云服务实现大数据

实验手册



华为技术有限公司

目录

[1 实验介绍 2](#_Toc69399502)

[1.1 实验说明 2](#_Toc69399503)

[1.2 实验目的 2](#_Toc69399504)

[1.3 实验概览 2](#_Toc69399505)

[2 实验环境准备 3](#_Toc69399506)

[2.1 实验介绍 3](#_Toc69399507)

[2.2 服务购买 3](#_Toc69399508)

[2.2.1 登录控制台 3](#_Toc69399509)

[2.2.2 购买OBS桶 4](#_Toc69399510)

[2.2.3 购买MRS服务 6](#_Toc69399511)

[2.2.4 购买弹性公网IP 10](#_Toc69399512)

[3 云服务实现大数据 12](#_Toc69399513)

[3.1 实验介绍 12](#_Toc69399514)

[3.1.1 实验说明 12](#_Toc69399515)

[3.1.2 实验需求 12](#_Toc69399516)

[3.1.3 实验流程 12](#_Toc69399517)

[3.2 实验任务 12](#_Toc69399518)

[3.2.1 程序及数据准备 12](#_Toc69399519)

[3.2.2 执行程序 17](#_Toc69399520)

[3.3 结果验证 20](#_Toc69399521)

[3.3.1 查看作业日志 20](#_Toc69399522)

[3.3.2 使用组件管理 20](#_Toc69399523)

[3.4 思考题 24](#_Toc69399524)

[4 资源清理释放 25](#_Toc69399525)

[4.1 释放MapReduce服务 25](#_Toc69399526)

[4.2 释放对象存储服务OBS 26](#_Toc69399527)

[4.3 释放弹性公网IP 28](#_Toc69399528)

# 实验介绍

## 实验说明

本实验手册适用于希望了解基于公有云大数据基础知识，掌握如何将大数据技术融合与具体实践的读者。

本实验主要指导用户如何基于公有云云服务实现大数据，并能够基于云上大数据环境运行一个经典的MapReduce示例程序。实验的数据存储在Object Storage Service（以下简称OBS）中，在MapReduce Service（以下简称MRS）作业中直接使用，运算结果也存储在OBS中，可以下载后查看。

