouse server 版本修订
timer_type	Enum8	计时器类型，取值 Real、CPU
thread_number	UInt32	线程标识符
query_id	String	查询标识符，可用于获取有关从 query_log 系统表中运行的查询的详细信息
trace	Array(UInt64)	采样时的堆栈跟踪，每个元素都是 ClickHouse server 进程内部的虚拟内存地址

示例：

```
SELECT * FROM system.trace_log LIMIT 1 \G;
```

```
Row 1:
──────
event_date:    2020-03-24
event_time:    2020-03-24 11:12:36
revision:      54427
timer_type:    CPU
thread_number: 47
query_id:      2063fbac-110b-4afd-b7b4-f6cdfe3ce507
trace:         [99887571,100858901,108095547,100858901,102292972,10
0858901,108755294,107833405,100858901,108095547,100858901,108095547
,100858901,100882409,100858901,100831379,100832522,55861763,5586940
0,55857811,126629183,140458940837317,140458935768941]

1 rows in set. Elapsed: 0.003 sec.
```

system.replicas

system.replicas 表包含本地服务器上复制表的信息和状态。该表可用于监视，该表为每个 `Replicated *` 表包含一行。

列信息：

列名	类型	描述
database	String	数据库名
table	String	表名
engine	String	表引擎名
is_leader	UInt8	副本是否为 leader
can_become_leader	UInt8	副本是否可以被选为 leader
is_readonly	UInt8	副本是否处于只读模式
is_session_expired	UInt8	与 ZooKeeper 的会话已过期。基本上与 is_readonly 相同
future_parts	UInt32	尚未完成的 INSERTs 或 merges 的结果数据分块数
parts_to_check	UInt32	队列中用于验证的数据分块的数量。如果怀疑分块可能已损坏，则将其放入验证队列
zookeeper_path	String	ZooKeeper 中表数据的路径
replica_name	String	ZooKeeper 中的副本名称，同一张表的不同副本具有不同的名称
replica_path	String	ZooKeeper 中复制数据的路径
columns_version	Int32	表结构的版本号，表示执行 ALTER 的次数
queue_size	UInt32	等待执行操作的队列的大小。操作包括插入数据块，合并及某些其他操作。它通常与 future_parts 一致
inserts_in_queue	UInt32	需要插入的数据块的插入数量。插入内容通常会很快复制，如果这个数字很大，则表示有问题
merges_in_queue	UInt32	等待进行的合并数。有时合并很长，因此该值可能长时间大于零
part_mutations_in_queue	UInt32	等待突变的数量
queue_oldest_time	DateTime	如果 queue_size 大于0，则显示何时将最早的操作添加到队列中
inserts_oldest_time	DateTime	见 queue_oldest_time
merges_oldest_time	DateTime	见 queue_oldest_time

列名	类型	描述
part_mutations_oldest_time	DateTime	见 queue_oldest_time
log_max_index	UInt64	通用活跃日志中的最大条目数
log_pointer	UInt64	副本复制到其执行队列的通用活跃日志中的最大条目数加1。如果 log_pointer 远小于 log_max_index，会出现问题
last_queue_update	DateTime	上次更新队列的时间
absolute_delay	UInt64	当前副本有几秒钟的延迟
total_replicas	UInt8	该表的已知副本总数
active_replicas	UInt8	在 ZooKeeper 中具有会话的该表的副本数（即正在运行的副本数）

一次只能有一个副本作为 leader。leader 负责选择要执行的后台 merges。**可以写入任何可用并在 ZK 中具有会话的副本，无论其是否为 leader。**

- 如果您请求所有列，则该表的工作可能会有点慢，因为对于每一行都会从 ZooKeeper 进行多次读取。
- 如果您无需请求最后4列（log_max_index、log_pointer、total_replicas、active_replicas），则该表可以快速运行。

例如，您可以检查所有内容是否正常运行，如下所示：

```
SELECT
database,
table,
is_leader,
is_readonly,
is_session_expired,
future_parts,
parts_to_check,
columns_version,
queue_size,
inserts_in_queue,
merges_in_queue,
log_max_index,
log_pointer,
total_replicas,
active_replicas
FROM system.replicas
WHERE
```

```
is_readonly
OR is_session_expired
OR future_parts > 20
OR parts_to_check > 10
OR queue_size > 20
OR inserts_in_queue > 10
OR log_max_index - log_pointer > 10
OR total_replicas < 2
OR active_replicas < total_replicas
```

如果此查询未返回任何内容，则表示一切正常。

```
SELECT
    database,
    table,
    is_leader,
    is_readonly,
    is_session_expired,
    future_parts,
    parts_to_check,
    columns_version,
    queue_size,
    inserts_in_queue,
    merges_in_queue,
    log_max_index,
    log_pointer,
    total_replicas,
    active_replicas
FROM system.replicas
WHERE is_readonly OR is_session_expired OR (future_parts > 20) OR (
parts_to_check > 10) OR (queue_size > 20) OR (inserts_in_queue > 10
) OR ((log_max_index - log_pointer) > 10) OR (total_replicas < 2) O
R (active_replicas < total_replicas)

Ok.

0 rows in set. Elapsed: 0.003 sec.
```

system.settings

该表包含有关当前正在使用的设置的信息，即用于执行您要从 system.settings 表中读取的查询。

列信息：

列名	类型	描述
name	String	设置名
value	String	设置值

列名	类型	描述
changed	UInt8	设置是在配置中显式定义还是显式更改

示例：

```
SELECT name,value,changed FROM system.settings WHERE changed;
```

```
┌─name──────────────────────┬─value──────┬─changed─┐
│ use_uncompressed_cache    │ 0          │       1 │
│ load_balancing            │ random     │       1 │
│ max_memory_usage          │ 10000000000 │       1 │
└───────────────────────────┴────────────┴─────────┘
```

system.table_engines

该表包含 server 支持的表引擎的描述及其功能支持信息。该表只有一列，表示表引擎的名称。

示例：

```
select * from system.table_engines limit 5;
```

```
┌─name──────────┐
│ JDBC          │
│ ODBC          │
│ HDFS          │
│ LiveView      │
│ Kafka         │
│ MaterializedView │
└───────────────┘
```

system.tables

该表包含 server 知道的每个表的元数据，分离的表未显示在 system.tables 中。

列信息：

列名	类型	描述
database	String	表所在的数据库的名称
name	String	表名

列名	类型	描述
engine	String	表引擎名称（不带参数）
is_temporary	UInt8	表示表是否为临时的标志
data_path	String	文件系统中表数据的路径
metadata_path	String	文件系统中表元数据的路径
metadata_modification_time	DateTime	表元数据的最新修改时间
dependencies_database	Array(String)	数据库依赖
dependencies_table	Array(String)	表依赖（基于当前表的 MaterializedView 表）
create_table_query	String	用于创建表的查询
engine_full	String	表引擎的参数
partition_key	String	表中指定的分区键表达式
sorting_key	String	表中指定的排序键表达式
primary_key	String	表中指定的主键表达式
sampling_key	String	表中指定的采样键表达式

system.tables 表用于 SHOW TABLES 查询实现。

system.zookeeper

如果未配置 ZooKeeper，则该表不存在，允许从配置中定义的 ZooKeeper 集群读取数据。该查询在 WHERE 子句中必须具有“path”相等条件。这是 ZooKeeper 中要获取其数据的子代的路径。

查询命令 `SELECT * FROM system.zookeeper WHERE path ='/ clickhouse';` 输出 `/ clickhouse` 节点上所有子级的数据。要输出所有根节点的数据，请写路径 `='/'` 。如果“path”中指定的路径不存在，则会抛出异常。

列信息：

列名	类型	描述
name	String	节点名称
path	String	节点路径

列名	类型	描述
value	String	节点值
dataLength	Int32	值的大小
numChildren	Int32	后代数
czxid	Int64	创建节点的事务的 ID
mzxid	Int64	最后更改节点的事务的 ID
pzxid	Int64	最后删除或添加后代的事务的 ID
ctime	DateTime	节点创建时间
mtime	DateTime	节点的最后修改时间
version	Int32	节点版本，节点被更改的次数
cversion	Int32	添加或删除后代的数量
aversion	Int32	ACL 的更改数量
ephemeralOwner	Int64	对于临时节点，拥有该节点的会话的 ID

system.mutations

该表包含有关 MergeTree 表的突变及其进度的信息，每个突变命令由单行表示。

列信息：

列名	类型	描述
database	String	应用了突变的数据库名称
table	String	应用了突变的表名称
mutation_id	String	突变 ID
command	String	突变命令
create_time	DateTime	该突变命令提交执行的时间
block_numbers.partition_id	Array(String)	分区 ID
block_numbers.number	Array(Int64)	块号

列名	类型	描述
parts_to_do	Int64	完成突变需要突变的数据分块的数量
is_done	UInt8	突变完成标识
latest_failed_part	String	不能突变的最近分块的名称
latest_fail_time	DateTime	最近一次分块突变失败的时间
latest_fail_reason	String	导致最近的分块突变失败的异常消息

system.disks

该表包含有关 server 配置中定义的磁盘的信息。

列信息：

列名	类型	描述
name	String	server 配置中的磁盘名称
path	String	文件系统中挂载点的路径
free_space	UInt64	磁盘上的可用空间（以字节为单位）
total_space	UInt64	磁盘卷（以字节为单位）
keep_free_space	UInt64	应该在磁盘上保持可用空间的磁盘空间量（以字节为单位）。在磁盘配置的 keep_free_space_bytes 参数中定义

system.storage_policies

该表包含有关 server 配置中定义的存储策略和卷的信息。

列信息：

列名	类型	描述
policy_name	String	存储策略名称
volume_name	String	存储策略中定义的卷名称
volume_priority	UInt64	配置中的卷优先级
disks	Array(String)	磁盘名称，在存储策略中定义
max_data_part_size	UInt64	可以存储在卷磁盘上的数据分块的大小（0 - 无穷）

列名	类型	描述
move_factor	Float32	可用磁盘空间比率。当比率超过配置参数的值时，ClickHouse 开始按顺序将数据移至下一个卷

如果存储策略包含一个以上的卷，则每个卷的信息都存储在表的单独一行中。

访问权限控制

最近更新时间：2021-07-01 10:51:03

用户访问权限通常配置在配置文件中，该配置文件默认名称为 user.xml。该文件通常包含三个子 session，分别是 users、profiles 和 quotas。

users 配置

用户名会记录在此配置文件中。在配置文件 user.xml 中，users session 中配置了用户访问权限，每一个用户在 users session 下有独立的子 session，以用户名命名，且包含如下内容：

- password：明文密码。
- password_sha256_hex：sha256 加密。
- password_double_sha1_hex：double sha1 加密。
- networks：配置访问源，可配置 IP、网段、支持通配符。
- profile：配置用的 profile 属性，控制用户资源使用。
- quota：配置 quota 属性，在一个周期内，通过多维度阈值，限制用户资源使用。

profiles 配置

profile 配置规定了资源使用限制。profile 配置在 user.xml 中的 profiles session 中。可以同时配置多个 profile，每一个 profile 在 profiles session 中单独一个子 session，其内容如下：

- max_memory_usage：单个查询最大使用内存量，单位 byte。
- use_uncompressed_cache：非压缩数据缓存开关，0表示关闭，1表示启用。
- load_balancing：采用分布式查询方式时，在多个 replicas 上查询时采取的访问策略，默认是 random。
- readonly：只读标志。1表示只读权限，0表示非只读权限。

quotas 配置

quota 允许用户在一个时间周期内（周期时长可配置），从多个维度设定阈值，达到资源访问控制的目的。这些维度包括：周期内查询次数、周期内异常查询次数、周期内查询结果的总行数、周期内在集群范围内查询读取行数（读取的列数因为过滤，通常要大于结果行数），以及查询执行时间（单位 s，wall time）。ClickHouse 允许同时配置多个 quota，每一个 quota 在 quotas session 中单独一个子 session，具体内容如下：

- duration：周期时长，单位 s。

- queries：周期内查询次数限制。
- errors：周期内查询异常次数限制。
- result_rows：周期内作为查询结果的行数限制。
- read_rows：周期内查询读取数据行数限制。
- execution_time：周期内查询执行时间，单位 s。

当用户在一个周期内，上述任何一个限制达到后，会有一个异常。如果上述限制值为0，表示没有限制。默认情况下，无限制。

附录

```
<?xml version="1.0"?>
<yandex>
<!-- Profiles of settings. -->
<profiles>
<!-- Default settings. -->
<default>
<!-- Maximum memory usage for processing single query, in bytes. -->
<max_memory_usage>10000000000</max_memory_usage>
<!-- Use cache of uncompressed blocks of data. Meaningful only for processing man
y of very short queries. -->
<use_uncompressed_cache>0</use_uncompressed_cache>
<!-- How to choose between replicas during distributed query processing.
random - choose random replica from set of replicas with minimum number of errors
nearest_hostname - from set of replicas with minimum number of errors, choose rep
lica
with minimum number of different symbols between replica's hostname and local hos
tname
(Hamming distance).
in_order - first live replica is chosen in specified order.
first_or_random - if first replica one has higher number of errors, pick a random
one from replicas with minimum number of errors.
-->
<load_balancing>random</load_balancing>
</default>
<!-- Profile that allows only read queries. -->
<readonly>
<readonly>1</readonly>
</readonly>
</profiles>
<!-- Users and ACL. -->
<users>
<!-- If user name was not specified, 'default' user is used. -->
```

```
<default>
<!-- Password could be specified in plaintext or in SHA256 (in hex format).
If you want to specify password in plaintext (not recommended), place it in 'pass
word' element.
Example: <password>qwerty</password>.
Password could be empty.
If you want to specify SHA256, place it in 'password_sha256_hex' element.
Example: <password_sha256_hex>65e84be33532fb784c48129675f9eff3a682b27168c0ea744b2
cf58ee02337c5</password_sha256_hex>
Restrictions of SHA256: impossibility to connect to ClickHouse using MySQL JS cli
ent (as of July 2019).
If you want to specify double SHA1, place it in 'password_double_sha1_hex' elemen
t.
Example: <password_double_sha1_hex>e395796d6546b1b65db9d665cd43f0e858dd4303</pass
word_double_sha1_hex>
How to generate decent password:
Execute: PASSWORD=$(base64 < /dev/urandom | head -c8); echo "$PASSWORD"; echo -n
"$PASSWORD" | sha256sum | tr -d '-'
In first line will be password and in second - corresponding SHA256.
How to generate double SHA1:
Execute: PASSWORD=$(base64 < /dev/urandom | head -c8); echo "$PASSWORD"; echo -n
"$PASSWORD" | openssl dgst -sha1 -binary | openssl dgst -sha1
In first line will be password and in second - corresponding double SHA1.
-->
<password></password>
<!-- List of networks with open access.
To open access from everywhere, specify:
<ip>::/0</ip>
To open access only from localhost, specify:
<ip>::1</ip>
<ip>127.0.0.1</ip>
Each element of list has one of the following forms:
<ip> IP-address or network mask. Examples: 213.180.204.3 or 10.0.0.1/8 or 10.0.0.
1/255.255.255.0
2a02:6b8::3 or 2a02:6b8::3/64 or 2a02:6b8::3/ffff:ffff:ffff:ffff::.
<host> Hostname. Example: server01.yandex.ru.
To check access, DNS query is performed, and all received addresses compared to p
eer address.
<host_regexp> Regular expression for host names. Example, ^server\d\d-\d\d-\d\.ya
ndex\.ru$
To check access, DNS PTR query is performed for peer address and then regexp is a
pplied.
Then, for result of PTR query, another DNS query is performed and all received ad
dresses compared to peer address.
Strongly recommended that regexp is ends with $
All results of DNS requests are cached till server restart.
-->
```

```
<networks incl="networks" replace="replace">
<ip>::/0</ip>
</networks>
<!-- Settings profile for user. -->
<profile>default</profile>
<!-- Quota for user. -->
<quota>default</quota>
<!-- For testing the table filters -->
<databases>
<test>
<!-- Simple expression filter -->
<filtered_table1>
<filter>a = 1</filter>
</filtered_table1>
<!-- Complex expression filter -->
<filtered_table2>
<filter>a + b &lt; 1 or c - d &gt; 5</filter>
</filtered_table2>
<!-- Filter with ALIAS column -->
<filtered_table3>
<filter>c = 1</filter>
</filtered_table3>
</test>
</databases>
</default>
<!-- Example of user with readonly access. -->
<!-- <readonly>
<password></password>
<networks incl="networks" replace="replace">
<ip>::1</ip>
<ip>127.0.0.1</ip>
</networks>
<profile>readonly</profile>
<quota>default</quota>
</readonly> -->
</users>
<!-- Quotas. -->
<quotas>
<!-- Name of quota. -->
<default>
<!-- Limits for time interval. You could specify many intervals with different li
mits. -->
<interval>
<!-- Length of interval. -->
<duration>3600</duration>
<!-- No limits. Just calculate resource usage for time interval. -->
<queries>0</queries>
```

```
<errors>0</errors>
<result_rows>0</result_rows>
<read_rows>0</read_rows>
<execution_time>0</execution_time>
</interval>
</default>
</quotas>
</yandex>
```

ClickHouse 数据迁移指引

最近更新时间：2021-10-29 09:48:04

功能说明

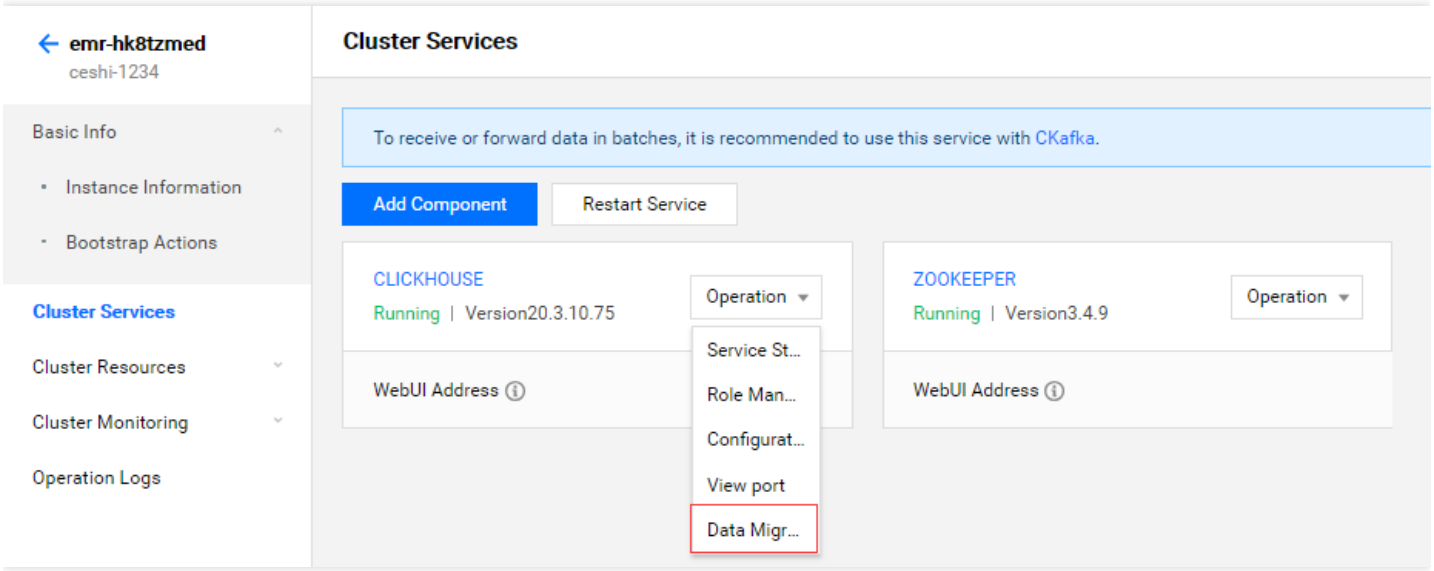
ClickHouse 集群由 N ≥ 1 个节点构成，在集群中用户可以通过配置定义出N≥1个虚拟集群（Cluster），当虚拟集群（Cluster）节点数量或存储量发生变化时，可以使用数据迁移功能，均衡数据分布，提升集群资源利用率，支持下线模式和均衡模式两种模式。

应用场景

- 均衡模式是将迁出节点中的数据按平均原则分摊到迁入节点中，只迁移分区表，不迁移非分区表；适用于集群扩容或节点负载均衡场景。
- 下线模式是将迁出节点中的数据全部迁移到迁入节点中，分区表和非分区表全部迁移；适用于销毁节点或数据备份场景。

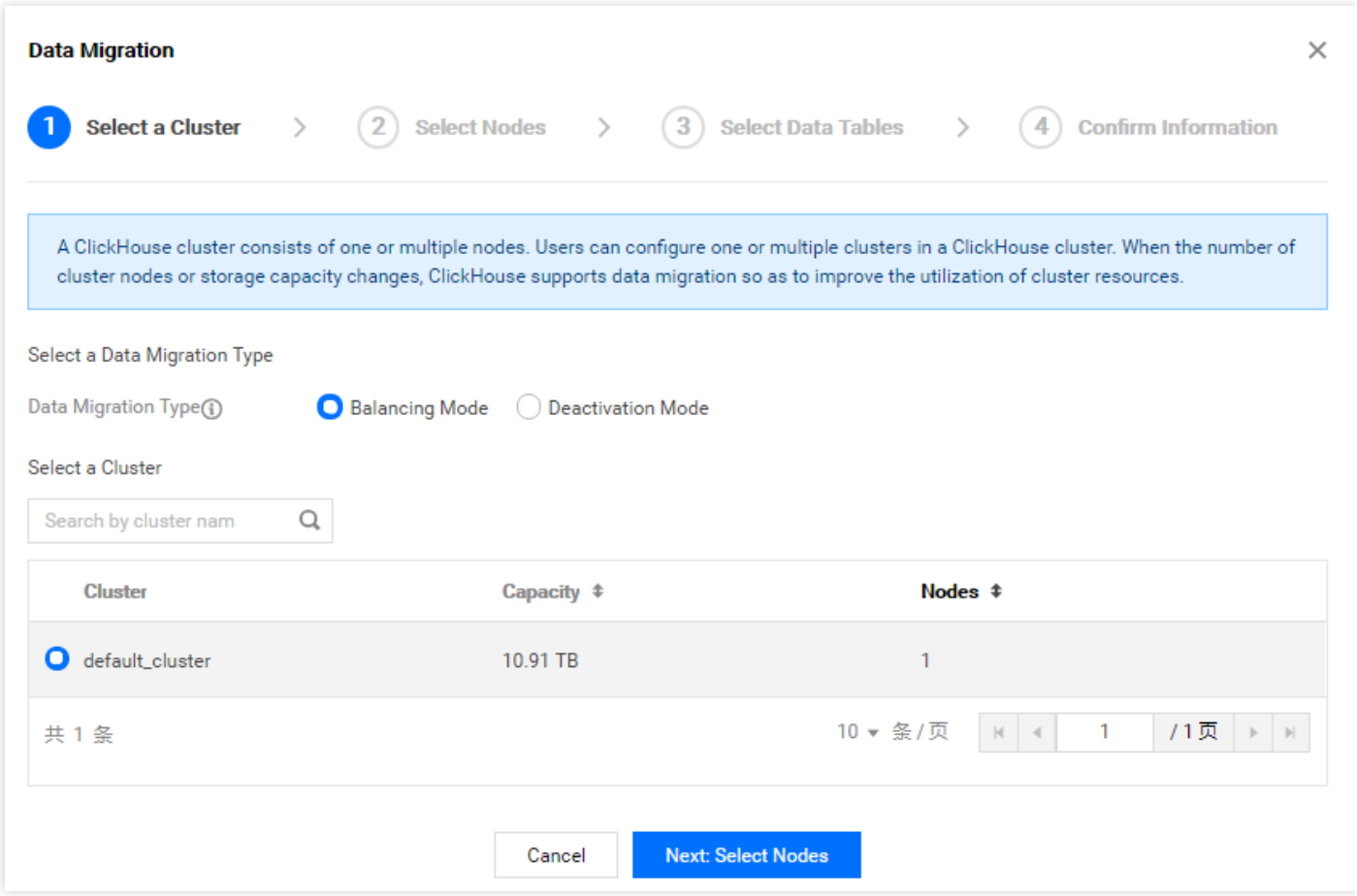
操作说明

登录 EMR 控制台，在集群列表中选择**CLICKHOUSE 集群>集群服务**，单击 CLICKHOUSE卡片中的**操作>数据迁移**。



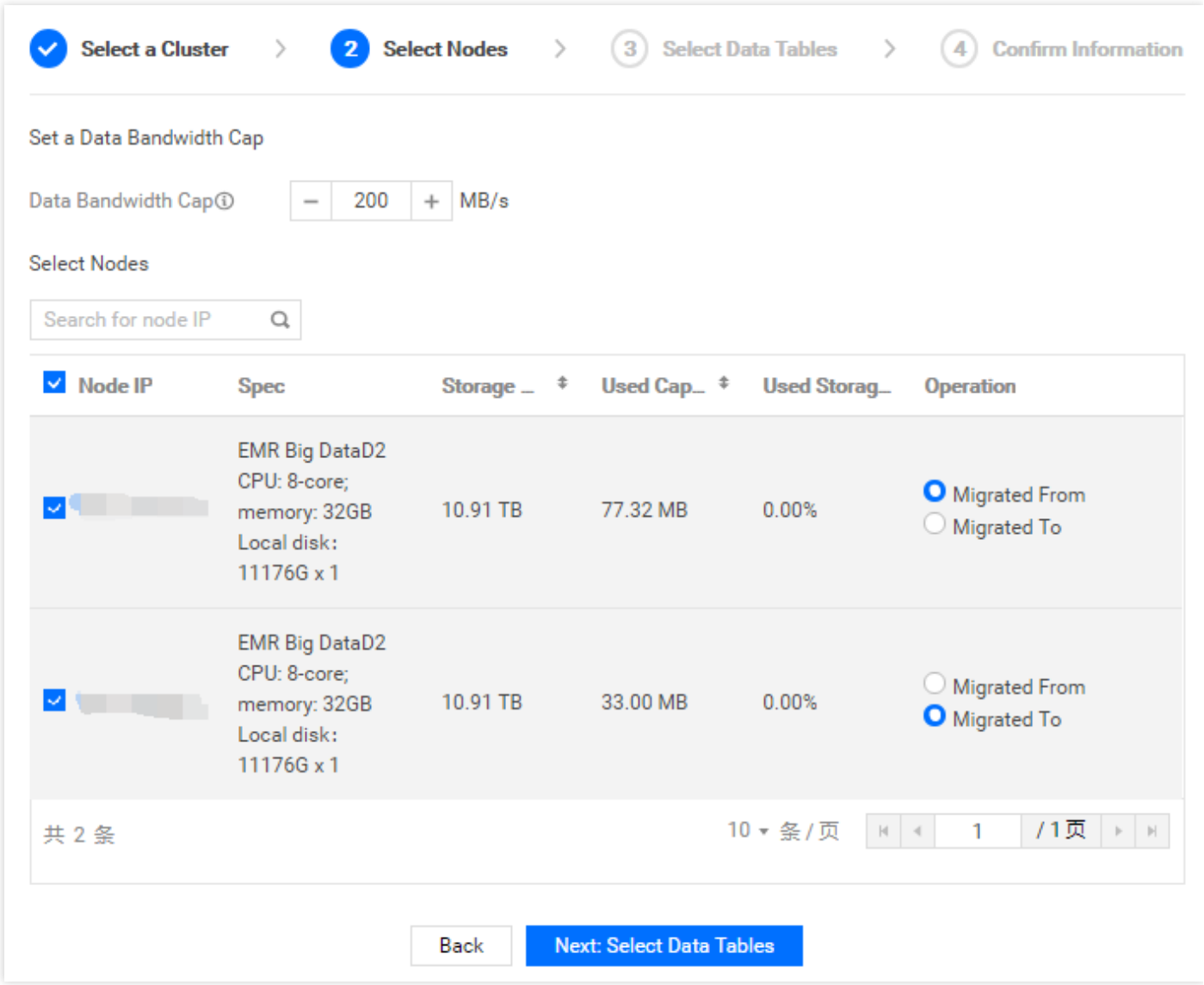
数据迁移步骤分为四步：

1. **选择 Cluster**，并选择迁移类型，默认为“均衡模式”，每次迁移时只能选择一个 Cluster。



2. **下一步：选择迁移节点**，设置数据迁移带宽上限值，系统默认推荐值200MB/S，可自定义调整。系统会默认勾选 Cluster 中所有节点列表，并自动标记迁出和迁入节点，可自行轻微调整设置，同一节点只能作为迁出节点或迁入

节点，且同一个 Cluster 中必须包含迁出节点和迁入节点。



3. **下一步：选择迁移数据表**，系统拉取所有迁出节点中的表，默认数据总量由高到低排列，单页只展示10张表，默认勾选10张表，可能根据阈值以及包含不包含条件进行搜索查询。

4. **下一步：信息确认**，确认最终迁移信息。

注意：

- 均衡模式：只迁移分区表。
- 下线模式：迁移分区表和非分区表。

Druid 开发指南

Druid 简介

最近更新时间：2021-07-01 10:56:19

Apache Druid 是一个分布式的、支持实时多维 OLAP 分析的数据处理系统，用于解决如何在大规模数据集下进行快速的、交互式的查询和分析的问题。

基本特点

Apache Druid 具有以下特点：

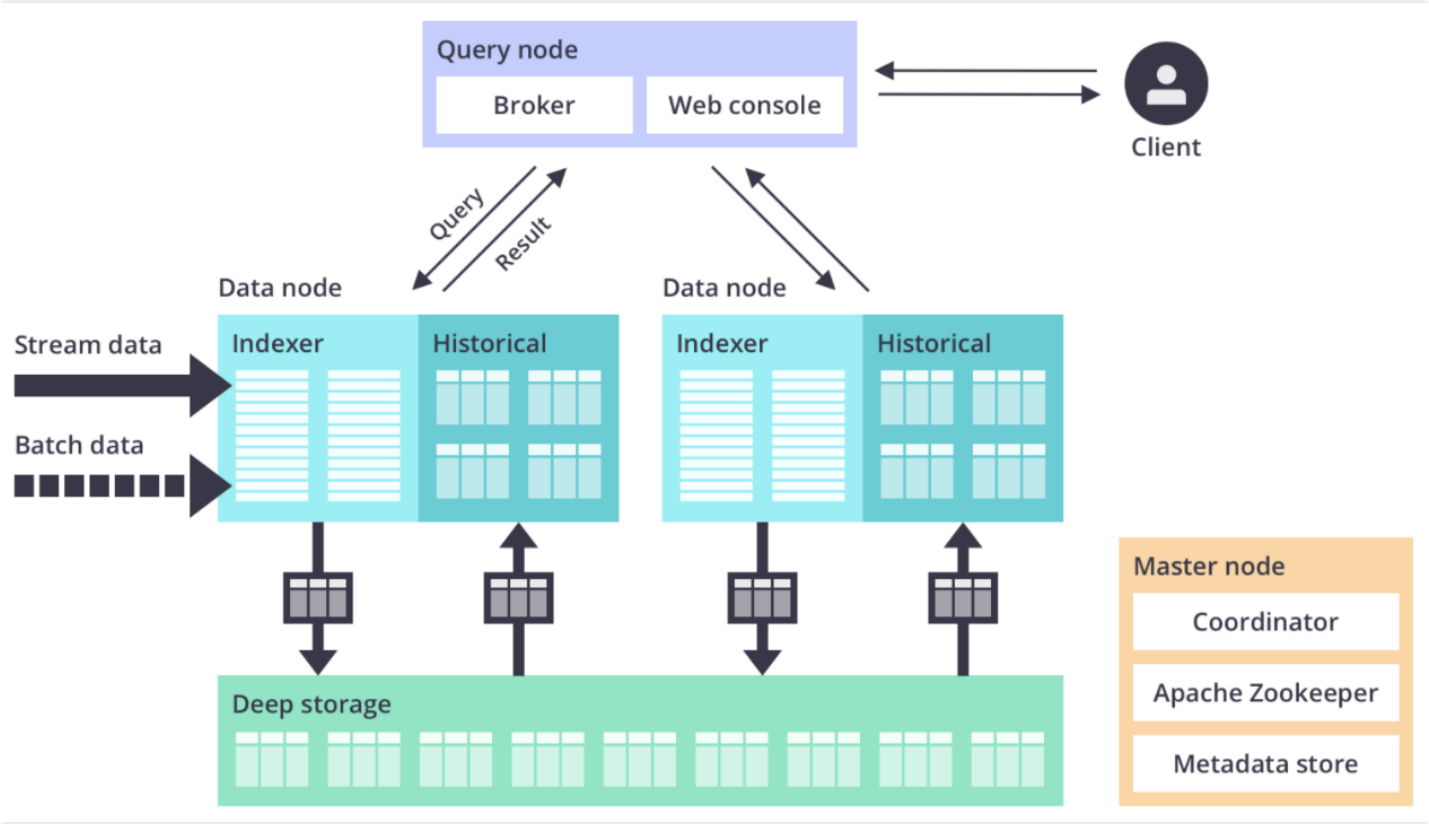
- 亚秒响应的交互式查询，支持较高；包括多维过滤、Ad-hoc 的属性分组、快速聚合数据等。
- 支持高并发、实时数据摄入，真正做到数据摄入实时、查询结果实时。
- 扩展性强，采用分布式 shared-nothing 的架构，支持 PB 级数据、千亿级事件快速处理，支持每秒数千查询并发。
- 支持多租户同时在线查询。
- 支持高可用，且支持滚动升级。

