实践报告：云服务实现大数据

学号：2212534

姓名：魏思诚

日期：2024/12/16

南开大学 软件学院

目 录

目录

[1 实验介绍 2](#_Toc185356197)

[1) 实验说明 2](#_Toc185356198)

[2) 实验概览 2](#_Toc185356199)

[2 实验目的 3](#_Toc185356200)

[1) 实验目标 3](#_Toc185356201)

[2) 实验要求 3](#_Toc185356202)

[3 实验流程 4](#_Toc185356203)

[1) 准备实验环境 4](#_Toc185356204)

[2) 服务购买 4](#_Toc185356205)

[3) 云服务实现大数据 11](#_Toc185356206)

[4) 实验任务 12](#_Toc185356210)

[5) 下载结果文件查看内容 21](#_Toc185356211)

[6) 思考题 23](#_Toc185356212)

[我们的现在使用的是MapReduce作业，使用Spark呢? 23](#_Toc185356213)

[4 实验结果或结论 24](#_Toc185356214)

[1) 结果 24](#_Toc185356215)

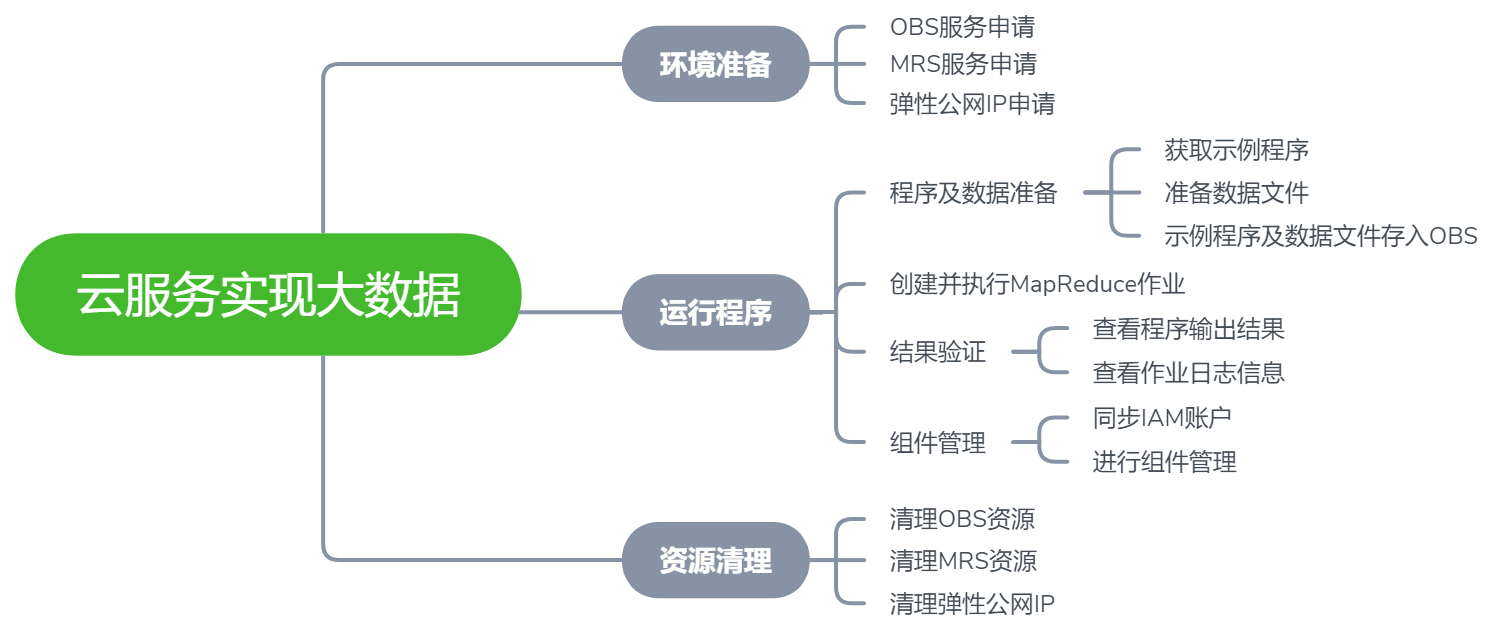
[2) 结论与收获 24](#_Toc185356216)

# 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。

## 实验概览

* 

# 实验目的

## 实验目标

* 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。

## 实验要求

* 使用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

# 实验流程

## 准备实验环境

* 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 | 3.1.5 | 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。