## 实验目的

理解云服务及其实践。

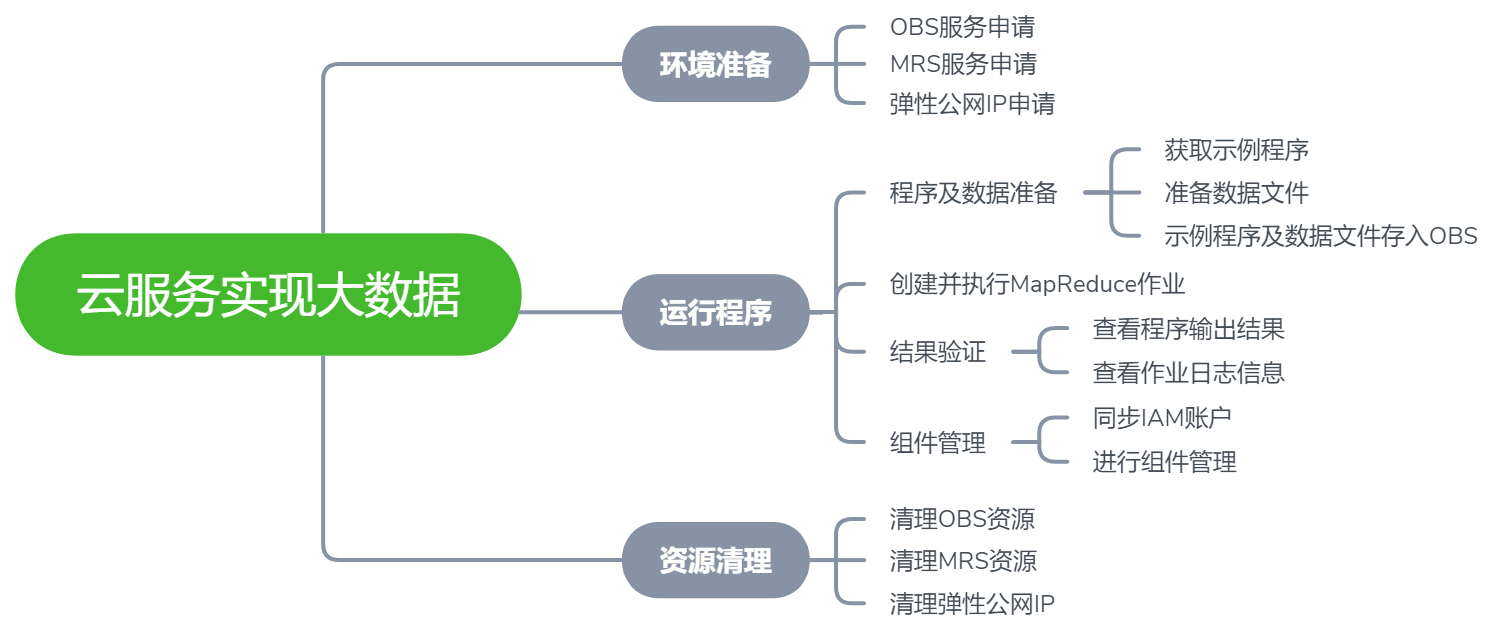
了解云服务的一般使用流程。

掌握基于云服务搭建大数据环境。

掌握基于MRS的作业管理及组件管理。

掌握云资源的清理。

## 实验概览



# 实验环境准备

## 实验介绍

本部分实验根据需要进行云服务的购买，为后续实验提供所需资源环境。

本实验需要购买OBS和MRS两个云服务以及一个弹性公网IP（EIP），具体规格如下：

实验设备配套关系

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 版本 | 规格 | 备注 |
| OBS | 无 | 多AZ | 标准存储 | 私有 |  |
| MRS | 1.9.2 | 分析集群 |  |
| EIP | 无 | 全动态BGP |按流量计费 |  |

## 服务购买

### 登录控制台

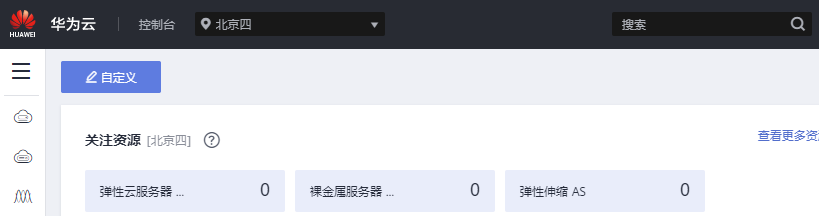
打开华为云官网首页（https://www.huaweicloud.com/），点击“登录”按钮后输入账号信息进行登录。



