

# 中国矿业大学计算机学院

## 2022级本科生课程设计报告

实验名称 实验二：云服务实现大数据

报告时间 2025 年 6 月

学生姓名 杨晓琦

学 号 08222213

专 业 计算机科学与技术

任课教师 徐东红

# 目 录

1 实验环境准备.....	3
1.1 服务购买 .....	3
1.1.1 登陆控制台 .....	3
1.1.2 购买 OBS 桶 .....	3
1.1.3 购买 MRS 服务 .....	5
1.1.4 购买弹性公网 IP .....	8
2 云服务实现大数据.....	9
2.1 试验任务 .....	9
2.1.1 程序及数据准备 .....	9
2.1.2 执行程序 .....	10
2.2 结果验证 .....	13
2.2.1 查看作业日志 .....	13
2.2.2 使用组件管理 .....	13
3 资源清理释放.....	15
3.1 释放 MapReduce 服务 .....	15
3.2 释放对象存储服务 OBS .....	16
3.3 释放弹性公网 IP .....	16

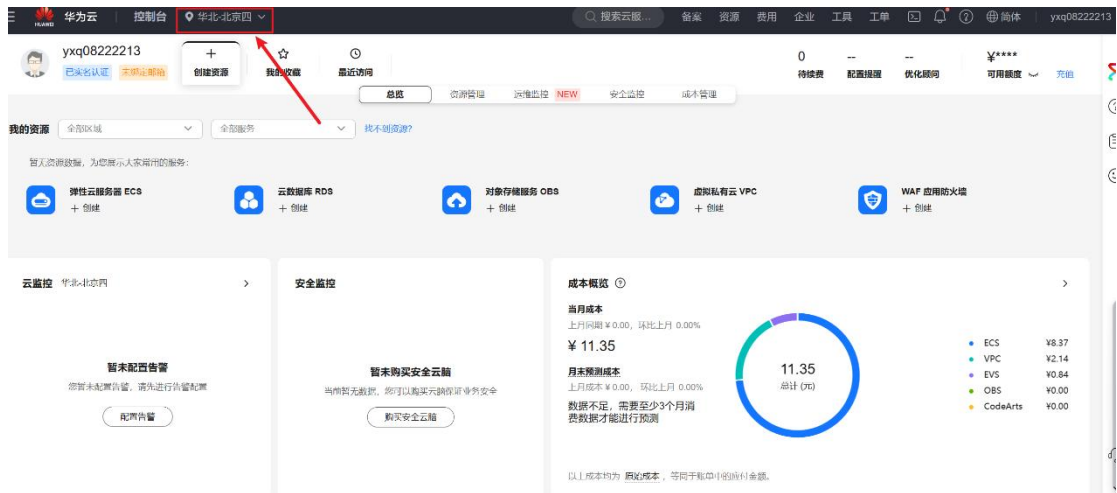
# 实验二：云服务实现大数据

## 1 实验环境准备

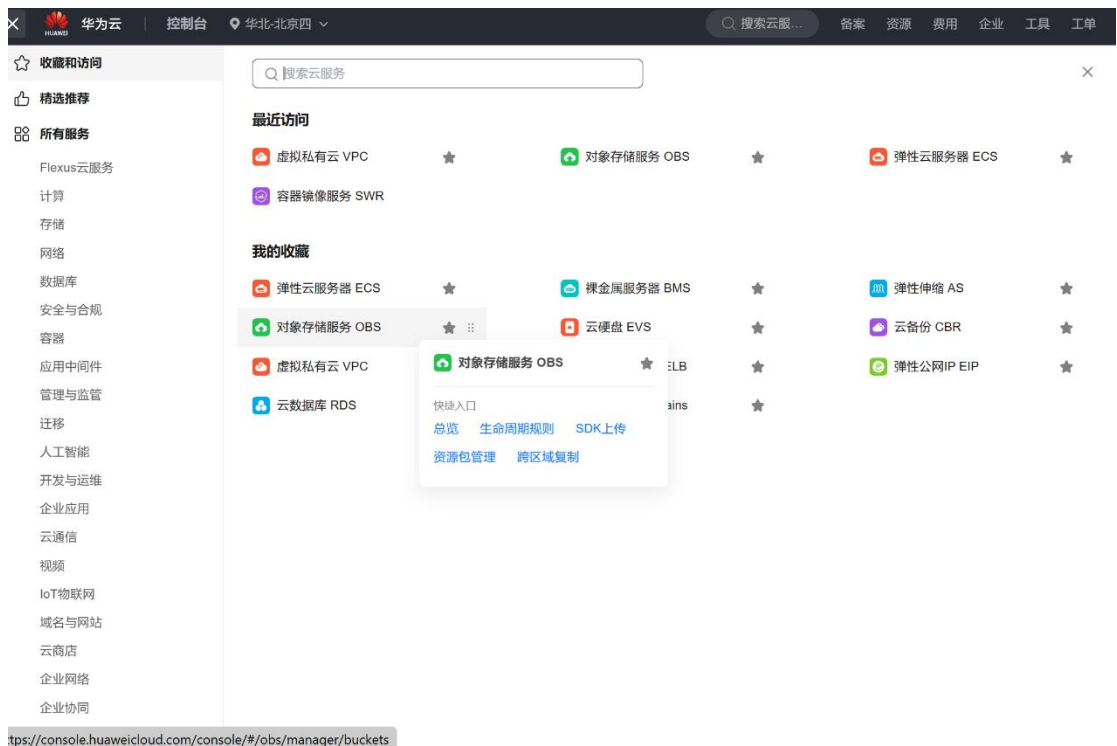
### 1.1 服务购买

#### 1.1.1 登陆控制台

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, 选择区域为“北京四”



#### 1.1.2 购买 OBS 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，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”

桶列表 / 创建桶

<  创建桶

桶配置

桶名称

obs-08222213yxq [查看命名规则](#)

桶访问域名: obs-08222213yxq.obs.cn-north-4.myhuaweicloud.com

数据冗余存储策略 [?](#)

多AZ存储

单AZ存储

创建成功后**不支持修改**。  
数据在同区域的多个可用区 (AZ) 中存储，当某个AZ数据不可用时，仍可保障数据正常访问，可用性更高。