应用场景

Druid 最常用的场景就是大数据背景下、灵活快速的多维 OLAP 分析。另外，Druid 还有一个关键的特点，它支持根据时间戳对数据进行预聚合摄入和聚合分析。因此也有用户经常在有时序数据处理分析的场景中用到它，例如广告平台、实时指标监控、推荐模型、搜索模型。

体系架构

Druid 是一个基于微服务的架构。Druid 中的每个核心服务均可以单独或联合部署在不同硬件上。



EMR 增强型 Druid

EMR Druid 基于 Apache Druid 做了大量的改进，包括与 EMR Hadoop 和云周边生态的集成、方便的监控与运维支持、易用的产品接口等，做到了即买即用和免运维。

EMR Druid 目前支持的特性如下所示：

- 方便和 EMR Hadoop 集群结合
- 方便快速弹性扩缩容
- 支持 HA
- 支持将 COS 作为 deep storage
- 支持将 COS 文件作为批量索引的数据源
- 支持将元数据存储到 TencentDB
- 集成了 Superset 等工具
- 丰富的监控指标和告警规则
- 故障迁移
- 高安全性

Druid 使用

最近更新时间：2023-07-14 10:57:58

EMR 支持将 E-MapReduce Druid 集群作为单独的集群类型，主要基于以下几方面的考虑：

- 使用场景：E-MapReduce Druid 可以脱离 Hadoop 使用来适配不同的业务应用场景。
- 资源抢占：E-MapReduce Druid 对内存要求比较高，尤其是 Broker节点和 Historical 节点。E-MapReduce Druid 本身资源使用不受 Hadoop YARN 统一调度，运行时容易发生资源抢夺。
- 集群规模：Hadoop 作为基础设施，其规模一般较大，而 E-MapReduce Druid 集群可能相对较小，部署在同一集群上，由于规模不一致可能造成资源浪费，单独部署则更加灵活。

购买建议

在创建 EMR 集群时选择 Druid 集群类型即可。Druid 集群自带了Hadoop HDFS 和 YARN 服务，并已经和 Druid 集群完成集成。但是建议仅用于测试，**对于线上生产环境，强烈推荐您采用专门的 Hadoop 集群**。
如果需要关闭 Druid 集群自带的 Hadoop 相关服务，可以在 EMR 控制台 的集群服务页，选择对应服务卡片，单击**操作 > 暂停服务**对服务进行暂停。

Hadoop 和 Druid 集群连通配置

本节介绍如何配置 Hadoop 集群和 Druid 集群的连通性。如果您使用 Druid 集群自带的 Hadoop 集群（生产环境不推荐这么做），则无须做额外设置即可正常连通使用，可以跳过此节。

如果您需要将索引数据存放在另外一个单独的 Hadoop 集群的 HDFS 上（生产环境推荐这种方式），则首先需要设置两个集群的连通性。具体步骤如下：

1. 确保 Druid 集群和 Hadoop 集群能够正常通信。
   两个集群在同一个 VPC 下，或两个集群在不同 VPC，但两个 VPC 之间能够正常通信（如通过云联网或者对等连接）。
2. 在 E-MapReduce Druid 集群的每个节点的 `/usr/local/service/druid/conf/druid/_common` 路径下，放置一份 Hadoop 集群中 `/usr/local/service/hadoop/etc/hadoop` 路径下的 core-site.xml、hdfs-site.xml、yarn-site.xml、mapred-site.xml 文件。

> 注意：
> Druid 集群由于自带 Hadoop 集群，因此 Druid 路径下已经提前创建了上述文件的相关软链接，需要先删除，再拷贝另一个 Hadoop 集群的配置过来。同时需要确保文件权限正确，能被 hadoop 用户正常访问。

3. 在 Druid 配置管理中修改 common.runtime.properties 配置文件。修改完成后，保存配置并重启 Druid 集群相关服务。

- druid.storage.type：默认为 hdfs，无需修改
- druid.storage.storageDirectory：

```
如果另一个 Hadoop 集群为非 HA：hdfs://{namenode_ip}:4007
如果另一个 Hadoop 集群为 HA：hdfs://HDFSXXXXX
请配置全路径，详细地址可以在目标 Hadoop 集群的 core-site.xml 文件的 fs.defaultFS 配置项里面找到。
```

使用 COS

E-MapReduce Druid 支持以 COS 作为 deep storage，本节介绍如何使用 COS 作为 Druid 集群的 deep storage。

首先您需要确保 Druid 集群和目标 Hadoop 均开启了 COS 服务，可以在购买 Druid 集群和 Hadoop 集群时开启，也可以购买后在 EMR 控制台进行后配置 COS。

1. 在 Druid 配置管理中修改 common.runtime.properties 配置文件：
   - druid.storage.type：仍然为 hdfs。
   - druid.storage.storageDirectory：`cosn://{bucket_name}/druid/segments`。
     可以到 COS 上预先创建并设置 segments 目录和权限。

2. 在 hdfs 配置管理中修改 core-site.xml 配置文件：
   - fs.cosn.impl ：修改为 `org.apache.hadoop.fs.CosFileSystem`。
   - 新增配置项 `fs.AbstractFileSystem.cosn.impl` ：修改为 `org.apache.hadoop.fs.CosN`。

3. 将 hadoop-cos 相关的 jar 包 （例如：cos_api-bundle-5.6.69.jar、hadoop-cos-2.8.5-8.1.6.jar）放到集群各个节点的/usr/local/service/druid/extensions/druid-hdfs-storage、/usr/local/service/druid/hadoopdependencies/hadoop-client/2.8.5、/usr/local/service/hadoop/share/hadoop/common/lib/目录下。

保存配置并重启 Druid 集群相关服务。

调整 Druid 参数

E-MapReduce Druid 在创建集群后会自动生成一套配置，不过建议您根据业务需求调整最优内存配置。要调整配置，您可以通过 配置管理 功能进行操作。

调整配置时，请确保调整正确：

```
MaxDirectMemorySize >= druid.processing.buffer.sizeByte *(druid.processing.numMer
geBuffers + druid.processing.numThreads + 1)
```

调整建议：

```
druid.processing.numMergeBuffers = max(2, druid.processing.numThreads / 4)
druid.processing.numThreads = Number of cores - 1 (or 1)
druid.server.http.numThreads = max(10, (Number of cores * 17) / 16 + 2) + 30
```

更多配置请参考 Druid 组件配置。

扩展 Router 作为查询节点

当前 Druid 集群默认部署 Broker 进程在 EMR Master 节点上，由于 Master 节点部署较多进程，进程之间影响可能出现内存不够的情况，影响查询效率；同时许多业务也希望查询节点和中心节点分离部署。在这些情况下您可以在控制台扩容一到多个 Router 节点并选择安装 Broker 进程，可以方便的扩展 Druid 集群的查询节点。

访问 Web

统一通过 console 访问 Druid 集群，端口开在主节点的18888端口，可以自行配置公网 IP，在安全组中开通18888的端口并设置带宽后，即可通过 `[http://{masterIp}:18888]()` 访问。

从 Hadoop 批量摄入数据

最近更新时间：2021-07-19 14:14:21

本节简单介绍如何从远程 Hadoop 集群中批量加载数据文件到 Druid 集群中。本文操作均是以 Hadoop 用户进行，请先在 Druid 集群和 Hadoop 集群上都切换到 Hadoop 用户。

批量加载数据到 Druid 集群

1. 在对应远程 hadoop 集群上，以 Hadoop 用户执行以下新建目录命令：

```
hdfs dfs -mkdir /druid
hdfs dfs -mkdir /druid/segments
hdfs dfs -mkdir /quickstart
hdfs dfs -chmod 777 /druid
hdfs dfs -chmod 777 /druid/segments
hdfs dfs -chmod 777 /quickstart
```

> 注意：
> 如果 Druid 集群和 Hadoop 集群是两个独立集群，则目录需要建立在对应 Hadoop 集群上（之后的操作类似，注意分辨正确操作对应的集群）；如果在测试环境下 Druid 集群和 Hadoop 集群是同一个集群，则在同集群操作即可。

2. 上传测试包
   Druid 集群下自带一个名为 Wikiticker 的数据集示例（默认路径 `/usr/local/service/druid/quickstart/tutorial/wikiticker-2015-09-12-sampled.json.gz`），将 Druid 集群内的数据集上传到对应远程 Hadoop 集群，**是在远程 Hadoop 集群上传**。

```
hdfs dfs -put wikiticker-2015-09-12-sampled.json.gz /quickstart/wikiticker-2015-09-12-sampled.json.gz
```

3. 编译索引文件
   准备一个索引文件，仍然使用 Druid 集群的样例文件 `/usr/local/service/druid/quickstart/tutorial/wikipedia-index-hadoop.json`，命令如下：

```
{
"type" : "index_hadoop",
```

```
"spec" : {
"dataSchema" : {
"dataSource" : "wikipedia",
"parser" : {
"type" : "hadoopyString",
"parseSpec" : {
"format" : "json",
"dimensionsSpec" : {
"dimensions" : [
"channel",
"cityName",
"comment",
"countryIsoCode",
"countryName",
"isAnonymous",
"isMinor",
"isNew",
"isRobot",
"isUnpatrolled",
"metroCode",
"namespace",
"page",
"regionIsoCode",
"regionName",
"user",
{ "name": "added", "type": "long" },
{ "name": "deleted", "type": "long" },
{ "name": "delta", "type": "long" }
]
},
"timestampSpec" : {
"format" : "auto",
"column" : "time"
}
}
},
"metricsSpec" : [],
"granularitySpec" : {
"type" : "uniform",
"segmentGranularity" : "day",
"queryGranularity" : "none",
"intervals" : ["2015-09-12/2015-09-13"],
"rollup" : false
}
},
"ioConfig" : {
"type" : "hadoop",
```

```
"inputSpec" : {
"type" : "static",
"paths" : "/quickstart/wikiticker-2015-09-12-sampled.json.gz"
}
},
"tuningConfig" : {
"type" : "hadoop",
"partitionsSpec" : {
"type" : "hashed",
"targetPartitionSize" : 5000000
},
"forceExtendableShardSpecs" : true,
"jobProperties" : {
"yarn.nodemanager.vmem-check-enabled" : "false",
"mapreduce.map.java.opts" : "-Duser.timezone=UTC -Dfile.encoding=UTF-8",
"mapreduce.job.user.classpath.first" : "true",
"mapreduce.reduce.java.opts" : "-Duser.timezone=UTC -Dfile.encoding=UTF-8",
"mapreduce.map.memory.mb" : 1024,
"mapreduce.reduce.memory.mb" : 1024
}
}
},
"hadoopDependencyCoordinates": ["org.apache.hadoop:hadoop-client:2.8.5"]
}
```

说明：

- hadoopDependencyCoordinates 为依赖的 Hadoop 版本。
- spec.ioConfig.inputSpec.paths 为输入文件路径。如果已经在 common.runtime.properties 配置中设置好集群连通性，可以使用相对路径（可参考 Druid 使用）。否则，应该根据情况使用以 `hdfs://` 或者 `cosn://` 开头的相对路径。
- tuningConfig.jobProperties 参数可以设置 mapreduce job 的相关参数。

iv. 提交索引任务

接下来可以在 Druid 集群上提交任务将数据摄入，在 Druid 目录下以 Hadoop 用户执行：

```
./bin/post-index-task --file quickstart/tutorial/wikipedia-index-hadoop.json
--url http://localhost:8090
```

成功后则会有如下类似的输出：

```
...
Task finished with status: SUCCESS
Completed indexing data for wikipedia. Now loading indexed data onto the clus
ter...
wikipedia loading complete! You may now query your data
```

数据查询

Druid 支持类 SQL 和原生 JSON 查询，下面将分别介绍，更多内容可参考 官方文档。

sql 方式查询

Druid 支持多种 SQL 查询方式：

- 在 Web UI 的 Query 菜单中查询。

```
SELECT page, COUNT(*) AS Edits
FROM wikipedia
WHERE TIMESTAMP '2015-09-12 00:00:00' <= "__time" AND "__time" < TIMESTAMP '201
5-09-13 00:00:00'
GROUP BY page
ORDER BY Edits DESC
LIMIT 10
```

- 在查询节点上使用命令行工具 `bin/dsql` 进行交互式查询。

```
[hadoop@172 druid]$ ./bin/dsql
Welcome to dsql, the command-line client for Druid SQL.
Connected to [http://localhost:8082/].
Type "\h" for help.
dsql> SELECT page, COUNT(*) AS Edits FROM wikipedia WHERE "__time" BETWEEN TIME
STAMP '2015-09-12 00:00:00' AND TIMESTAMP '2015-09-13 00:00:00' GROUP BY page O
RDER BY Edits DESC LIMIT 10;
┌──────────────────────────────────────────────────────────┬───────┐
| page | Edits |
├──────────────────────────────────────────────────────────┼───────┤
| Wikipedia:Vandalismusmeldung | 33 |
| User:Cyde/List of candidates for speedy deletion/Subpage | 28 |
| Jeremy Corbyn | 27 |
| Wikipedia:Administrators' noticeboard/Incidents | 21 |
| Flavia Pennetta | 20 |
| Total Drama Presents: The Ridonculous Race | 18 |
| User talk:Dudeperson176123 | 18 |
| Wikipédia:Le Bistro/12 septembre 2015 | 18 |
| Wikipedia:In the news/Candidates | 17 |
| Wikipedia:Requests for page protection | 17 |
└──────────────────────────────────────────────────────────┴───────┘
Retrieved 10 rows in 0.06s.
```

- 用 HTTP 服务查询 SQL。

```
curl -X 'POST' -H 'Content-Type:application/json' -d @quickstart/tutorial/wikip
edia-top-pages-sql.json http://localhost:18888/druid/v2/sql
```

格式化后的输出结果：


```
[
{
"page":"Wikipedia:Vandalismusmeldung",
"Edits":33
},
{
"page":"User:Cyde/List of candidates for speedy deletion/Subpage",
"Edits":28
},
{
"page":"Jeremy Corbyn",
"Edits":27
},
{
"page":"Wikipedia:Administrators' noticeboard/Incidents",
"Edits":21
},
{
"page":"Flavia Pennetta",
"Edits":20
},
{
"page":"Total Drama Presents: The Ridonculous Race",
"Edits":18
},
{
"page":"User talk:Dudeperson176123",
"Edits":18
},
{
"page":"Wikipédia:Le Bistro/12 septembre 2015",
"Edits":18
},
{
"page":"Wikipedia:In the news/Candidates",
"Edits":17
},
{
"page":"Wikipedia:Requests for page protection",
"Edits":17
```

```
}
]
```

原生 JSON 查询

- 在 Web UI 上 Query 菜单直接输入 json 查询。

```
{
"queryType" : "topN",
"dataSource" : "wikipedia",
"intervals" : ["2015-09-12/2015-09-13"],
"granularity" : "all",
"dimension" : "page",
"metric" : "count",
"threshold" : 10,
"aggregations" : [
{
"type" : "count",
"name" : "count"
}
]
}
```

- 在查询节点上 druid 方式目录下用 HTTP 提交。

```
curl -X 'POST' -H 'Content-Type:application/json' -d @quickstart/tutorial/wikip
edia-top-pages.json http://localhost:18888/druid/v2?pretty
```

输出结果：

```
[ {
"timestamp" : "2015-09-12T00:46:58.771Z",
"result" : [ {
"count" : 33,
"page" : "Wikipedia:Vandalismusmeldung"
}, {
"count" : 28,
"page" : "User:Cyde/List of candidates for speedy deletion/Subpage"
}, {
"count" : 27,
"page" : "Jeremy Corbyn"
}, {
"count" : 21,
"page" : "Wikipedia:Administrators' noticeboard/Incidents"
}, {
"count" : 20,
"page" : "Flavia Pennetta"
```

```
}, {
"count" : 18,
"page" : "Total Drama Presents: The Ridonculous Race"
}, {
"count" : 18,
"page" : "User talk:Dudeperson176123"
}, {
"count" : 18,
"page" : "Wikipédia:Le Bistro/12 septembre 2015"
}, {
"count" : 17,
"page" : "Wikipedia:In the news/Candidates"
}, {
"count" : 17,
"page" : "Wikipedia:Requests for page protection"
} ]
} ]
```

从 Kafka 实时摄入数据

最近更新时间：2021-07-01 11:00:18

本文介绍如何使用 Apache Druid Kafka Indexing Service 实时消费 Kafka 数据。开始本节前，类似 Hadoop 集群，需要确保 Kafka 集群和 Druid 集群之间能够正常通信。

说明：

- 两个集群在同一个 VPC 下，或两个集群在不同 VPC，但两个 VPC 之间能够正常通信（如通过云联网或者对等连接）。
- 如有必要需要将 Kafka 集群的 Host 信息配置到 Druid 集群中。

命令行方式

1. 首先在 Kafka 集群启动 kafka broker。

```
./bin/kafka-server-start.sh config/server.properties
```

2. 创建一个kafka topic，名为 mytopic。

```
./bin/kafka-topics.sh --create --zookeeper {kafka_zk_ip}:2181 --replication-fac
tor 1 --partitions 1 --topic mytopic
输出：
Created topic "mytopic".
```

`{kafka_zk_ip}:2181` 为 kafka 集群的 zookeeper 地址。

3. 在 Druid 集群上准备一个数据描述文件 kafka-mytopic.json。

```
{
"type": "kafka",
"dataSchema": {
"dataSource": "mytopic-kafka",
"parser": {
"type": "string",
"parseSpec": {
"timestampSpec": {
"column": "time",
"format": "auto"
},
"dimensionsSpec": {
"dimensions": ["url", "user"]
```

```
},
"format": "json"
}
},
"granularitySpec": {
"type": "uniform",
"segmentGranularity": "hour",
"queryGranularity": "none"
},
"metricsSpec": [{
"type": "count",
"name": "views"
},
{
"name": "latencyMs",
"type": "doubleSum",
"fieldName": "latencyMs"
}
]
},
"ioConfig": {
"topic": "mytopic",
"consumerProperties": {
"bootstrap.servers": "{kafka_ip}:9092",
"group.id": "kafka-indexing-service"
},
"taskCount": 1,
"replicas": 1,
"taskDuration": "PT1H"
},
"tuningConfig": {
"type": "kafka",
"maxRowsInMemory": "100000"
}
}
```

`{kafka_ip}:9092` 为您 Kafka 集群的 bootstrap.servers IP 和端口。

4. 在 Druid 集群的 Master 节点上添加 Kafka supervisor。

```
curl -XPOST -H 'Content-Type: application/json' -d @kafka-mytopic.json http://
{druid_master_ip}:8090/druid/indexer/v1/supervisor
输出：
{"id":"mytopic-kafka"}
```

`{druid_master_ip}:8090` 为 overlord 进程部署的节点，一般是 Master 节点。

5. 在 Kafka 集群上开启一个 console producer。

```
./bin/kafka-console-producer.sh --broker-list {kafka_ip}:9092 --topic mytopic
```

`{kafka_ip}:9092` 为您 Kafka 集群的 bootstrap.servers IP 和端口。

6. 在 druid 集群准备一个查询文件，命名为 query-mytopic.json。

```
{
"queryType" : "search",
"dataSource" : "mytopic-kafka",
"intervals" : ["2020-03-13T00:00:00.000/2020-03-20T00:00:00.000"],
"granularity" : "all",
"searchDimensions": [
"url",
"user"
],
"query": {
"type": "insensitive_contains",
"value": "roni"
}
}
```

7. 在 kafka 上实时输入一些数据。

```
{"time": "2020-03-19T09:57:58Z", "url": "/foo/bar", "user": "brozo", "latencyM
s": 62}
{"time": "2020-03-19T16:57:59Z", "url": "/", "user": "roni", "latencyMs": 15}
{"time": "2020-03-19T17:50:00Z", "url": "/foo/bar", "user": "roni", "latencyMs"
: 25}
```

时间戳生成命令：

```
python -c 'import datetime; print(datetime.datetime.utcnow().strftime("%Y-%m-%d
T%H:%M:%SZ"))'
```

8. 在 Druid 集群上查询。

```
curl -XPOST -H 'Content-Type: application/json' -d @query-mytopic.json http://
{druid_ip}:8082/druid/v2/?pretty
```

`{druid_ip}:8082` 为您 Druid 集群的 broker 节点，一般在 Master 或 Router 节点上。

查询结果：

```
[ {
"timestamp" : "2020-03-19T16:00:00.000Z",
"result" : [ {
"dimension" : "user",
```

```
"value" : "roni",
"count" : 2
} ]
} ]
```

Web 可视化方式

您可通过 Druid Web UI 控制台可视化方式，从 Kafka 集群摄入数据并查询，详细可参考 通过 data loader 加载 Kafka 数据。

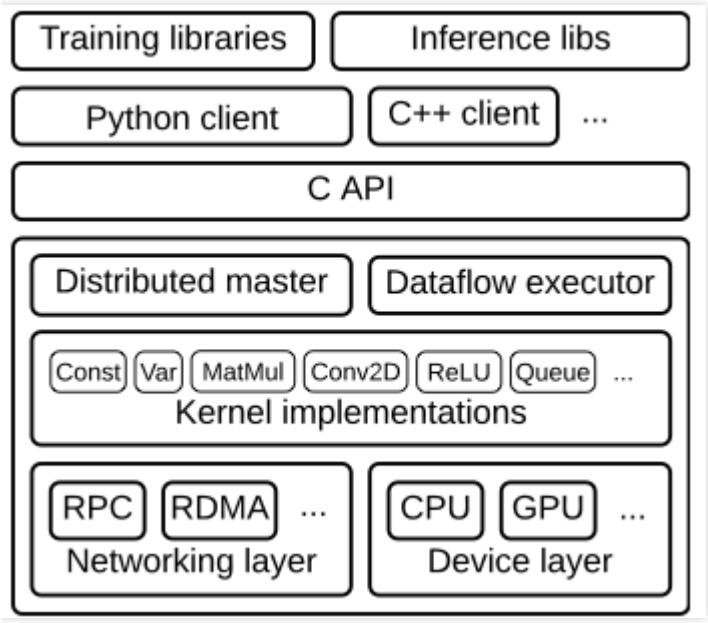
Tensorflow 开发指南
Tensorflow 简介

最近更新时间：2021-07-01 15:23:23

TensorFlow 是一个端到端开源机器学习平台。它拥有一个全面而灵活的生态系统，其中包含各种工具、库和社区资源，可助力研究人员推动先进机器学习技术的发展，并使开发者能够轻松地构建和部署由机器学习提供支持的应用。

- 轻松地构建模型
  在即刻执行环境中使用 Keras 等直观的高阶 API 轻松地构建和训练机器学习模型，此环境使我们能够快速迭代模型并轻松地调试模型。
- 随时随地进行可靠的机器学习生产
  无论您使用哪种语言，都可以在云端、本地、浏览器中或设备上轻松地训练和部署模型。
- 强大的研究实验
  一个简单而灵活的架构，可以更快地将新想法从概念转化为代码，然后创建出先进的模型，并最终对外发布。

Tensorflow 架构



- 客户端（Client）
  将计算过程定义为数据流图。使用 `_Session_` 初始化数据流图的执行。

- 分布式主控端（Master）
  修剪图中的某些特殊子图，即 `Session.run()` 中所定义的参数。将子图划分为在不同进程和设备中运行的多个部分。将图分发给不同的工作进程。由工作进程初始化子图的计算。
- 工作进程（Worker service）（每个任务的）
  使用内核实现调度图操作并在合适的硬件（CPU、GPU 等）执行。向其他工作进程发送或从其接收操作的结果。
- 内核实现
  执行一个独立的图操作计算。

EMR 支持 Tensorflow

- Tensorflow 版本：v1.14.0
- 目前 Tensorflow 只支持运行在 CPU 机型，暂不支持 GPU 机型
- 支持 tensorflow on spark 做分布式训练

Tensorflow 开发示例

首先需要安装 Tensorflow，切换到 root 用户下，密码为创建 EMR 集群时设置的密码，先安装 python-pip 工具再安装依赖包：

```
[hadoop@172 hbase]$ su
Password: ********
[root@172 hbase]# yum install python-pip
[root@172 hbase]# pip install Tensorflow
```

编写代码：`test.py`

```
import tensorflow as tf
hello = tf.constant('Hello, TensorFlow!')
sess = tf.Session()
print sess.run(hello)
a = tf.constant(10)
b = tf.constant(111)
print sess.run(a+b)
exit()
```

执行如下命令：

```
python test.py
```

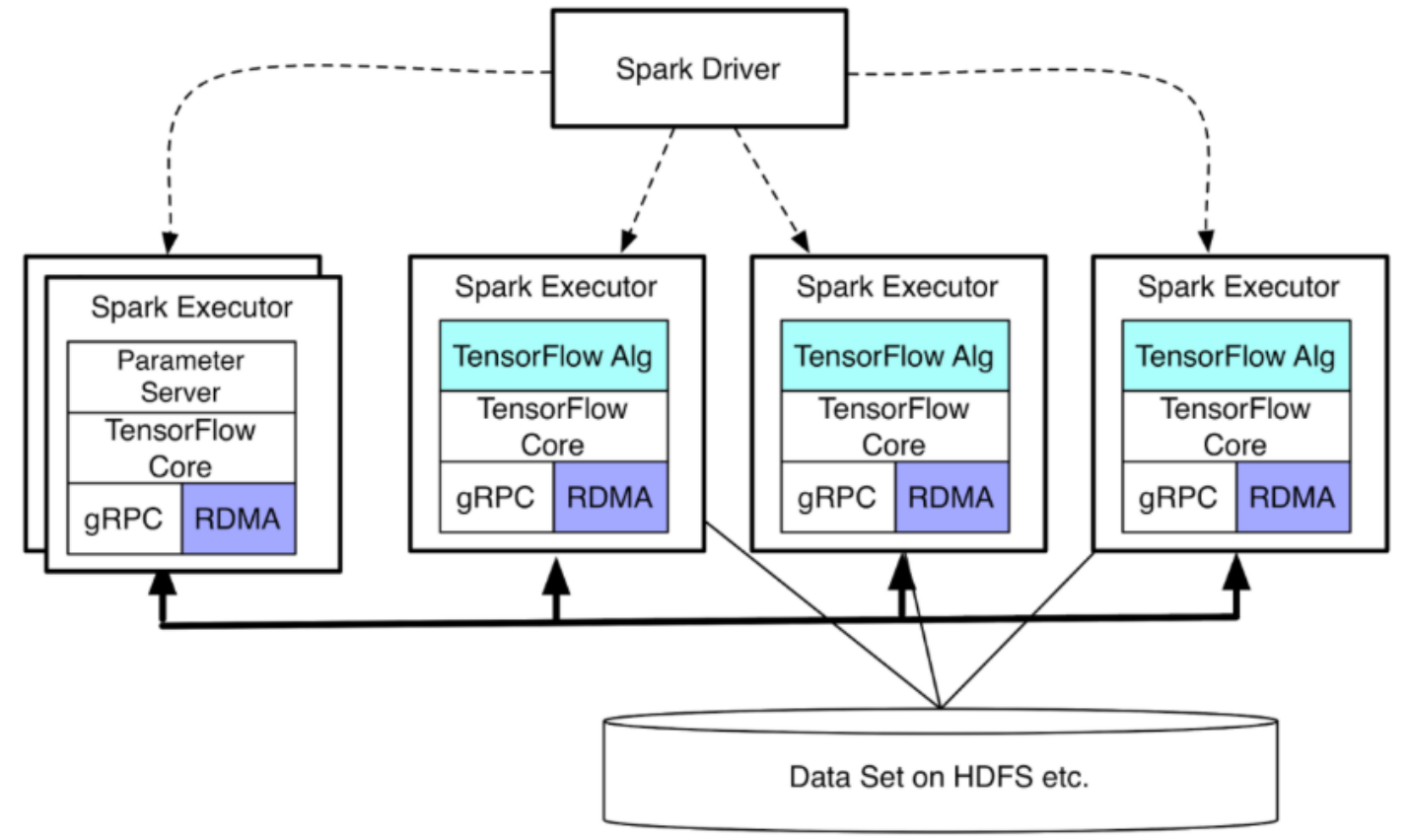
更多用法请参考 Tensorflow 官网。

TensorFlowOnSpark 简介

最近更新时间：2022-04-18 15:28:42

TensorFlowOnSpark 为 Apache Hadoop 和 Apache Spark 集群提供了可扩展的深度学习，TensorFlowOnSpark 支持所有类型的 TensorFlow 程序，可以实现异步/同步的训练和推理，同时也支持模型并行性和数据的并行处理。详情可查阅 TensorFlowOnSpark 官网。

TensorFlowOnSpark 架构图



TensorFlowOnSpark 支持 TensorFlow 进程（计算节点和参数服务节点）之间的直接张量通信。过程到过程的直接通信机制使 TensorFlowOnSpark 程序能够在增加的机器上很轻松的进行扩展。TensorFlowOnSpark 不涉及张量通信中的 Spark 驱动程序，因此实现了与独立 TensorFlow 集群类似的可扩展性。

安装 TensorFlowOnSpark

1. 进入 EMR 购买页，选择产品 EMR-2.3.0 版本及以上版本。

2. 在【可选组件】列表中，勾选 tensorflowonspark 1.4.4 组件。

3. tensorflowonspark 默认安装在 `/usr/local/service/tensorflowonspark` 目录下。

> 注意：
> tensorflowonspark 依赖的组件包含 hive 和 spark，在 tensorflowonspark 的同时也会安装 hive 和 spark 组件。

使用示例

在安装好的 tensorflowonspark 组件目录下，已经有完整的 example 代码，可以按如下操作步骤：

- 下载测试数据
  使用 hadoop 用户，在 `/usr/local/service/tensorflowonspark` 目录下，执行命令：

```
sh mnist_download.sh
cat mnist_download.sh
mkdir ${HOME}/mnist
pushd ${HOME}/mnist >/dev/null
curl -O "http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz"
curl -O "http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz"
curl -O "http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz"
curl -O "http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz"
zip -r mnist.zip *
popd >/dev/null
```

- 上传原始数据和依赖包

```
hdfs dfs -mkdir -p /mnist/tools/
hdfs dfs -put ~/mnist/mnist.zip /mnist/tools
hdfs dfs -mkdir /tensorflow
hdfs dfs -put TensorFlowOnSpark/tensorflow-hadoop-1.10.0.jar /tensorflow
```

- 特征数据准备

```
sh prepare_mnist.sh
```

可以看到特征数据已准备就绪：

```
hdfs dfs -ls /user/hadoop/mnist
Found 2 items
drwxr-xr-x - hadoop supergroup 0 2020-05-21 11:40 /user/hadoop/mnist/csv
drwxr-xr-x - hadoop supergroup 0 2020-05-21 11:41 /user/hadoop/mnist/tfr
```

