

大数据交互式查询实验



华为技术有限公司

目录

1 理论知识储备	3
1.1 Presto 简介	3
1.2 Presto 架构	3
1.3 Presto 数据模型	4
1.4 Presto 多实例	4
1.5 总结	5
2 实验环境介绍	5
2.1 实验介绍	5
2.1.1 关于本实验	5
2.1.2 实验组网介绍	5
2.1.3 实验设备介绍	5
3 大数据交互式查询	7
3.1 实验介绍	7
3.1.1 关于本实验	7
3.1.2 实验目的	7
3.1.3 实验流程	7
3.1.4 实验数据格式说明	7
3.2 实验任务	8
3.2.1 集群搭建	8
3.2.2 使用 SSH 远程工具连接 MRS 的主节点	18
3.2.3 数据准备	20
3.2.4 Hive 查询	31
3.2.5 Presto 查询	33
3.3 思考题	34
4 附录	35
4.1 释放 MapReduce 服务 MRS	35
4.2 释放对象存储服务 OBS	36



4.3 释放 VPC 相关资源 39

1 理论知识储备

1.1 Presto 简介

以 MapReduce 为底层计算框架的 Hive 是专门为批处理设计的，随着数据量的增大，已经不能满足大数据快速实时 adhoc 查询计算的性能要求，Facebook 2012 年开发了 Presto，并于 2013 年正式宣布开源。

Presto 是一个 facebook 开源的用户交互式分析查询的 SQL 查询引擎，用于针对各种大小的数据源进行交互式分析查询。其主要应用于海量结构化数据/半结构化数据分析、海量多维数据聚合/报表、ETL、Ad-Hoc 查询等场景。

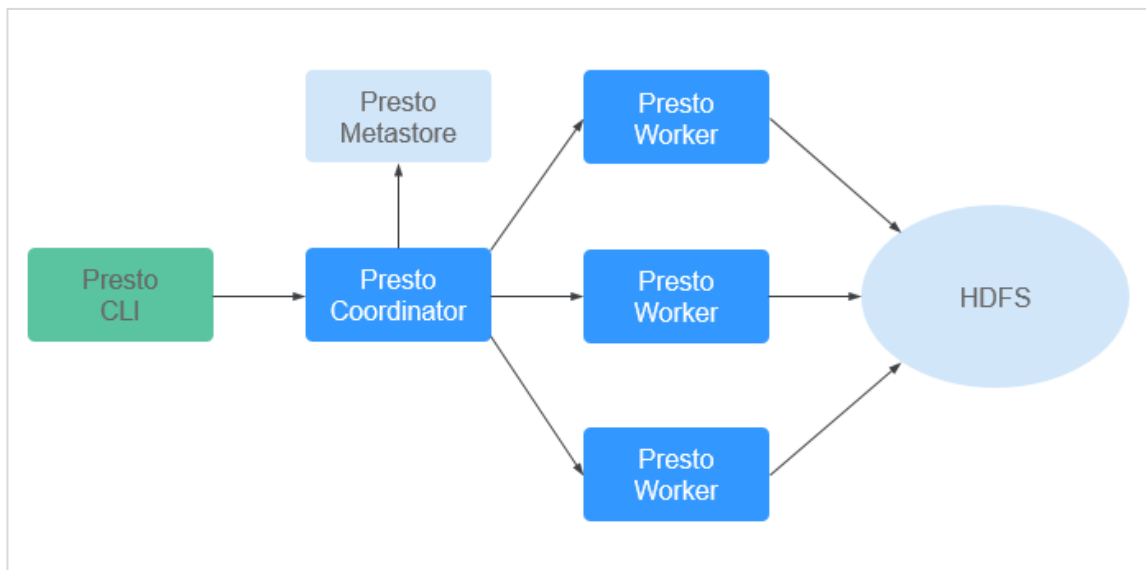
Presto 允许查询的数据源包括 Hadoop 分布式文件系统 (HDFS)，Hive，HBase，Cassandra，关系数据库甚至专有数据存储。一个 Presto 查询可以组合不同数据源，执行跨数据源的数据分析。

Presto 官网：<https://prestodb.io/>

1.2 Presto 架构

Presto 的架构是由关系型数据库的架构演化而来的，它的查询引擎是一个典型的 Master-Slave 的模型。

Presto 分布式地运行在一个集群中，包含一个 Coordinator 和多个 Worker 进程，查询从客户端（例如 CLI）提交到 Coordinator，Coordinator 进行 SQL 的解析和生成执行计划，然后分发到多个 Worker 进程上执行。



1.3 Presto 数据模型

Presto 采取三层结构管理数据：

- **Catalog**：对应数据源。例如 Hive 是一类数据源，MySQL 是一类数据源。一个 Catalog 可以包含多个 Schema。
- **Schema**：对应数据库实例。一个 Schema 包含多张 Table。
- **Table**：对应数据表。就是一般意义上的数据库表。

在 Presto 中通过表的完全限定名来定位表：catalog.schema.table，例如：

hive.webdata.user，表示 hive 为 catalog，webdata 为 schema，user 为 table。

1.4 Presto 多实例

在 MRS(MapReduce Service)中支持为大规格的集群默认安装 Presto 多实例，即一个 Core/Task 节点上安装多个 Worker 实例，分别为 Worker1，Worker2，Worker3...，多个 Worker 实例共同与 Coordinator 交互执行计算任务，相比较单实例，能够大大提高节点资源的利用率和计算效率。

Presto 多实例仅作用于 ARM 架构规格，当前单节点最多支持 4 个实例。

1.5 总结

Presto 是一个分布式 SQL 查询引擎，数据量支持 GB 到 PB 字节，其本身已经内置了多个常见数据库类型的连接，并且可以在多个不同类型的数据源之间进行关联 JOIN 查询，其查询速度相对于 Hive 来说快很多。但由于 Presto 查询基于内存，所以对于多个大表的 join 操作来说可能比较慢，这种情景使用 Hive 更加合适。

2 实验环境介绍

2.1 实验介绍

2.1.1 关于本实验

本实验指导书适用于希望了解大数据知识，掌握如何将大数据技术融合与具体实践的读者。

本实验指导书主要指导用户如何基于已有的 YouTube 视频统计的数据集，根据业务需求对数据进行交互式查询。

本次实验的数据是关于 YouTube 视频信息的数据(本实验选取 CA 国家数据)，我们需要使用 Presto 结合 Hive 按照业务需求进行查询。

2.1.2 实验组网介绍

本实验环境基于公有云实现。

2.1.3 实验设备介绍

为了满足本实验需要，使用以下云服务及相关资源或工具：

表2-1 实验设备配套关系

名称	版本	说明
MRS	2.1.0	MapReduce服务
OBS	无	对象存储服务
PuTTY	无	SSH远程连接工具



3 大数据交互式查询

3.1 实验介绍

3.1.1 关于本实验

本实验通过对 YouTube 视频数据集，使用大数据组件和技术来处理视频数据，让学生能够学习到大数据交互式分析的步骤和流程。

本实验通过大数据公有云服务(MapReduce Service)MRS 来构建 Presto+Hive 交互式查询环境，按照业务需求进行相关的交互式分析并显示结果。