存储类别

创建桶时选择的存储类别会作为上传对象的默认存储类别。 [了解存储类别差异](#)



**标准存储**  
适用于需要频繁访问大量热点文件（平均一个月多次）的场景。访问时延迟低，吞吐量高，多AZ存储可靠性更高，单AZ存储成本更低。  
云应用 数据分享 内容分享 热点对象



**低频访问存储**  
适合高可靠，低成本，较少访问场景，可靠、较低成本的实时访问存储服务。  
网盘应用 企业备份 活跃归档 监控数据



**归档存储**  
适用于很少访问（平均一年访问一次）数据的业务场景。安全、持久且成本低，数据恢复时间数分钟到数小时不等。  
档案数据 医疗影像 视频素材

桶策略 [?](#)

私有

公共读 

公共读写 

[复制桶策略](#)

桶的拥有者拥有完全控制权限，其他用户在未经授权的情况下均无访问权限。

功能配置

归档数据直读 [?](#)

☐

未开启

关闭归档数据直读，归档存储类别的数据要先恢复才能访问。归档存储数据恢复和访问会**收取相应的费用**。 [价格详情](#) [?](#)

服务端加密 [?](#)

☐

未开启

开启服务端加密后，上传到当前桶的对象会被加密。您也可以在建桶完成之后在桶概览页面调整服务端加密配置。  
建议开启加密，核心数据更安全，如果您使用KMS加密模式，超过免费配额会**收取相应费用**。 [价格详情](#) [?](#)

WORM

☐

未开启

使用WORM（一次写入多次读取）功能，帮助您防止对象版本在指定时间段内被删除或覆盖。  
启用WORM会自动启用版本控制，且不可关闭版本控制。  
WORM功能开启后无法关闭，永久允许锁定此桶中对象。 [了解更多](#)

## OBS 桶创建成功

控制台

搜索云服务、资源(IP/名称/ID)、快捷操作... 备案 资源 费用 企业 工具 工单   简体中文 yqx08222213

桶列表 [了解详情](#) [账单查询](#) [工单反馈](#) [帮助中心](#) [购买资源包](#) [创建桶](#)

桶列表

桶名称 

桶名称

特色功能

存储类型

区域

数据冗余...

存储容量

桶策略

对象数量

标签

创建时间

操作

obs-08222213yxq

标准存储

华北-北京四

多AZ存储

0 byte

私有桶

0

--

2025/06/20 0...