- 基于 InputMode.SPARK 模型训练

```
sh mnist_train_with_spark_cpu.sh
```

查看模型训练好的模型：

```
[hadoop@10 tensorflow-on-spark]$ hdfs dfs -ls /user/hadoop/mnist_model
Found 10 items
-rw-r--r-- 1 hadoop supergroup 128 2020-05-21 11:46 /user/hadoop/mnist_model/chec
kpoint
-rw-r--r-- 1 hadoop supergroup 243332 2020-05-21 11:46 /user/hadoop/mnist_model/e
vents.out.tfevents.1590032704.10.0.0.114
-rw-r--r-- 1 hadoop supergroup 164619 2020-05-21 11:45 /user/hadoop/mnist_model/g
raph.pbtxt
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 11:45 /user/hadoop/mnist_model/m
odel.ckpt-0.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 11:45 /user/hadoop/mnist_model/mode
l.ckpt-0.index
-rw-r--r-- 1 hadoop supergroup 64658 2020-05-21 11:45 /user/hadoop/mnist_model/mo
del.ckpt-0.meta
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 11:46 /user/hadoop/mnist_model/m
odel.ckpt-595.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 11:46 /user/hadoop/mnist_model/mode
l.ckpt-595.index
-rw-r--r-- 1 hadoop supergroup 64658 2020-05-21 11:46 /user/hadoop/mnist_model/mo
del.ckpt-595.meta
drwxr-xr-x - hadoop supergroup 0 2020-05-21 11:46 /user/hadoop/mnist_model/train
```

- 基于 InputMode.SPARK 模型预测

```
sh mnist_inference_with_spark_cpu.sh
```

查看预测结果：

```
hdfs dfs -cat /user/hadoop/predictions/part-00000 |more
2020-05-21T11:49:56.561506 Label: 7, Prediction: 7
2020-05-21T11:49:56.561535 Label: 2, Prediction: 2
2020-05-21T11:49:56.561541 Label: 1, Prediction: 1
2020-05-21T11:49:56.561545 Label: 0, Prediction: 0
2020-05-21T11:49:56.561550 Label: 4, Prediction: 4
2020-05-21T11:49:56.561555 Label: 1, Prediction: 1
2020-05-21T11:49:56.561559 Label: 4, Prediction: 4
2020-05-21T11:49:56.561564 Label: 9, Prediction: 9
2020-05-21T11:49:56.561568 Label: 5, Prediction: 6
2020-05-21T11:49:56.561573 Label: 9, Prediction: 9
2020-05-21T11:49:56.561578 Label: 0, Prediction: 0
2020-05-21T11:49:56.561582 Label: 6, Prediction: 6
2020-05-21T11:49:56.561587 Label: 9, Prediction: 9
2020-05-21T11:49:56.561603 Label: 0, Prediction: 0
2020-05-21T11:49:56.561608 Label: 1, Prediction: 1
2020-05-21T11:49:56.561612 Label: 5, Prediction: 5
```

- 基于 InputMode.TENSORFLOW 训练模型

```
sh mnist_train_with_tf_cpu.sh
```

查看模型：

```
hdfs dfs -ls mnist_model
Found 25 items
-rw-r--r-- 1 hadoop supergroup 265 2020-05-21 14:58 mnist_model/checkpoint
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:53 mnist_model/events.out.tfevent
s.1590044017.10.0.0.144
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:57 mnist_model/events.out.tfevent
s.1590044221.10.0.0.144
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:57 mnist_model/events.out.tfevent
s.1590044227.10.0.0.144
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:57 mnist_model/events.out.tfevent
s.1590044232.10.0.0.144
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:57 mnist_model/events.out.tfevent
s.1590044238.10.0.0.144
-rw-r--r-- 1 hadoop supergroup 40 2020-05-21 14:58 mnist_model/events.out.tfevent
s.1590044303.10.0.0.114
-rw-r--r-- 1 hadoop supergroup 198078 2020-05-21 14:58 mnist_model/graph.pbtxt
drwxr-xr-x - hadoop supergroup 0 2020-05-21 14:58 mnist_model/inference
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 14:57 mnist_model/model.ckpt-238
.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 14:57 mnist_model/model.ckpt-238.in
```

```
dex
-rw-r--r-- 1 hadoop supergroup 76255 2020-05-21 14:57 mnist_model/model.ckpt-238.
meta
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 14:57 mnist_model/model.ckpt-277
.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 14:57 mnist_model/model.ckpt-277.in
dex
-rw-r--r-- 1 hadoop supergroup 76255 2020-05-21 14:57 mnist_model/model.ckpt-277.
meta
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 14:57 mnist_model/model.ckpt-315
.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 14:57 mnist_model/model.ckpt-315.in
dex
-rw-r--r-- 1 hadoop supergroup 76255 2020-05-21 14:57 mnist_model/model.ckpt-315.
meta
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 14:57 mnist_model/model.ckpt-354
.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 14:57 mnist_model/model.ckpt-354.in
dex
-rw-r--r-- 1 hadoop supergroup 76255 2020-05-21 14:57 mnist_model/model.ckpt-354.
meta
-rw-r--r-- 1 hadoop supergroup 814168 2020-05-21 14:58 mnist_model/model.ckpt-393
.data-00000-of-00001
-rw-r--r-- 1 hadoop supergroup 375 2020-05-21 14:58 mnist_model/model.ckpt-393.in
dex
-rw-r--r-- 1 hadoop supergroup 76255 2020-05-21 14:58 mnist_model/model.ckpt-393.
meta
drwxr-xr-x - hadoop supergroup 0 2020-05-21 14:53 mnist_model/train
```

- 基于 InputMode.TENSORFLOW 模型预测

```
sh mnist_train_with_tf_cpu.sh
```

查看预测结果：

```
hdfs dfs -cat predictions/part-00000 |more
9 4
9 9
4 4
1 1
4 4
8 8
9 9
2 2
3 5
```

```
6 6
9 9
2 2
6 6
0 0
7 7
5 5
3 3
```

Jupyter 开发指南
Jupyter Notebook 简介

最近更新时间：2022-04-18 15:35:06

Jupyter Notebook 简介

Jupyter Notebook 是基于网页的用于交互计算的应用程序。其可被应用于全过程计算：开发、文档编写、运行代码和展示结果。详情可查看 Jupyter Notebook 官方介绍。

简而言之，Jupyter Notebook 是以网页的形式打开，可以在网页页面中直接编写代码和运行代码，代码的运行结果也会直接在代码块下显示。如在编程过程中需要编写说明文档，可在同一个页面中直接编写，便于作及时的说明和解释。

组成部分

- 网页应用
  网页应用即基于网页形式的、结合了编写说明文档、数学公式、交互计算和其他富媒体形式的工具。简言之，网页应用是可以实现各种功能的工具。
- 文档
  Jupyter Notebook 中所有交互计算、编写说明文档、数学公式、图片以及其他富媒体形式的输入和输出，都是以文档的形式体现的。这些文档是保存为后缀名为 `.ipynb` 的 `JSON` 格式文件，不仅便于版本控制，也方便与他人共享。此外，文档还可以导出为 HTML、LaTeX、PDF 等格式。

Jupyter Notebook 的主要特点

1. 编程时具有语法高亮 、缩进、tab 补全的功能。
2. 可直接通过浏览器运行代码，同时在代码块下方展示运行结果。
3. 以富媒体格式展示计算结果。富媒体格式包括 HTML、LaTeX、PNG、SVG 等。
4. 对代码编写说明文档或语句时，支持 Markdown 语法。
5. 支持使用 LaTeX 编写数学性说明。

安装 jupyter

进入 EMR 购买页。

- 选择产品版本：EMR-V2.3.0。

- 在【可选组件】列表中，选择【tensorflowonspark 1.4.4】后就会默认安装 Jupyter，安装目录位于 `/usr/local/service/jupyter` ；jupyter 不会启动任何服务，如果您没有安装 tensorflowonspark，那默认的安装路径位于 `/usr/local/service/apps/jupyter` 。

使用 jupyter

初始化 jupyter 配置

```
Usage: init.sh [password] [port]
# 示例
./init.sh 123456 10086
```

一路回车，会出现提示：

```
[hadoop@10 jupyter]$ ./init.sh 123456 10086
Your password is: 123456
Your signature is: sha1:139fa061bae6:bcdc6a7870878458c7c14594fe65dd21f85f84a4
Generating a 4096 bit RSA private key
...............................++
.....................................................................................
++
writing new private key to '/usr/local/service/jupyter/conf/jkey.key'
-----
You are about to be asked to enter information that will be incorporated
into your certificate request.
What you are about to enter is what is called a Distinguished Name or a DN.
There are quite a few fields but you can leave some blank
For some fields there will be a default value,
If you enter '.', the field will be left blank.
-----
Country Name (2 letter code) [XX]:
State or Province Name (full name) []:
Locality Name (eg, city) [Default City]:
Organization Name (eg, company) [Default Company Ltd]:
Organizational Unit Name (eg, section) []:
Common Name (eg, your name or your server's hostname) []:
Email Address []:
Jupyter config has already generated at /usr/local/service/jupyter/conf/jupyter_n
otebook_config.py
Now, you can execute the following command to start jupyter:
jupyter notebook --config=/usr/local/service/jupyter/conf/jupyter_notebook_confi
g.py --allow-root
```

这里生成了 jupyter 配置，您也可以修改生成的配置文件 `jupyter_notebook_config.py` 中的相关参数，参考 jupyter 官网即可。

> 注意：
> 最后一行是启动命令，复制这行命令即可启动 jupyter。

启动 jupyter notebook

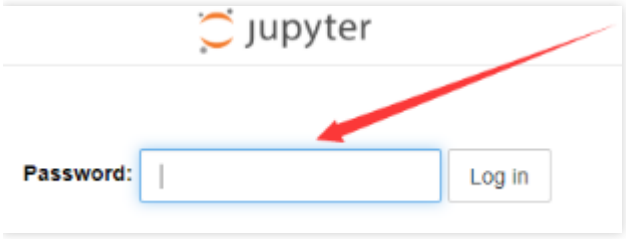
```
jupyter notebook --config=/usr/local/service/jupyter/conf/jupyter_notebook_config.py --allow-root
[I 10:47:46.972 NotebookApp] Writing notebook server cookie secret to /home/hadoop/.local/share/jupyter/runtime/notebook_cookie_secret
[I 10:47:47.748 NotebookApp] Serving notebooks from local directory: /usr/local/service/jupyter
[I 10:47:47.749 NotebookApp] The Jupyter Notebook is running at:
[I 10:47:47.749 NotebookApp] https://(10.0.0.7 or 127.0.0.1):10086/
[I 10:47:47.749 NotebookApp] Use Control-C to stop this server and shut down all kernels (twice to skip confirmation).
```

访问 jupyter

在 web 页面上打开 jupyter（在此之前可能需要去安全组打开 jupyter 端口）：

```
https://IP:10086/
```

这里端口10086是上面初始化 `init.sh` 的参数。

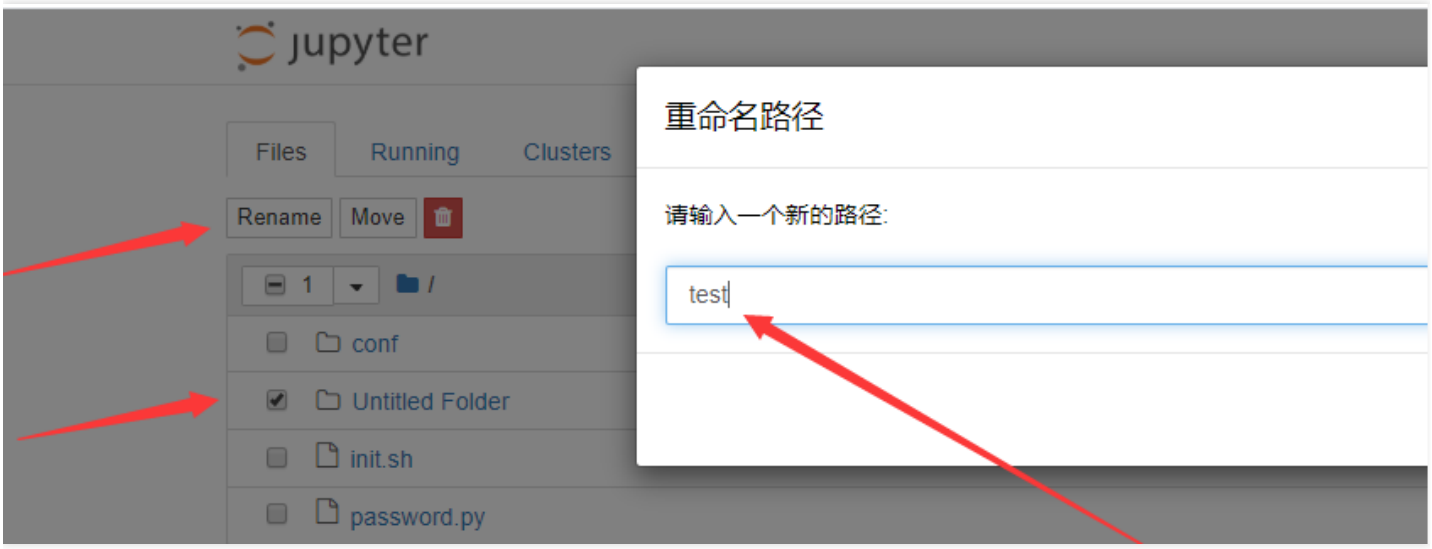


在这里输入刚设置的密码后即可进入 jupyter 主页。

使用 jupyter 进行开发操作

创建目录

rename 目录



编写 tensorflow 代码

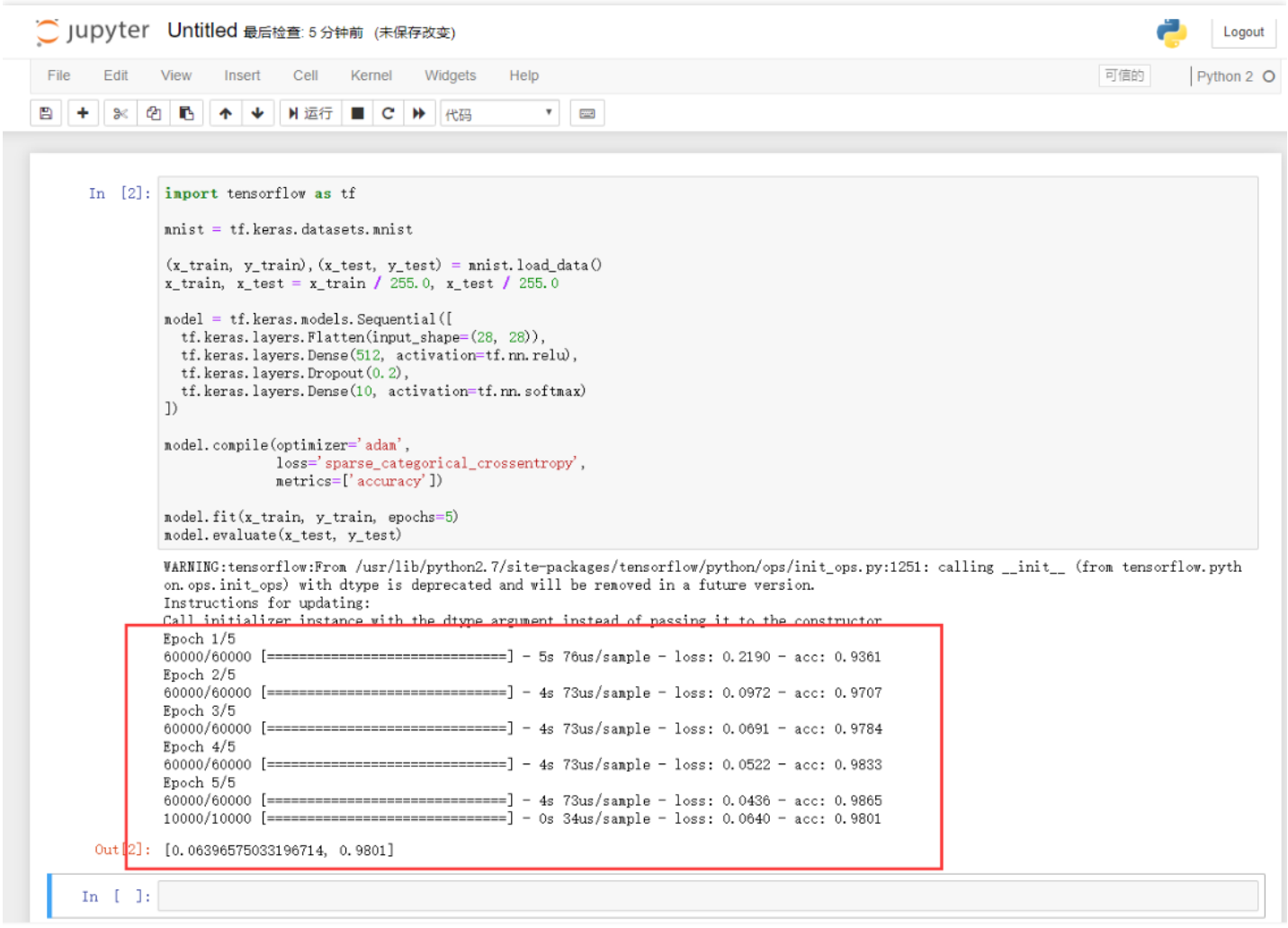
可参考 tensorflow 官网。

说明：
这里需下载数据集，国内网速会比较慢。

```
import tensorflow as tf
mnist = tf.keras.datasets.mnist
(x_train, y_train),(x_test, y_test) = mnist.load_data()
x_train, x_test = x_train / 255.0, x_test / 255.0
model = tf.keras.models.Sequential([
tf.keras.layers.Flatten(input_shape=(28, 28)),
tf.keras.layers.Dense(512, activation=tf.nn.relu),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.Dense(10, activation=tf.nn.softmax)
])
model.compile(optimizer='adam',
loss='sparse_categorical_crossentropy',
metrics=['accuracy'])
model.fit(x_train, y_train, epochs=5)
model.evaluate(x_test, y_test)
```

运行代码

重新在 jupyter 上执行模型训练。



停止 jupyter 服务

```
./stop_jupy.sh [jupyter_port]
```

Kudu 开发指南
Kudu 简介

最近更新时间：2022-12-02 14:52:30

Apache Kudu 是一个分布式，可水平扩展的列式存储系统，它完善了 Hadoop 的存储层，可对快速变化数据进行快速分析。

Kudu 基本特点

- 高效处理类 OLAP 负载。
- 与 MapReduce、Spark 以及 Hadoop 生态系统中其他组件进行友好集成。
- 可与 Impala 集成，替代目前 Impala 常用的 HDFS + Parquet 组合。
- 灵活的一致性模型。
- 顺序写和随机写并存的场景下，仍能达到良好的性能。
- 高可用，使用 Raft 协议保证数据高可靠存储。
- 结构化数据模型。

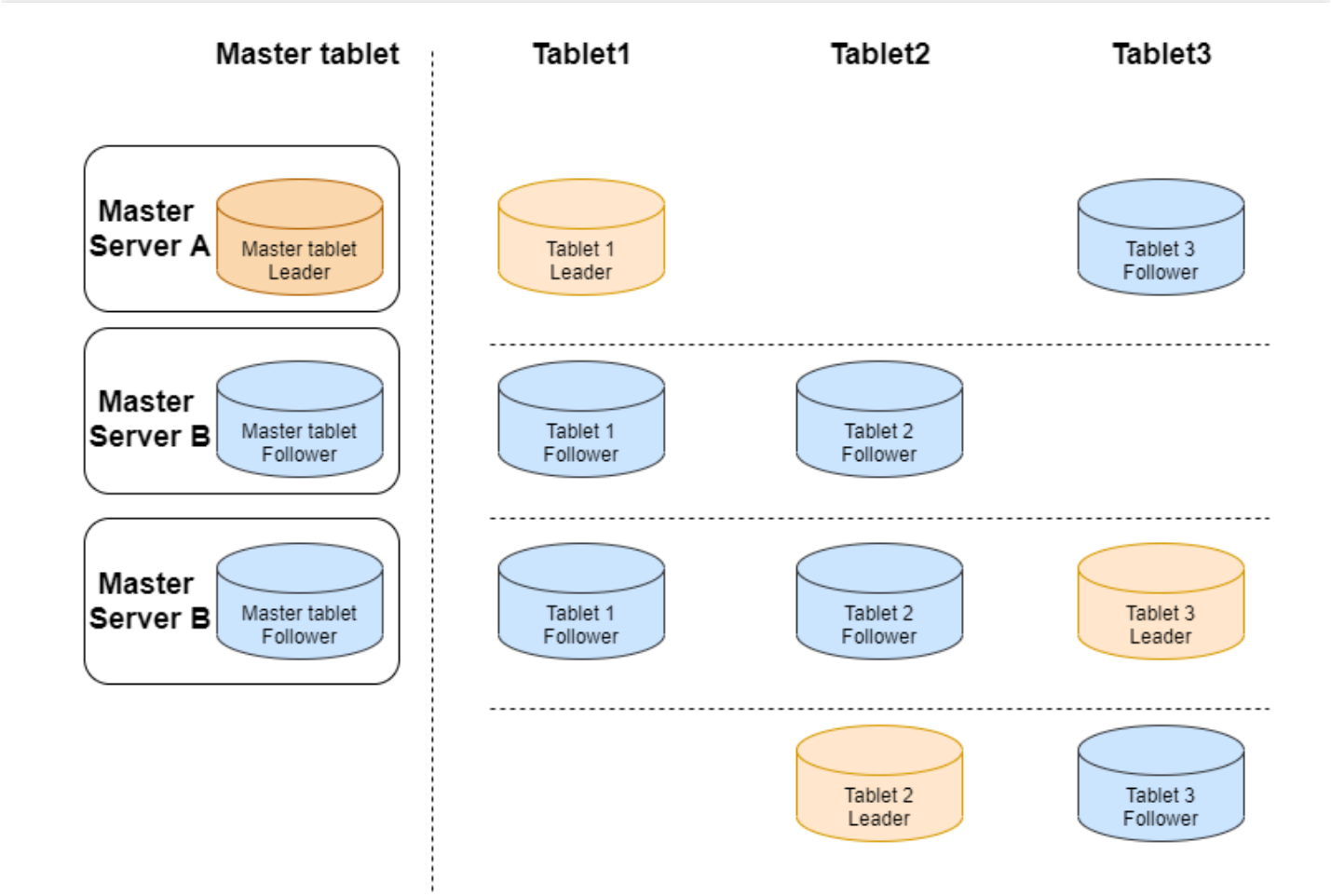
Kudu 使用场景

- 适用于那些既有随机访问，也有批量数据扫描的复合场景。
- 高计算量的场景。
- 实时预测模型的应用，支持根据所有历史数据周期地更新模型。
- 支持数据更新，避免数据反复迁移。
- 支持跨地域的实时数据备份和查询。

Kudu 基本架构

Kudu 包含如下两种类型的组件：

- master 主要负责管理元数据信息、监听 server，当 server 宕机后负责 tablet 的重分配。
- tserver 主要负责 tablet 的存储与和数据的增删改查。



Kudu 使用

EMR-2.4.0版本以上支持了 Kudu 组件。在创建 Hadoop 集群时勾选 Kudu 组件，即会创建 Kudu 集群。默认情况下 Kudu 集群包含3个 Kudu Master 服务并开启 HA。

> 说明：
> 以下所用到的 IP 为内网 IP。

- Impala 与 Kudu 集成

```
[172.30.0.98:27001] > CREATE TABLE t2(id BIGINT,name STRING,PRIMARY KEY(id))PARTI
TION BY HASH PARTITIONS 2 STORED AS KUDU TBLPROPERTIES (
'kudu.master_addresses' = '172.30.0.240,172.30.1.167,172.30.0.96,172.30.0.94,172.
```

```
30.0.214',
'kudu.num_tablet_replicas' = '1');
Query: create TABLE t2 (id BIGINT,name STRING,PRIMARY KEY(id)) PARTITION BY HASH
PARTITIONS 2 STORED AS KUDU TBLPROPERTIES (
'kudu.master_addresses' = '172.30.0.240,172.30.1.167,172.30.0.96,172.30.0.94,172.
30.0.214',
'kudu.num_tablet_replicas' = '1')
Fetched 0 row(s) in 0.12s
[hadoop@172 root]$ /usr/local/service/kudu/bin/kudu table list 172.30.0.240,172.3
0.1.167,172.30.0.96,172.30.0.94,172.30.0.214
impala::default.t2
```

- 数据插入

```
[172.30.0.98:27001] > insert into t2 values(1, 'test');
Query: insert into t2 values(1, 'test')
Query submitted at: 2020-08-10 20:07:21 (Coordinator: http://172.30.0.98:27004)
Query progress can be monitored at: http://172.30.0.98:27004/query_plan?query_id=
b44fe203ce01254d:b055e98200000000
Modified 1 row(s), 0 row error(s) in 5.63s
```

- 基于 Impala 查询数据

```
[172.30.0.98:27001] > select * from t2;
Query: select * from t2
Query submitted at: 2020-08-10 20:09:47 (Coordinator: http://172.30.0.98:27004)
Query progress can be monitored at: http://172.30.0.98:27004/query_plan?query_id=
ec4c9706368f135d:f20ccb6e00000000
+----+------+
| id | name |
+----+------+
| 1 | test |
+----+------+
Fetched 1 row(s) in 0.20s
```

- 其他命令

  i. 集群健康检测

```
[hadoop@172 root]$ /usr/local/service/kudu/bin/kudu cluster ksck 172.30.0.240,17
2.30.1.167,172.30.0.96,172.30.0.94,172.30.0.214
```

ii. 创建表

```
[hadoop@172 root]$ /usr/local/**service**/**kudu**/bin/kudu table create '172.30.0.240,172.30.1.167,172.30.0.96,172.30.0.94,172.30.0.214' '{"table_name":"test","schema":{"columns":[{"column_name":"id","column_type":"INT32","default_value":"1"},{"column_name":"key","column_type":"INT64","is_nullable":false,"comment":"range key"},{"column_name":"name","column_type":"STRING","is_nullable":false,"comment":"user name"}],"key_column_names":["id","key"]},"partition":{"hash_partitions":[{"columns":["id"],"num_buckets":2,"seed":100}],"range_partition":{"columns":["key"],"range_bounds":[{"upper_bound":{"bound_type":"inclusive","bound_values":["2"]}},{"lower_bound":{"bound_type":"exclusive","bound_values":["2"]},"upper_bound":{"bound_type":"inclusive","bound_values":["3"]}}]}},"extra_configs":{"configs":{"kudu.table.history_max_age_sec":"3600"}},"num_replicas":1}'
```

iii. 查询创建的 test 表

```
[hadoop@172 root]$ /usr/local/service/kudu/bin/kudu table list 172.30.0.240,172.30.1.167,172.30.0.96,172.30.0.94,172.30.0.214
test
```

iv. 查看表结构

```
[hadoop@172 root]$ /usr/local/service/kudu/bin/kudu table describe 172.30.0.240,172.30.1.167,172.30.0.96,172.30.0.94,172.30.0.214 test
TABLE test (
id INT32 NOT NULL,
key INT64 NOT NULL,
name STRING NOT NULL,
PRIMARY KEY (id, key)
)
HASH (id) PARTITIONS 2 SEED 100,
RANGE (key) (
PARTITION VALUES < 3,
PARTITION 3 <= VALUES < 4
)
REPLICAS 1
```

Kudu 节点缩容数据搬迁指南

最近更新时间：2023-02-01 15:51:18

本文档主要对 Kudu 集群类型 Core 节点 tserver 缩容时，需进行的数据搬迁进行介绍。

> 说明：
> Kudu 集群类型支持 Core 节点缩容，该功能未默认开放，如有需要请您 提交工单 申请开通。

在进行 tserver 节点下线前，可以使用 rebalance tool 做数据迁移。请注意，一次只能下线一台 tserver，如果下线多台，需重复执行下述步骤。

Kudu 基于 rebalance tool 迁移

1. 确保集群状态 ok。

```
/usr/local/service/kudu/bin/kudu cluster ksck 10.0.1.29:7051,10.0.1.16:7051,10.0.1.36:7051
```

```
==================
Warnings:
==================
Some masters have unsafe, experimental, or hidden flags set
Some tablet servers have unsafe, experimental, or hidden flags set

OK
[hadoop@10 bin]$
```

2. 使用步骤1的 ksck 命令，获取下线的节点 uid。

```
Tablet Server Summary
                UUID                |     Address     | Status  | Location | Tablet Leaders | Active Scanners
------------------------------------+-----------------+---------+----------+----------------+-----------------
 20681b1d6b9942cbab95dded905406ec   | 10.0.1.37:7050  | HEALTHY | <none>   |       9        |        0
 6929daf14f8647c89fb8cc51db5d70b6   | 10.0.1.15:7050  | HEALTHY | <none>   |       5        |        0
 b53b28bfad2c41d38d6f08a261ceb486   | 10.0.1.40:7050  | HEALTHY | <none>   |       2        |        0
 be018287364d4443a48ad1bba248c87f   | 10.0.1.9:7050   | HEALTHY | <none>   |       0        |        0
 fb9afb1b2989456cac5800bf6990dfea   | 10.0.1.45:7050  | HEALTHY | <none>   |       0        |        0
```

以 fb9afb1b2989456cac5800bf6990dfea 节点为例子。

3. 将 fb9afb1b2989456cac5800bf6990dfea 节点进入维护模式。

```
/usr/local/service/kudu/bin/kudu tserver state enter_maintenance 10.0.1.29:7051,10.0.1.16:7051,10.0.1.36:7051 fb9afb1b2989456cac5800bf6990dfea
```

4. 执行 rebalance 命令

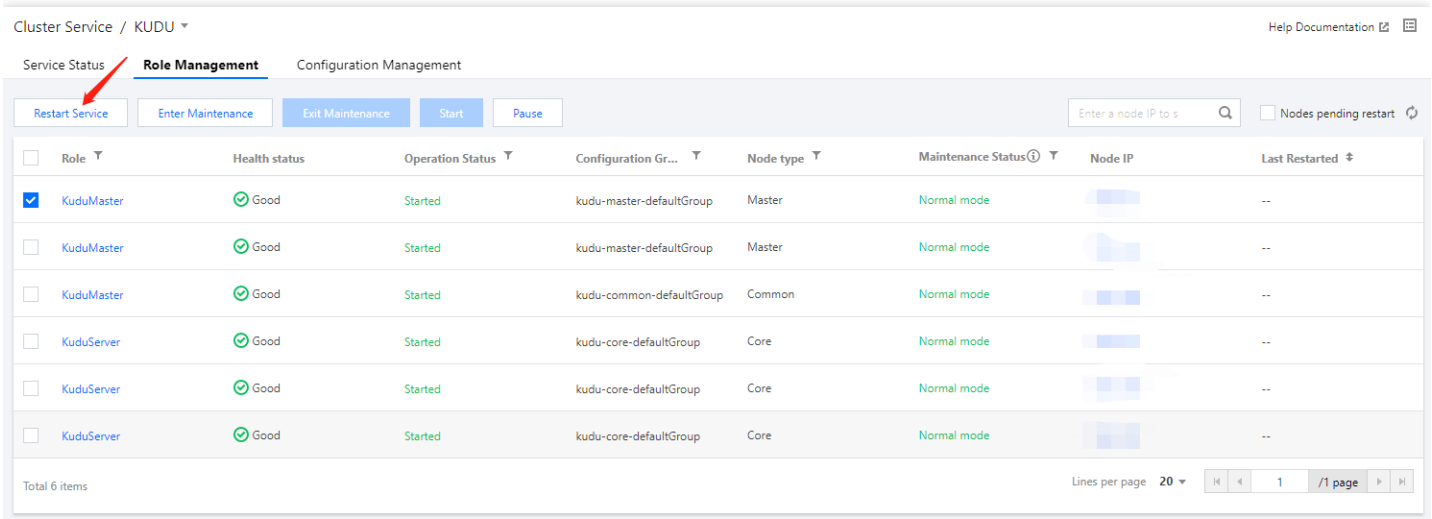
```
/usr/local/service/kudu/bin/kudu cluster rebalance 10.0.1.29:7051,10.0.1.16:7051,10.0.1.36:7051 --ignored_tservers fb9afb1b2989456cac5800bf6990dfea --move_replicas_from_ignored_tservers
```