登录成功后点击“控制台”。



进入控制台后，选择区域为“北京四”。



### 购买OBS桶

鼠标放到控制台左上角三根横线处，弹出服务列表，点击“存储”分类里的“对象存储服务OBS”。



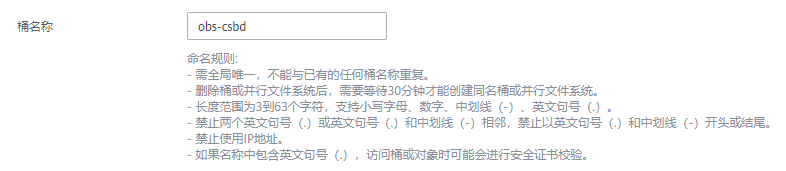
在对象存储界面点击“创建桶”。



区域为“华北-北京四”，冗余策略为“多AZ存储”。



按照规则输入桶名称“obs-csbd”（桶名称为全局唯一，请自行确定）。



存储类别“标准存储”，桶策略“私有”，默认加密、归档数据直读“关闭”。



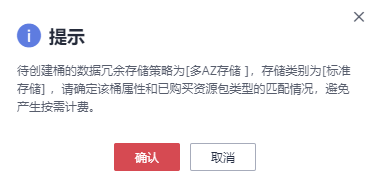
标签无需处理。



点击“立即创建”。



弹出对话框点击“确认”。



OBS桶创建成功。



### 购买MRS服务

在服务列表中点击“EI企业智能”分类下的“MapReduce服务”。



在现有集群界面点击“购买集群”。



选择“自定义购买”，区域选择“华北-北京四”。



集群名称“mrs\_csbd”（可自定义），版本“1.9.2”，类型为“分析集群”，组件默认勾选Hadoop即可。



点击“下一步”进入硬件配置。



计费模式“按需计费”，可用区、虚拟私有云、子网默认，安全组“自动创建”，弹性公网IP“暂不绑定”。

注：若无虚拟私有云，则点击后面的“查看虚拟私有云”进行创建。



CPU架构选择“鲲鹏计算”，集群节点默认。



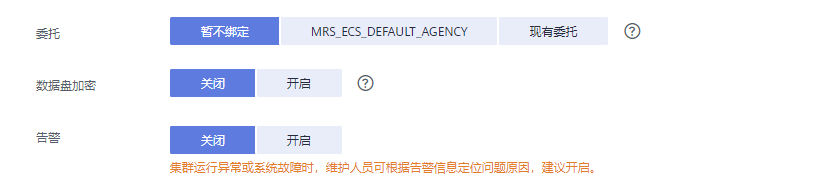
点击“下一步”进入高级配置。



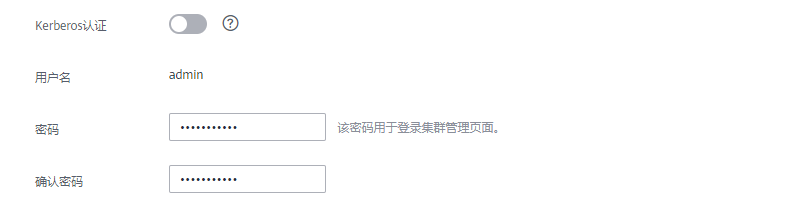
标签、弹性伸缩、引导操作默认。



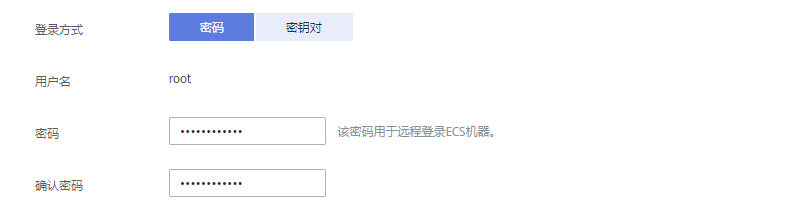
委托、数据盘加密默认，告警“关闭”。



关闭Kerberos认证（“Kerberos认证”关闭时，普通用户可使用MRS集群的所有功能，建议单用户场景下使用），输入密码（该密码用于登录集群管理页面）。