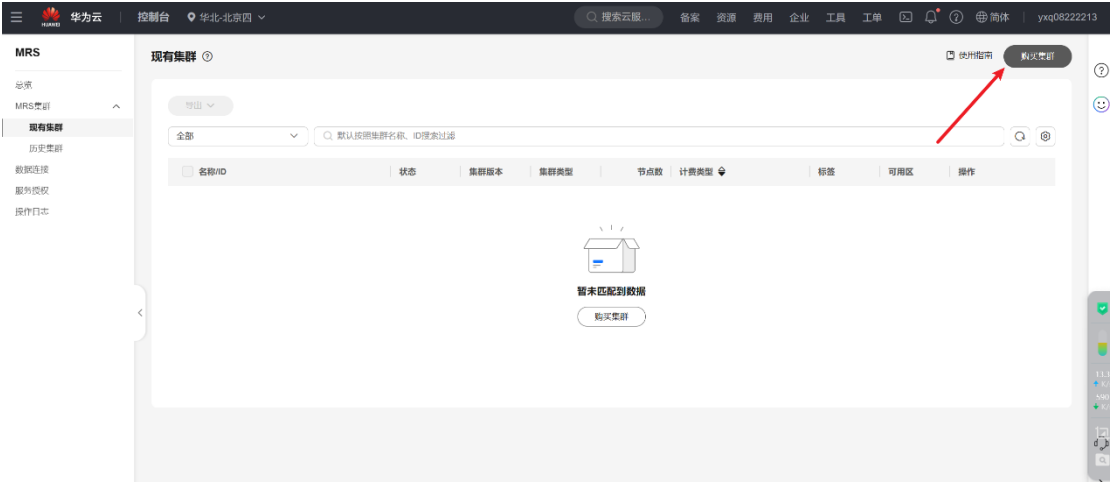
[修改名称](#) [删除](#)