等待命令执行结束，再次用 ksck 检查，状态为 ok，继续后面步骤。

5. 暂停 fb9afb1b2989456cac5800bf6990dfea 对应节点10.0.1.45的 tserver 进程。注意此时，使用ksck命令，集群状态不健康，需要重启 tmaster。

```
Tablet Server Summary
              UUID              |     Address     |   Status    | Location | Tablet Leaders | Active Scanners
--------------------------------+-----------------+-------------+----------+----------------+-----------------
 20681b1d6b9942cbab95dded905406ec | 10.0.1.37:7050 | HEALTHY     | <none>   |       9        |        0
 6929daf14f8647c89fb8cc51db5d70b6 | 10.0.1.15:7050 | HEALTHY     | <none>   |       5        |        0
 b53b28bfad2c41d38d6f08a261ceb486 | 10.0.1.40:7050 | HEALTHY     | <none>   |       2        |        0
 be018287364d4443a48ad1bba248c87f | 10.0.1.9:7050  | HEALTHY     | <none>   |       0        |        0
 fb9afb1b2989456cac5800bf6990dfea | 10.0.1.45:7050 | UNAVAILABLE | <none>   | n/a            | n/a
```

6. 在 EMR 控制台 重启 master。注意需要手动一台一台的重启（不建议使用控制台的滚动重启）。重启结束后，使用ksck命令，确保集群状态健康。



Ranger 开发指南
Ranger 简介

最近更新时间：2023-07-12 10:35:46

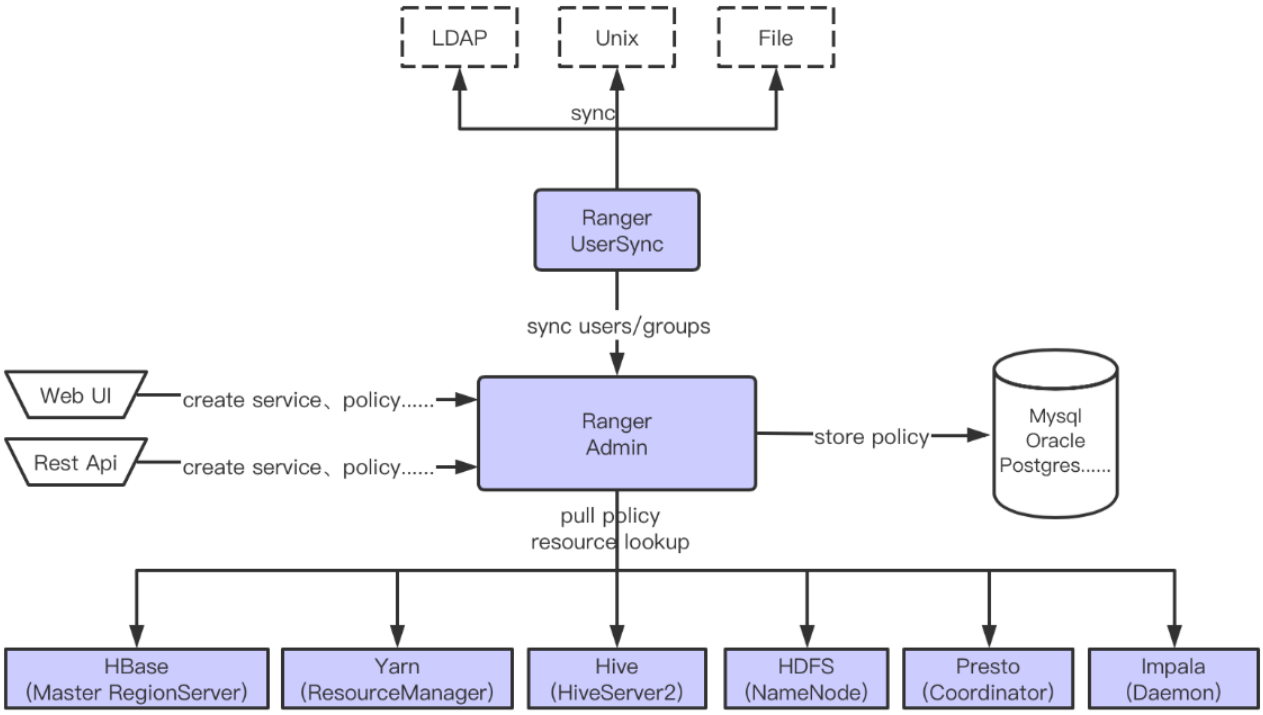
Ranger 简介

Ranger 是大数据领域的一个集中式安全管理框架，实现对 Hadoop 生态组件的集中式安全管理。用户可以通过 Ranger 实现对集群中数据的安全访问，它主要是对 Hadoop 平台组件进行监管、启动服务以及资源访问进行控制，其核心思想为：

- 用户可以使用 Ranger 提供的 REST API 或者使用 Ranger 提供 Web UI 对大数据组件进行集中化管理。
- 可以针对大数据组件进行基于角色、属性进行授权。
- 针对大数据组件所涉及安全的审计进行集中管理。

Ranger 架构

Ranger 主要是由 Ranger Admin、Ranger UserSync、Ranger Plugin 三个组件构成的，其中 Ranger Admin、Ranger UserSync 都是一个单独的 JVM 进程，而 Ranger Plugin 需要根据不同组件安装在不同节点上。



- Ranger Admin：管理用户配置好的策略及创建的服务、审计日志及 Report、将配置好的策略及创建的服务持久化到数据库中并提供给 Plugin 定期查询。
- Ranger UserSync：将 LDAP、File、Unix 的相关信息同步到 Ranger Admin 中，例如，将用户的 LDAP 目录访问系统或者 Unix 中用户的信息及组信息进行同步，同步 Unix 用户及组信息需要开启 unixAuthenticationService 进程，同时对同步过来的信息进行持久化。
- Ranger Plugin：Plugin 会被部署在需要的服务节点中，并会定期到 Ranger Admin 同步策略信息。

组件集成 Ranger 请参照如下表格：

Service	install Node	EMR 版本
HDFS	NameNode	EMR-V 2.0.1 及以上版本
Hbase	Master、RegionServer	EMR-V 2.0.1 及以上版本
Hive	HiveServer2	EMR-V 2.0.1 及以上版本
Yarn	ResourceManager	EMR-V 2.0.1 及以上版本
Presto	All Coordinator	EMR-V 2.0.1 及以上版本
Impala	All Daemon	EMR-V 2.2.0 及以上版本

Service	install Node	EMR 版本
Kudu	All Master	EMR-V 3.2.0 版本

Ranger 使用指南
HDFS 集成 Ranger

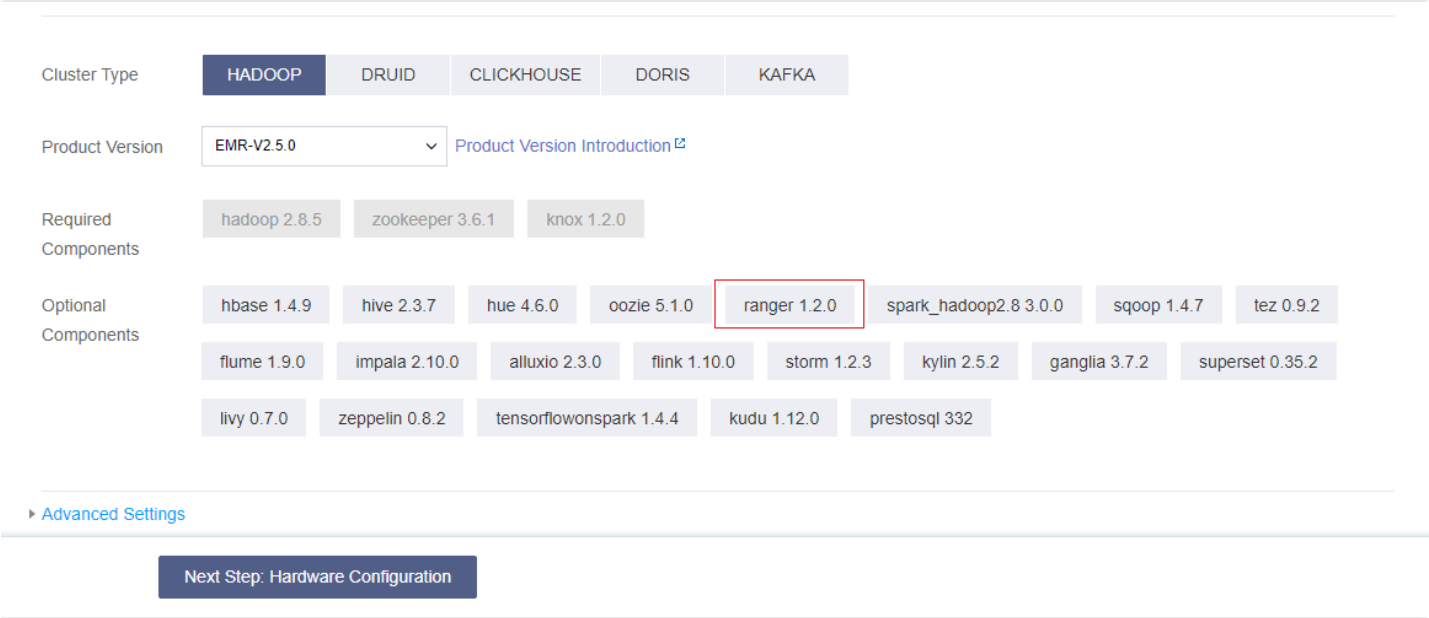
最近更新时间：2023-06-19 16:56:39

使用准备

仅支持在购买集群时选择了可选组件 Ranger 的集群，若是在已创建的集群上新增 Ranger 组件，可能会出现 Web UI 无法访问的情况。默认 Ranger 安装时，Ranger Admin、Ranger UserSync 都是部署在 Master 节点上，Ranger Plugin 是部署嵌入组件主守护进程节点上。

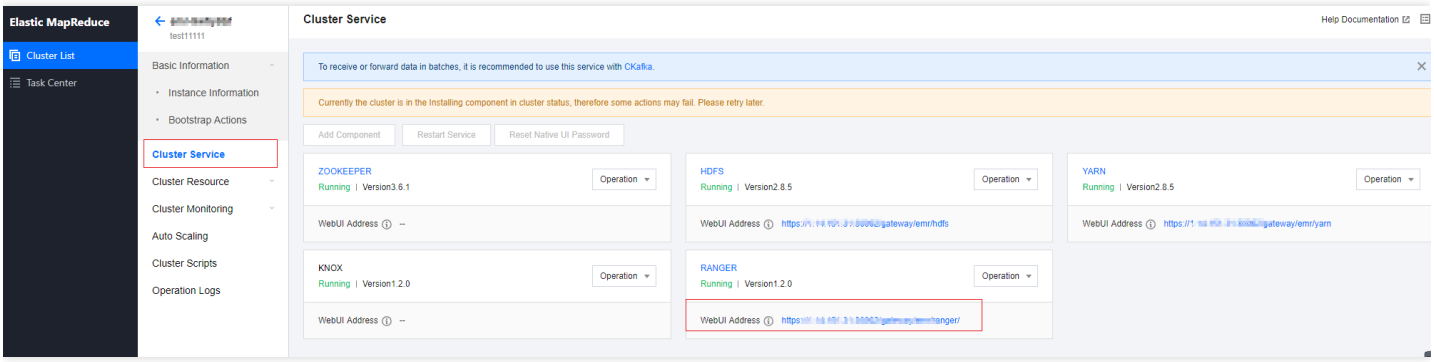
创建集群时，在选择集群类型为 Hadoop 时可以在可选组件中选择 Ranger，Ranger 的版本根据您选择的 EMR 版本不同而存在差异。

> 说明：
> 集群类型为 Hadoop 且选择了可选组件 Ranger 时，EMR-Ranger 默认会为 HDFS、YARN 创建服务并设置默认策略。

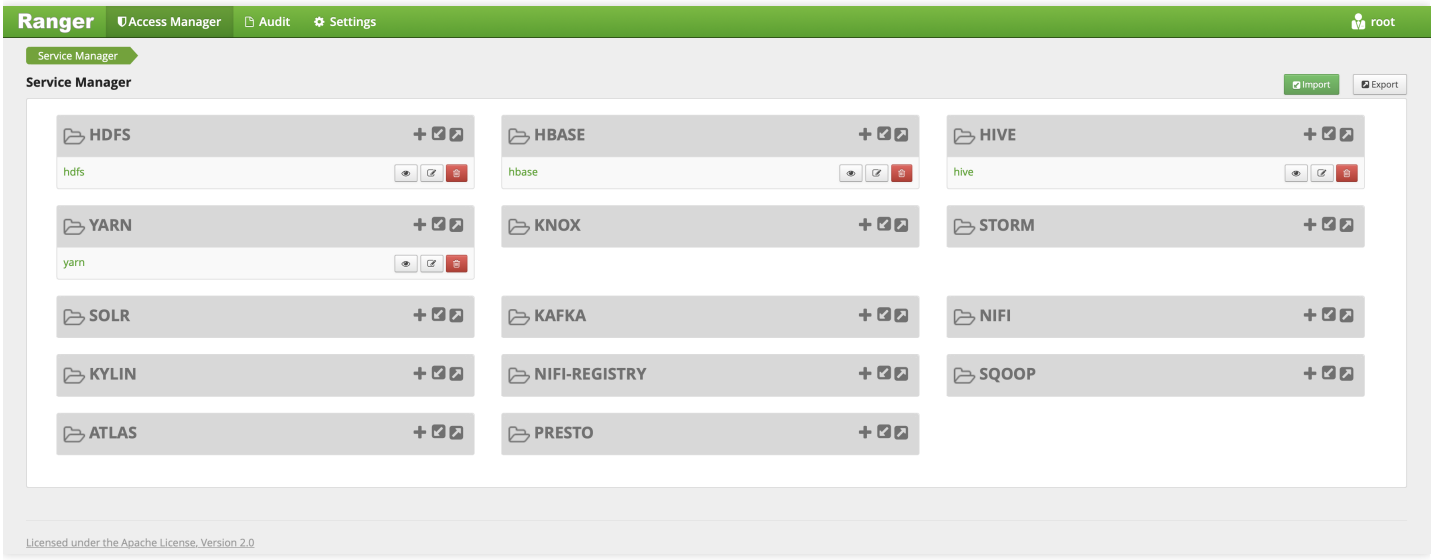


Ranger Web UI

在访问 Ranger Web UI 之前，请务必确认当前所购买的集群是否配置了公网 IP，然后在集群服务中单击 Ranger 组件的 Web UI 地址链接。



Web UI 地址链接跳转后，会提示输入用户名及密码，即在购买集群时设置的用户名及密码。

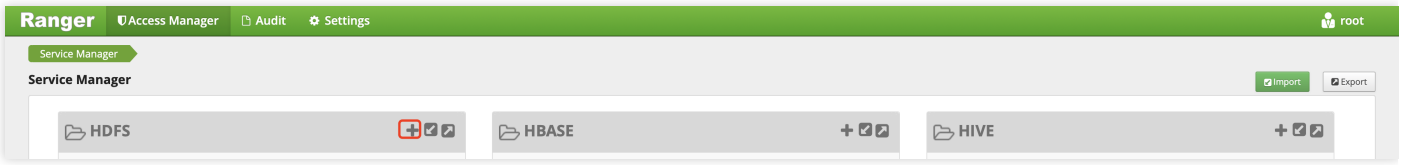


HDFS 集成 Ranger

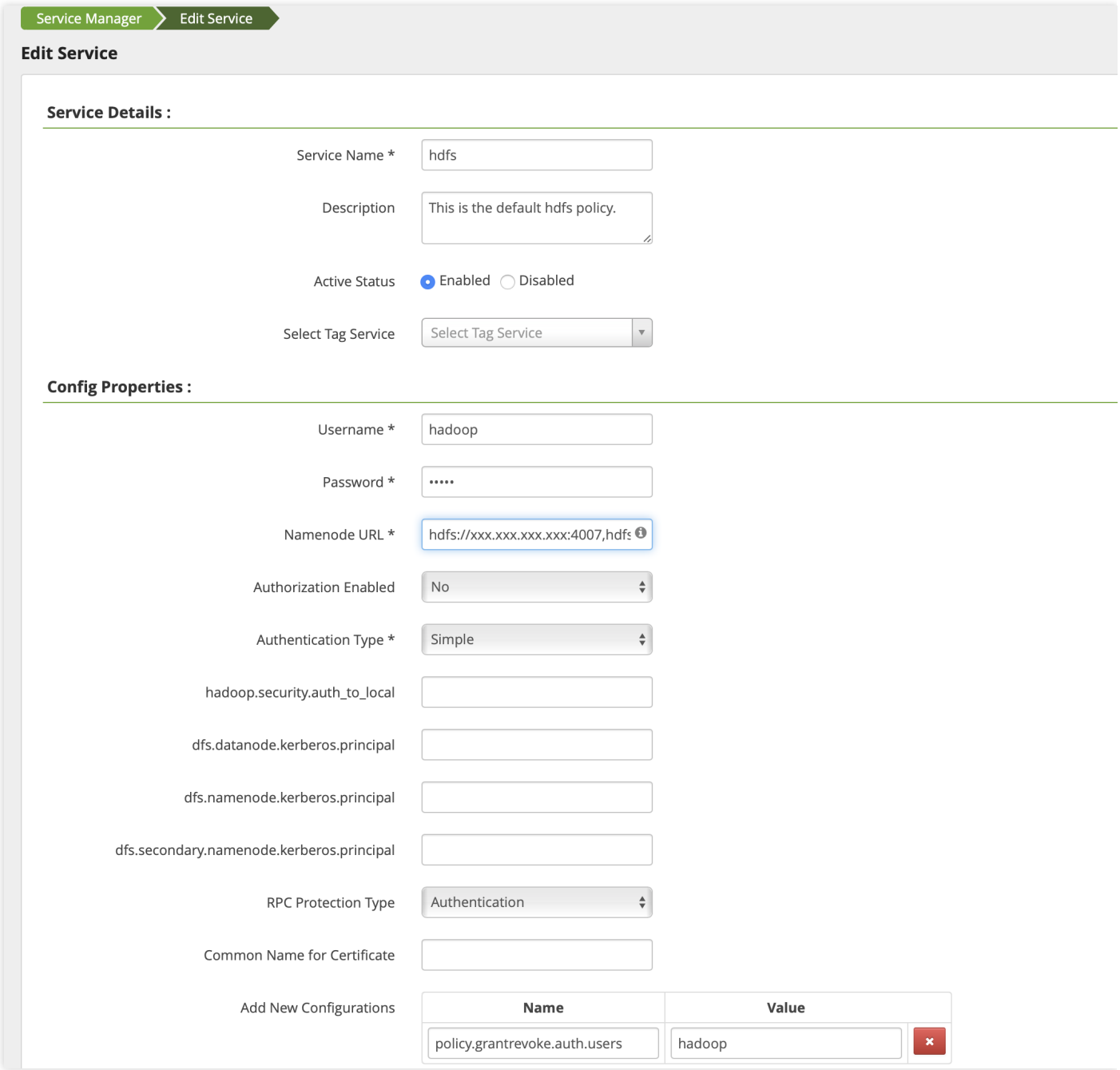
注意：

请确保 HDFS 相关服务运行正常并且当前集群已安装 Ranger。

1. 使用 EMR Ranger Web UI 页面添加 EMR Ranger HDFS 服务。



2. 配置 EMR Ranger HDFS Service 相关参数。



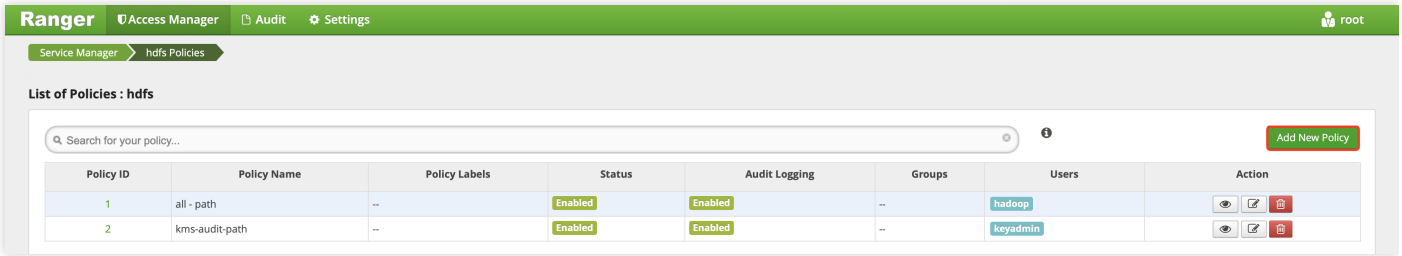
参数名	是否必填项	解释

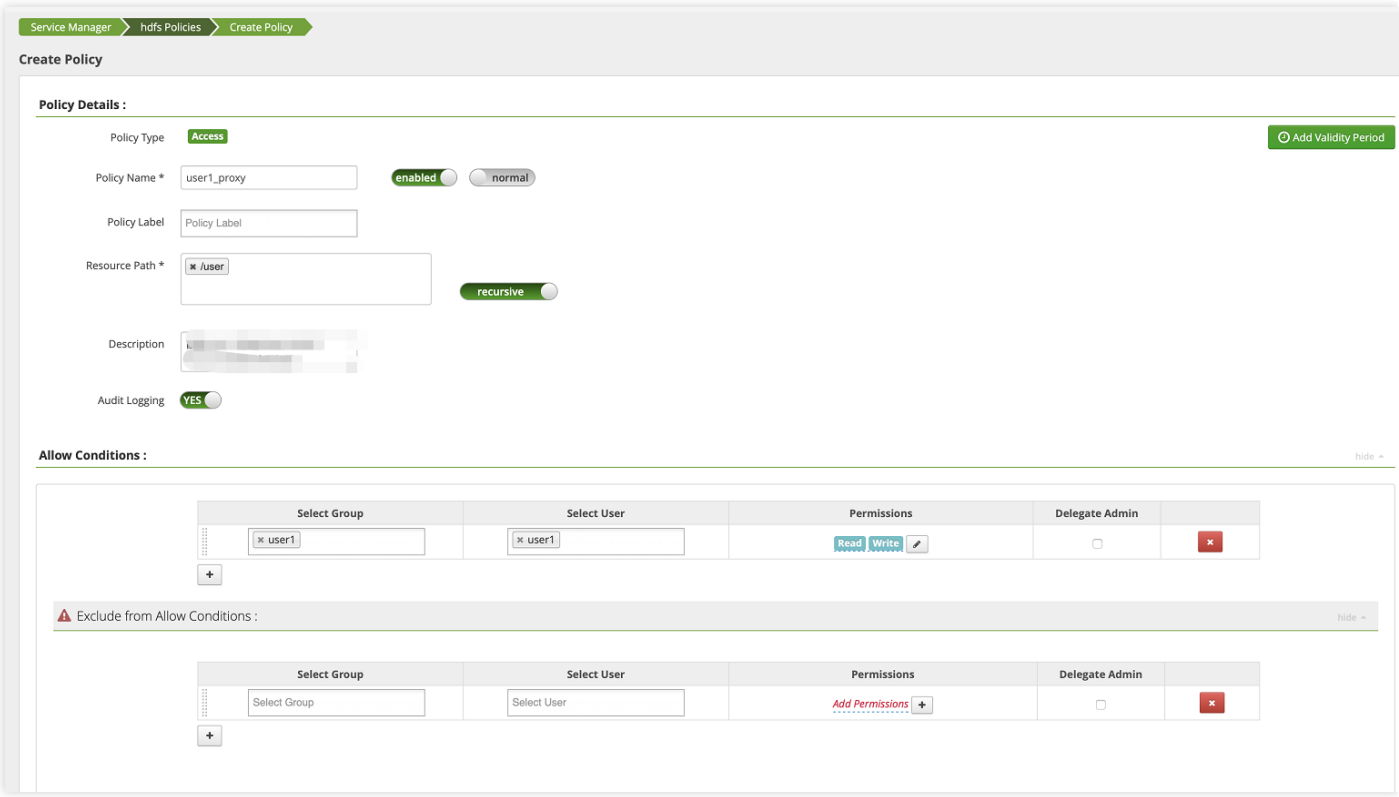
参数名	是否必填项	解释
Service Name	是	服务名称，Ranger Web UI 主 HDFS 组件显示服务名称
Description	否	服务描述信息
Active Status	默认	服务启用状态，默认启用
UserName	是	资源使用的用户名
Password	是	用户密码
NameNode URL	是	HDFS 地址
Authorization Enable	默认	标准集群选择 No；高安全集群选择 Yes
Authorization Type	是	Simple：标准集群；Kerberos：高安全集群

3. EMR Ranger HDFS 资源权限配置。

- 单击配置好的 EMR Ranger HDFS Service

- 配置 Policy





4. 添加完 Policy 后，稍等约半分钟等待 Policy 生效。生效后使用 user1 就可以对 HDFS 文件系统的 /user 进行读写操作。

YARN 集成 Ranger

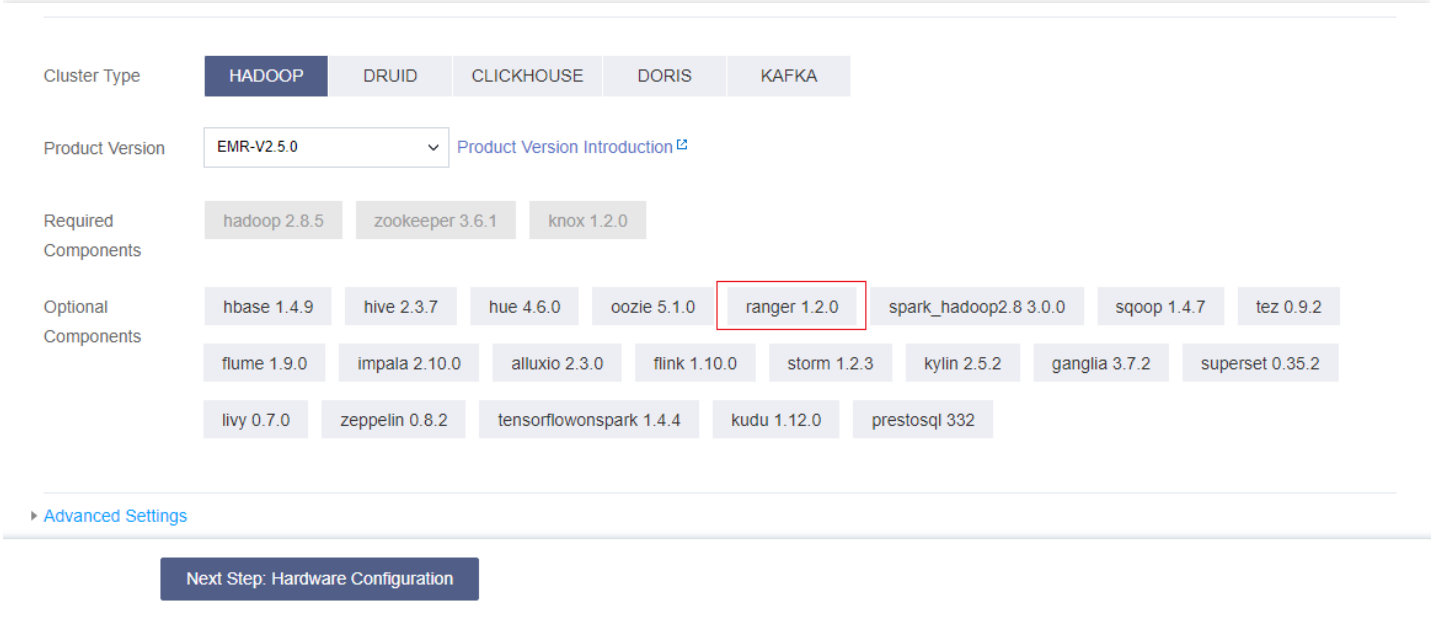
最近更新时间：2023-06-19 16:35:44

使用准备

仅支持在购买集群时选择了可选组件 Ranger 的集群，若是在已创建的集群上新增 Ranger 组件，可能会出现 Web UI 无法访问的情况。默认 Ranger 安装时，Ranger Admin、Ranger UserSync 都是部署在 Master 节点上，Ranger Plugin 是部署嵌入组件主守护进程节点上。

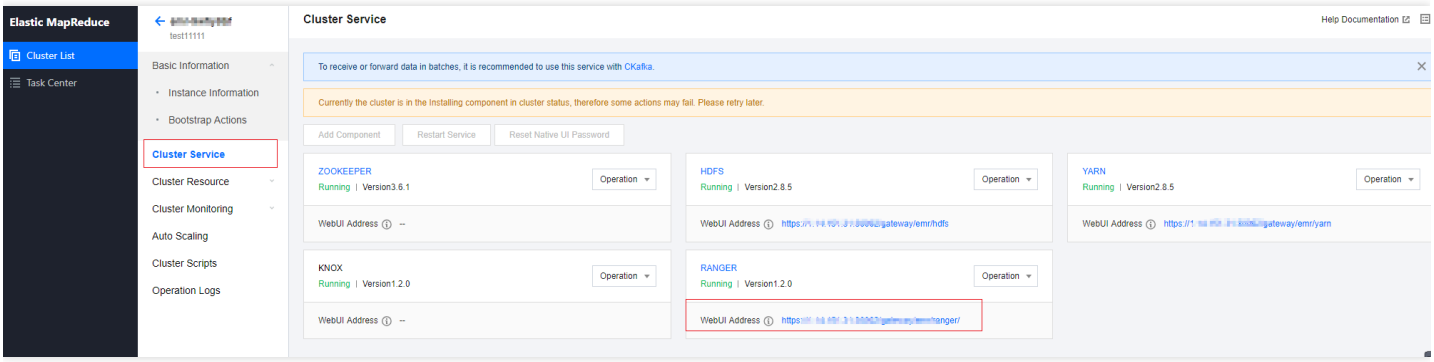
创建集群时，在选择集群类型为 Hadoop 时可以在可选组件中选择 Ranger，Ranger 的版本根据您选择的 EMR 版本不同而存在差异。