登录方式“密码”，输入密码（该密码用于远程登录集群节点的ECS机器）。



勾选“确认授权”。



点击“立即购买”。



点击“返回集群列表”。



等待集群创建。



集群状态变为“运行中”，创建成功。



### 购买弹性公网IP

在服务列表中点击“网络”分类下的“弹性公网IP EIP”。



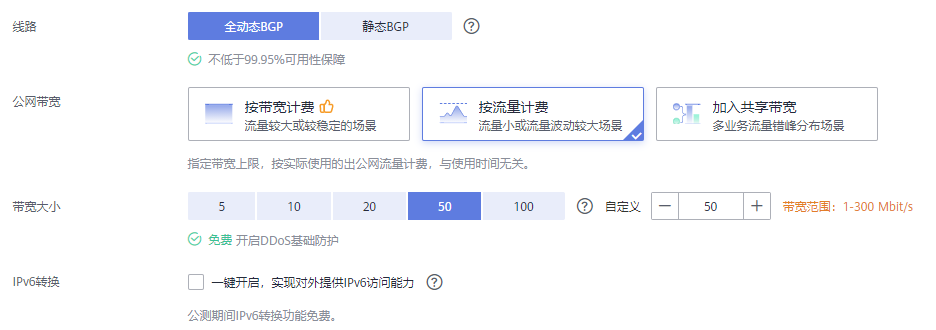
点击“购买弹性公网IP”。



选择“按需计费”，区域“华北-北京四”。



线路“全动态BGP”，公网带宽“按流量计费”，带宽大小“50”。



购买数量“1”，其他默认。



点击“立即购买”，进入确认页面点击“提交”。



回到弹性公网IP页面后点击刷新按钮，可以看到已经购买的弹性公网IP。



# 云服务实现大数据

## 实验介绍

### 实验说明

本部分实验基于前面已经搭建好的大数据环境，运行一个MapReduce程序，读取OBS中的数据进行计算，结果存储在OBS中。

### 实验需求

实验程序选用Hadoop自带的WordCount示例程序。

数据集文件及程序输出结果文件存放到OBS中。

使用MRS构建大数据环境、作业运行示例程序及组件管理。

### 实验流程

把待处理数据集存储到OBS中

把示例程序存储到OBS中

创建MapReduce类型作业执行示例程序

下载程序输出文件并查看结果

使用组件管理查看作业执行日志

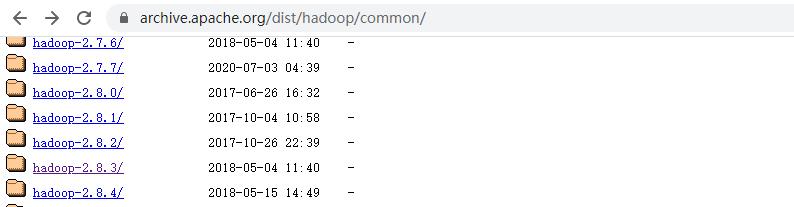
## 实验任务

### 程序及数据准备

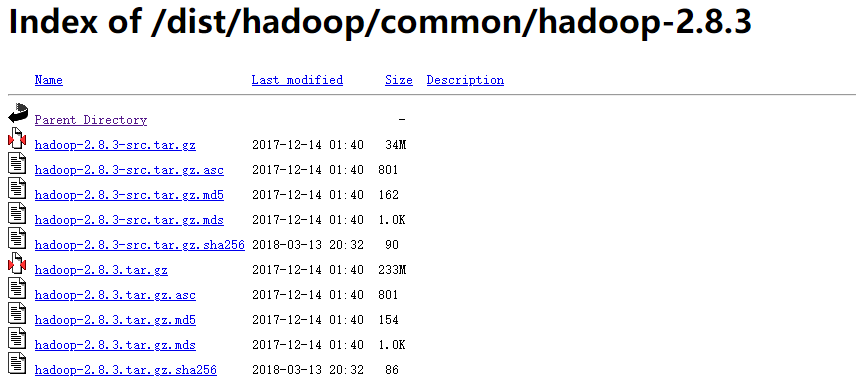
准备示例程序

下载hadoop安装包

# 进入apache的发布存档网页（https://archive.apache.org/dist/hadoop/common/），向下滚动页面，找到对应版本（2.8.3）点击进入



# 点击要下载的包进行下载



# 得到hadoop的安装包



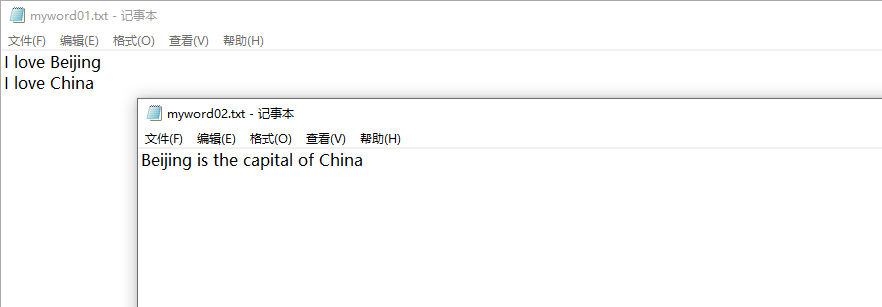