总条数: 1 | 已读: 0

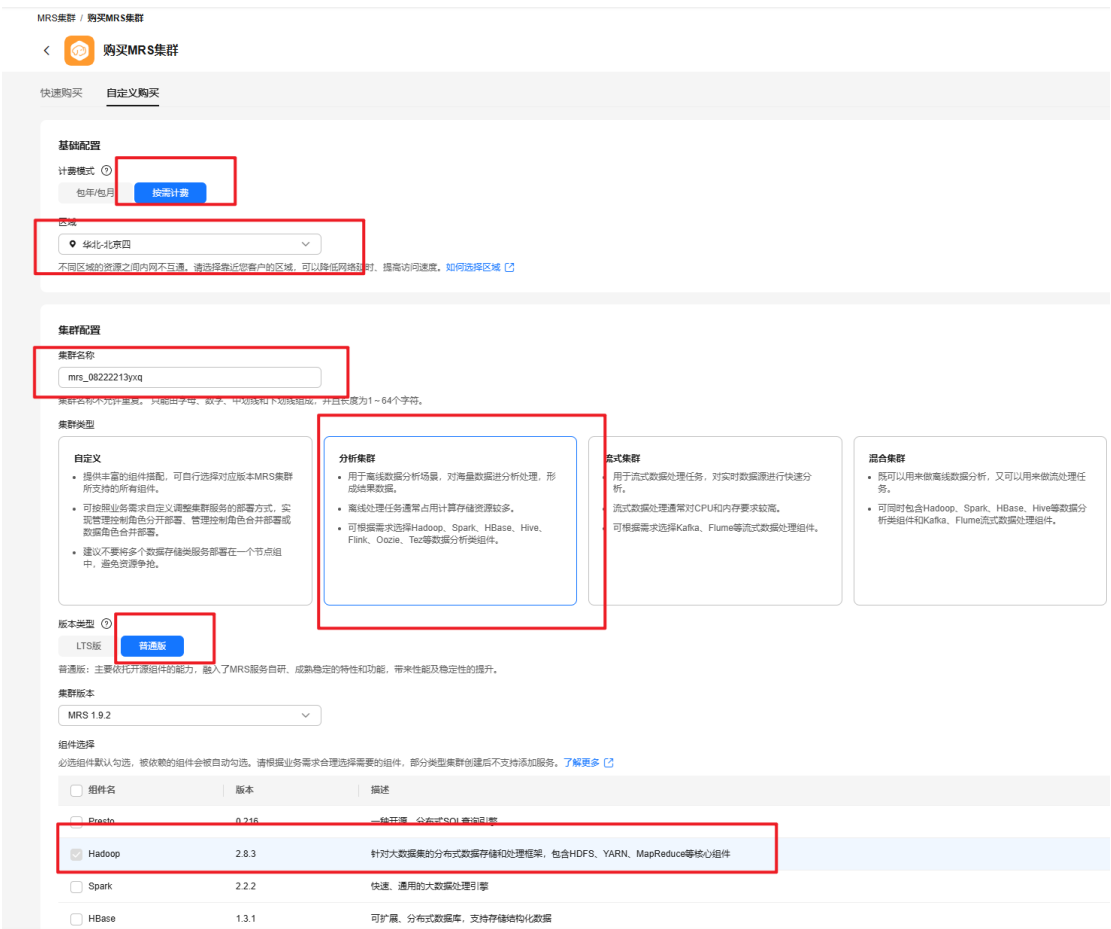
10 < 1 >

### 1.1.3 购买 MRS 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”，版本“1.9.2”，类型为“分析集群”，组件默认勾选 Hadoop



关闭 Kerberos 认证，输入密码

登录凭证

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

登录方式

密码

密钥对

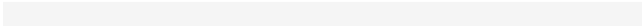
用户名

root

密码

该密码用于远程登录ECS机器或BMS机器。

确认密码



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”，CPU 架构选择“鲲鹏计算”，集群节点默认

网络配置

可用区 ?

可用区7

虚拟私有云 ?

vpc-08222213

查看虚拟私有云

子网 ?

subnet-b036 (192.168.0.0/24)

查看子网

安全组 ?

自动创建 ×

管理安全组

弹性公网IP ?

暂不绑定

创建弹性公网IP

节点配置

CPU架构

x86计算

鲲鹏计算

鲲鹏CPU架构采用精简指令集 (RISC)，它能够以更快的速度执行操作，相对于x86 CPU架构具有更加均衡的性能功耗比。

集群节点

搜索云服务、资源(IP / 名称 / ID)、快捷操作...