> 说明：
> 集群类型为 Hadoop 且选择了可选组件 Ranger 时，EMR-Ranger 默认会为 HDFS、YARN 创建服务并设置默认策略。

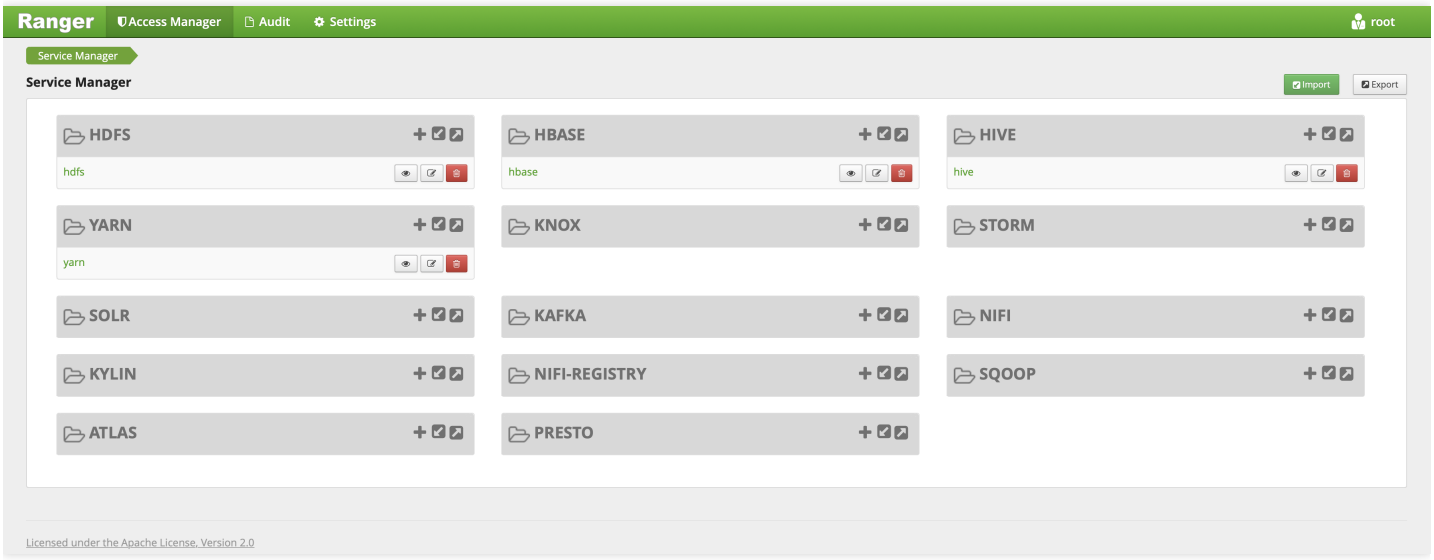


Ranger Web UI

在访问 Ranger Web UI 之前，请务必确认当前所购买的集群是否配置了公网 IP，然后在集群服务中单击 Ranger 组件的 Web UI 地址链接。



Web UI 地址链接跳转后，会提示输入用户名及密码，即在购买集群时设置的用户名及密码。



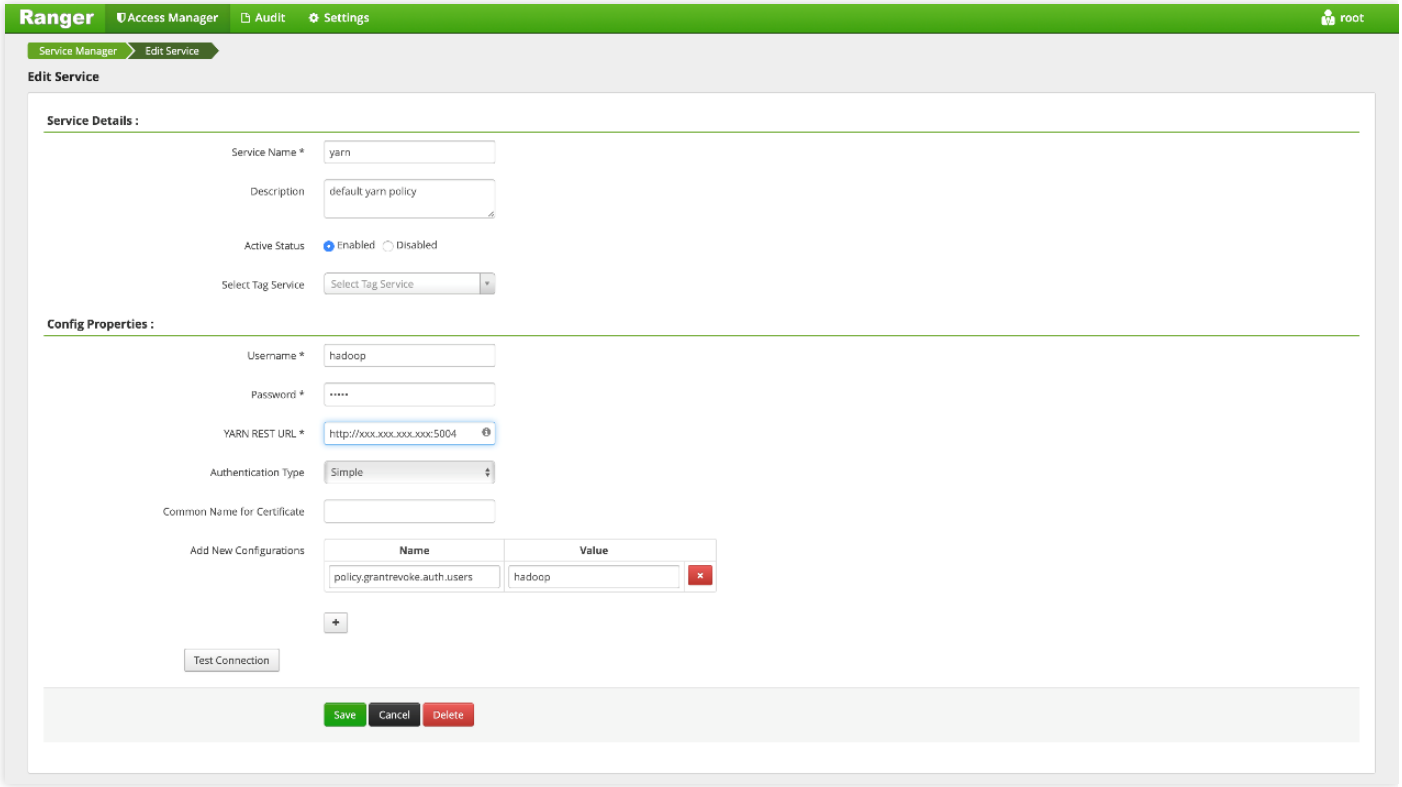
YARN 集成 Ranger

> 注意：
> 请确保 YARN 相关服务运行正常并且当前集群已安装 Ranger。

EMR Ranger YARN 目前仅支持 Capacity Scheduler 队列的 ACL，不支持 Fair Scheduler 队列的 ACL。Ranger YARN 队列 ACL 与 YARN 自带的 Capacity Scheduler 配置共同生效，且优先级低于 Capacity Scheduler 配置，只有在 YARN 自带的 Capacity Scheduler 配置拒绝校验时才会校验 Ranger YARN 权限。**建议不要在配置文件设置 ACL，而是使用 Ranger 设置 ACL**。

1. 使用 EMR Ranger Web UI 页面添加 EMR Ranger YARN 服务。

2. 配置 EMR Ranger YARN Service 相关参数。



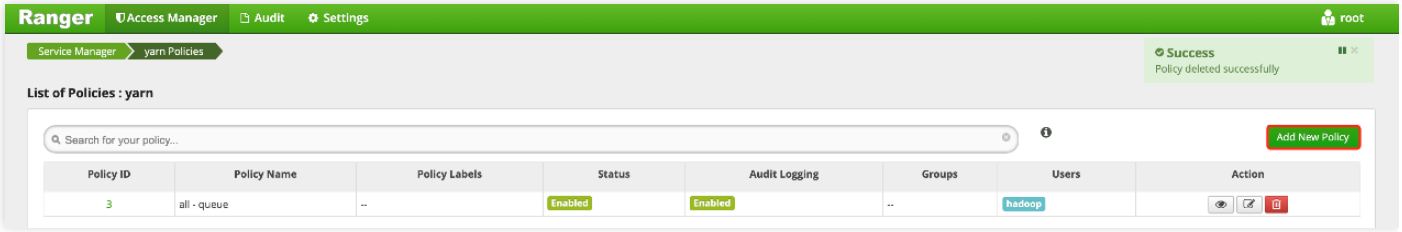
参数名	是否必填项	解释
Service Name	是	服务名称，Ranger Web UI 主 YARN 组件显示服务名称
Description	否	服务描述信息
Active Status	默认	服务启用状态，默认启用
UserName	是	资源使用的用户名
Password	是	用户密码
NameNode URL	是	YARN 地址

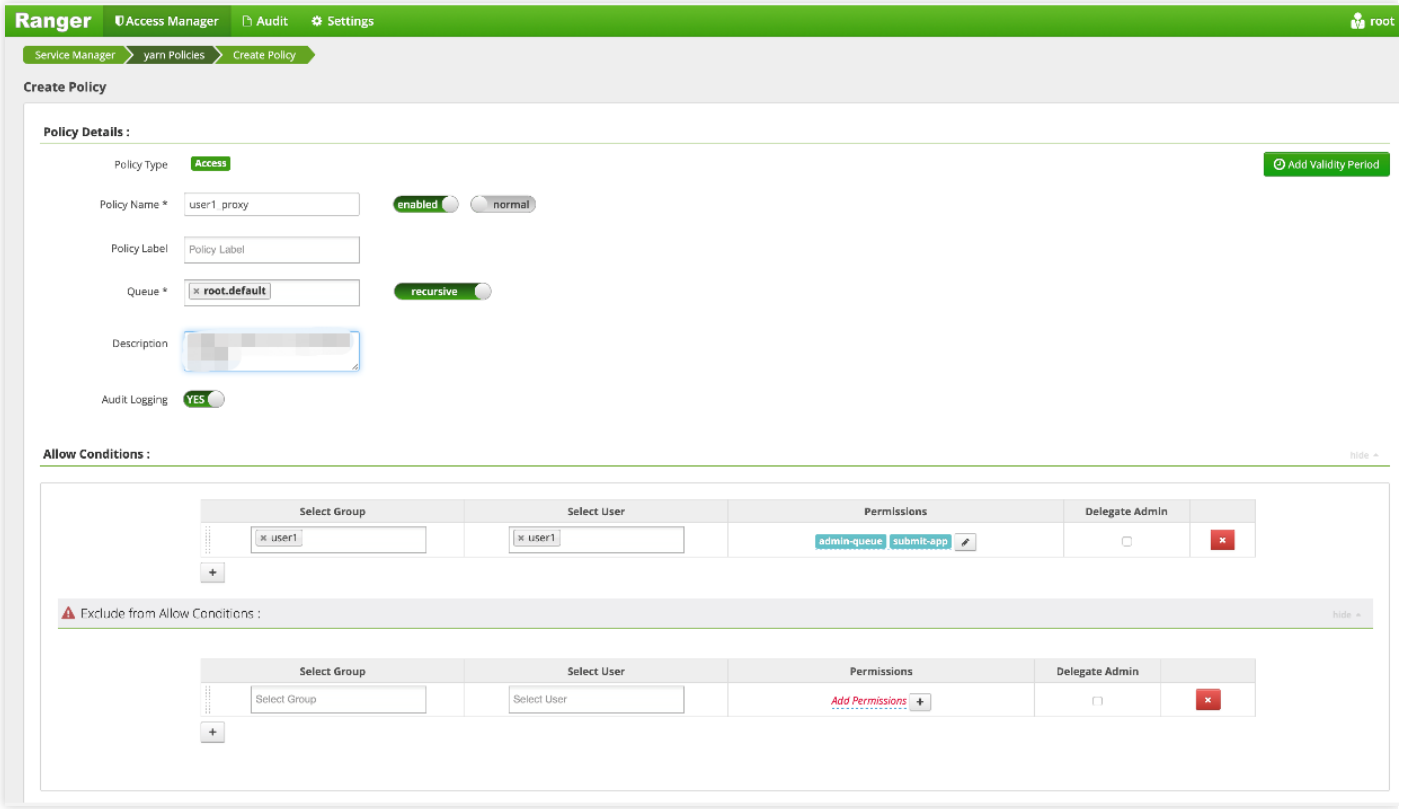
参数名	是否必填项	解释
Authorization Enable	默认	标准集群选择 No；高安全集群选择 Yes
Authorization Type	是	Simple：标准集群；Kerberos：高安全集群

3. EMR Ranger YARN 资源权限配置。

- 单击配置好的 EMR Ranger HDFS Service

- 配置 Policy





4. 添加完 Policy 后，稍等约半分钟等待 Policy 生效。生效后使用 user1 就可以向 YARN 的 root.default 队列中提交、删除、查询作业等操作。

> 注意：
> 在配置 Ranger YARN Service 以及 Policy 时请务必确保期间没有 YARN 作业，否则会出现某些用户作业提交权限问题。

Hbase 集成 Ranger

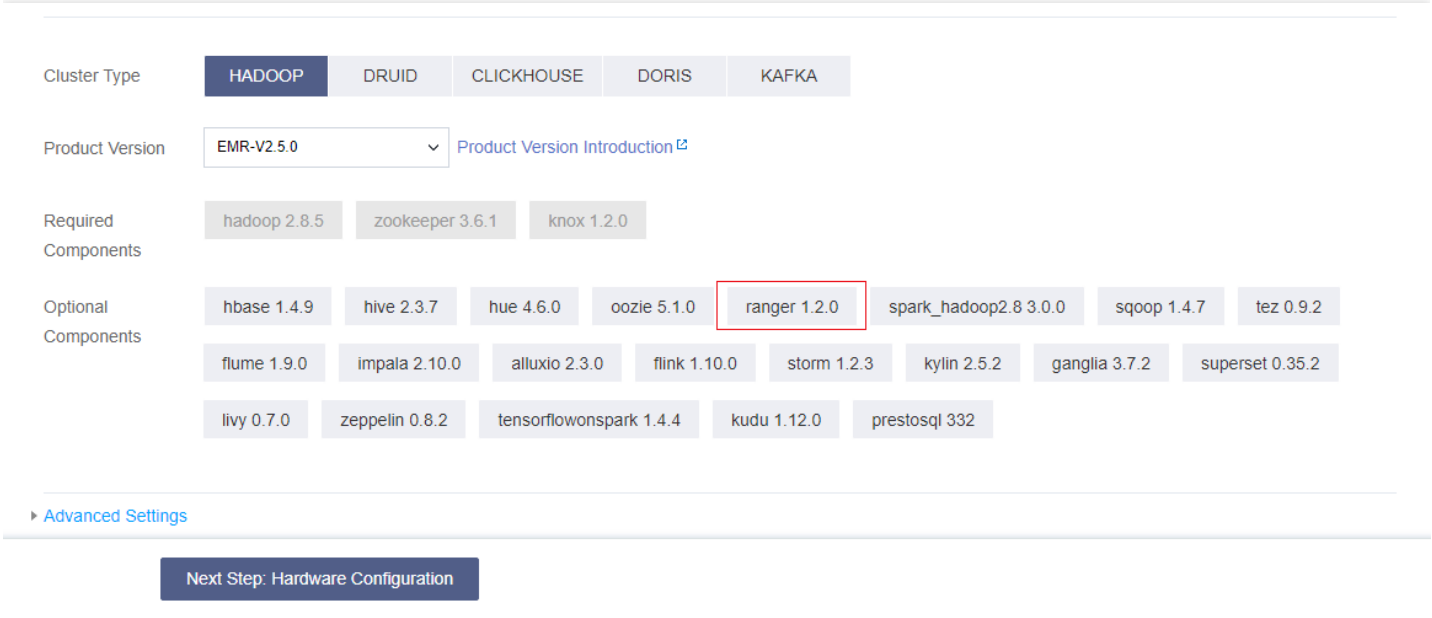
最近更新时间：2023-06-19 16:33:05

使用准备

仅支持在购买集群时选择了可选组件 Ranger 的集群，若是在已创建的集群上新增 Ranger 组件，可能会出现 Web UI 无法访问的情况。默认 Ranger 安装时，Ranger Admin、Ranger UserSync 都是部署在 Master 节点上，Ranger Plugin 是部署嵌入组件主守护进程节点上。

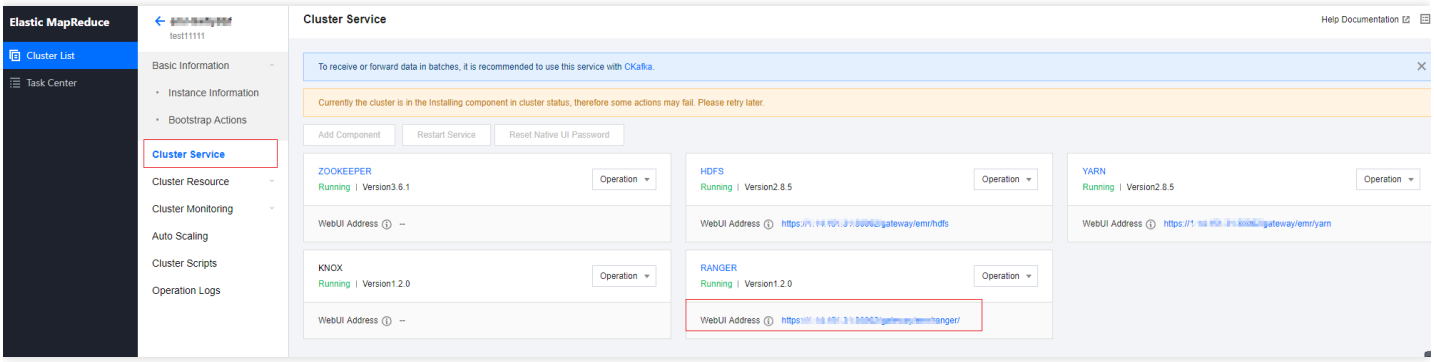
创建集群时，在选择集群类型为 Hadoop 时可以在可选组件中选择 Ranger，Ranger 的版本根据您选择的 EMR 版本不同而存在差异。

> 说明：
> 集群类型为 Hadoop 且选择了可选组件 Ranger 时，EMR-Ranger 默认会为 HDFS、YARN 创建服务并设置默认策略。

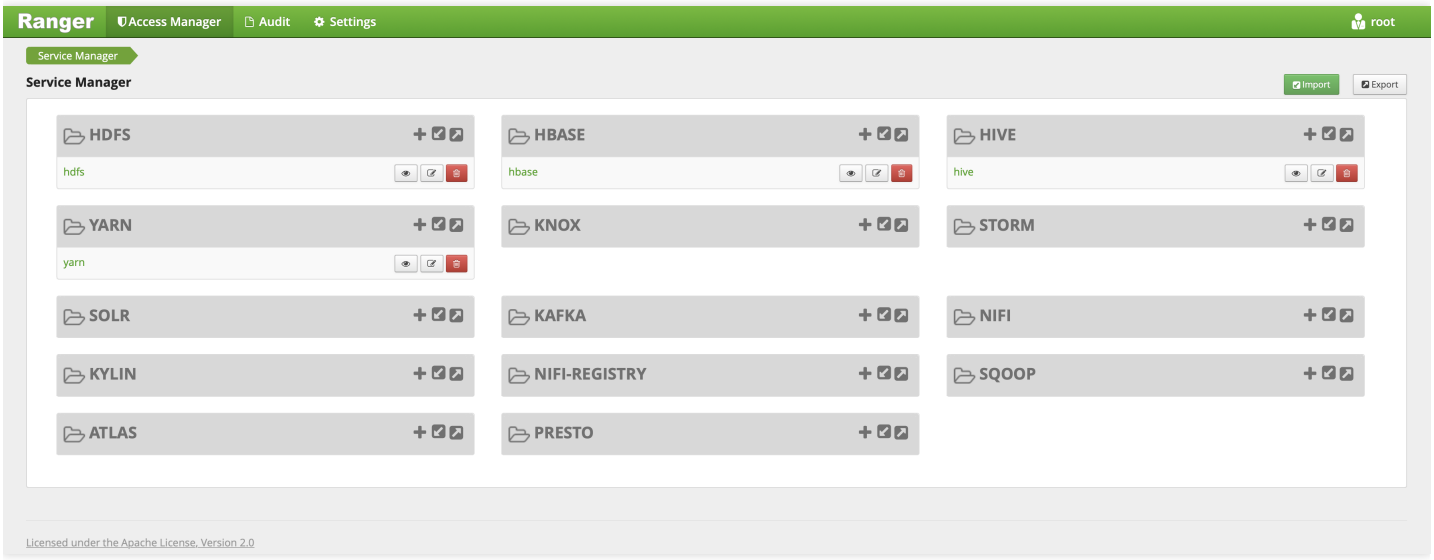


Ranger Web UI

在访问 Ranger Web UI 之前，请务必确认当前所购买的集群是否配置了公网 IP，然后在集群服务中单击 Ranger 组件的 Web UI 地址链接。



Web UI 地址链接跳转后，会提示输入用户名及密码，即在购买集群时设置的用户名及密码。

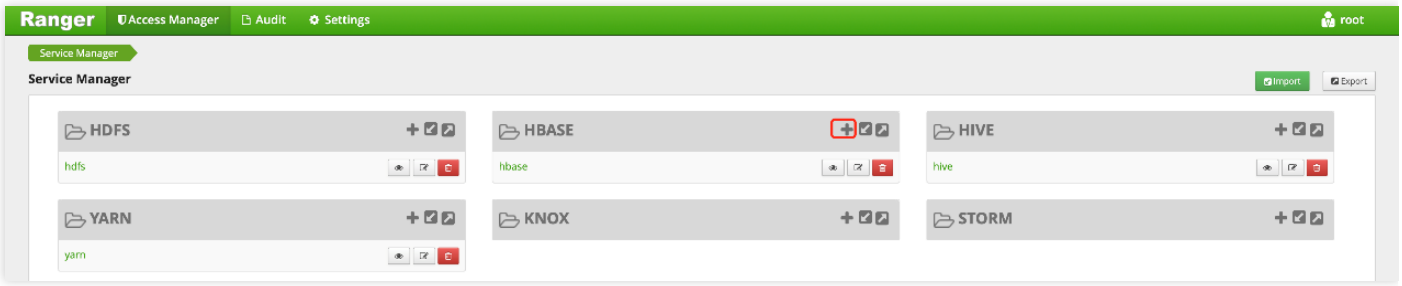


Hbase 集成 Ranger

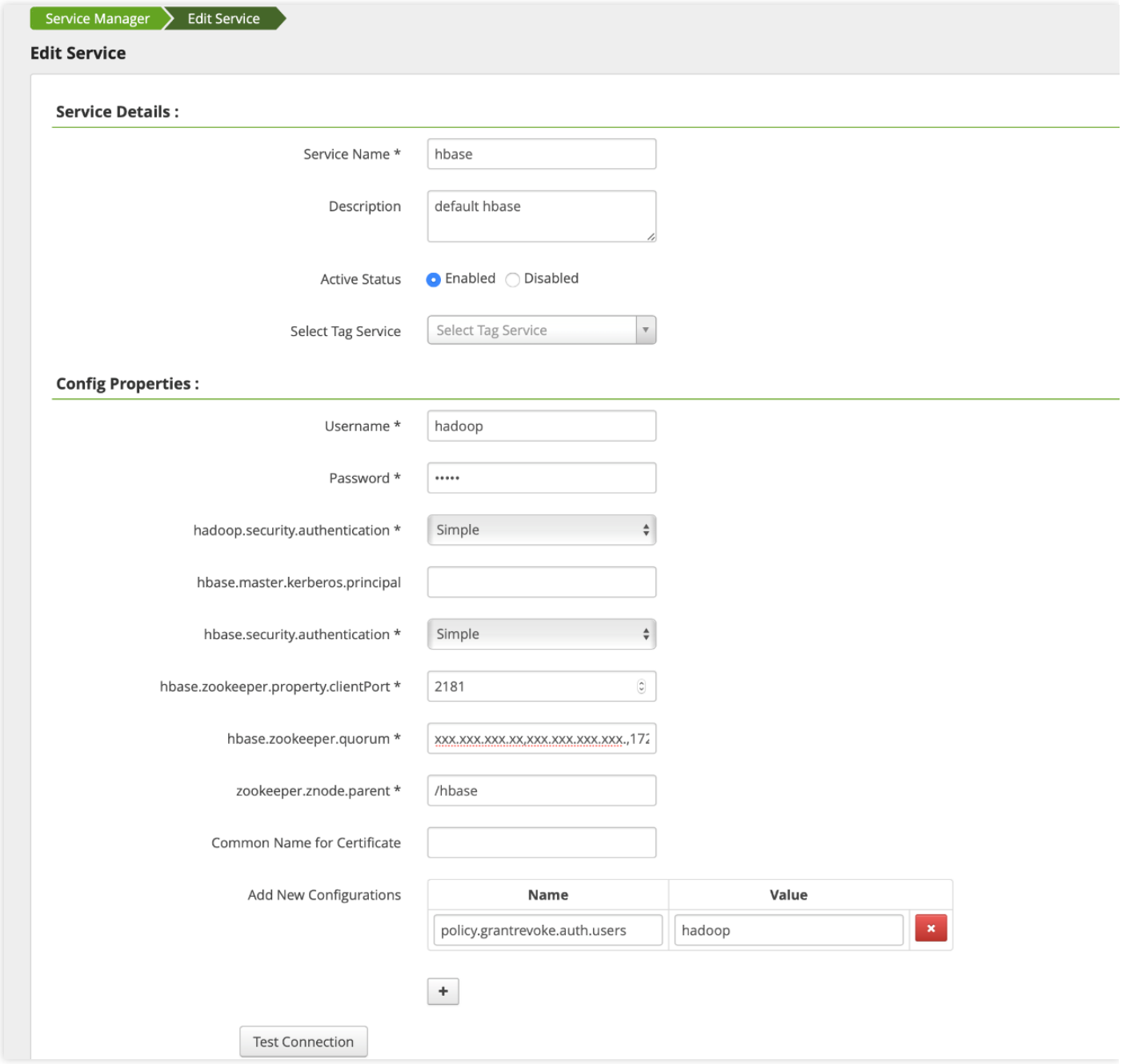
注意：

请确保 HBase 相关服务运行正常并且当前集群已安装 Ranger。

1. 使用 EMR Ranger Web UI 页面添加 EMR Ranger Hbase 服务。



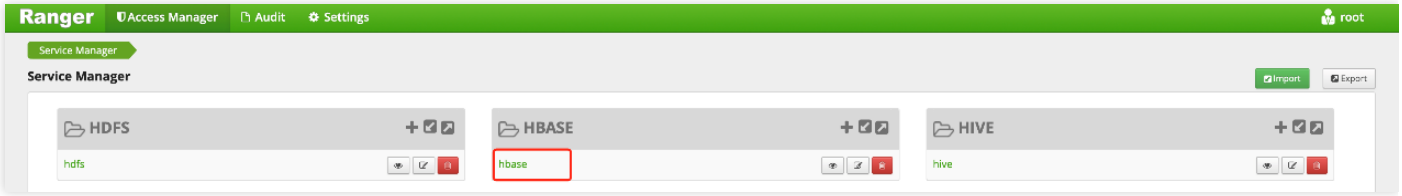
2. 配置 EMR Ranger Hbase Service 相关参数。



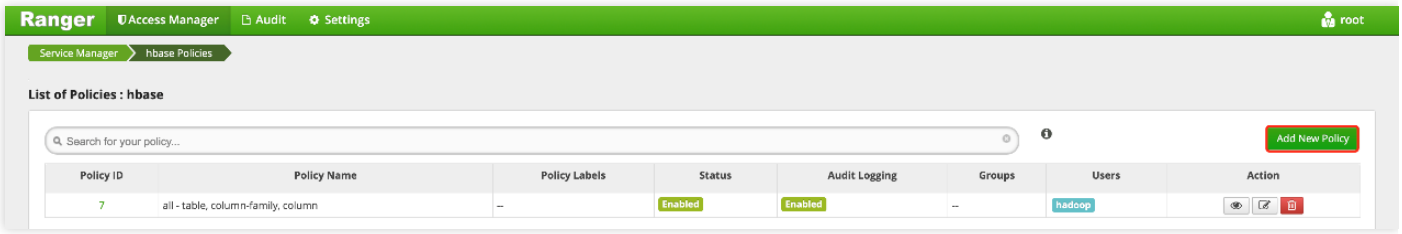
参数名	是否必填项	解释
Service Name	是	服务名称，Ranger Web UI 主 HBase 组件显示服务名称
Description	否	服务描述信息
Active Status	默认	服务启用状态，默认启用
UserName	是	资源使用的用户名
Password	是	用户密码
Hbase.zookeeper.property.clientPort	是	ZK 客户端请求端口
Hbase.zookeeper.quorum	是	ZK 集群 IP
Zookeeper.znode.parent	是	ZK 节点信息

3. EMR Ranger Hbase 资源权限配置。

- 单击配置好的 EMR Ranger Hbase Service

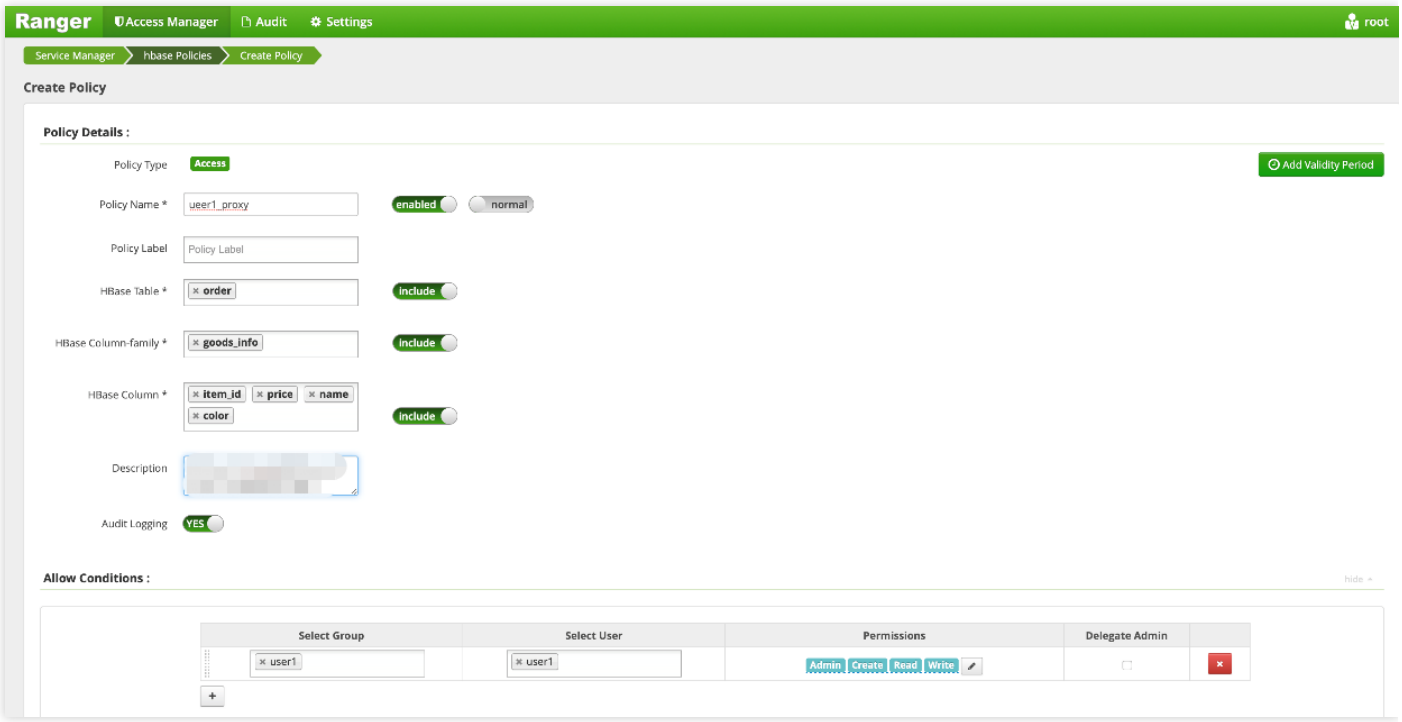


- 配置 Policy



上图中的 Users 为 Hbase，它的 Policy Name 是 all-table、column-family、cloumn，也就是 Hbase 用户具有

Region Balance、MemeStore Fluh、Compaction、Split 权限。**请确保创建的 Service 是有些这些权限的。**



参数	是否必选项	解释
HBase Table	是	HBase 表名
HBase Column-family	是	HBase 表中的列簇
HBase Column	是	HBase 表中的列簇下的限定符

4. 添加完 Policy 后，稍等约半分钟等待 Policy 生效。生效后使用 user1 就可以对 Hbase 的 order 表相关列簇和限定符进行相关操作。

Presto 集成 Ranger

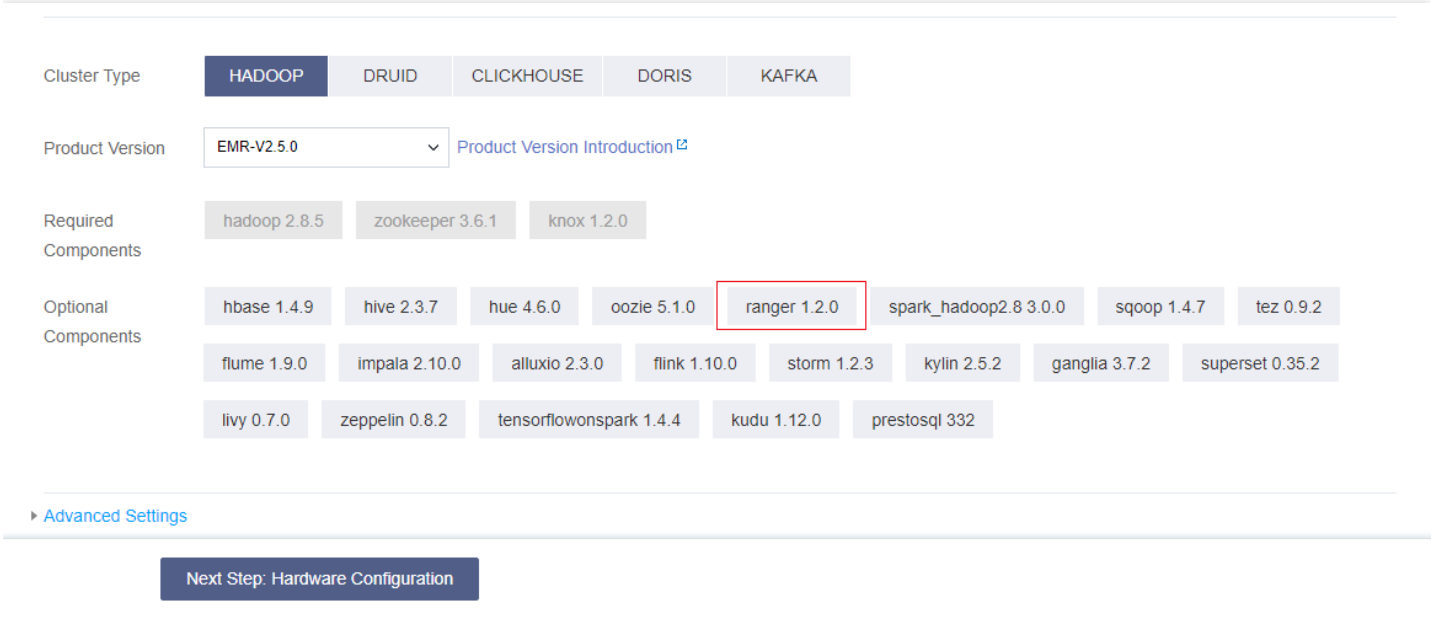
最近更新时间：2023-06-19 16:27:08

使用准备

仅支持在购买集群时选择了可选组件 Ranger 的集群，若是在已创建的集群上新增 Ranger 组件，可能会出现 Web UI 无法访问的情况。默认 Ranger 安装时，Ranger Admin、Ranger UserSync 都是部署在 Master 节点上，Ranger Plugin 是部署嵌入组件主守护进程节点上。