3.1.2 实验目的

- 理解交互式查询及其实践。
- 掌握 OBS 对象存储。
- 掌握基于云服务搭建大数据环境。
- 掌握 Presto+Hive 的交互式查询。
- 掌握基于云服务的组件管理。

3.1.3 实验流程

- 把待处理数据集存储到 OBS 中；
- 把 OBS 中的数据集导入到 MRS 集群的 HDFS 上；
- 在 Hive 中创建表并把数据加载到表中；
- 使用 Presto 集成 Hive；
- 通过 Presto 客户端根据业务需求完成查询。

3.1.4 实验数据格式说明

表3-1 实验数据格式

字段	含义	备注
----	----	----

video_id	视频ID	唯一的ID, 11位字符串
trending_date	榜单日期	字符串
title	标题	字符串
channel_title	频道标题	视频所属频道的标题, 字符串
category_id	类别ID	视频所属类别编号, 整型
publish_time	发布时间	视频的发布时间, 时间类型
tags	标签	视频的标签, 字符串
views	观看数	视频被观看的次数, 整型
likes	点赞数	视频被点赞的次数, 整型
dislikes	被踩数	视频被踩的次数, 整型
comment_count	评论数	视频被评论的次数, 整型
thumbnail_link	缩略图	视频的缩略图链接, 字符串
comments_disabled	评论禁用	视频评论功能是否被禁用, 布尔型
ratings_disabled	打分禁用	视频打分功能是否被禁用, 布尔型
video_error_or_removed	错误或删除	视频是否出错或者被删除
description	描述	视频的详情

3.2 实验任务

3.2.1 集群搭建

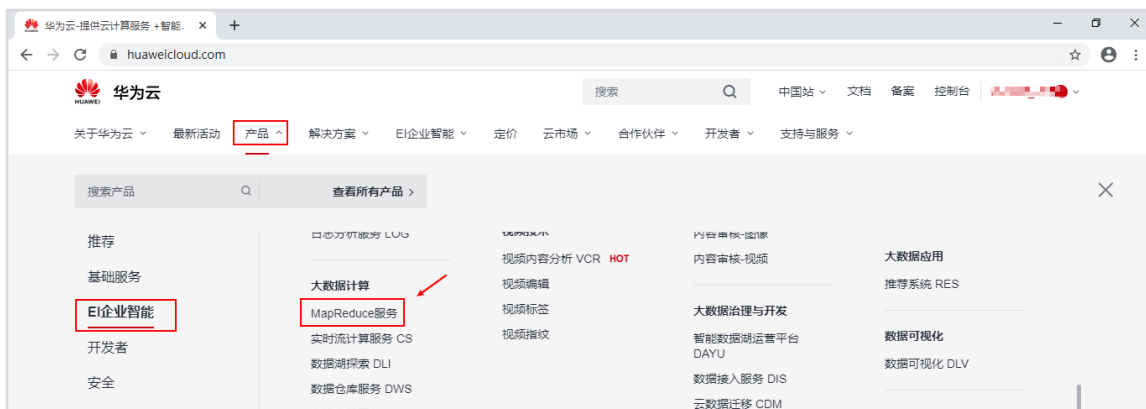
步骤 1 购买 MRS 服务

购买并配置 MRS

打开华为云官网首页, 点击右上角的“登录”完成登录



登录后依次点击“产品 > EI 企业智能 > MapReduce 服务”



选择“立即购买”



进入购买页面后选择“自定义购买”，选择区域、填写集群名称并选择集群版本号“MRS 2.1.0”

MapReduce服务

快速购买

自定义购买

1 软件配置

2 硬件配置

3 高级配置

区域

华北-北京四

不同区域的资源之间内网不互通。请选择靠近您客户的区域，可以降低网络延时、提高访问速度。[如何选择区域](#)

集群名称

mrs_presto

?

集群版本

MRS 2.1.0

集群类型选择“分析集群”，并选择必要的组件(Presto、Hive、Tez)

集群类型

分析集群

流式集群

混合集群

必选组件默认勾选，被依赖的组件会被自动勾选。

分析组件

<input type="checkbox"/>	组件名	版本	描述
<input checked="" type="checkbox"/>	Presto	308	一种开源、分布式SQL查询引擎。
<input checked="" type="checkbox"/>	Hadoop	3.1.1	针对大数据集的分布式数据存储和处理框架，包含HDFS、YARN、MapRe...
<input type="checkbox"/>	Spark	2.3.2	快速、通用的大数据处理引擎
<input type="checkbox"/>	HBase	2.1.1	可扩展、分布式数据库，支持存储结构化数据
<input checked="" type="checkbox"/>	Hive	3.1.0	提供数据汇聚和即席查询的数据仓库
<input type="checkbox"/>	Hue	3.11.0	提供hadoop UI能力，让用户通过浏览器分析处理Hadoop集群数据。
<input type="checkbox"/>	Loader	2.0.0	Loader是基于开源Sqoop 1.99.7开发，专为Apache Hadoop和结构化数据...
<input checked="" type="checkbox"/>	Tez	0.9.1	一个支持有向无环图的分布式计算框架。
<input type="checkbox"/>	Flink	1.7.0	一个分布式大数据处理引擎，可对有限数据流和无限数据流进行有状态计...

关闭 kerberos 认证，输入集群管理密码

Kerberos认证

☐

?

用户名

admin

密码

.....

该密码用于登录集群管理页面。

确认密码

.....

点击右下角的“下一步”

下一步

进入新页面继续配置

选择“按需计费”，选择可用区，选择虚拟私有云(没有可以新建)和子网，安全组选择“自动创建”，弹性公网IP选择“暂不绑定”



MapReduce服务 快速购买 自定义购买

① 软件配置 ② 硬件配置 ③ 高级配置

计费模式: **按需计费** 包年/包月

可用区: 可用区1 **可用区2** 可用区3 ?

虚拟私有云: vpc-c719 ? 查看虚拟私有云 ?

子网: subnet-c73f(192.168.0.0/24) ?

安全组: 自动创建 ? 管理安全组 ?

弹性公网IP: 暂不绑定 ? 管理弹性公网IP ?