获取示例程序

# 解压下载的安装包，在share\hadoop\mapreduce中可以找到文件hadoop-mapreduce-examples-2.8.3.jar即为示例程序



准备数据文件

格式无特殊要求，准备两个文本文件即可，输入如下内容：



存储到OBS

登录华为云控制台，进入到对象存储服务OBS，在对象存储界面点击前面创建的桶名称。



# 点击“对象”，然后点击“新建文件夹”



# 输入文件夹名“program”并点击确定



# 同理再创建input文件夹



存入程序和数据文件

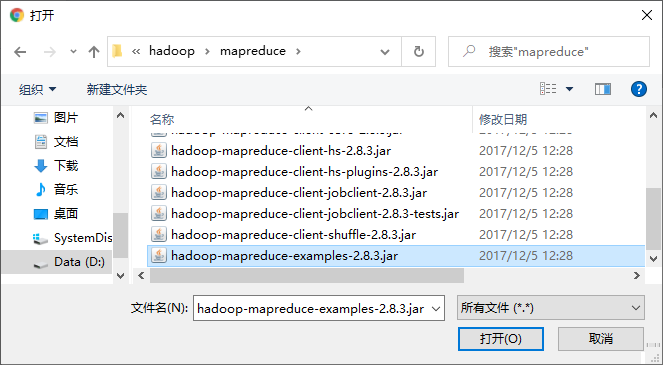
# 点击文件夹program名称进入，然后点击“上传对象”



# 点击“添加文件”



# 选择前面准备好的示例程序文件hadoop-mapreduce-examples-3.1.3.jar，点击“打开”



# 点击“上传”按钮



# 上传成功，示例程序存储到OBS桶的program目录中



# 同理把前面准备好的数据文件上传到input目录中



### 执行程序

创建并运行作业

# 进入控制台MRS现有集群页面，点击前面创建的集群名称



# 选择“作业管理”标签，单击“添加”按钮

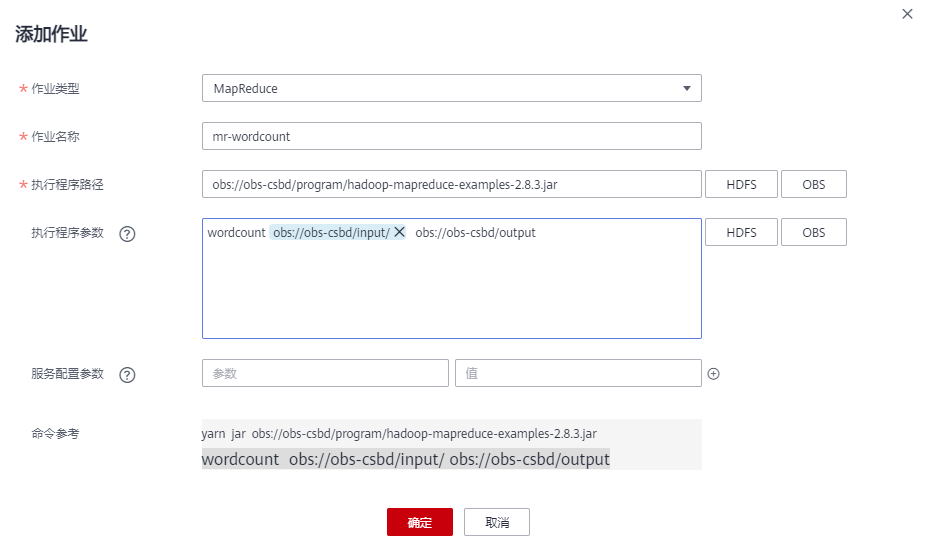


# 作业类型选择“MapReduce”，输入作业名称“mr-wordcount”

# 点击程序执行路径后面的“OBS”按钮，在弹出对话框中点击桶名称、文件夹名称、文件名，直到选择到hadoop样例程序，点击“是”



# 执行程序参数配置为“wordcount **obs://obs-csbd/input/** ​obs://obs-csbd/output”（其中input为输入目录，由后面的OBS按钮选择，output为输出路径，手动输入一个不存在的目录即可），服务配置参数不配置，点击“确定”