备案

资源

✓

任务提交成功

您已成功购买MRS集群，您的mrs\_08222213yxq1已开始创建。

进入MRS集群列表

集群状态变为“运行中”，创建成功



### 1.1.4 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”



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





## 2 云服务实现大数据

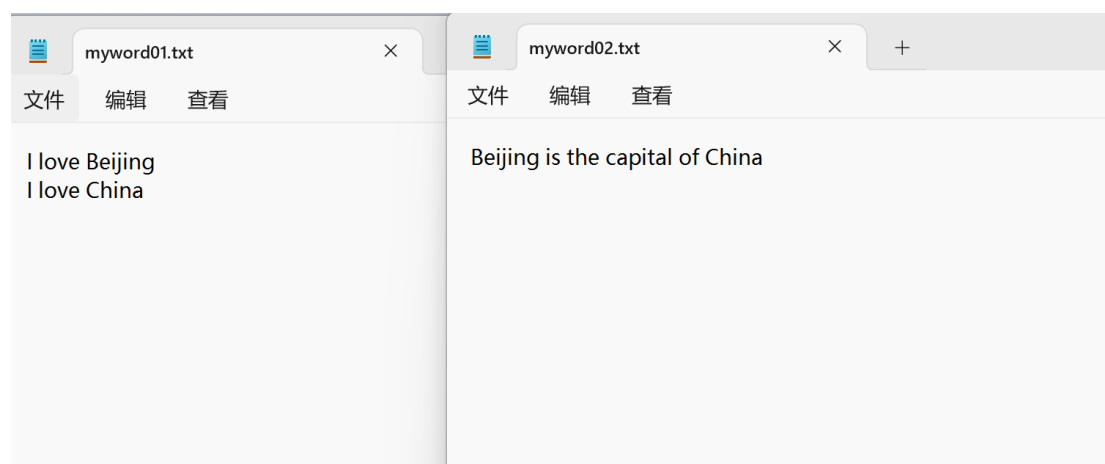
### 2.1 试验任务

#### 2.1.1 程序及数据准备

步骤 1 准备示例程序 `hadoop-mapreduce-examples-2.8.3.jar`

步骤 2 准备数据文件

准备两个文本文件，输入如下内容



步骤 3 存储到 OBS

登录华为云控制台，进入到对象存储服务 OBS；

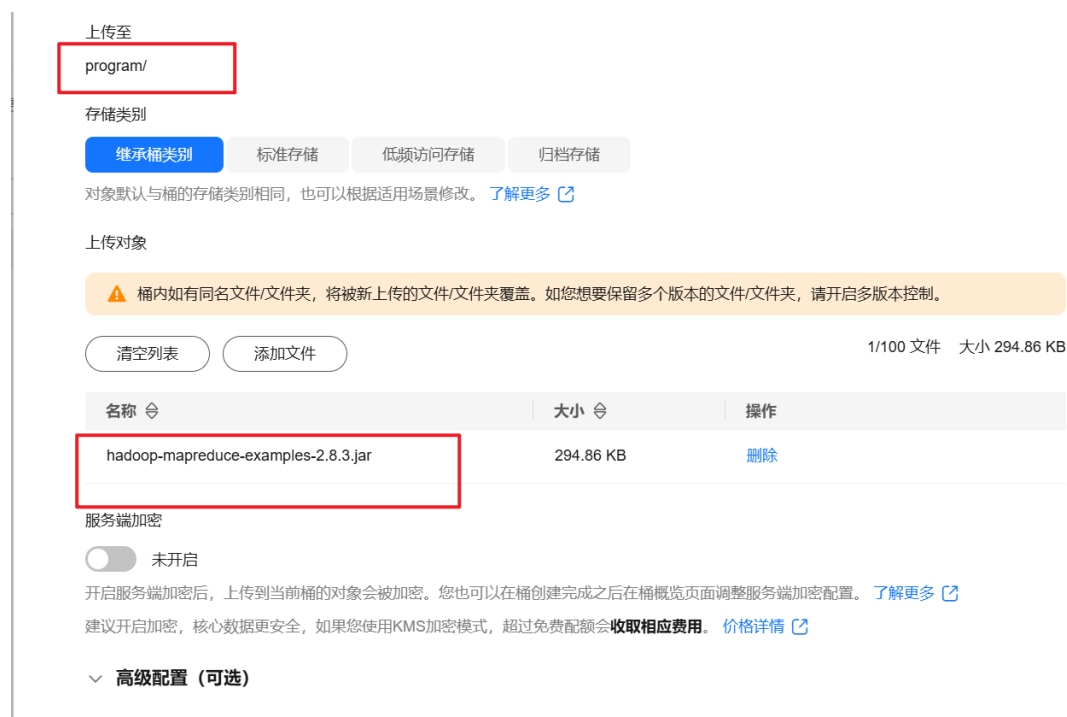
在对象存储界面点击前面创建的桶名称，点击“对象”；

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”



上传成功，示例程序存储到 OBS 桶的 program 目录中



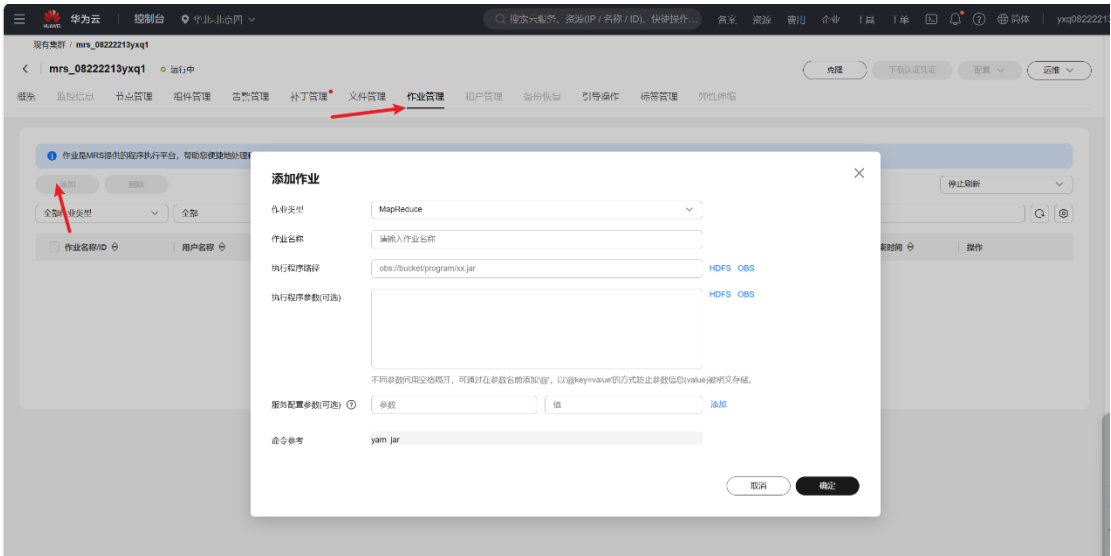
把前面准备好的数据文件上传到 input 目录中



### 2.1.2 执行程序

步骤 1 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



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”