CPU架构选择“鲲鹏计算”，关闭“集群高可用”，分析Core节点默认选择“3”



CPU架构: x86计算 **鲲鹏计算**

集群节点	节点类型	计费模式	实例规格	实例数量
Master节点 ?	Master节点 ?	按需计费	鲲鹏通用计算增强型 4 vCPUs 16 GB kc1.xlarge.4 系统盘 高IO 100 GB x 1 数据盘 高IO 200 GB x 1	1 集群高可用 <input type="checkbox"/>
分析Core节点 ?	分析Core节点 ?	按需计费	鲲鹏通用计算增强型 4 vCPUs 16 GB kc1.xlarge.4 系统盘 高IO 100 GB x 1 数据盘 高IO 100 GB x 1	<input type="button" value="-"/> 3 <input type="button" value="+"/>
分析Task节点 ?	分析Task节点 ?	按需计费		<input type="button" value="+"/>

登录方式选择“密码”并输入密码

登录方式

密码

密钥对

用户名

root

密码

该密码用于远程登录ECS机器。

确认密码

点击右下角的“下一步”

上一步

下一步

进入新的配置页面继续配置

标签、弹性伸缩、引导操作均默认

MapReduce服务

快速购买

自定义购买

① 软件配置

② 硬件配置

③ 高级配置

标签

你还可以添加10个标签。

弹性伸缩

请先返回上一步设置Task节点规格后再设置弹性伸缩策略。

引导操作

名称	执行节点	执行时机	操作
<div>添加</div>			

引导操作添加的脚本个数不能超过18个。

委托、数据盘加密默认，警告选择“关闭”

委托

暂不绑定

MRS_ECS_DEFAULT_AGENCY

现有委托

数据盘加密

关闭

开启

告警

关闭

开启

集群运行异常或系统故障时，维护人员可根据告警信息定位问题原因，建议开启。

点击“立即购买”

上一步

立即购买

查看创建的集群

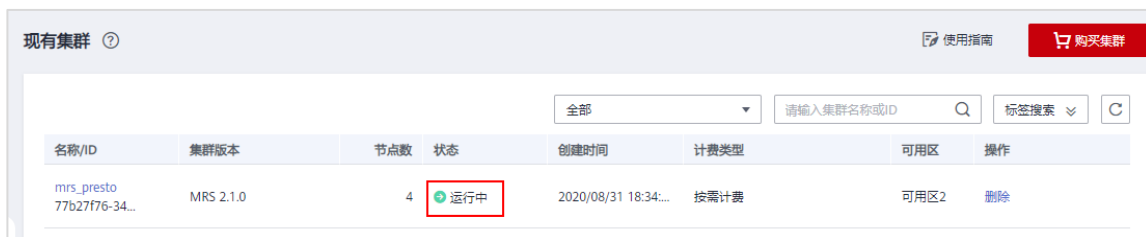
点击“返回集群列表”



可以看到刚购买的集群正在创建中



集群创建需要一个过程，等待状态变为“运行中”即可使用



步骤 2 集群管理

绑定公网 IP

在“现有集群”列表中，点击前面创建的集群名称



在集群页面点击“节点管理” (在此可以点击不同的标签进行相关管理)



展开 “master_node_default_group” 分组，点击 master 的节点名称



在节点内点击 “弹性公网 IP”，然后点击 “查看弹性公网 IP”



在弹出的网络控制台中点击右上角的 “购买弹性公网 IP”



选择 “按需计费”，选择区域(同一个实验中购买的所有产品区域要选择同一个)，线路选择 “全动态 BGP”，公网带宽选择 “按流量计费”，带宽大小默认

购买弹性公网IP

计费模式

包年/包月

按需计费

区域

华北-北京四

弹性公网IP仅支持绑定在处于相同区域的云资源上。购买后不能更换区域，请谨慎选择。

线路

全动态BGP

静态BGP

不低于99.95%可用性保障

公网带宽

按带宽计费

流量较大或较稳定的场景

按流量计费

流量小或流量波动较大场景

加入共享带宽

多业务流量错峰分布场景

指定带宽上限，按实际使用的出公网流量计费，与使用时间无关。

带宽大小

5

10

20

50

100

自定义

带宽范围：1-300 Mbit/s

免费开启DDoS基础防护

默认或指定带宽名称，监控默认，购买量为“1”

带宽名称

bandwidth-bc24

高级配置

标签

监控

默认开启基础监控

免费

免费提供分钟级粒度的流量监控

监控带宽流量波动、出入网带宽速率等指标详情

购买量

1

一次最多可以购买20个弹性公网IP。您还可以购买20个弹性公网IP，如需申请更多配额请点击[申请扩大配额](#)。

点击右下角的“购买”

立即购买

确认详情，点击右下角的“提交”

详情

产品类型	产品规格	计费模式	数量	价格
弹性公网IP	<div>区域</div> <div>北京四</div> <div>类型</div> <div>全动态BGP</div> <div>标签</div> <div>--</div>	按需计费	1	¥0.02/小时
带宽	<div>带宽名称</div> <div>bandwidth-bc24</div> <div>带宽类型</div> <div>独享带宽</div> <div>计费方式</div> <div>按流量计费</div> <div>带宽大小</div> <div>5 Mbit/s</div>	按需计费	1	¥0.80/GB

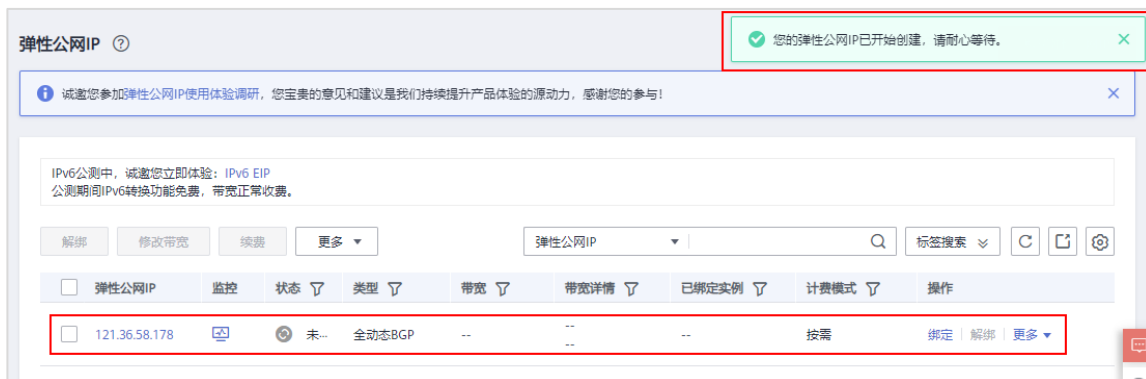
弹性公网IP费用 ¥0.02/小时 + 公网流量费用 ¥0.80/GB

【绑定实例后不收取IP费用】 参考价格，具体扣费请以账单为准。 [了解计费详情](#)

上一步

提交

购买成功，关闭此页面



回到节点页面，点击“绑定弹性公网 IP”



弹出对话框中，选择刚购买的弹性公网 IP，点击“确定”



绑定成功



点击右上角刷新按钮可以看到绑定成功的公网 IP 以及 IP 地址

77b27f76-3420-4966-97ea-a42f3b349865_node_master1Cyxq

开机 关机 重启 远程登录 更多 C

基本信息 云硬盘 网卡 安全组 弹性公网IP 监控 标签

绑定弹性公网IP 查看弹性公网IP

121.36.58.178 | 192.168.0.230 解绑

弹性公网IP 121.36.58.178 流量详情

ID 09a245f1-dd7b-45c7-9384-5cd4d550d49b

类型 全动态BGP

创建时间 2020/08/31 19:12:09 GMT+08:00

状态 绑定

带宽名称 bandwidth-bc24

已绑定私有IP 192.168.0.230

带宽大小 5 Mbit/s

订单详情 --

带宽类型 独享

带宽ID 0480c0a8-8bbf-455a-92b5-b2f61999f976 流量详情

到期时间 --

计费模式 按需计费

添加组策略

点击节点中的“安全组”，然后展开安全组信息，点击右侧的“更改安全组规则”