**由OBS选择**

**由OBS选择**

# 作业创建完成，状态开始为“已接受”，不需要用户手动执行作业



# 等待作业状态变为“已完成”则作业执行完毕（若出现错误可以查看日志）



查看程序输出结果

下载结果文件（结果文件会存入指定的output目录）

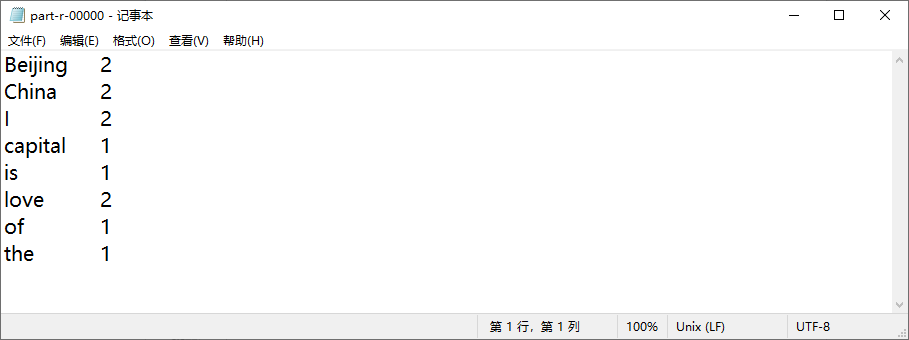
# 进入OBS控制台，点击桶名称obs-csbd进入桶中可以看到已经生成的output输出目录



# 点击进入output目录中可以看到输出文件，点击结果文件“part-r-00000”后面的“下载”，把文件下载到本地



查看结果内容

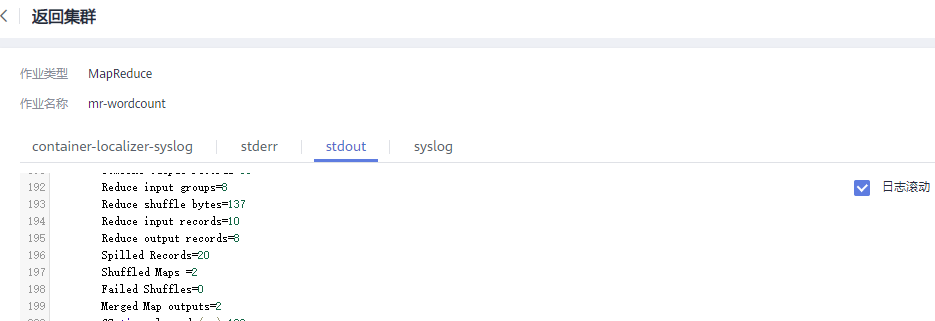
# 使用notepad（或其他任意文本编辑器)打开下载的结果文件），可以看到输出结果内容 