选择OBS文件

OBS / obs-08222213yxq / program / **hadoop-mapreduce-examples-2.8.3.jar**

请输入文件名

文件名

存储类别

文件大小

修改时间

..

hadoop-mapreduce-examples-...

标准存储

294.86 KB

2025/06/20 10:27:37 GMT+0...

只能选择存储类别为标准存储和低频访问储存的文件，所选脚本将在集群对应节点被执行，同时请在OBS侧管控好脚本的访问操作权限。

☒

我已确认所选脚本安全，了解可能存在的风险，并接受对集群可能造成的异常或影响。

取消

确定

执行程序参数配置为“wordcount    obs://obs-08222213yxq/input/    obs://obs-008222213yxqoutput”

添加作业

作业类型

MapReduce

作业名称

mr-wordcount

执行程序路径

obs://obs-08222213yxq/program/hadoop-mapreduce-examples-2.8.3.jar

HDFS   OBS

执行程序参数(可选)

wordcount   **obs:///obs-08222213yxq/input/obs://obs-csbd/output**

HDFS   OBS

不同参数间用空格隔开，可通过在参数名前添加 '@'，以 '@key=value' 的方式防止参数信息(value)被明文存储。

服务配置参数(可选) ?

参数

值

添加

命令参考

yarn jar obs://obs-08222213yxq/program/hadoop-mapreduce-examples-2.8.3.jar  
wordcount   obs:///obs-08222213yxq/input/obs://obs-csbd/output

取消

确定

作业创建完成，状态开始为“已接受”