77b27f76-3420-4966-97ea-a42f3b349865_node_master1Cyxq

开机 关机 重启 远程登录 更多 C

基本信息 云硬盘 网卡 安全组 弹性公网IP 监控 标签

更改安全组

mrs_mrs_presto_Fssz NIC1: 192.168.0.230

出方向规则 1 入方向规则 9 ID 9d4202fa-cda9-4e93-a55a-03a563acb908

更改安全组规则

方向	类型	协议	端口范围/ICMP类型	远端
入方向	IPv4	Any	Any	192.168.0.75/32
入方向	IPv4	Any	Any	192.168.0.108/32
入方向	IPv4	TCP	9022	100.125.1.45/32

在弹出的页面上选择“入方向规则”，然后点击“添加规则”

mrs_mrs_presto_Fssz

导入规则 导出规则

基本信息 入方向规则 出方向规则 关联实例

添加规则 快速添加规则 删除 一键放通 入方向规则: 9 教我设置

协议端口	类型	源地址	描述	操作
全部	IPv4	192.168.0.108/32	--	修改 复制 删除
全部	IPv4	192.168.0.75/32	--	修改 复制 删除
全部	IPv4	198.19.32.0/19	MRS 默认安全组规则	修改 复制 删除

在弹出窗口中点击小三角，选择“基本协议 > 全部方通”，点击“确定”

添加方向规则 教我设置

安全组 mrs_mrs_presto_Fssz

如您要添加多条规则，建议单击导入规则以进行批量导入。

协议端口 ?	源地址 ?	描述	操作
<div>TCP</div> <div>基本协议</div> <div>常用协议端口</div> <div>全部放通</div> <div>全部TCP</div> <div>全部UDP</div> <div>自定义TCP</div> <div>自定义UDP</div> <div>ICMP</div>	<div>IP地址</div> <div>0.0.0.0/0</div> <div>增加1条规则</div>		复制 删除

确定

取消

规则添加成功，关闭此页面和安全组页面

mrs_mrs_presto_Fssz

导入规则 导出规则

基本信息

入方向规则

出方向规则

关联实例

添加规则

快速添加规则

删除

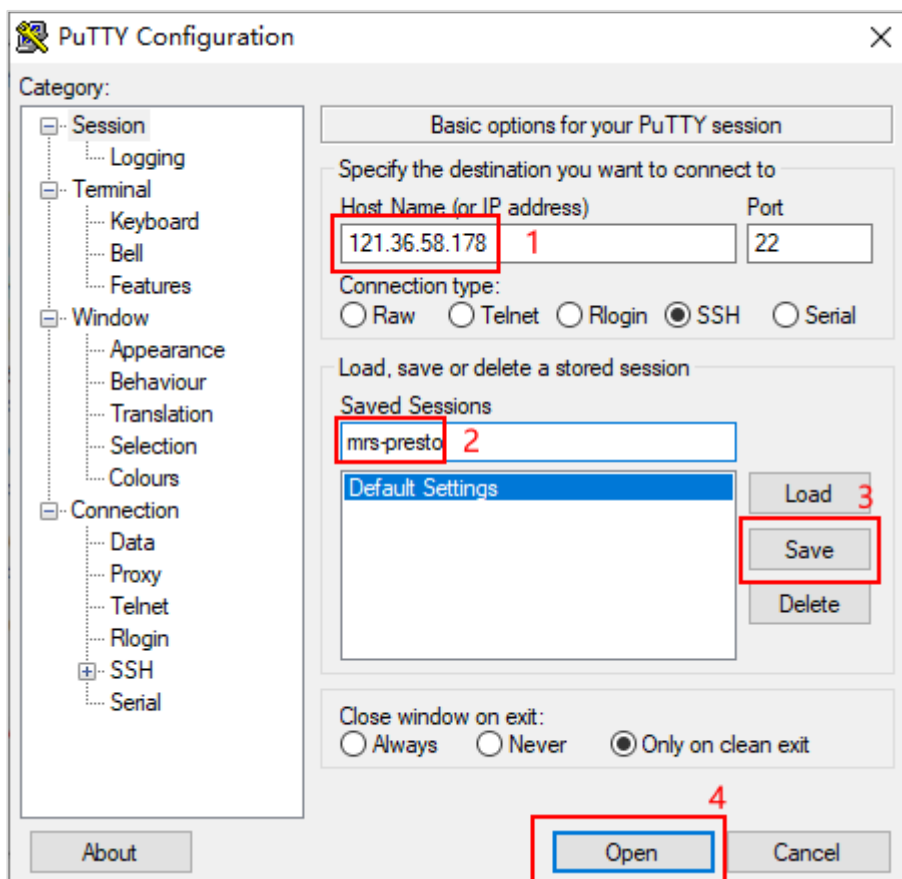
一键放通

入方向规则: 10 教我设置

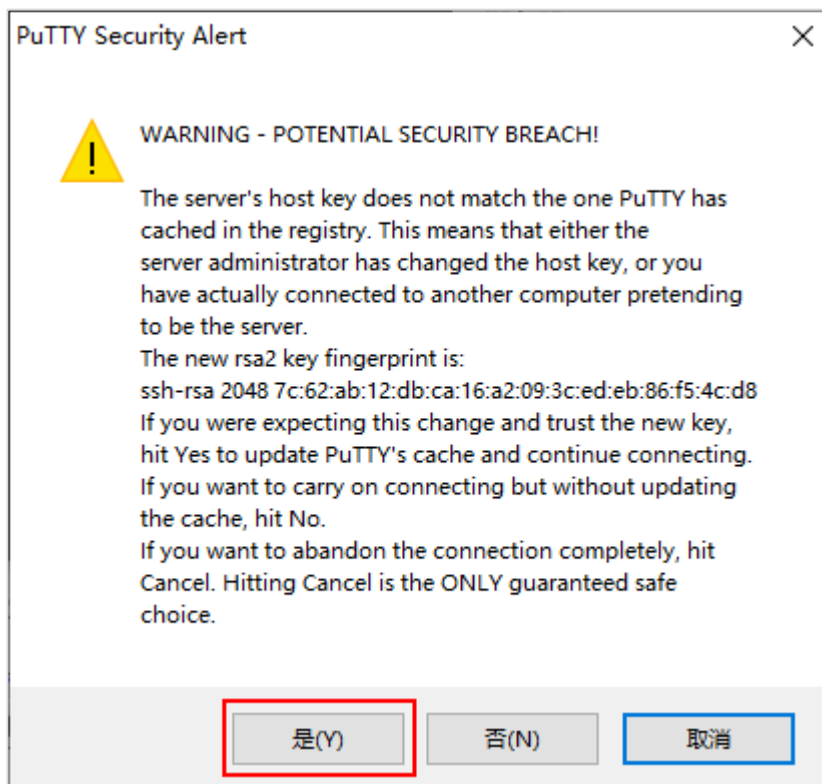
协议端口 ?	类型	源地址 ?	描述	操作
<input type="checkbox"/> 全部	IPv4	0.0.0.0/0	--	修改 复制 删除
<input type="checkbox"/> 全部	IPv4	192.168.0.108/32	--	修改 复制 删除
<input type="checkbox"/> 全部	IPv4	192.168.0.75/32	--	修改 复制 删除