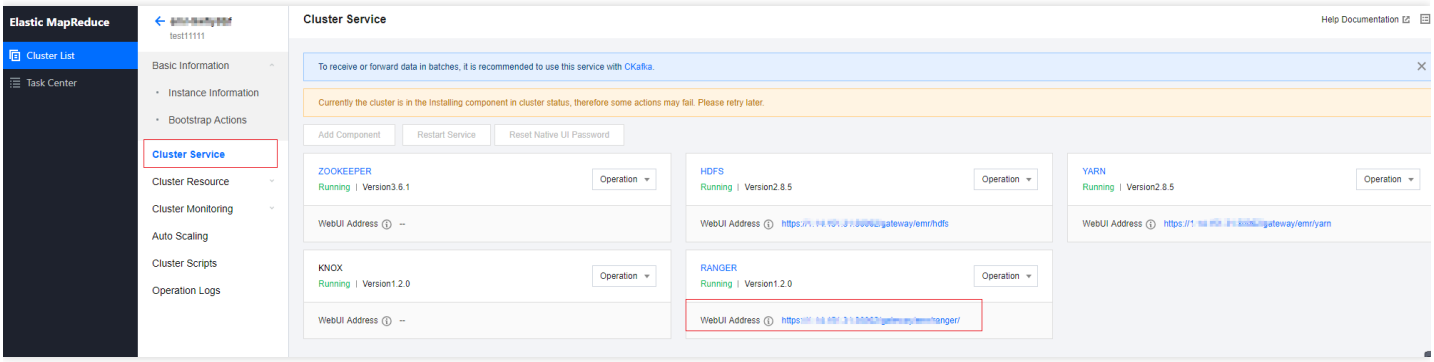
创建集群时，在选择集群类型为 Hadoop 时可以在可选组件中选择 Ranger，Ranger 的版本根据您选择的 EMR 版本不同而存在差异。

> 说明：
> 集群类型为 Hadoop 且选择了可选组件 Ranger 时，EMR-Ranger 默认会为 HDFS、YARN 创建服务并设置默认策略。

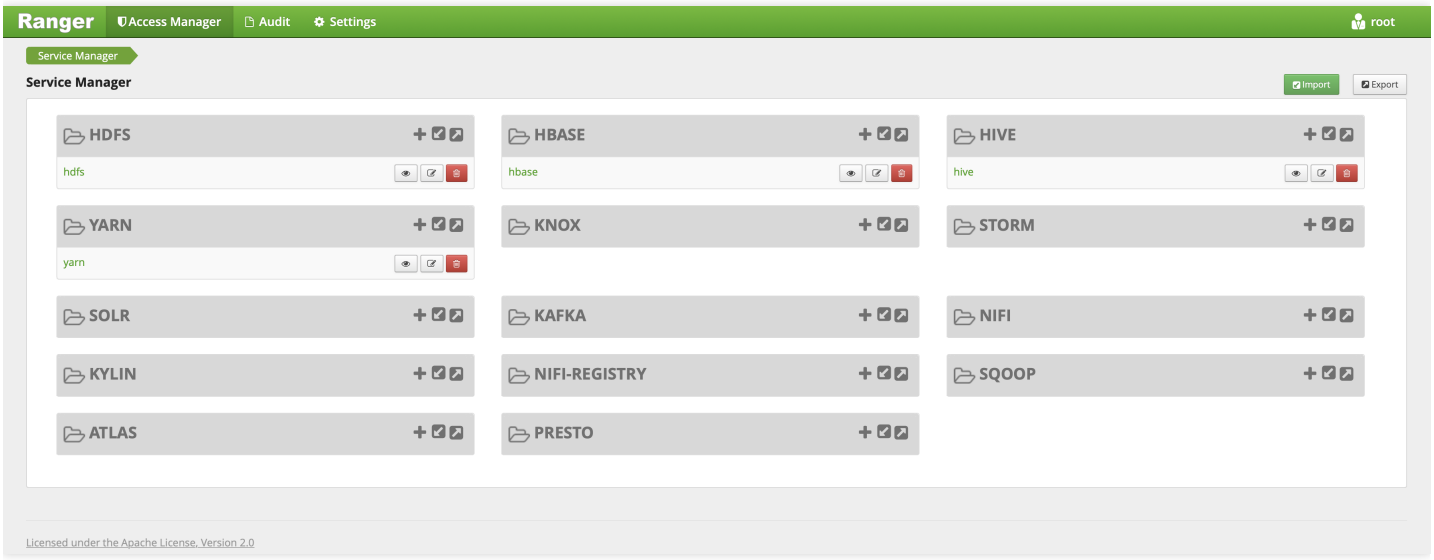


Ranger Web UI

在访问 Ranger Web UI 之前，请务必确认当前所购买的集群是否配置了公网 IP，然后在集群服务中单击 Ranger 组件的 Web UI 地址链接。



Web UI 地址链接跳转后，会提示输入用户名及密码，即在购买集群时设置的用户名及密码。

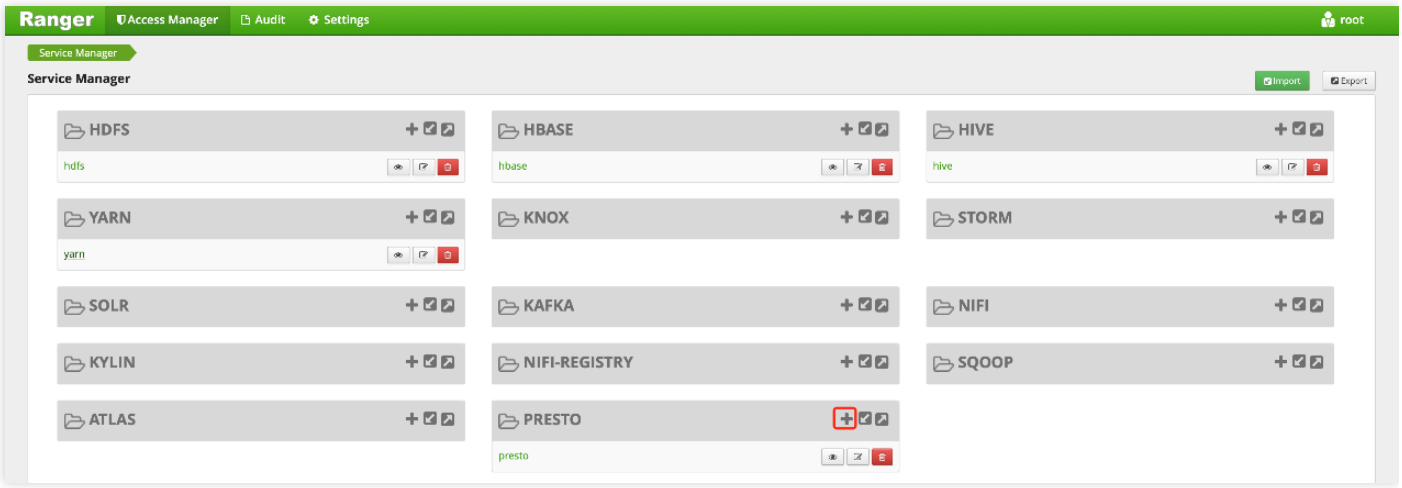


Presto 集成 Ranger

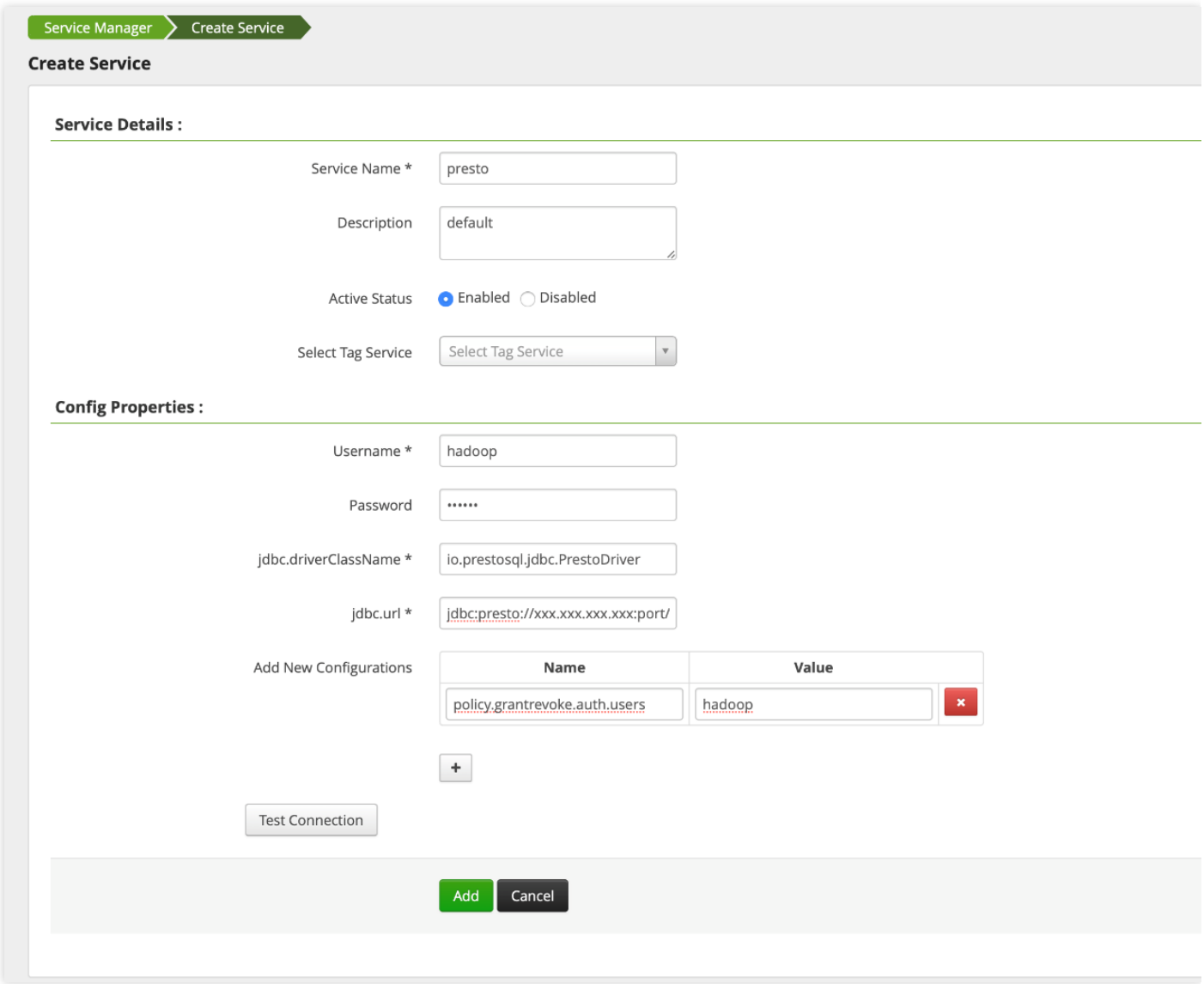
注意：

请确保 Presto 相关服务运行正常并且当前集群已安装 Ranger。

1. 使用 EMR Ranger Web UI 页面添加 EMR Ranger Presto 服务。



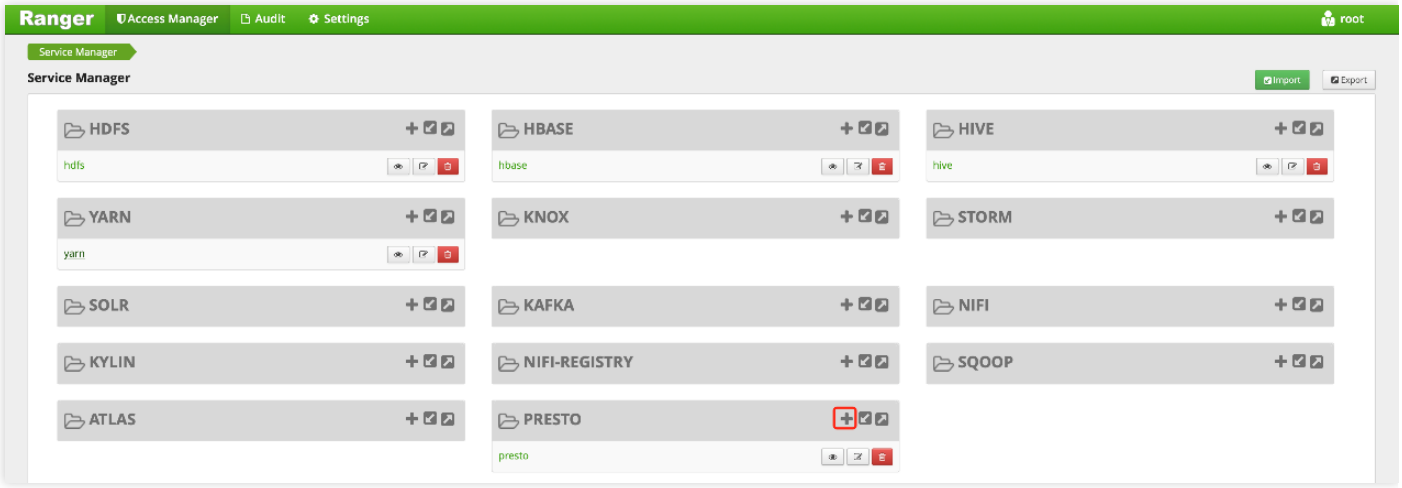
2. 配置 EMR Ranger Presto Service 相关参数。



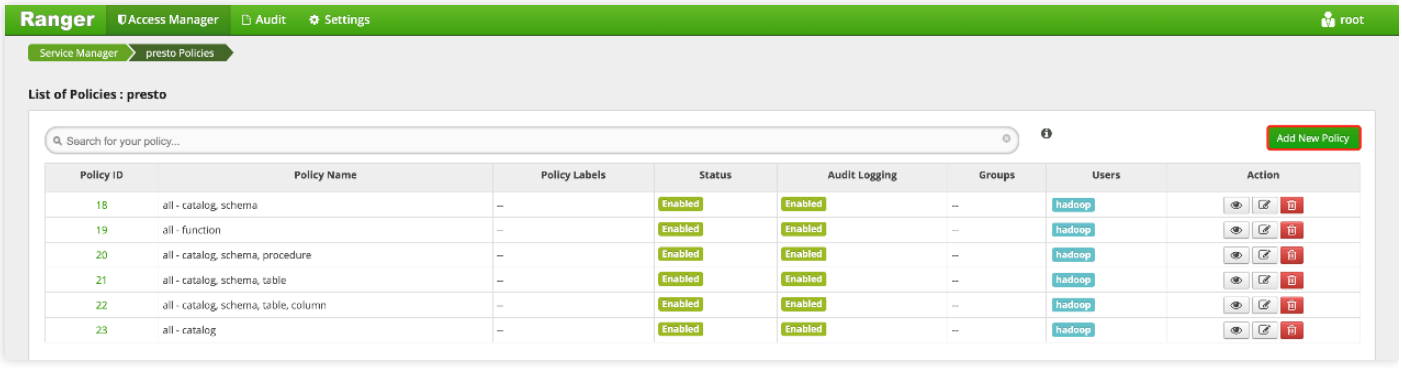
参数	是否必填项	解释
Service Name	是	服务名称，Ranger Web UI 主 Predo 组件显示服务名称
Description	否	服务描述信息
Active Status	默认	服务启用状态，默认启用
UserName	是	资源使用的用户名
Password	是	用户密码
JDBC.driverClassName	是	驱动的类名全路径
jdbc.url	是	Presto jdbc 连接形式的地址，例如 jdbc:presto://ip/hostname:port

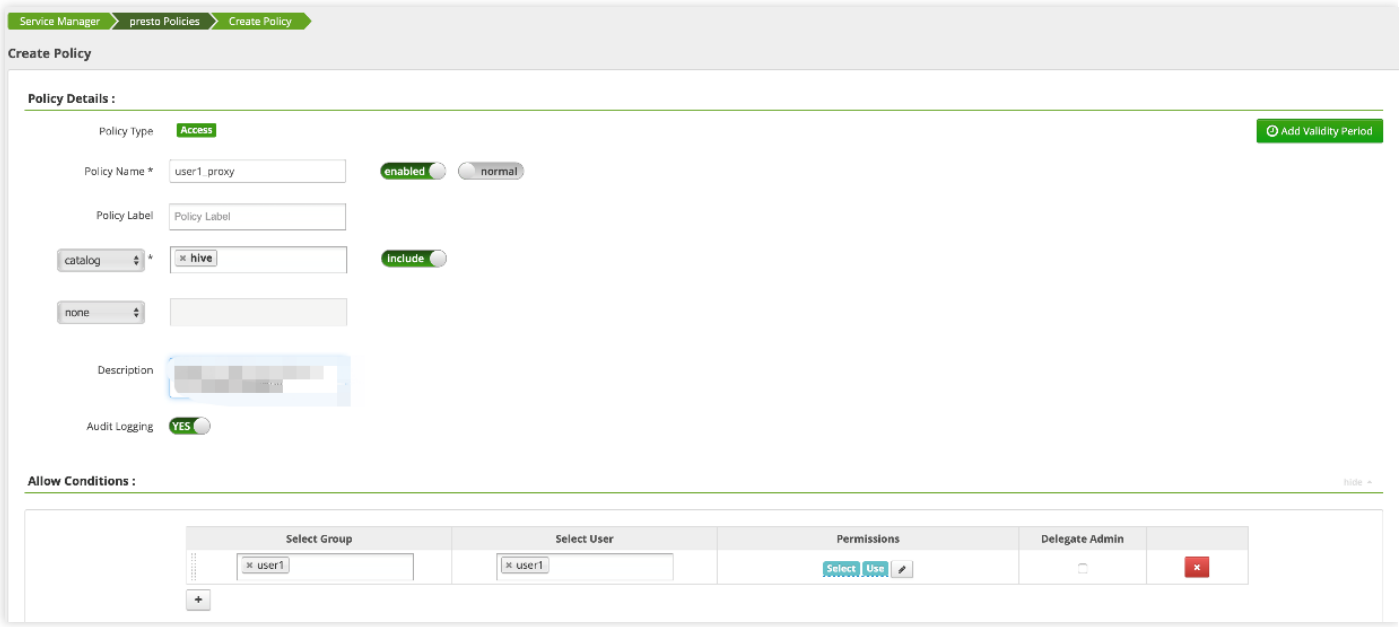
3. EMR Ranger Presto 资源权限配置

- 单击配置好的 EMR Ranger Presto Service



- 配置 Policy





4. 添加完 Policy 后，稍等约半分钟等待 Policy 生效。生效后使用 user1 就可以对 Presto 的 catalog 进行查看和使用。

Doris 开发指南

Doris 简介

最近更新时间：2021-06-07 17:25:20

腾讯云弹性 MapReduce（EMR）- Doris 提供开源 MPP 分析型数据库 Doris 的云上半托管服务，提供了便捷的 Doris 集群部署、配置修改、监控告警等功能。Doris 支持标准 SQL 语言、兼容 MySQL 协议、支持对 PB 级的海量数据进行高并发查询，可以满足多种数据分析需求，如离线数据分析、实时数据分析、交互式数据分析和探索式数据分析等。

Doris 功能特点

MySQL 协议兼容

Doris 提供兼容 MySQL 协议的连接接口，用户无需单独部署新的客户端库或者工具，可直接使用 MySQL 的相关库或者工具。提供了 MySQL 接口，可便捷的与上层应用兼容。用户学习曲线降低，方便用户上手使用。

大查询高吞吐

利用 MPP 架构的优势，使得查询能够分布式的在多个节点并行执行，充分利用集群整体计算资源，提高大查询的吞吐能力。

高并发小查询

通过使用分区裁剪、预聚合、谓词下推、向量化执行、异步 RPC 等技术，Doris 可以支持高并发点查询场景。

数据更新

Doris 支持按主键删除和更新数据。能够方便的从 MySQL 等事务数据库中同步实时更新的数据。

高可用和高可靠

Doris 中的数据和元数据都默认使用3副本存储（BE 节点需大于等于3）。在少数节点宕机的情况下，依然可以保证数据的可靠性。Doris 会自动检查和修复损坏的数据，并将查询请求自动路由到健康的节点，7×24 小时保证数据的可用性。

水平扩展和数据均衡

FE 节点和 BE 节点都可以进行横向扩展。用户可以根据计算和存储需要，灵活的对节点进行扩展。其中 BE 节点在扩展后，Doris 会自动根据节点间的负载情况，进行数据分片的自动均衡，无需人工干预。

物化视图和预聚合引擎

Doris 支持通过物化视图或上卷表的形式对数据预聚合计算后的结果进行存储，从而加速部分聚合类场景的查询效率。同时，Doris 能够保证物化视图和基础表之间的数据一致性，从而使得物化视图会查询和导入完全透明。Doris 内部会自动根据用户的查询语句，选择合适的物化视图进行数据摄取。

高效的列式存储引擎

Doris 采用自研的列式存储格式来提升 OLAP 领域的查询效率。存储采用字典编码、RLE 等多种编码方式，配合列式存储的特点，提供了非常高的数据压缩比，帮助用户节省存储空间。同时，存储格式上提供包括 Min/Max 智能索引、稀疏索引、布隆过滤器、bitmap 倒排索引等多种查询加速技术，进一步提升了查询效率。

在线表结构修改

支持在已导入数据的情况下修改表结构，包括增加列、删除列、修改列类型和改变列顺序等操作。变更操作不会影响当前数据库的查询和写入操作。

使用场景

大数据分析、统计

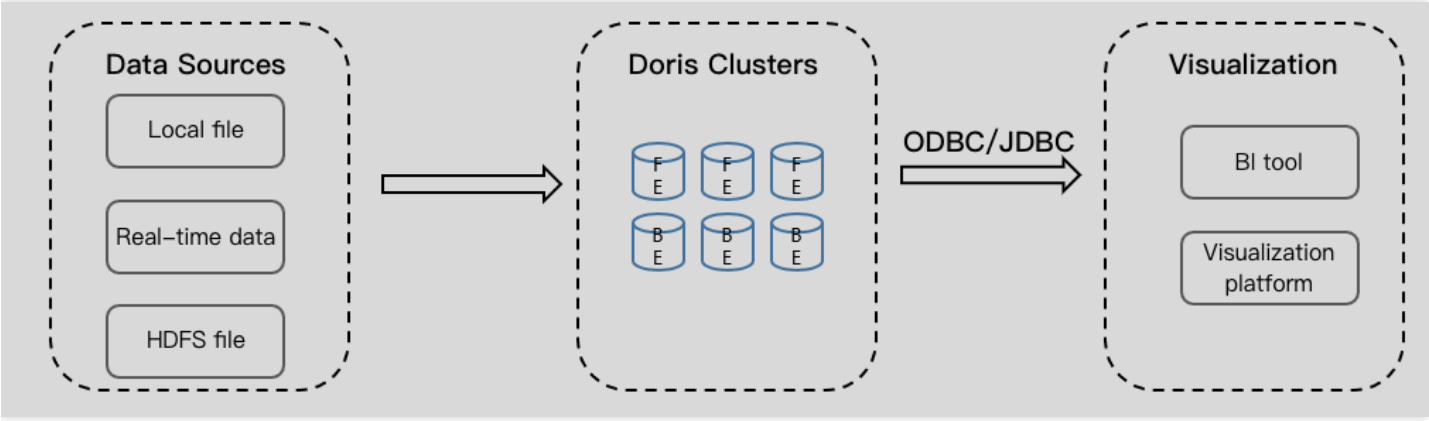
数据分析大体上可以分为两大类场景：一种偏向于报表类的，另一种偏向于多维分析的。

报表

报表类数据分析，数据分析以及查询的模式相对比较固定，而且后台 SQL 的模式都是确定的。针对此类应用场景，选择使用 MySQL 存结果数据，用户可从界面选择执行批处理以及发送邮件。在 Doris 平台中，报表类查询时延一般在秒级以下。

多维分析

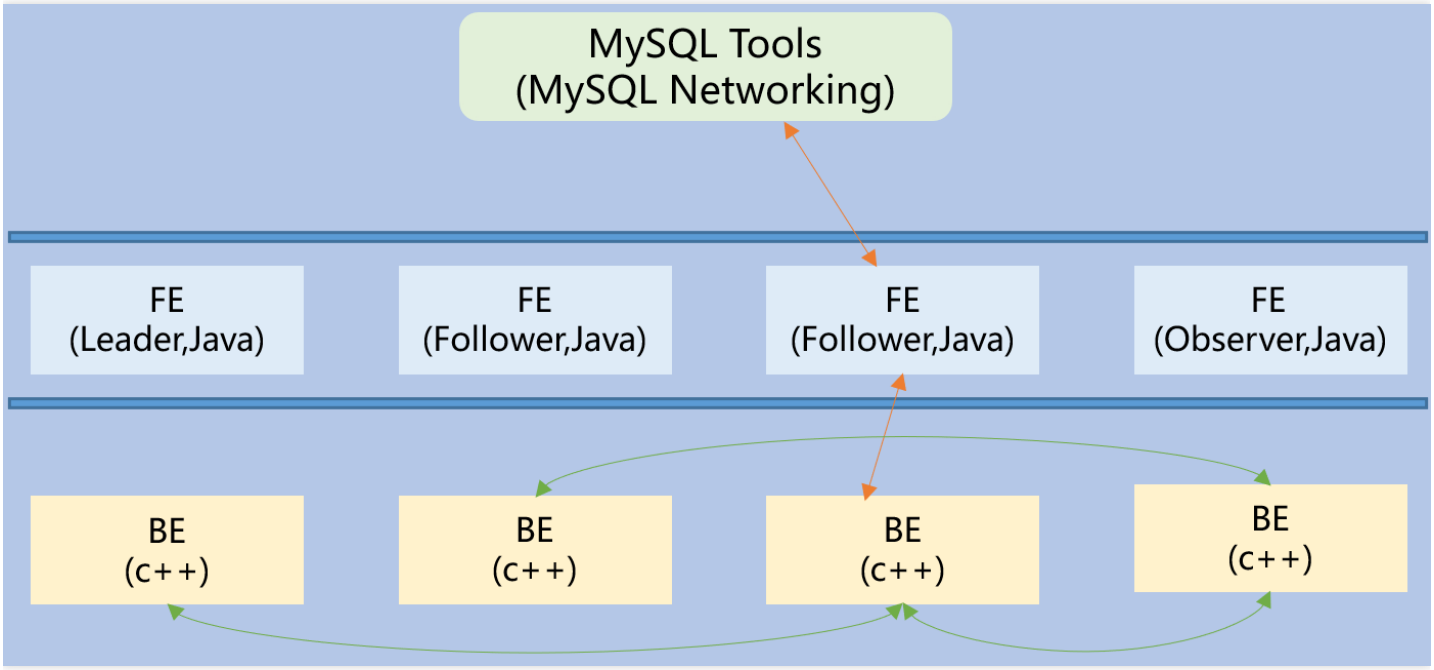
多维分析要求数据是结构化的，适用于查询相对灵活的场景，例如数据分析条件以及聚合维度等方面不是很确定，一般将此类数据分析定义为多维分析。



Doris 架构

Doris 主要有三个组件：

- FE（Frontend）是 Doris 的前端节点。主要负责接收和返回客户端请求、元数据以及集群管理、查询计划生成等工作。
- BE（Backend）是 Doris 的后端节点。主要负责数据存储与管理、查询计划执行等工作。
- Broker 是 Doris 集群中一种可选进程，主要用于支持 Doris 读写远端存储上的文件和目录，如 HDFS、腾讯云对象存储 COS 等。



基础使用指南
创建用户

最近更新时间：2021-06-01 17:27:01

Doris 采用 MySQL 协议进行通信，用户可通过 MySQL client 或者 MySQL JDBC 连接到 Doris 集群。选择 MySQL client 版本时建议采用5.1之后的版本，因为5.1之前版本不能支持长度超过16个字符的用户名。本文以 MySQL client 为例，通过一个完整的流程向用户展示 Doris 的基本使用方法。

> 注意：
> 重置原生 UI 密码仅更改 Knox 密码，Doris 本身的 UI 认证密码与集群密码一致。

Root 用户登录与密码修改

Doris 内置 root 和 admin 用户，密码与集群密码一致。启动完 Doris 程序后，可通过 root 或 admin 用户连接到 Doris 集群。使用下面命令即可登录 Doris：

```
mysql -h FE_HOST -P9030 -uroot
```

参数说明：

- FE_HOST 是任一 FE 节点的 IP 地址。
- 9030 是 fe.conf 中的 query_port 配置内容。

登录后，可通过以下命令修改 root 密码：

```
SET PASSWORD FOR 'root' = PASSWORD('your_password');
```

创建新用户

通过下面的命令创建一个普通用户：

```
CREATE USER 'test' IDENTIFIED BY 'test_passwd';
```

后续登录时即可通过如下连接命令登录：

```
mysql -h FE_HOST -P9030 -utest -ptest_passwd
```

新创建的普通用户默认没有任何权限。权限授予操作可参考 账户授权。

创建数据表并导入数据

最近更新时间：2022-07-08 11:50:45

创建数据库

初始可通过 root 或 admin 用户创建数据库，命令如下：

```
CREATE DATABASE example_db;
```

所有命令都可以使用 `HELP command;` 查看到详细的语法帮助，例如 `HELP CREATE DATABASE;` 。

如果不清楚命令的全名，可使用"help 命令某一字段"进行模糊查询。如键入 `HELP CREATE` ，可以匹配到 `CREATE DATABASE` 、 `CREATE TABLE` 、 `CREATE USER` 等命令。

数据库创建完成后，可以通过 `SHOW DATABASES;` 查看数据库信息。

```
MySQL> SHOW DATABASES;
+--------------------+
| Database |
+--------------------+
| example_db |
| information_schema |
+--------------------+
2 rows in set (0.00 sec)
```

information_schema 是为了兼容 MySQL 协议而存在，实际中信息可能不是很准确，所以关于具体数据库的信息建议通过直接查询相应数据库而获得。

账户授权

example_db 创建完成后，可通过 root/admin 账户将 example_db 读写权限授权给普通账户，例如 test。授权后，即可通过 test 账户登录并操作 example_db 数据库。

```
GRANT ALL ON example_db TO test;
```

建表

使用 CREATE TABLE 命令建立一个表（Table），更多详细参数可输入 `HELP CREATE TABLE;` 命令查看。

切换数据库命令如下：

```
USE example_db;
```

Doris 支持单分区和复合分区两种建表方式。

在复合分区中：

- 第一级称为 Partition，即分区。用户可以指定某一维度列作为分区列（当前只支持整型和时间类型的列），并指定每个分区的取值范围。
- 第二级称为 Distribution，即分桶。用户可以指定一个或多个维度列以及桶数对数据进行 HASH 分布。

以下场景推荐使用复合分区：

- 有时间维度或类似带有有序值的维度，可以这类维度列作为分区列。分区粒度可以根据导入频次、分区数据量等进行评估。
- 历史数据删除需求：如有删除历史数据的需求（例如仅保留最近 N 天的数据）。使用复合分区，可通过删除历史分区来达到目的。也可以通过在指定分区内发送 DELETE 语句进行数据删除。
- 解决数据倾斜问题：每个分区可以单独指定分桶数量。如按天分区，当每天的数据量差异很大时，可以通过指定分区的分桶数，合理划分不同分区的数据，分桶列建议选择区分度大的列。
- 用户也可以不使用复合分区，即使用单分区。则数据只做 HASH 分布。