## 结果验证

### 查看作业日志

# 在集群的“作业管理”界面，点击对应作业后面的“查看日志”



# 可以看到作业的日志信息（点击不同的标签可以查看不同的日志信息） 

### 使用组件管理

同步IAM账户（同步后才能使用组件管理）

# 进入mrs\_test集群页面，选择“概览”标签，点击“IAM用户同步”后面的“点击同步”



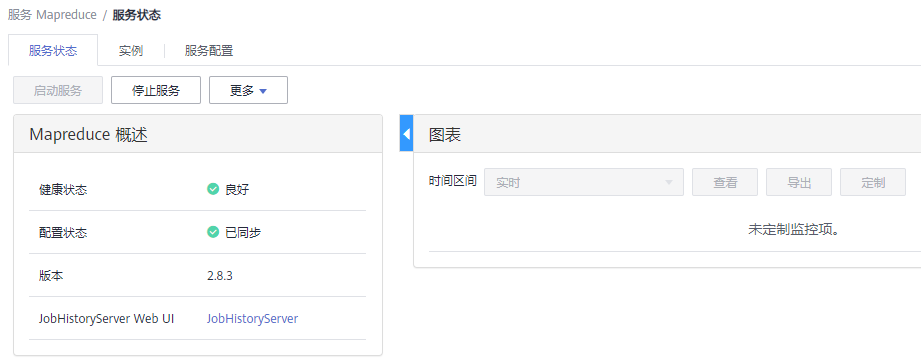
# 稍后刷新页面，“组件管理”等标签变的可用



组件管理

# 在集群页面点击“组件管理”，可以看到正在运行的组件服务，点击“MapReduce” 

# 在MapReduce页面可以对组件进行相关管理，点击“JobHistoryServer”



# 弹出访问MRS管理页面对话框，选择前面购买的弹性公网IP，其他默认，点击“确定”



# 若弹出私密连接提示，则点击“高级”后点击“继续前往”（因浏览器而异）



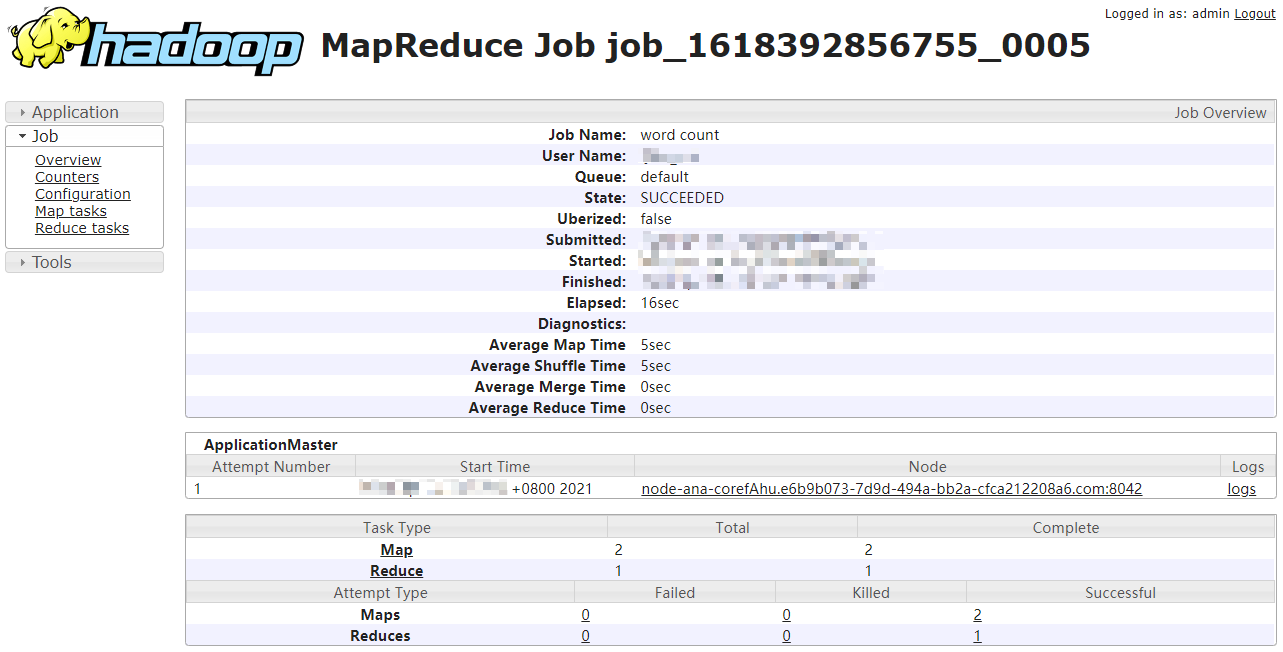
# 进入登录页面，输入MRS集群管理的用户名和密码(密码为创建集群时设置)