3.2.2 使用 SSH 远程工具连接 MRS 的主节点

启动 PuTTY，选择 Session，输入 MRS 集群 master 节点绑定的弹性公网 IP 的 IP 地址，输入 Sessions 的名字，点击“Save”，保存会话后点击“Open”



弹出对话框选择“是”



输入用户名和密码后登录成功(用户名 root, 密码为远程登录 ECS 的密码)

```
login as: root
root@121.36.58.178's password:
Last login: Thu Jan  1 08:00:10 1970
[root@node-master1Cyxq ~]#
```

执行命令配置环境变量

```
[root@node-master1Cyxq ~]# source /opt/client/bigdata_env
```

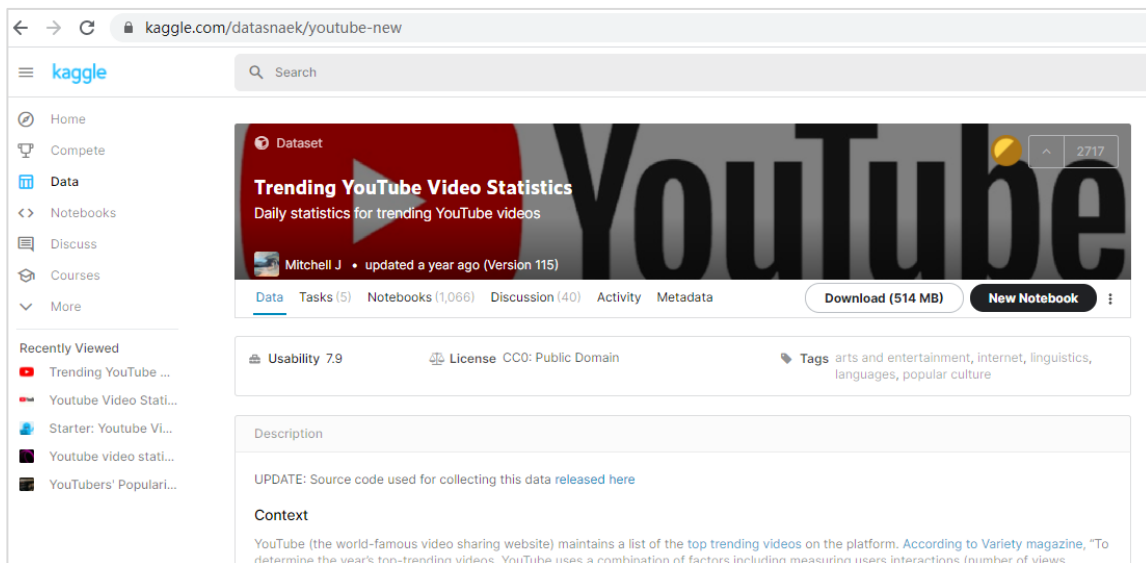
使用 HDFS 测试

```
[root@node-master1Cyxq ~]# hadoop fs -ls /
2020-08-31 19:42:47,277 INFO obs.OBSFileSystem: This Filesystem GC-ful, clear resource.
Found 9 items
drwxrwxrwx - hdfs  hadoop          0 2020-08-31 18:28 /app-logs
drwxrwxrwx - hive   hive            0 2020-08-31 18:31 /apps
drwxrwxrwx - hdfs  hadoop          0 2020-08-31 18:28 /ats
drwxr-xr-x - hdfs  hadoop          0 2020-08-31 18:28 /datasets
drwxr-xr-x - hdfs  hadoop          0 2020-08-31 18:28 /datastore
drwxrwxrwx - mapred hadoop          0 2020-08-31 18:28 /mr-history
drwxrwxrwt - spark  hadoop          0 2020-08-31 18:28 /sparkJobHistory
drwxrwxrwx - hdfs  hadoop          0 2020-08-31 18:31 /tmp
drwxrwxrwx - hdfs  hadoop          0 2020-08-31 18:31 /user
[root@node-master1Cyxq ~]#
```

3.2.3 数据准备

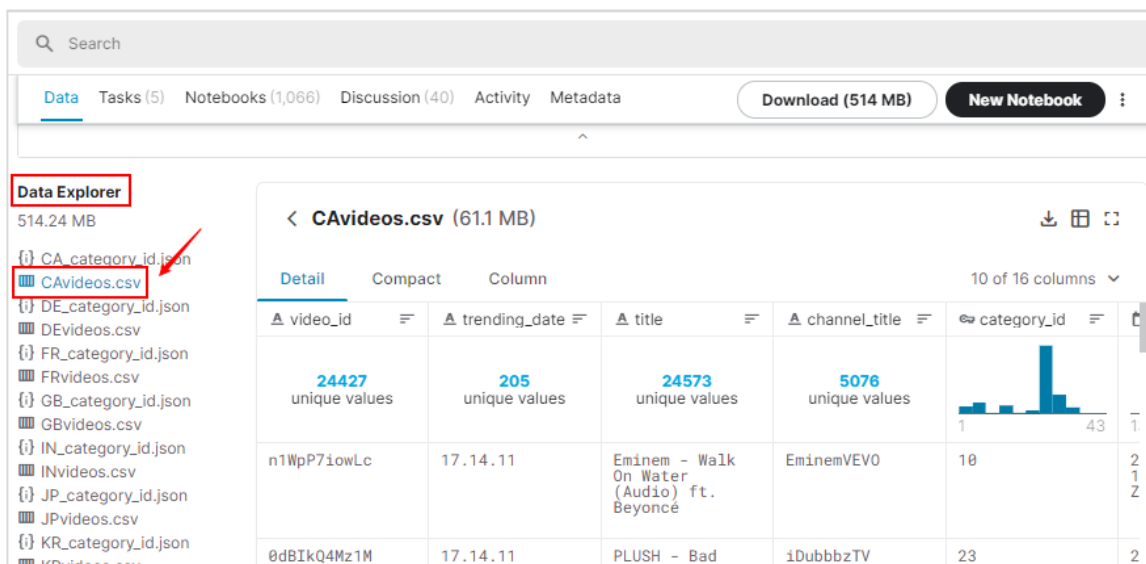
步骤 1 数据集下载

打开 kaggle 网站的数据集所在页面(<https://www.kaggle.com/datasnaek/youtube-new>), 可以看到数据集的相关信息