等待作业状态变为“已完成”则作业执行完毕

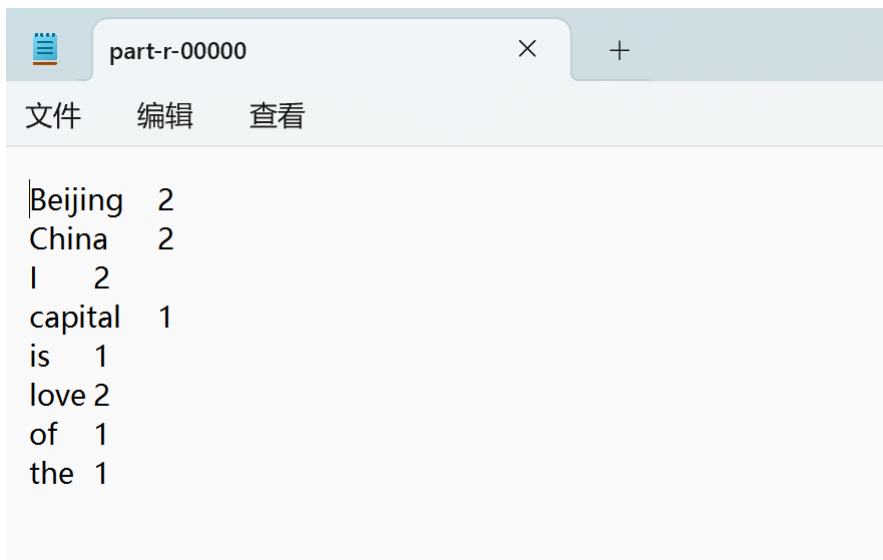


步骤 2 查看程序输出结果

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



## 2.2 结果验证

### 2.2.1 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”

< | 返回集群

作业类型MapReduce

作业名称mr-mycount

container-localizer-syslogstderrstdoutsyslog

183Total megabyte-milliseconds taken by all map tasks=23457792

184Total megabyte-milliseconds taken by all reduce tasks=18883584

185Map-Reduce Framework

186Map input records=3

187Map output records=12

188Map output bytes=108

189Map output materialized bytes=137

190Input split bytes=210

191Combine input records=12

192Combine output records=10

193Reduce input groups=8

194Reduce shuffle bytes=137

195Reduce input records=10

196Reduce output records=8

197Spilled Records=20

198Shuffled Maps =2

199Failed Shuffles=0

200Merged Map outputs=2

201GC time elapsed (ms)=161

202CPU time spent (ms)=3340

203Physical memory (bytes) snapshot=1223606272

204Virtual memory (bytes) snapshot=12296949760

205Total committed heap usage (bytes)=961019904

206Shuffle Errors

207BAD\_ID=0

208CONNECTION=0

209IO\_ERROR=0

210WRONG\_LENGTH=0

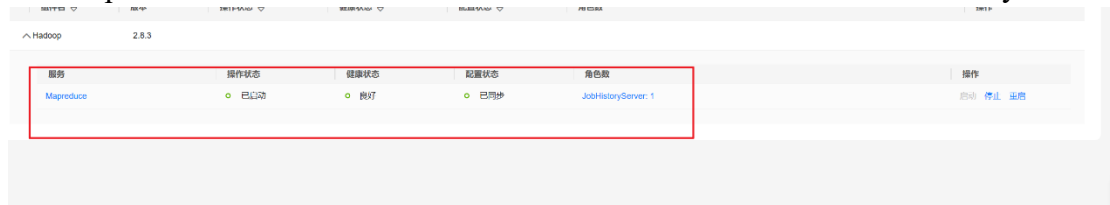