# 登录成功后进到MapReduce的Web管理界面，可以看到刚执行的MapReduce任务，点击前面执行的Job ID



# 可以看到Job的具体执行信息



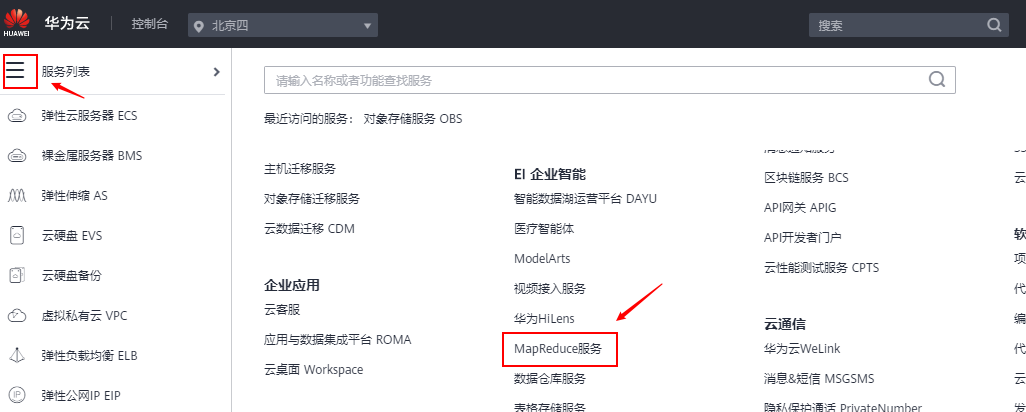
## 思考题

我们的现在使用的是MapReduce作业，如何使用Spark作业呢？

# 资源清理释放

## 释放MapReduce服务

# 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（集群名称以实际实验中为准），点击后面的“删除”链接进行删除



# 弹出的对话款中勾选复选框后点击“是”



# 弹出提示信息，集群状态变为“删除中”，一段时间后集群被则被删除



## 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（桶名称以实际实验中为准），点击桶的名字进入桶（桶为空才可以删除）



# 在桶内选择“对象”，勾选里面所有的对象，点击“删除”按钮



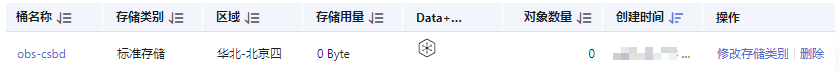
# 弹出的对话框选择“是”



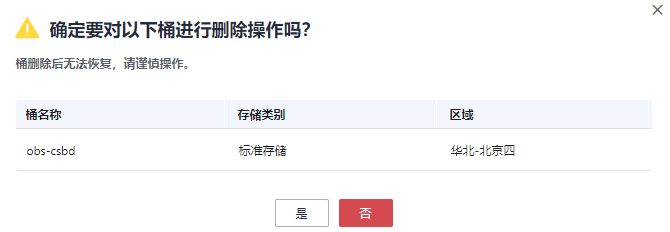
# 对象删除成功后点击“对象存储”返回桶列表



# 点击桶后面的“删除”

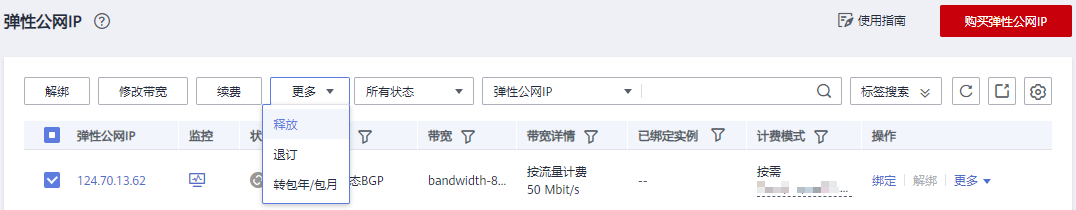


# 弹出对话框选择“是”，稍等桶则被删除



## 释放弹性公网IP

进入弹性公网IP控制台，勾选要释放的弹性公网IP，点击更多中的“释放”。



弹出对话框点击“是”。