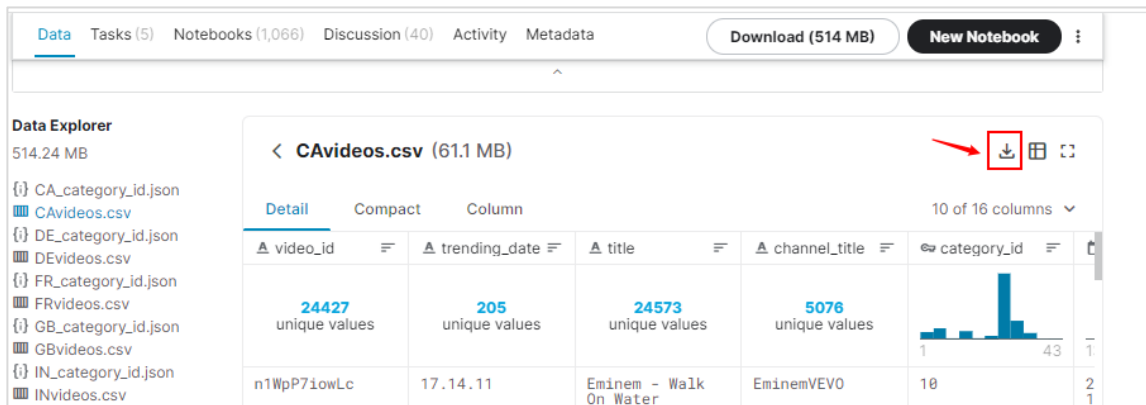
The screenshot shows the Kaggle dataset page for 'Trending YouTube Video Statistics' by Mitchell J. The page includes a sidebar with navigation links (Home, Compete, Data, Notebooks, Discuss, Courses, More), a search bar, and a list of recently viewed datasets. The main content area displays the dataset title, description, and download options. The dataset is described as 'Daily statistics for trending YouTube videos' and has a download size of 514 MB. The license is CC0: Public Domain. The tags include 'arts and entertainment, internet, linguistics, languages, popular culture'.

下拉滚动条到 “Data Explorer” ， 点击 “CAvideos.csv” 可以看到加拿大的数据集



The screenshot shows the 'Data Explorer' view for the 'CAvideos.csv' dataset (61.1 MB). The left sidebar lists various category ID JSON files and CSV files. The main area displays a table of video data with columns: video_id, trending_date, title, channel_title, and category_id. The table shows unique values for each column and a histogram for the category_id. The download button is highlighted with a red box and a red arrow.

点击右侧的下载按钮可以下载此数据集（点击顶部的 Download 可以下载所有的数据集）

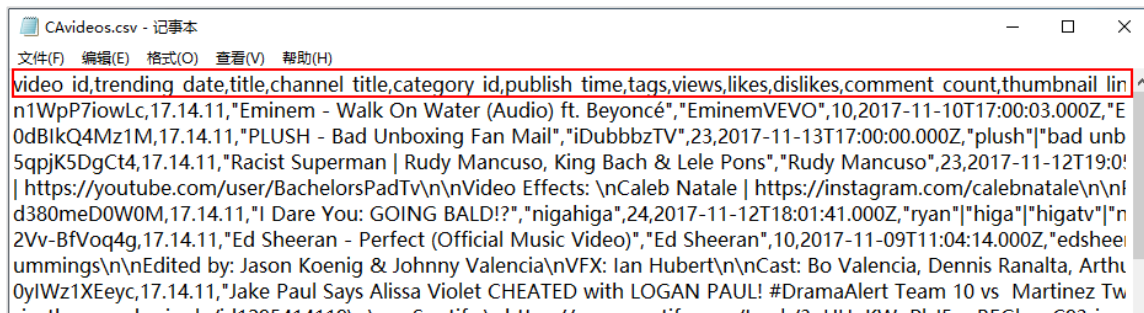


This screenshot is similar to the previous one, showing the 'Data Explorer' view for the 'CAvideos.csv' dataset. The download button is highlighted with a red box and a red arrow, indicating the action to download the dataset.

解压下载的压缩包，可以看到数据文件



打开文件可以看到数据集的数据，第一行是各字段名称



步骤 2 上传数据集到 OBS

创建 OBS 桶

在华为云首页依次点击 “产品 > 基础服务 > 对象存储服务 OBS”



在 OBS 页面点击 “管理控制台”



进入 OBS 控制台，选择右上角的 “创建桶”



选择区域，选择“多 AZ 存储”，输入桶的名称



其他选项默认，点击“立即创建”



弹出对话框选择“确定”



桶创建成功



上传数据集到桶

点击桶的名字进入桶管理页面



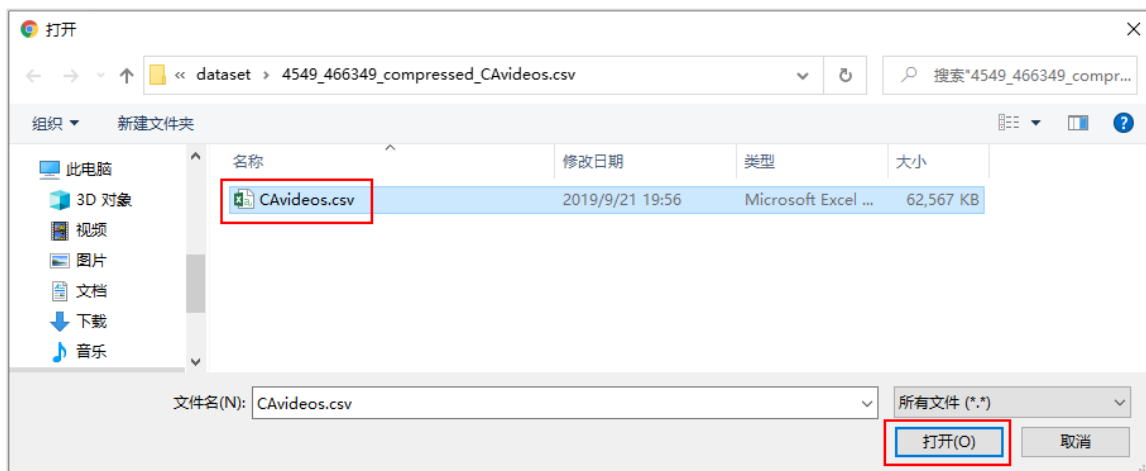
选择“对象”，然后点击“上传对象”



选择“标准存储”，点击“添加文件”



选中前面下载的数据集文件, 点击“打开”



点击“上传”

上传对象 超过5GB如何上传?