下面以聚合模型为例，分别演示两种分区的建表语句。

单分区

建立一个名字为 table1 的逻辑表。分桶列为 siteid，桶数为10。这个表的 schema 如下：

- siteid：类型是 INT（4字节），默认值为10。
- citycode：类型是 SMALLINT（2字节）。
- username：类型是 VARCHAR，最大长度为32，默认值为空字符串。
- pv：类型是 BIGINT（8字节），默认值是0。这是一个指标列，Doris 内部会对指标列做聚合操作，这个列的聚合方法是求和（SUM）。

建表语句如下：

```
CREATE TABLE table1
(
siteid INT DEFAULT '10',
citycode SMALLINT,
username VARCHAR(32) DEFAULT '',
pv BIGINT SUM DEFAULT '0'
```

```
)
AGGREGATE KEY(siteid, citycode, username)
DISTRIBUTED BY HASH(siteid) BUCKETS 10
PROPERTIES("replication_num" = "1");
```

复合分区

建立一个名字为 table2 的逻辑表。这个表的 schema 如下：

- event_day：类型是 DATE，无默认值。
- siteid：类型是 INT（4字节），默认值为10。
- citycode：类型是 SMALLINT（2字节）。
- username：类型是 VARCHAR，最大长度为32，默认值为空字符串。
- pv：类型是 BIGINT（8字节），默认值是0。这是一个指标列，Doris 内部会对指标列做聚合操作，这个列的聚合方法是求和（SUM）。

我们使用 event_day 列作为分区列，建立3个分区 p201706、p201707、p201708。

- p201706：范围为 [最小值, 2017-07-01)。
- p201707：范围为 [2017-07-01, 2017-08-01)。
- p201708：范围为 [2017-08-01, 2017-09-01)。

> 注意：
> 区间为左闭右开。

每个分区使用 siteid 进行哈希分桶，桶数为10。建表语句如下：

```
CREATE TABLE table2
(
event_day DATE,
siteid INT DEFAULT '10',
citycode SMALLINT,
username VARCHAR(32) DEFAULT '',
pv BIGINT SUM DEFAULT '0'
)
AGGREGATE KEY(event_day, siteid, citycode, username)
PARTITION BY RANGE(event_day)
(
PARTITION p201706 VALUES LESS THAN ('2017-07-01'),
PARTITION p201707 VALUES LESS THAN ('2017-08-01'),
PARTITION p201708 VALUES LESS THAN ('2017-09-01')
)
```

```
DISTRIBUTED BY HASH(siteid) BUCKETS 10
PROPERTIES("replication_num" = "1");
```

表建完后，可查看 example_db 中表的信息：

```
MySQL> SHOW TABLES;
+---------------------+
| Tables_in_example_db |
+---------------------+
| table1 |
| table2 |
+---------------------+
2 rows in set (0.01 sec)

MySQL> DESC table1;
+----------+-------------+------+-------+---------+-------+
| Field | Type | Null | Key | Default | Extra |
+----------+-------------+------+-------+---------+-------+
| siteid | int(11) | Yes | true | 10 | |
| citycode | smallint(6) | Yes | true | N/A | |
| username | varchar(32) | Yes | true | | |
| pv | bigint(20) | Yes | false | 0 | SUM |
+----------+-------------+------+-------+---------+-------+
4 rows in set (0.00 sec)

MySQL> DESC table2;
+-----------+-------------+------+-------+---------+-------+
| Field | Type | Null | Key | Default | Extra |
+-----------+-------------+------+-------+---------+-------+
| event_day | date | Yes | true | N/A | |
| siteid | int(11) | Yes | true | 10 | |
| citycode | smallint(6) | Yes | true | N/A | |
| username | varchar(32) | Yes | true | | |
| pv | bigint(20) | Yes | false | 0 | SUM |
+-----------+-------------+------+-------+---------+-------+
5 rows in set (0.00 sec)
```

注意事项

> 说明：
> Doris 使用中更多语法说明，可参考 数据表的创建与数据导入。

1. 上述表通过设置 replication_num 建的都是单副本的表，Doris 建议用户采用默认的3副本设置，以保证高可用。

2. 可以对复合分区表动态的增删分区。
3. 数据导入可以导入指定的 Partition。
4. 可以动态修改表的 Schema。
5. 可以对 Table 增加上卷表（Rollup）以提高查询性能。
6. 表的列的 Null 属性默认为 true，会对查询性能有一定的影响。

导入数据

Doris 支持多种数据导入方式，下文以使用流式导入和 Broker 导入为例进行说明。

流式导入

流式导入通过 HTTP 协议向 Doris 传输数据，可以不依赖其他系统或组件直接导入本地数据。

示例一

以 "table1_20170707" 为 Label，使用本地文件 table1_data 导入 table1 表。

```
curl --location-trusted -u test:test -H "label:table1_20170707" -H "column_separa
tor:," -T table1_data http://FE_HOST:8030/api/example_db/table1/_stream_load
```

1. FE_HOST 是任一 FE 所在节点 IP，8030 为 fe.conf 中的 http_port。
2. 可以使用任一 BE 的 IP，以及 be.conf 中的 webserver_port 进行导入，例如 BE_HOST:8040。

本地文件 table1_data，以英文逗号作为数据之间的分隔，具体内容如下：

```
1,1,jim,2
2,1,grace,2
3,2,tom,2
4,3,bush,3
5,3,helen,3
```

示例二

以 "table2_20170707" 为 Label，使用本地文件 table2_data 导入 table2 表。

```
curl --location-trusted -u test:test -H "label:table2_20170707" -H "column_separa
tor:|" -T table2_data http://127.0.0.1:8030/api/example_db/table2/_stream_load
```

本地文件 table2_data，以 | 作为数据之间的分隔，具体内容如下：

```
2017-07-03|1|1|jim|2
2017-07-05|2|1|grace|2
```

```
2017-07-12|3|2|tom|2
2017-07-15|4|3|bush|3
2017-07-12|5|3|helen|3
```

注意事项

1. 采用流式导入建议文件大小限制在10GB以内，过大的文件会导致失败重试代价变大。
2. 每一批导入数据都需要取一个 Label，Label 最好是一个和一批数据有关的字符串，方便阅读和管理。Doris 基于 Label 保证在一个Database 内，同一批数据只可导入成功一次。失败任务的 Label 可以重用。
3. 流式导入是同步命令。命令返回成功则表示数据已经导入，返回失败表示这批数据没有导入。

Broker 导入

Broker 导入通过部署的 Broker 进程，读取外部存储上的数据进行导入。

以 "table1_20170708" 为 Label，将 HDFS 上的文件导入 table1 表。

```
LOAD LABEL table1_20170708
(
DATA INFILE("hdfs://your.namenode.host:port/dir/table1_data")
INTO TABLE table1
)
WITH BROKER hdfs
(
"username"="hdfs_user",
"password"="hdfs_password"
)
PROPERTIES
(
"timeout"="3600",
"max_filter_ratio"="0.1"
);
```

Broker 导入是异步命令。以上命令执行成功只表示提交任务成功。导入是否成功需要通过 `SHOW LOAD;` ，命令如下：

```
SHOW LOAD WHERE LABEL = "table1_20170708";
```

返回结果中，State 字段为 FINISHED 则表示导入成功。

异步的导入任务在结束前可以取消，命令如下：

```
CANCEL LOAD WHERE LABEL = "table1_20170708";
```

查询数据

最近更新时间：2021-06-01 17:50:51

简单查询

```
MySQL> SELECT * FROM table1 LIMIT 3;
+--------+----------+----------+------+
| siteid | citycode | username | pv |
+--------+----------+----------+------+
| 2 | 1 | 'grace' | 2 |
| 5 | 3 | 'helen' | 3 |
| 3 | 2 | 'tom' | 2 |
+--------+----------+----------+------+
3 rows in set (0.01 sec)
MySQL> SELECT * FROM table1 ORDER BY citycode;
+--------+----------+----------+------+
| siteid | citycode | username | pv |
+--------+----------+----------+------+
| 2 | 1 | 'grace' | 2 |
| 1 | 1 | 'jim' | 2 |
| 3 | 2 | 'tom' | 2 |
| 4 | 3 | 'bush' | 3 |
| 5 | 3 | 'helen' | 3 |
+--------+----------+----------+------+
5 rows in set (0.01 sec)
```

Join 查询

```
MySQL> SELECT SUM(table1.pv) FROM table1 JOIN table2 WHERE table1.siteid = table
2.siteid;
+--------------------+
| sum(`table1`.`pv`) |
+--------------------+
| 12 |
+--------------------+
1 row in set (0.20 sec)
```

子查询

```
MySQL> SELECT SUM(pv) FROM table2 WHERE siteid IN (SELECT siteid FROM table1 WHER
E siteid > 2);
+-----------+
| sum(`pv`) |
+-----------+
| 8 |
+-----------+
1 row in set (0.13 sec)
```

Kafka 开发指南

Kafka 简介

最近更新时间：2021-06-07 17:25:21

腾讯云弹性 MapReduce（EMR）- Kafka 提供开源 Kafka 的云上托管服务，提供了便捷的 Kafka 集群部署、配置修改、监控告警等功能，为企业及用户提供安全稳定的 OLAP 解决方案。Kafka 数据管道是流计算系统中最常用的数据源（Source）和数据目的（Sink）。用户可以把流数据导入到 Kafka 的某个 Topic 中，通过 Flink 算子进行处理后，输出到相同或不同 Kafka 示例的另一个 Topic。Kafka 支持同一个 Topic 多分区读写，数据可以从多个分区读入，也可以写入到多个分区，以提供更高的吞吐量，减少数据倾斜和热点。

架构

支持单节点、多节点架构。根据业务需求，灵活选择。

运维

在控制台提供了开箱即用的监控、日志检索、参数调整等服务。

功能

- 收发解耦：有效解耦生产者、消费者之间的关系。在确保同样的接口约束的前提下，允许独立扩展或修改生产者/消费者间的处理过程。
- 削峰填谷：Kafka 集群能够抵挡突增的访问压力，不会因为突发的超负荷的请求而完全崩溃，有效提升系统健壮性。
- 顺序读写：Kafka集群能够保证一个 Partition 内消息的有序性。和大部分的消息队列一致，Kafka 集群可以保证数据按照顺序进行处理，极大提升磁盘效率。
- 异步通信：在业务无需立即处理消息的场景下，消息队列 Kafka 集群提供了消息的异步处理机制，访问量高时仅将消息放入队列中，在访问量降低后再对消息进行处理，缓解系统压力。

优势

100%兼容开源，轻松迁移

- Kafka 集群兼容开源 Kafka 1.1.1 版本。

- Kafka 集群业务系统基于现有的开源 Apache Kafka 生态的代码，无需任何改造，即可迁移上云，享受到腾讯云提供的高性能消息队列 Kafka 服务。

高性能

- 腾讯云专业团队对服务性能进一步调优，免除复杂的参数配置，提供更高性能。
- 界面化升降配能力，高性能 IaaS 层支撑。

高可用性

- 依托腾讯技术工程多年监控平台的技术积累，对集群全方位多角度监控，更有专业运维团队7 × 24小时处理告警保障 Kafka 集群服务的高可用性。
- 支持同地域自定义多可用区部署，提升容灾能力。

高可靠性

- 磁盘高可靠，即使服务器坏盘50%也不影响业务。
- 默认2副本，支持3副本，副本越多可靠性越高。

应用场景

最近更新时间：2022-01-05 11:00:33

网页行为分析

Kafka 集群通过实时处理网站活动（PV、搜索、用户其他活动等），并根据类型发布到 Topic 中，这些信息流可以被用于实时监控或离线统计分析等。

由于每个用户的 page view 中会生成许多活动信息，因此网站活动跟踪需要很高的吞吐量，Kafka 集群可以完美满足高吞吐、离线处理等要求。

日志聚合

Kafka 集群的低延迟处理特性，易于支持多个数据源和分布式的数据处理（消费）。相比于中心化的日志聚合系统，消息队列 Kafka 可以在提供同样性能的条件下，实现更强的持久化保证以及更低的端到端延迟。

Kafka 集群的特性决定它非常适合作为日志收集中心。多台主机/应用可以将操作日志批量异步地发送到 Kafka 集群，而无需保存在本地或者 DB 中。Kafka 集群可以批量提交消息/压缩消息，对于生产者而言，几乎感觉不到性能的开支。此时消费者可以使用 Hadoop 等其他系统化的存储和分析系统，对拉取日志进行统计分析。

在线/离线分析

在一些大数据相关的业务场景中，需要对大量并发数据进行处理和汇总，此时对集群的处理性能和扩展性都有很高的要求。Kafka 集群在实现上的数据分发机制，包括磁盘存储空间的分配、消息格式的处理、服务器选择以及数据压缩等方面，也决定了其适合处理海量的实时消息，并能汇总分布式应用的数据，方便系统运维。

在具体的大数据场景中，Kafka 集群能够很好地支持离线数据、流式数据的处理，并能够方便地进行数据聚合、分析等操作。

Kafka 使用

最近更新时间：2021-06-07 17:25:21

生成数据

java 代码方式

```
@Component
@Slf4j
public class KafkaProducer {
@Autowired
private KafkaTemplate<String, Object> kafkaTemplate;
//自定义topic
public static final String TOPIC_TEST = "topic.test";
//
public static final String TOPIC_GROUP1 = "topic.group1";
//
public static final String TOPIC_GROUP2 = "topic.group2";
public void send(Object obj) {
String obj2String = JSONObject.toJSONString(obj);
log.info("准备发送消息为：{}", obj2String);
//发送消息
ListenableFuture<SendResult<String, Object>> future = kafkaTemplate.send(TOPIC_TE
ST, obj);
future.addCallback(new ListenableFutureCallback<SendResult<String, Object>>() {
@Override
public void onFailure(Throwable throwable) {
//发送失败的处理
log.info(TOPIC_TEST + " - 生产者 发送消息失败：" + throwable.getMessage());
}
@Override
public void onSuccess(SendResult<String, Object> stringObjectSendResult) {
//成功的处理
log.info(TOPIC_TEST + " - 生产者 发送消息成功：" + stringObjectSendResult.toString
());
}
});
}
}
```

命令方式

```
bin/kafka-console-producer.sh --broker-list node86:9092 --topic t_cdr
```

消费数据

java 代码方式

```
@Component
@Slf4j
public class KafkaConsumer {
@KafkaListener(topics = KafkaProducer.TOPIC_TEST, groupId = KafkaProducer.TOPIC_G
ROUP1)
public void topic_test(ConsumerRecord<?, ?> record, Acknowledgment ack, @Header(K
afkaHeaders.RECEIVED_TOPIC) String topic) {
Optional message = Optional.ofNullable(record.value());
if (message.isPresent()) {
Object msg = message.get();
log.info("topic_test 消费了： Topic:" + topic + ",Message:" + msg);
ack.acknowledge();
}
}
@KafkaListener(topics = KafkaProducer.TOPIC_TEST, groupId = KafkaProducer.TOPIC_G
ROUP2)
public void topic_test1(ConsumerRecord<?, ?> record, Acknowledgment ack, @Header
(KafkaHeaders.RECEIVED_TOPIC) String topic) {
Optional message = Optional.ofNullable(record.value());
if (message.isPresent()) {
Object msg = message.get();
log.info("topic_test1 消费了： Topic:" + topic + ",Message:" + msg);
ack.acknowledge();
}
}
}
```

命令方式

```
bin/kafka-console-consumer.sh --zookeeper node01:2181 --topic t_cdr --from-beginn
ing
```

新增 topic（命令方式）

```
bin/kafka-topics.sh --zookeeper node01:2181 --create --topic t_cdr --partitions 3
0 --replication-factor 2
```

详细使用可参考 kafka 官方文档。

Iceberg 开发指南

最近更新时间：2022-11-25 16:06:47

Iceberg 简介

Apache Iceberg 是一种新型的用于大规模数据分析的开源表格式。它被设计用于存储移动缓慢的大型表格数据。它旨在改善 Hive、Trino（PrestoSQL）和 Spark 中内置的事实上的标准表布局。Iceberg 可以屏蔽底层数据存储格式上的差异，向上提供统一的操作 API，使得不同的引擎可以通过其提供的 API 接入。

Apache Iceberg 具备以下能力：

- 模式演化（Schema evolution）：支持 Add（添加）、Drop（删除）、Update（更新）、Rename（重命名）和 Reorder（重排）表格式定义。
- 分区布局演变（Partition layout evolution）：可以随着数据量或查询模式的变化而更新表的布局。
- 隐式分区（Hidden partitioning）：查询不再取决于表的物理布局。通过物理和逻辑之间的分隔，Iceberg 表可以随着数据量的变化和时间的推移发展分区方案。错误配置的表可以得到修复，无需进行昂贵的迁移。
- 时光穿梭（Time travel）：支持用户使用完全相同的快照进行重复查询，或者使用户轻松检查更改。
- 版本回滚（Version rollback）：使用户可以通过将表重置为良好状态来快速纠正问题。

在可靠性与性能方面，Iceberg 可在生产中应用到数十 PB 的数据表，即使没有分布式 SQL 引擎，也可以读取这些巨大规模的表：

- 扫描速度快，无需使用分布式 SQL 引擎即可读取表或查找文件。
- 高级过滤，基于表元数据，使用分区和列级统计信息对数据文件以进行裁剪。

Iceberg 被设计用来解决最终一致的云对象存储中的正确性问题：

- 可与任何云存储一起使用，并且通过避免调用 list 和 rename 来减少 HDFS 的 NameNode 拥塞。
- 可序列化的隔离，表更改是原子性的，用户永远不会看到部分更改或未提交的更改。
- 多个并发写入使用乐观锁机制进行并发控制，即使写入冲突，也会重试以确保兼容更新成功。

Iceberg 设计为以快照（Snapshot）的形式来管理表的各个历史版本数据。快照代表一张表在某个时刻的状态。每个快照中会列出表在某个时刻的所有数据文件列表。Data 文件存储在不同的 Manifest 文件中，Manifest 文件存储在一个 Manifest List 文件中，Manifest 文件可以在不同的 Manifest List 文件间共享，一个 Manifest List 文件代表一个快照。

- Manifest list 文件是元数据文件，其中存储的是 Manifest 文件的列表，每个 Manifest 文件占据一行。
- Manifest 文件是元数据文件，其中列出了组成某个快照的数据文件列表。每行都是每个数据文件的详细描述，包括数据文件的状态、文件路径、分区信息、列级别的统计信息（例如每列的最大最小值、空值数等）、文件的大

小以及文件中数据的行数等信息。
- Data 文件是 Iceberg 表真实存储数据的文件，一般是在表的数据存储目录的 data 目录下。

使用示例

更多示例可参考 Iceberg 官网示例。
本文以 EMR- V3.3.0中的 Iceberg0.11.0版本为示例，不同EMR版本相关jar包名称可能有所差异，请您根据路径下实际名称取用。

1. 登录 master 节点，切换为 hadoop 用户。
2. Iceberg 相关的包放置在 `/usr/local/service/iceberg/` 下面。
3. 使用计算引擎查询数据。

- Spark 引擎
  - Spark-SQL 交互式命令行

```
spark-sql --master local[*] --conf spark.sql.extensions=org.apache.iceberg.spark.extensions.IcebergSparkSessionExtensions --conf spark.sql.catalog.local=org.apache.iceberg.spark.SparkCatalog --conf spark.sql.catalog.local.type=hadoop --conf spark.sql.catalog.local.warehouse=/usr/hive/warehouse --jars /usr/local/service/iceberg/iceberg-spark3-runtime-0.11.0.jar
```

  - 插入和查询数据

```
CREATE TABLE local.default.t1 (id int, name string) USING iceberg;
INSERT INTO local.default.t1 values(1, "tom");
SELECT * from local.default.t1;
```

- Hive 引擎
  - 使用 beeline

```
beeline -u jdbc:hive2://[hiveserver2_ip:hiveserver2_port] -n hadoop --hiveconf hive.input.format=org.apache.hadoop.hive.ql.io.HiveInputFormat --hiveconf hive.stats.autogather=false
```

  - 查询数据

```
ADD JAR /usr/local/service/iceberg/iceberg-hive-runtime-0.11.0.jar;
CREATE EXTERNAL TABLE t1 STORED BY 'org.apache.iceberg.mr.hive.HiveIcebergStorageHandler' LOCATION '/usr/hive/warehouse/default/t1' TBLPROPERTIES ('iceberg.catalog'='location_based_table');
select count(*) from t1;
```

- Flink 引擎
  - 根据 Flink 和 Hive 版本在 Maven 仓库 下载相应版本 flink-sql-connector-hive 包，以 Flink standalone 模式为例，并使用 Flink shell 交互式命令行。

```
wget https://repo1.maven.org/maven2/org/apache/flink/flink-sql-connector-hive-3.1.2_2.11/1.12.1/flink-sql-connector-hive-3.1.2_2.11-1.12.1.jar
/usr/local/service/flink/bin/start-cluster.sh
sql-client.sh embedded -j /usr/local/service/iceberg/iceberg-flink-runtime-0.11.0.jar -j flink-sql-connector-hive-3.1.2_2.11-1.12.1.jar shell
```

  - 查询数据

```
CREATE CATALOG hive_catalog WITH ('type'='iceberg','catalog-type'='hive','uri'='hivemetastore_ip:hivemetastore_port','clients'='5','property-version'='1','warehouse'='hdfs:///usr/hive/warehouse/');
CREATE DATABASE hive_catalog.iceberg_db;
CREATE TABLE hive_catalog.iceberg_db.t1 (id BIGINT COMMENT 'unique id',data STRING);
INSERT INTO hive_catalog.iceberg_db.t1 values(1, 'tom');
SELECT count(*) from hive_catalog.iceberg_db.t1;
```

StarRocks 开发指南

StarRocks 简介

最近更新时间：2022-05-16 12:53:04

StarRocks 是什么

- StarRocks 是新一代极速全场景 MPP 数据库。
- StarRocks 充分吸收关系型 OLAP 数据库和分布式存储系统在大数据时代的优秀研究成果，在业界实践的基础上，进一步改进优化、升级架构，并增添了众多全新功能，形成了全新的企业级产品。
- StarRocks 致力于构建极速统一分析体验，满足企业用户的多种数据分析场景，支持多种数据模型(明细模型、聚合模型、更新模型)，多种导入方式（批量和实时），支持导入多达10000列的数据，可整合和接入多种现有系统(Spark、Flink、Hive、 ElasticSearch)。
- StarRocks 兼容 MySQL 协议，可使用 MySQL 客户端和常用 BI 工具对接 StarRocks 来进行数据分析。
- StarRocks 采用分布式架构，对数据表进行水平划分并以多副本存储。集群规模可以灵活伸缩，能够支持10PB 级别的数据分析; 支持 MPP 框架，并行加速计算; 支持多副本，具有弹性容错能力。
- StarRocks 采用关系模型，使用严格的数据类型和列式存储引擎，通过编码和压缩技术，降低读写放大；使用向量化执行方式，充分挖掘多核 CPU 的并行计算能力，从而显著提升查询性能。

StarRocks 特性

StarRocks 的架构设计融合了 MPP 数据库，以及分布式系统的设计思想，具有以下特性：

架构精简

StarRocks 内部通过 MPP 计算框架完成 SQL 的具体执行工作。MPP 框架本身能够充分的利用多节点的计算能力，整个查询并行执行，从而实现良好的交互式分析体验。 StarRocks 集群不需要依赖任何其他组件，易部署、易维护，极简的架构设计，降低了 StarRocks 系统的复杂度和维护成本，同时也提升了系统的可靠性和扩展性。 管理员只需要专注于 StarRocks 系统，无需学习和管理任何其他外部系统。

全面向量化引擎

StarRocks 的计算层全面采用了向量化技术，将所有算子、函数、扫描过滤和导入导出模块进行了系统性优化。通过列式的内存布局、适配 CPU 的 SIMD 指令集等手段，充分发挥了现代 CPU 的并行计算能力，从而实现亚秒级别的多维分析能力。

智能查询优化

StarRocks 通过 CBO 优化器(Cost Based Optimizer)可以对复杂查询自动优化。无需人工干预，就可以通过统计信息合理估算执行成本，生成更优的执行计划，极大提高了 Adhoc 和 ETL 场景的数据分析效率。

联邦查询

StarRocks 支持使用外表的方式进行联邦查询，当前可以支持 Hive、MySQL、Elasticsearch 三种类型的外表，用户无需通过数据导入，可以直接进行数据查询加速。

高效更新

StarRocks 支持多种数据模型，其中更新模型可以按照主键进行 upsert/delete 操作，通过存储和索引的优化可以在并发更新的同时实现高效的查询优化，更好的服务实时数仓的场景。

智能物化视图

StarRocks 支持智能的物化视图。用户可以通过创建物化视图，预先计算生成预聚合表用于加速聚合类查询请求。StarRocks 的物化视图能够在数据导入时自动完成汇聚，与原始表数据保持一致。并且在查询的时候，用户无需指定物化视图，StarRocks 能够自动选择最优的物化视图来满足查询请求。

标准 SQL

StarRocks 支持标准的 SQL 语法，包括聚合、JOIN、排序、窗口函数和自定义函数等功能。StarRocks 可以完整支持 TPC-H 的22个 SQL 和 TPC-DS 的99个 SQL。此外，StarRocks还兼容 MySQL 协议语法，可使用现有的各种客户端工具、BI 软件访问 StarRocks，对 StarRocks 中的数据进行拖拽式分析。

流批一体

StarRocks 支持实时和批量两种数据导入方式，支持的数据源有 Kafka、HDFS、本地文件，支持的数据格式有 ORC、Parquet 和 CSV 等，支持导入多达10000列的数据。StarRocks 可以实时消费 Kafka 数据来完成数据导入，保证数据不丢不重（exactly once）。StarRocks 也可以从本地或者远程（HDFS）批量导入数据。

高可用易扩展

StarRocks 的元数据和数据都是多副本存储，并且集群中服务有热备，多实例部署，避免了单点故障。集群具有自愈能力，可弹性恢复，节点的宕机、下线、异常都不会影响 StarRocks 集群服务的整体稳定性。 StarRocks 采用分布式架构，存储容量和计算能力可近乎线性水平扩展。StarRocks 单集群的节点规模可扩展到数百节点，数据规模可达到10PB 级别。 扩缩容期间无需停服，可以正常提供查询服务。 另外StarRocks 中表模式热变更，可通过一条简单 SQL 命令动态地修改表的定义，例如增加列、减少列、新建物化视图等。同时，处于模式变更中的表也可也正常导入和查询数据。

StarRocks 适合什么场景

StarRocks 可以满足企业级用户的多种分析需求，包括 OLAP 多维分析、定制报表、实时数据分析和 Ad-hoc 数据分析等。具体的业务场景包括：

- OLAP 多维分析用户行为分析。
  - 用户画像、标签分析、圈人。
  - 高维业务指标报表。
  - 自助式报表平台。
  - 业务问题探查分析。
  - 跨主题业务分析。
  - 财务报表。
  - 系统监控分析。
- 实时数据分析电商大促数据分析。
  - 教育行业的直播质量分析。
  - 物流行业的运单分析。
  - 金融行业绩效分析、指标计算。
  - 广告投放分析。
  - 管理驾驶舱。
  - 探针分析 APM（Application Performance Management）。
- 高并发查询广告主报表分析。
  - 零售行业渠道人员分析。
  - SaaS 行业面向用户分析报表。
  - Dashbroad 多页面分析。
- 统一分析通过使用一套系统解决多维分析、高并发查询、预计算、实时分析、Adhoc查询等场景，降低系统复杂度和多技术栈开发与维护成本。

基础使用指南

最近更新时间：2022-05-16 12:52:26

目前 StarRocks 支持多种方式连接，下面简介使用 mysql 客户端连接 StarRocks 进行开发。

Root 用户登录

使用 MySQL 客户端连接某一个 FE 实例的 query_port(9030), StarRocks 内置 root 用户，密码默认与集群密码相同：

```
mysql -h fe_host -P9030 -u root -p
```

清理环境：

```
mysql > drop database if exists example_db;
mysql > drop user test;
```

查看部署节点

1. 查看 FE 节点。

```
mysql> SHOW PROC '/frontends'\G
*************************** 1. row ***************************
Name: 172.16.139.24_9010_1594200991015
IP: 172.16.139.24
HostName: starrocks-sandbox01
EditLogPort: 9010
HttpPort: 8030
QueryPort: 9030
RpcPort: 9020
Role: FOLLOWER
IsMaster: true
ClusterId: 861797858
Join: true
Alive: true
ReplayedJournalId: 64
LastHeartbeat: 2020-03-23 20:15:07
IsHelper: true
ErrMsg:
1 row in set (0.03 sec)
```

Role 为 FOLLOWER 说明这是一个能参与选主的 FE；IsMaster 为 true，说明该 FE 当前为主节点。

2. 查看 BE 节点。

```
mysql> SHOW PROC '/backends'\G
*************************** 1. row ***************************
BackendId: 10002
Cluster: default_cluster
IP: 172.16.139.24
HostName: starrocks-sandbox01
HeartbeatPort: 9050
BePort: 9060
HttpPort: 8040
BrpcPort: 8060
LastStartTime: 2020-03-23 20:19:07
LastHeartbeat: 2020-03-23 20:34:49
Alive: true
SystemDecommissioned: false
ClusterDecommissioned: false
TabletNum: 0
DataUsedCapacity: .000
AvailCapacity: 327.292 GB
TotalCapacity: 450.905 GB
UsedPct: 27.41 %
MaxDiskUsedPct: 27.41 %
ErrMsg:
Version:
1 row in set (0.01 sec)
```

如果 isAlive 为 true，则说明 BE 正常接入集群。如果 BE 没有正常接入集群，请查看 log 目录下的 be.WARNING 日志文件确定原因。

3. 查看 Broker 节点。

```
MySQL> SHOW PROC "/brokers"\G
*************************** 1. row ***************************
Name: broker1
IP: 172.16.139.24
Port: 8000
Alive: true
LastStartTime: 2020-04-01 19:08:35
LastUpdateTime: 2020-04-01 19:08:45
ErrMsg:
```

```
1 row in set (0.00 sec)
```

Alive 为 true 代表状态正常。

创建新用户

通过下面的命令创建一个普通用户：

```
mysql > create user 'test' identified by '123456';
```

创建数据库

StarRocks 中 root 账户才有权建立数据库，使用 root 用户登录，建立 example_db 数据库:

```
mysql > create database example_db;
```

数据库创建完成之后，可以通过 show databases 查看数据库信息：

```
mysql > show databases;
+--------------------+
| Database |
+--------------------+
| example_db |
| information_schema |
+--------------------+
2 rows in set (0.00 sec)
```

information_schema 是为了兼容 mysql 协议而存在，实际中信息可能不是很准确，所以关于具体数据库的信息建议通过直接查询相应数据库而获得。

账户授权

example_db 创建完成之后，可以通过 root 账户 example_db 读写权限授权给 test 账户，授权之后采用 test 账户登录就可以操作 example_db 数据库了：

```
mysql > grant all on example_db to test;
```

退出 root 账户，使用 test 登录 StarRocks 集群：

```
mysql > exit
mysql -h 127.0.0.1 -P9030 -utest -p123456
```

建表

StarRocks 支持支持单分区和复合分区两种建表方式。
在复合分区中：

- 第一级称为 Partition，即分区。用户可以指定某一维度列作为分区列（当前只支持整型和时间类型的列），并指定每个分区的取值范围。
- 第二级称为 Distribution，即分桶。用户可以指定某几个维度列（或不指定，即所有 KEY 列）以及桶数对数据进行 HASH 分布。

以下场景推荐使用复合分区：

- 有时间维度或类似带有有序值的维度：可以以这类维度列作为分区列。分区粒度可以根据导入频次、分区数据量等进行评估。
- 历史数据删除需求：如有删除历史数据的需求（例如仅保留最近 N天的数据）。使用复合分区，可以通过删除历史分区来达到目的。也可以通过在指定分区内发送 DELETE 语句进行数据删除。
- 解决数据倾斜问题：每个分区可以单独指定分桶数量。如按天分区，当每天的数据量差异很大时，可以通过指定分区的分桶数，合理划分不同分区的数据,分桶列建议选择区分度大的列。