### 2.2.2 使用组件管理

步骤 1 同步 IAM 账户

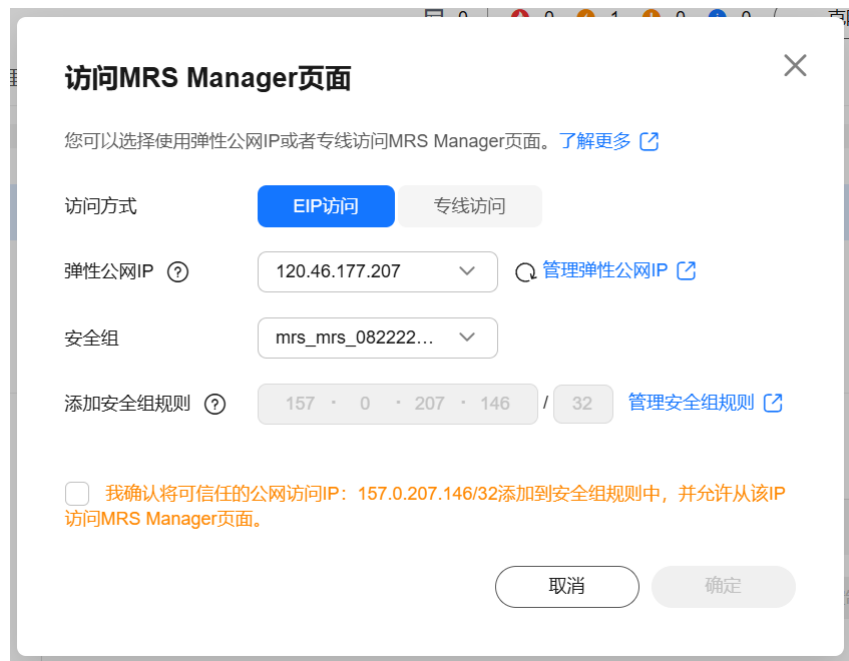


## 步骤 2 组件管理

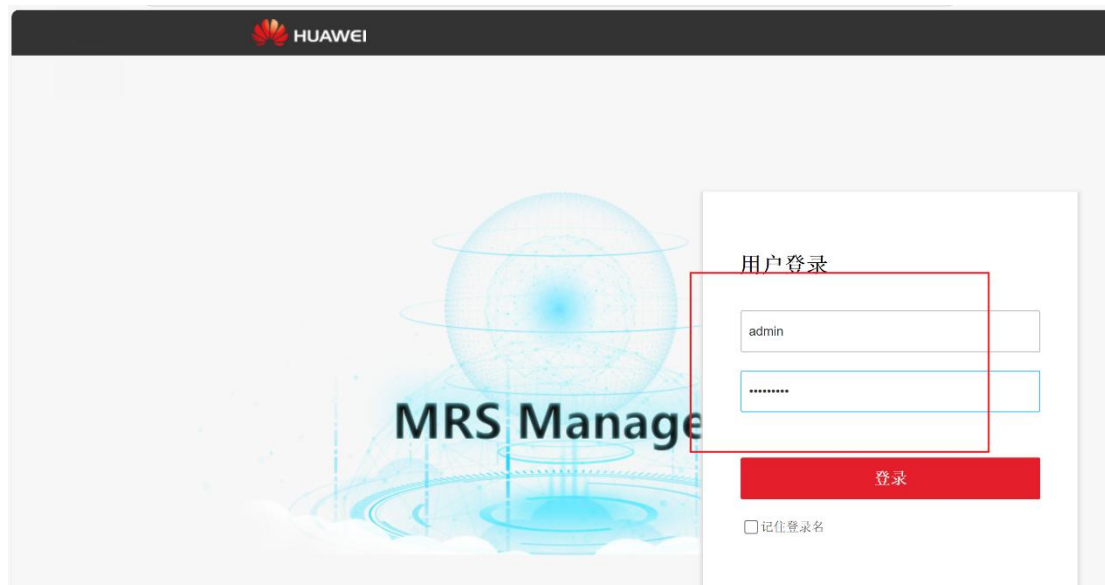
在 MapReduce 页面可以对组件进行相关管理，点击“JobHistoryServer”



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”



进入登录页面，输入 MRS 集群管理的用户名和密码



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

**MapReduce Job job\_1750385375974\_0003**

Job Overview

Job Name: word count  
 User Name: yxq08222213  
 Queue: default  
 State: SUCCEEDED  
 Uberized: false  
 Submitted: Fri Jun 20 10:42:08 +0800 2025  
 Started: Fri Jun 20 10:42:16 +0800 2025  
 Finished: Fri Jun 20 10:42:34 +0800 2025  
 Elapsed: 18sec  
 Diagnostics:  
 Average Map Time: 5sec  
 Average Shuffle Time: 5sec  
 Average Merge Time: 0sec  
 Average Reduce Time: 0sec

ApplicationMaster		Start Time	Node	Logs
Attempt Number	1	Fri Jun 20 10:42:12 +0800 2025	node-ana-coreFimZ.mrs-os2s.com:8042	logs

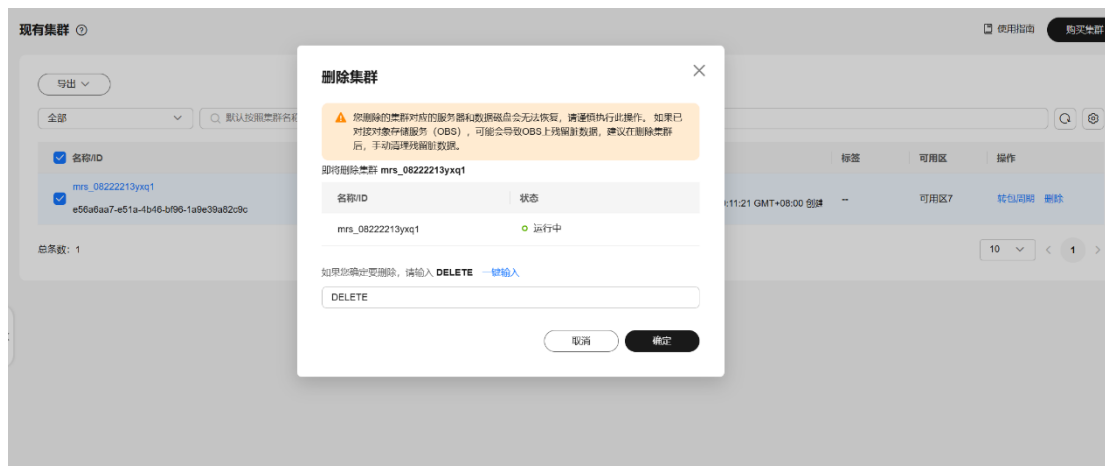
Task Type	Total	Complete
Map	2	2
Reduce	1	1

Attempt Type	Failed	Killed	Successful
Maps	0	0	2
Reduces	0	0	1

## 3 资源清理释放

### 3.1 释放 MapReduce 服务

在“现有集群”中可以看到购买的集群，点击后面的“删除”链接进行删除



### 3.2 释放对象存储服务 OBS



### 3.3 释放弹性公网 IP