标准存储

低频访问存储

归档存储

适用于有大量热点文件或小文件，且需要频繁访问（平均一个月多次）并快速获取数据的业务场景。

对象默认与桶的存储类别相同，也可以根据适用场景修改。 [了解更多](#)

上传对象 注意：桶内如有同名文件/文件夹，将被新上传的文件/文件夹覆盖。

清空列表

添加文件

1/100 文件 大小 61.09 MB

名称	大小	操作
CAvideos.csv	61.09 MB	移除

加密 将文件加密成密文存储，加密后的文件不能修改加密状态。

☐ KMS加密

上传

取消

数据集文件已经上传到桶

对象 | 已删除对象 | 碎片

对象是数据存储的基本单位，在OBS中文件和文件夹都是对象。您可以上传任何类型（文本、图片、视频等）的文件，并在桶中对这些文件进行管理。 [了解更多](#)

上传对象

新建文件夹

恢复

删除

修改存储类别

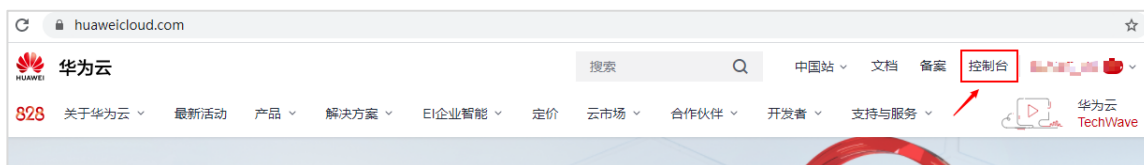
输入对象名前缀搜索

<input type="checkbox"/>	名称	存储类别	大小	加密状态	恢复状态	最后修改时间	操作
<input type="checkbox"/>	CAvideos.csv	标准存储	61.09 MB	未加密	--	2020/08/31 19:53:54...	下载 分享 更多

步骤 3 数据集上传到 HDFS

在 HDFS 上新建文件夹

点击华为云网页顶部的“控制台”进入到控制台



在控制台中展开左上角的服务列表，点击“EI 企业智能”下的“MapReduce 服务”



进到 MRS 控制台，在现有集群里点击集群名称



进入 mrs_presto 集群信息页面，依次点击“文件管理 > 新建”



输入文件夹名称，点击“确定”

×

新建文件夹

文件夹名称

youtubedata

命名规则

1. 不能为空
2. 不能以"."开头或结尾
3. 不能包括下列符号：/:?* "<"> \; &!{}[]\$%+
4. 不能超过255个字节
5. 开头和末尾的空格会被忽略。

确定

取消

上传数据

点击新建的文件夹名 youtubedata 进入文件夹

HDFS文件列表

文件操作记录

/ user /

新建

导入数据

导出数据

C

文件名	文件大小	修改时间	操作
-			
hive	--	2020/08/31 18:28:50 GMT+08:00	删除
mapred	--	2020/08/31 18:28:50 GMT+08:00	删除
omm	--	2020/08/31 18:31:46 GMT+08:00	删除
youtubedata	--	2020/08/31 20:05:04 GMT+08:00	删除

点击“导入数据”

←

mrs_presto

使用指南

下载认证凭证

管理操作

配置

运维

概览

节点管理

组件管理

告警管理

补丁管理

文件管理

作业管理

租户管理

备份恢复

引导操作

标签管理

MapReduce服务支持数据导入、导出，查看数据导入、导出进度，并将数据存储在指定目录中，目前只支持导入对象存储上的数据。[了解更多](#)

HDFS文件列表

文件操作记录

/ user / youtubedata /

新建

导入数据

导出数据

C

文件名	文件大小	修改时间	操作
..			

弹出对话框中点击“浏览”

从OBS导入数据至HDFS

OBS路径 ?

浏览

HDFS路径

/user/youtubedata/

浏览

确定

取消

点击刚建立的桶 “obs-youtube”

选择OBS文件

OBS

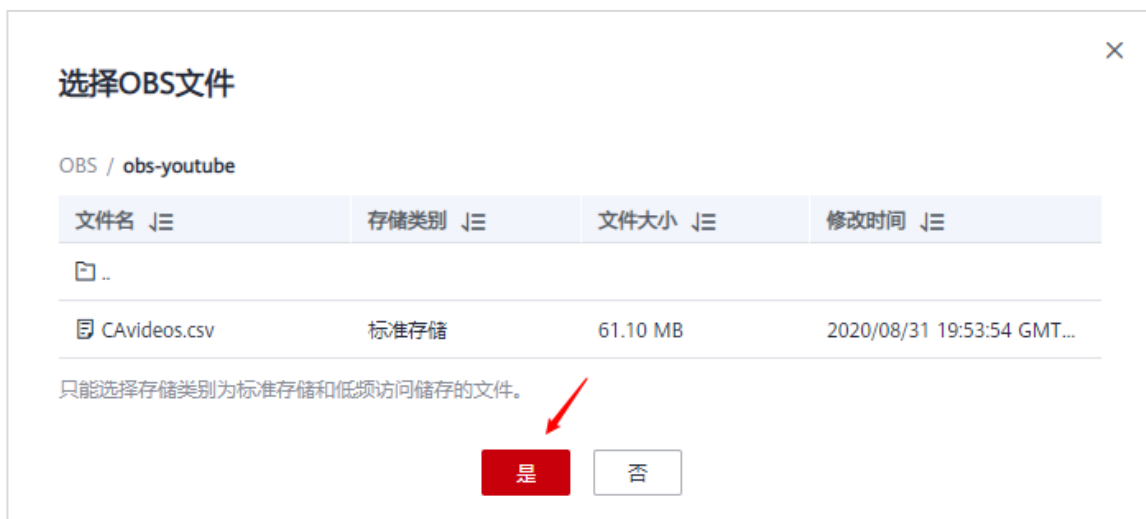
OBS桶 𠄎	存储类别 𠄎	文件大小 𠄎	修改时间 𠄎
obs-youtube	--	--	2020/08/31 19:50:52 GMT...

只能选择存储类别为标准存储和低频访问储存的文件。

是

否

#点击 “是”



查看信息无误，点击“确定”



提示“提交成功”，点击“确认”



稍后刷新即可看到数据集文件



3.2.4 Hive 查询

步骤 1 创建数据表

进入 beeline 工具

再次使用 PuTTY 工具连接 MRS 的 master 节点

```
login as: root
root@121.36.58.178's password:
Last login: Mon Aug 31 19:34:34 2020 from 119.3.119.19
[root@node-master1Cyxq ~]#
```

执行 Hive 组件的客户端命令 beeline --silent=true (--silent=true 为安静模式不打印额外信息)

```
[root@node-master1Cyxq ~]# beeline --silent=true
0: jdbc:hive2://192.168.0.154:2181,192.168.0.154>
```

使用 show tables 命令查看 Hive 中现有的表

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.154> show tables;
+-----+
| tab_name |
+-----+
+-----+
```

根据数据集的字段创建表 videos(建表语句可以写在一行)

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.154> create table videos(
0: jdbc:hive2://192.168.0.154:2181,192.168.0.154> video_id string,
```



```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> trending_date string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> title string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> channel_title string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> category_id int,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> publish_time string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> tags string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> views int,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> likes int,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> dislikes int,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> comment_count int,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> thumbnail_link string,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> comments_disabled boolean,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> ratings_disabled boolean,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> video_error_or_removed boolean,
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> description string)
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> row format delimited fields terminated by ",";
```