用户也可以不使用复合分区，即使用单分区。则数据只做 HASH 分布。
下面分别演示两种分区的建表语句：

1. 首先切换数据库：mysql > use example_db;
2. 建立单分区表建立一个名字为 table1的逻辑表。使用全 hash 分桶，分桶列为 siteid，桶数为10。这个表的 schema 如下：
   - siteid：类型是 INT（4字节）, 默认值为10。
   - city_code：类型是 SMALLINT（2字节）。
   - username：类型是 VARCHAR, 最大长度为32, 默认值为空字符串。
   - pv：类型是 BIGINT（8字节）, 默认值是0; 这是一个指标列, StarRocks 内部会对指标列做聚合操作, 这个列的聚合方法是求和（SUM）。这里采用了聚合模型，除此之外StarRocks 还支持明细模型和更新模型，具体参考数据模型介绍。
     建表语句如下:

```
mysql >
CREATE TABLE table1
(
siteid INT DEFAULT '10',
citycode SMALLINT,
username VARCHAR(32) DEFAULT '',
pv BIGINT SUM DEFAULT '0'
)
AGGREGATE KEY(siteid, citycode, username)
DISTRIBUTED BY HASH(siteid) BUCKETS 10
PROPERTIES("replication_num" = "1");
```

3. 建立复合分区表

   建立一个名字为 table2的逻辑表。这个表的 schema 如下：

   - event_day：类型是 DATE，无默认值。
   - siteid：类型是 INT（4字节）, 默认值为10。
   - city_code：类型是 SMALLINT（2字节）。
   - username：类型是 VARCHAR, 最大长度为32, 默认值为空字符串。
   - pv：类型是 BIGINT（8字节）, 默认值是0; 这是一个指标列, StarRocks 内部会对指标列做聚合操作, 这个列的聚合方法是求和（SUM）。

     我们使用 event_day 列作为分区列，建立3个分区: p1, p2, p3
   - p1：范围为 [最小值, 2017-06-30)。
   - p2：范围为 [2017-06-30, 2017-07-31)。
   - p3：范围为 [2017-07-31, 2017-08-31)。

   每个分区使用 siteid 进行哈希分桶，桶数为10。

   建表语句如下:

```
CREATE TABLE table2
(
event_day DATE,
siteid INT DEFAULT '10',
citycode SMALLINT,
username VARCHAR(32) DEFAULT '',
pv BIGINT SUM DEFAULT '0'
)
AGGREGATE KEY(event_day, siteid, citycode, username)
PARTITION BY RANGE(event_day)
(
PARTITION p1 VALUES LESS THAN ('2017-06-30'),
PARTITION p2 VALUES LESS THAN ('2017-07-31'),
```

```
PARTITION p3 VALUES LESS THAN ('2017-08-31')
)
DISTRIBUTED BY HASH(siteid) BUCKETS 10
PROPERTIES("replication_num" = "1");
```

表建完之后，可以查看 example_db 中表的信息:

```
mysql> show tables;
+------------------------+
| Tables_in_example_db |
+------------------------+
| table1 |
| table2 |
+------------------------+
2 rows in set (0.01 sec)

mysql> desc table1;
+---------+-------------+------+-------+---------+-------+
| Field | Type | Null | Key | Default | Extra |
+---------+-------------+------+-------+---------+-------+
| siteid | int(11) | Yes | true | 10 | |
| citycode | smallint(6) | Yes | true | N/A | |
| username | varchar(32) | Yes | true | | |
| pv | bigint(20) | Yes | false | 0 | SUM |
+---------+-------------+------+-------+---------+-------+
4 rows in set (0.00 sec)

mysql> desc table2;
+-----------+-------------+------+-------+---------+-------+
| Field | Type | Null | Key | Default | Extra |
+-----------+-------------+------+-------+---------+-------+
| event_day | date | Yes | true | N/A | |
| siteid | int(11) | Yes | true | 10 | |
| citycode | smallint(6) | Yes | true | N/A | |
| username | varchar(32) | Yes | true | | |
| pv | bigint(20) | Yes | false | 0 | SUM |
+-----------+-------------+------+-------+---------+-------+
5 rows in set (0.00 sec)
```

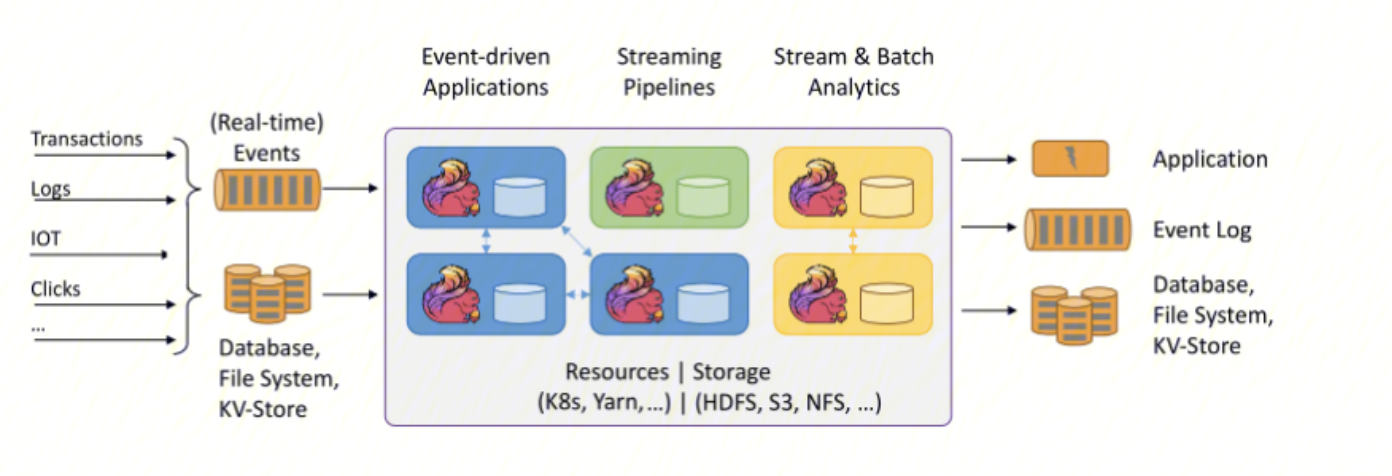
Flink 开发指南

Flink 简介

最近更新时间：2023-06-19 17:07:59

Flink 核心是一个开源的分布式、高性能、高可用、准确的数据流执行引擎，其针对数据流的分布式计算提供了数据分布、数据通信以及容错机制等功能。基于流执行引擎，Flink 提供了更高抽象层的 API 以便您编写分布式任务。

- 分布式：表示 Flink 程序可以运行在多台机器上。
- 高性能：表示 Flink 处理性能比较高。
- 高可用：表示 Flink 支持程序的自动重启机制。
- 准确的：表示 Flink 可以保证处理数据的准确性。



下图中左边是数据源，从这里可以看出来，这些数据是实时生产的一些日志，或者是数据库，文件系统，kv 存储系统中的数据。中间是 Flink，负责对数据进行梳理。右边是目的地，Flink 可以将计算好的数据输出到其它应用系统中，或者存储系统中。Flink 的三大核心组件如下：

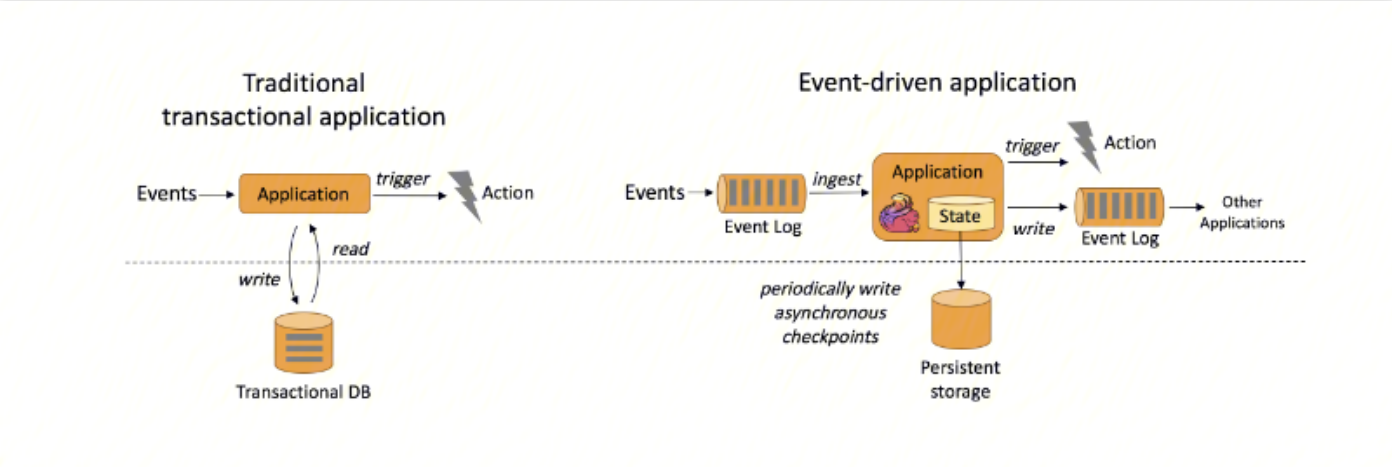
- Data Source：也就是图中左边的数据源。
- Transformations：算子（负责对数据进行处理）。
- Data Sink：输出组件（负责把计算好的数据输出到其它应用系统中）。

使用场景

Flink 主要有以下三种应用场景：

1. 事件驱动型应用
   事件驱动型应用是一类具有状态的应用，它从一个或多个事件流提取数据，并根据到来的事件触发计算、状态更

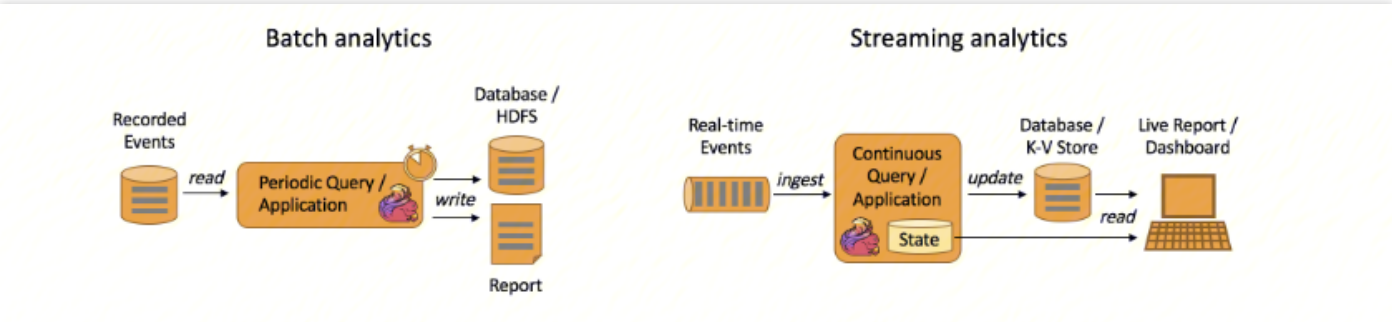
新或其他外部动作。



在传统架构中（图左），我们需要读写远程事务型数据库，例如 MySQL。在事件驱动应用中数据和计算不会分离，应用只需访问本地（内存或磁盘）即可获取数据，所以具有更高的吞吐和更低的延迟。

- Flink 的以下特性完美的支持了事件驱动型应用。
- 高效的状态管理，Flink 自带的 State Backend 可以很好的存储中间状态信息。
- 丰富的窗口支持，Flink 支持包含滚动窗口、滑动窗口及其他窗口。
- 多种时间语义，Flink 支持 Event Time、Processing Time 和 Ingestion Time。
  不同级别的容错，Flink 支持 At Least Once 或 Exactly Once 容错级别。

2. 实时数据分析应用：
   数据分析任务需要从原始数据中提取有价值的信息和指标。传统的分析方式通常是利用批查询，或将事件记录下来并基于此有限数据集构建应用来完成。为了得到最新数据的分析结果，必须先将它们加入分析数据集并重新执行查询或运行应用，随后将结果写入存储系统或生成报告。

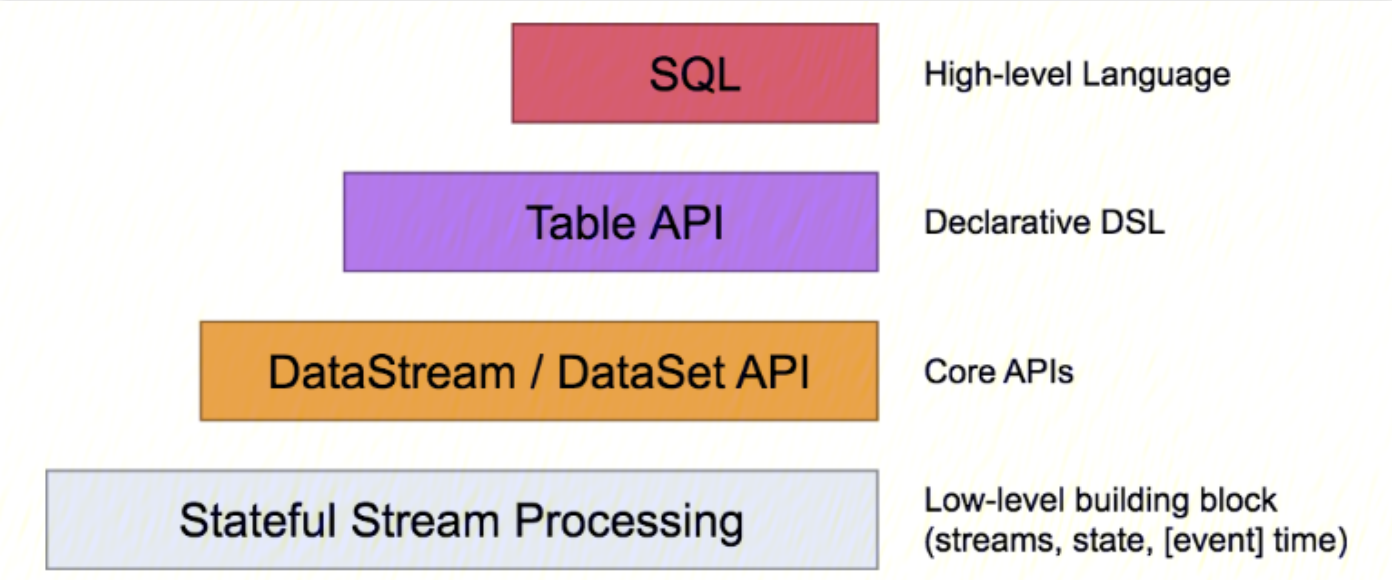


3. 实时数据仓库和 ETL
   ETL（Extract-Transform-Load）的目的是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程。传统的离线数据仓库将业务数据集中进行存储后，以固定的计算逻辑定时进行 ETL 和其他建模后产出报表等应用。离线数据仓库主要是构建 T+1 的离线数据，通过定时任务每天拉取增量数据，然后创建各个业务相关的主题

维度数据，对外提供 T+1 的数据查询接口。

上图展示了离线数据仓库 ETL 和实时数据仓库的差异，可以看到离线数据仓库的计算和数据的实时性均较差。数据本身的价值随着时间的流逝会逐步减弱，因此数据发生后必须尽快的达到用户的手中，实时数仓的构建需求也应运而生。

相关分层 API 请参考以下文档：



- Table API & SQL ：Table API 一般与 DataSet 或者 DataStream 紧密关联，可以通过一个 DataSet 或者 DataStream 创建出一个 Table，然后再使用类似 flilter、sum、join、select 等这种操作。最近还可以将一个 Table 对象转换成 DataSet 或者 DataStream。SQL API 的底层是基于 Apache Calcite，Apache Calcite 实现了标准 SQL，使用起来比其它 API 更加灵活，因为可以直接使用 SQL 语句。Table API 和 SQL API 可以很容易地结合在一块使用。因为它们都返回 Table 对象。
- DataStream API & DataSet API ：主要提供针对流数据和批数据的处理，是对低级 API 进行了一些封装，提供了 flilter、sum、max、min等高阶函数，简单易用，所以这些 API 在实际生产中应用还是比较广泛的。
- Stateful Stream Processing ：提供了对时间和状态的细粒度控制，简洁性和易用性较差，主要应用在一些复杂事件处理逻辑上。

环境信息

- Flink 默认部署在集群的 Master、Core 节点，安装了 Flink 组件的集群其功能都是开箱即用的。
- 登录机器后使用命令 su hadoop 切换到 hadoop 用户进行一些Flink的本地测试。
- Flink 软件路径在 /usr/local/service/flink 下。
- 相关日志路径在 /data/emr/flink/logs 下。

更多详细资料请参考 社区文档。

Flink 分析 COS 上的数据

最近更新时间：2023-02-01 15:59:58

Flink 擅长处理无界和有界数据集 精确的时间控制和状态化使得 Flink 的运行（runtime）能够运行任何处理无界流的应用。有界流则由一些专为固定大小数据集特殊设计的算法和数据结构进行内部处理，产生了出色的性能。
以下针对于在对象存储上的有界或者无界数据集的数据集进行演示操作，这里为了更好的观察作业的运行情况使用 yarn-session 模式来提交任务。Flink On Yarn 支持 Session Mode 和 Application Mode 两种形式，具体可参考 社区文档。
本教程演示的是提交的任务为 wordcount 任务即统计单词个数，提前需要在集群中上传需要统计的文件。

开发准备

1. 由于任务中需要访问腾讯云对象存储（COS），所以需要在 COS 中先 创建存储桶。
2. 确认您已开通腾讯云，且已创建一个 EMR 集群。在创建 EMR 集群的时候需要在软件配置界面选择 Flink 组件，并且在基础配置页面开启对象存储的授权。
3. 集群购买完成后，可以使用 HDFS 访问对象存储确保其基础功能可用。具体命令如：

```
[hadoop@10 ~]$ hdfs dfs -ls cosn://$BUCKET_NAME/path
Found 1 items
-rw-rw-rw- 1 hadoop hadoop 27040 2022-10-28 15:08 cosn://$BUCKET_NAME/path/LICENS
E
```

示例

```
# -n 表示申请1个容器，这里指的就是多少个taskmanager
# -tm 表示每个TaskManager的内存大小
# -s 表示每个TaskManager的slots数量
# -d 表示以后台程序方式运行，后面接的session名字
[hadoop@10 ~]$ yarn-session.sh -jm 1024 -tm 1024 -n 1 -s 1 -nm wordcount-example
-d
```

```
/usr/local/service/flink/bin/flink run -m yarn-cluster /usr/local/service/flink/e
xamples/batch/WordCount.jar --input cosn://$BUCKET_NAME/path/LICENSE -output cos
n://$BUCKET_NAME/path/wdp_test
[hadoop@10 ~]$ hdfs dfs -ls cosn://$BUCKET_NAME/path/wdp_test
```

```
-rw-rw-rw- 1 hadoop hadoop 7484 2022-11-04 00:47 cosn://$BUCKET_NAME/path/wdp_tes
t
```

Maven 工程示例

在本次演示中，不再采用系统自带的演示程序，而是自己建立工程编译打包之后上传到 EMR 集群运行。推荐您使用 Maven 来管理您的工程。Maven 是一个项目管理工具，能够帮助您方便的管理项目的依赖信息，即它可以通过 pom.xml 文件的配置获取 jar 包，而不用去手动添加。

1. 首先下载并安装 Maven，配置 Maven 的环境变量。如果您使用 IDE，请在 IDE 中设置 Maven 相关配置。
2. 在本地 shell 下进入您想要新建工程的目录，例如 D://mavenWorkplace 中，输入如下命令新建一个 Maven 工程：

```
mvn archetype:generate -DgroupId=$yourgroupID -DartifactId=$yourartifactID -Darch
etypeArtifactId=maven-archetype-quickstart
```

其中 yourgroupID 即为您的包名。yourartifactID 为您的项目名称，而 maven-archetype-quickstart 表示创建一个 Maven Java 项目。工程创建过程中需要下载一些文件，请保持网络通畅。
创建成功后，在 D://mavenWorkplace 目录下就会生成一个名为 $yourartifactID 的工程文件夹。其中的文件结构如下所示：

```
simple
---pom.xml        核心配置，项目根下
---src
---main
---java         Java 源码目录
---resources    Java 配置文件目录
---test
---java         测试源码目录
---resources    测试配置目录
```

其中我们主要关心 pom.xml 文件和 main 下的 Java 文件夹。pom.xml 文件主要用于依赖和打包配置，Java 文件夹下放置您的源代码。
首先在 pom.xml 中添加 Maven 依赖：

```
<properties>
<scala.version>2.12</scala.version>
<flink.version>1.14.3</scala.version>
</properties>
<dependencies>
<dependency>
<groupId>org.apache.flink</groupId>
```

```
<artifactId>flink-java</artifactId>
<version>1.14.3</version>
<scope>provided</scope>
</dependency>
<dependency>
<groupId>org.apache.flink</groupId>
<artifactId>flink-streaming-scala_${scala.version}</artifactId>
<version>${flink.version}</version>
<scope>provided</scope>
</dependency>
<dependency>
<groupId>org.apache.flink</groupId>
<artifactId>flink-clients_${scala.version}</artifactId>
<version>${flink.version}</version>
<scope>provided</scope>
</dependency>
</dependencies>
```

说明：

scala.version 和 flink.version 请根据自身使用 EMR 版本中的组件版本保持一致。

继续在 pom.xml 中添加打包和编译插件：

```
<build>
<plugins>
<plugin>
<groupId>org.apache.maven.plugins</groupId>
<artifactId>maven-compiler-plugin</artifactId>
<configuration>
<source>1.8</source>
<target>1.8</target>
<encoding>utf-8</encoding>
</configuration>
</plugin>
<plugin>
<artifactId>maven-assembly-plugin</artifactId>
<configuration>
<descriptorRefs>
<descriptorRef>jar-with-dependencies</descriptorRef>
</descriptorRefs>
</configuration>
<executions>
```

```
<execution>
<id>make-assembly</id>
<phase>package</phase>
<goals>
<goal>single</goal>
</goals>
</execution>
</executions>
</plugin>
</plugins>
</build>
```

如果您的 Maven 配置正确并且成功的导入了依赖包，那么整个工程应该没有错误可以直接编译。在本地命令行模式下进入工程目录，执行下面的命令对整个工程进行打包：

```
mvn package
```

运行过程中可能还需要下载一些文件，直到出现 build success 表示打包成功。然后您可以在工程目录下的 target 文件夹中看到打好的 jar 包。

数据准备

首先需要把压缩好的 jar 包上传到 EMR 集群中，使用 scp 或者 sftp 工具来进行上传。在本地命令行模式下运行：

```
scp $localfile root@公网IP地址:$remotefolder
```

其中，是您的本地文件的路径加名称；为服务器用户名；公网可以在控制台的节点信息中或者在云服务器控制台查看；localfile 是您的本地文件的路径加名称；root 为 CVM 服务器用户名；公网 IP 可以在 EMR 控制台的节点信息中或者在云服务器控制台查看；remotefolder 是您想存放文件的 CVM 服务器路径。上传完成后，在 EMR 命令行中即可查看对应文件夹下是否有相应文件。
需要处理的文件需要事先上传到 COS 中。如果文件在本地则可以通过 COS 控制台直接上传。如果文件在 EMR 集群上，可以使用 Hadoop 命令上传。指令如下：

```
[hadoop@10 hadoop]$ hadoop fs -put $testfile cosn://BUCKET_NAME/
```

运行样例

首先需要登录 EMR 集群中的任意机器，最好是登录到 Master 节点。登录 EMR 的方式请参考 录 Linux 实例。这里我们可以选择使用 WebShell 登录。单击对应云服务器右侧的登录，进入登录界面，用户名默认为 root，密码为创建 EMR 时用户自己输入的密码。输入正确后，即可进入命令行界面。
在 EMR 命令行先使用以下指令切换到 Hadoop 用户：

```
[root@172 ~]# su hadoop
[hadoop@172 ~]$ flink run -m yarn-cluster -c com.tencent.flink.CosWordcount ./fli
nk-example-1.0-SNAPSHOT.jar cosn://$BUCKET_NAME/test/data.txt cosn://$BUCKET_NAM
E/test/result
[hadoop@172 ~]$ hdfs dfs -cat cosn://becklong-cos/test/result
(Flink,8)
(Hadoop,3)
(Spark,7)
(Hbase,3)
```

RSS 开发指南

最近更新时间：2023-02-23 14:17:19

RSS 简介

Apache Uniffle 是用于计算引擎的统一远程 Shuffle 服务（RSS，Remote Shuffle Service），它具有在远程服务器上聚合和存储 Shuffle 数据的能力，从而极大地提高了大型作业的性能和可靠性。

背景介绍

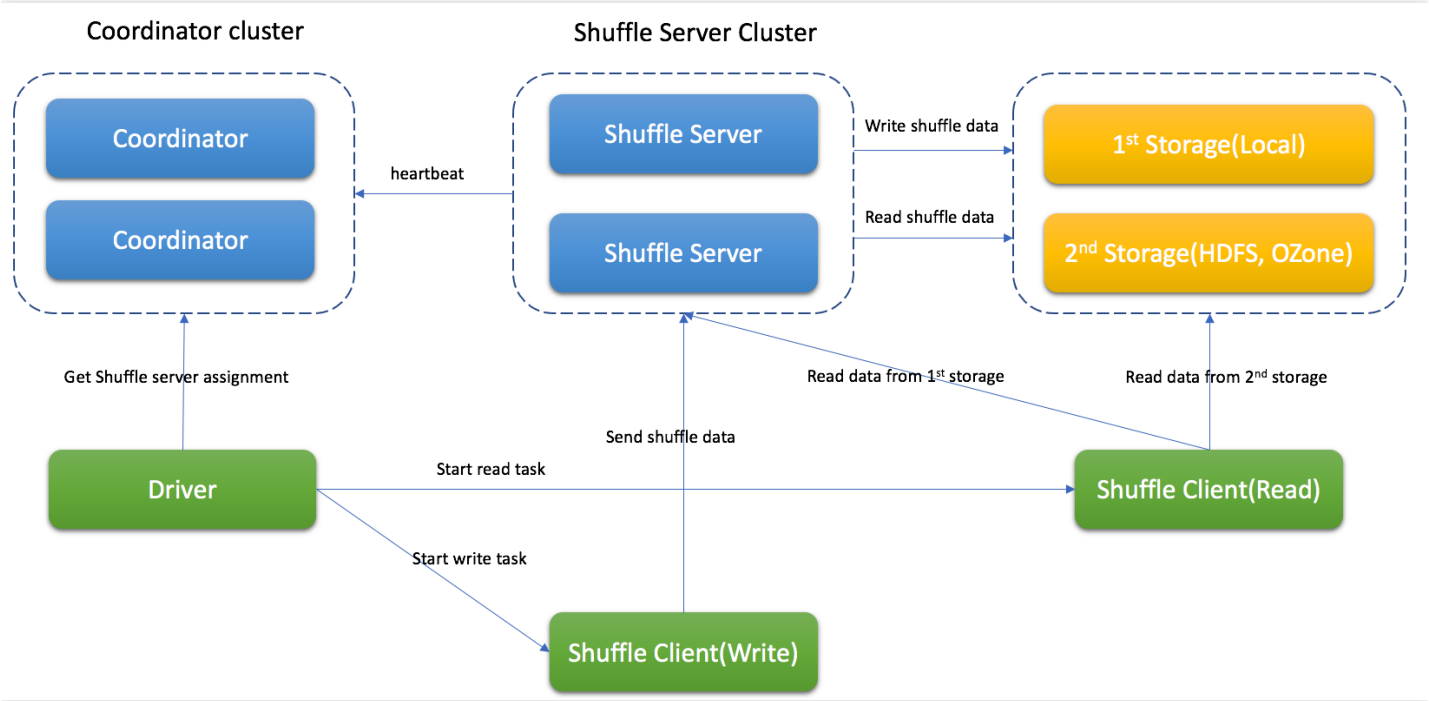
现有 Shuffle 方案存在的问题：

- 现有 Shuffle 无法使计算做到 serverless，在机器资源被抢占时会导致 Shuffle 数据的重新计算。
- Shuffle Partition 会导致大量的 Shuffle 连接和网络小包，大规模场景下极容易发生超时问题。
- Shuffle 过程存在写放大和随机 IO 问题，当 Shuffle 数据过大时，会严重影响集群性能和稳定性。
- Spark Shuffle 服务和 NodeManager 在同一进程，IO 负载较高时极易导致 NodeManager 重启，影响 Yarn 调度。

RSS 基本特点

1. 将 Shuffle 数据存储在远程服务器，支持计算存储分离，计算存储混布等集群部署模式。
2. 支持 Shuffle 数据聚合和数据缓存机制，最大化利用内存资源，降低对于磁盘的随机访问。
3. 支持 Shuffle 数据的多种存储方式，如本地文件，HDFS 文件及混合模式等。
4. 支持 Shuffle 数据的正确性校验，过滤无效数据的同时，保证任务计算过程中的数据正确性。
5. 采用主从架构和主备多活模式，提升了集群资源利用率和服务的稳定性。

RSS 基本架构



其中，各个组件的功能如下：

- Coordinator，基于心跳机制管理 Shuffle Server，存储 Shuffle Server 的资源使用等元数据信息，还承担任务分配职责，根据 Shuffle Server 的负载，分配合适的 Shuffle Server 给 Spark 应用处理不同的 Partition 数据。
- Shuffle Server，主要负责接收 Shuffle 数据，聚合后再写入存储中，基于不同的存储方式，还能用来读取 Shuffle 数据（如 LocalFile 存储模式）。
- Shuffle Client，主要负责和 Coordinator 和 Shuffle Server 通讯，发送 Shuffle 数据的读写请求，保持应用和 Coordinator 的心跳等。

RSS 使用

RSS 集群提供远程 Shuffle 服务，无法单独使用，需要 Spark 容器集群关联后使用。
关联方法为：单击 **Spark 集群 > 集群信息 > 关联 RSS 集群**，然后选取处于运行中的 RSS 集群即可。

RSS 配置

您可在**集群服务 > 配置管理**页面查看 RSS 不同角色的配置文件和常见配置项内容。
Coordinator 角色的配置文件有 coordinator.conf 和 log4j.properties，Shuffle Server 角色有 server.conf 和 log4j.properties 文件。
coordinator.conf 常见配置项，不建议修改：

参数	默认值	描述
rss.coordinator.app.expired	60000	Application 过期时
rss.coordinator.dynamicClientConf.enabled	false	是否开启动态客户 spark 客户端获取
rss.coordinator.exclude.nodes.file.path	file:///usr/local/service/rss/conf/exclude_nodes	排除节点的配置文
rss.coordinator.server.heartbeat.timeout	30000	如果无法从 shuffl 跳，则超时
rss.coordinator.shuffle.nodes.max	1000	分配时最大 Shuff
rss.jetty.http.port	19998	Coordinator 的 H
rss.rpc.server.port	19999	Coordinator 的 RI
rss.storage.type	MEMORY_LOCALFILE	RSS 存储类型，支 MEMORY_ONLY MEMORY_LOCA MEMORY_LOCA 由于无可用的 Ha 前默认 MEMORY

server.conf 常见配置项，不建议修改：

参数	默认值	描述
rss.coordinator.quorum	rss-coordinator-rss-<集群Id>-0:19999,rss-coordinator-rss-<集群Id>-1:19999	Coordinator 地址信息
rss.jetty.http.port	19998	Shuffle Server 的 HTTP 端口
rss.rpc.server.port	19999	Shuffle Server 的 RPC 端口
rss.server.buffer.capacity	内存 Limit 值 * 0.75 * 0.6	Shuffle Server 缓冲区管理器的最大内存
rss.server.disk.capacity	单个数据盘容量*0.9	Shuffle Server 可以使用的磁盘容量
rss.server.flush.thread.alive	数据盘数量	将数据刷新到文件的线程数

参数	默认值	描述
rss.server.flush.threadPool.size	数据盘数量*2	将数据刷新到文件的线程池大小
rss.server.heartbeat.interval	10000	到 Coordinator 的心跳间隔
rss.server.heartbeat.timeout	60000	心跳超时时间
rss.server.read.buffer.capacity	内存 Limit 值 * 0.75 * 0.2	读取数据的最大缓冲区大小
rss.storage.basePath	/data1/rssdata,/data2/rssdata... 路径个数默认等于数据盘数量	Shuffle 数据写入数据盘的路径
rss.storage.type	MEMORY_LOCALFILE	RSS存储类型，需与 Coordinator 保持一致

更多信息可参考 incubator-uniffle。