再次使用 show tables 命令查看 Hive 中现有的表

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> show tables;
+-----+
| tab_name |
+-----+
| videos   |
+-----+
```

步骤 2 加载数据到 Hive

使用 load data 命令把先前加入到 HDFS 中的数据载入到表中

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> load data inpath '/user/youtubedata/CAvideos.csv' into table
videos;
```

查看表中的记录数量，数据已经成功导入

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> select count(*) from videos;
+-----+
| _c0   |
+-----+
| 45802 |
+-----+
```

使用 !quit 命令退出 Hive

```
0: jdbc:hive2://192.168.0.154:2181,192.168.0.> !quit
2020-08-31 20:20:50,716 Thread-1 WARN Unable to register Log4j shutdown hook because JVM is shutting
down. Using SimpleLogger
[root@node-master1Cyxq ~]#
```

3.2.5 Presto 查询

步骤 1 使用 Presto 客户端

在 PuTTY 中直接输入 presto_cli.sh 进入 MRS 提供的 Presto 客户端

```
[root@node-master1Cyxq ~]# presto_cli.sh
--server http://192.168.0.230:7520
presto>
```

使用 show 命令显示所有的 catalog

```
presto> show catalogs;
Catalog
-----
hive
jmx
system
tpcds
tpch
(5 rows)
```

#使用 show 命令显示 hive 中的所有 schema

```
presto> show schemas from hive;
Schema
-----
default
information_schema
mrs_reserved
(3 rows)
```

使用 show 命令显示 default 中的所有 table

```
presto> show tables from hive.default;
Table
-----
videos
(1 row)
```

步骤 2 业务需求实现

统计出视频观看数的 TOP10

```
presto> select video_id,views from hive.default.videos order by views desc limit 10;
video_id | views
-----+-----
FlsCjmMhFmw | 137843120
```

```
FlsCjmMhFmw | 125431369
FlsCjmMhFmw | 113876217
FlsCjmMhFmw | 100911567
VYOjWnS4cMY | 98938809
6ZfuNTqbHE8 | 89930713
6ZfuNTqbHE8 | 87450245
VYOjWnS4cMY | 85092067
6ZfuNTqbHE8 | 84281319
7C2z4GqqS5E | 80738011
(10 rows)

Query 20200831_122228_00001_3n2hb, FINISHED, 3 nodes
Splits: 19 total, 19 done (100.00%)
0:06 [45.8K rows, 61.1MB] [7.56K rows/s, 10.1MB/s]
```

3.3 思考题

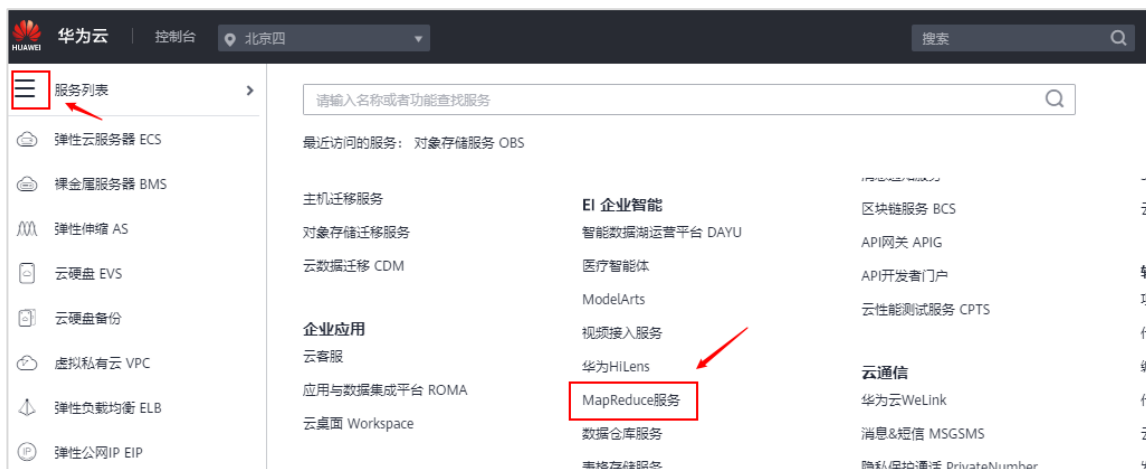
我们的现在使用的数据并没有经过清洗，应该如何处理呢？

业务：统计出视频观看数量最高的前 10 个视频的所属类别，如何实现呢？

4 附录

4.1 释放 MapReduce 服务 MRS

在控制台里点开服务列表，然后点击“MapReduce 服务”进入 MRS 的控制台



在“现有集群”中可以看到我们购买的集群，点击后面的“删除”链接进行删除



弹出的对话框中点击“是”



弹出提示信息，集群状态变为“删除中”



一段时间后集群被删除



4.2 释放对象存储服务 OBS

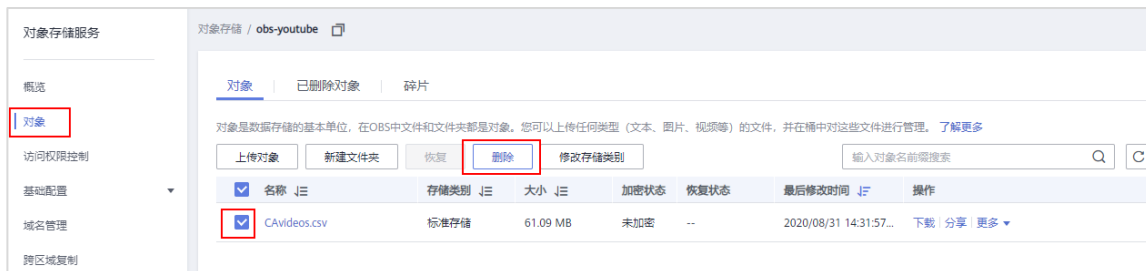
在控制台里点开服务列表，然后点击“对象存储服务 OBS”进入 OBS 的控制台



在 OBS 控制台的“对象存储”中，找到要删除的桶，点击桶的名字进入桶



在桶内选择“对象”，勾选里面所有的对象，点击“删除”按钮删除



弹出的对话框选择“是”



对象删除成功后点击“对象存储”返回桶列表



点击桶后面的“删除”



弹出对话框选择“是”，稍等桶则被删除



4.3 释放 VPC 相关资源

步骤 1 进入 VPC 控制台

在控制台里点开服务列表，然后点击“虚拟私有云 VPC”



进入到网络控制台



步骤 2 删除安全组

展开“访问控制”，点击“安全组”，然后点击对应安全组名字后面的“更多”里的“删除”



弹出对话框选择“是”



看到安全组删除成功的提示



步骤 3 删除弹性公网 IP

展开“弹性公网 IP 和带宽”，点击“弹性公网 IP”，然后点击对应弹性公网 IP 后面的“更多”里的“释放”



弹出对话框选择“是”



稍等可以看到释放成功的消息