进入控制台后，选择区域为“北京四”。



* 购买OBS桶

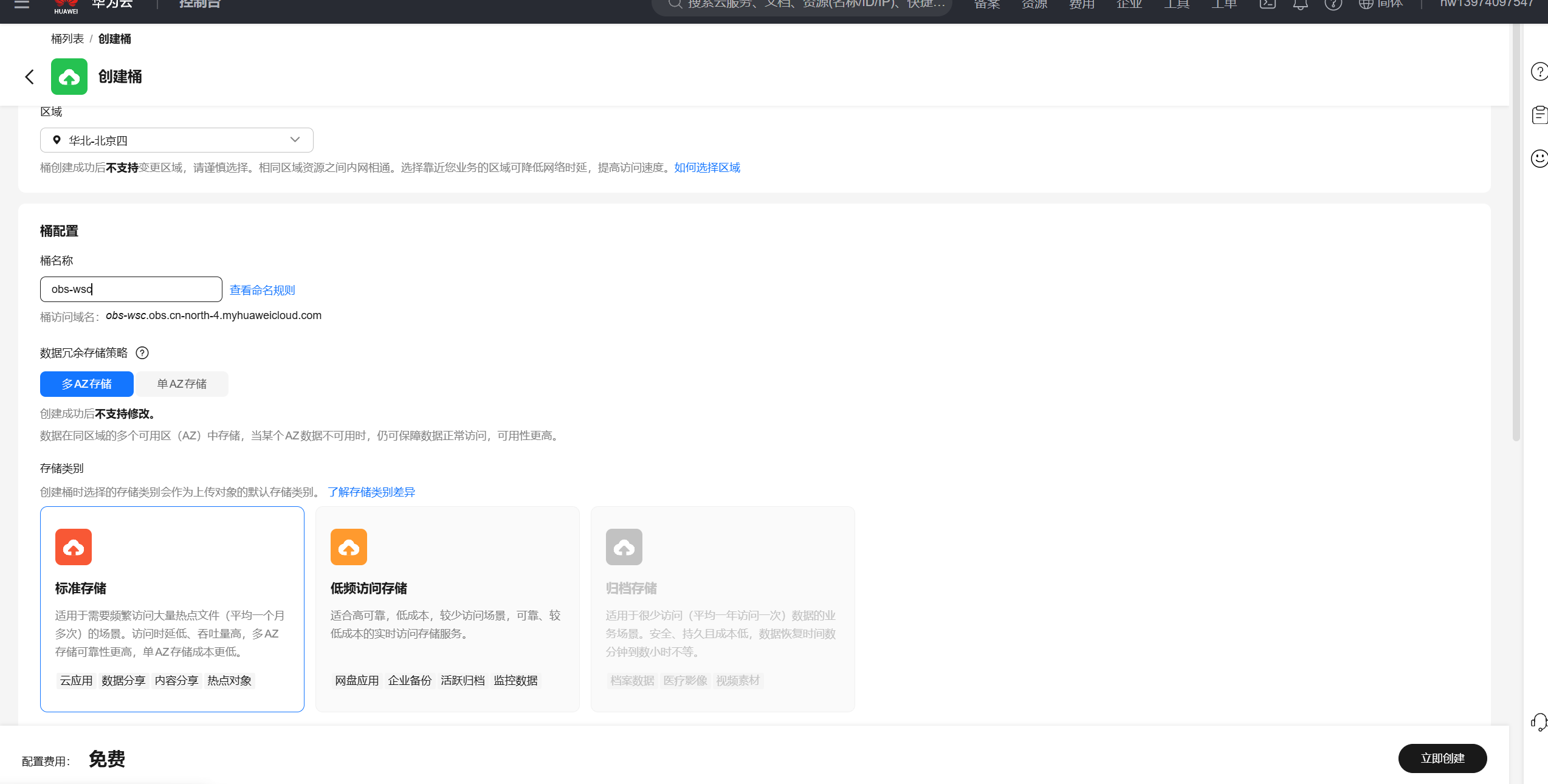
鼠标放到控制台左上角三根横线处，弹出服务列表，点击“存储”分类里的“对象存储服务OBS”



在对象存储界面点击“创建桶”。



区域为“华北-北京四”



立即创建



* 购买MRS服务

在服务列表中点击“MapReduce服务MRS” 或者大数据分类下的“MapReduce服务MRS”。



在现有集群界面点击“购买集群”。

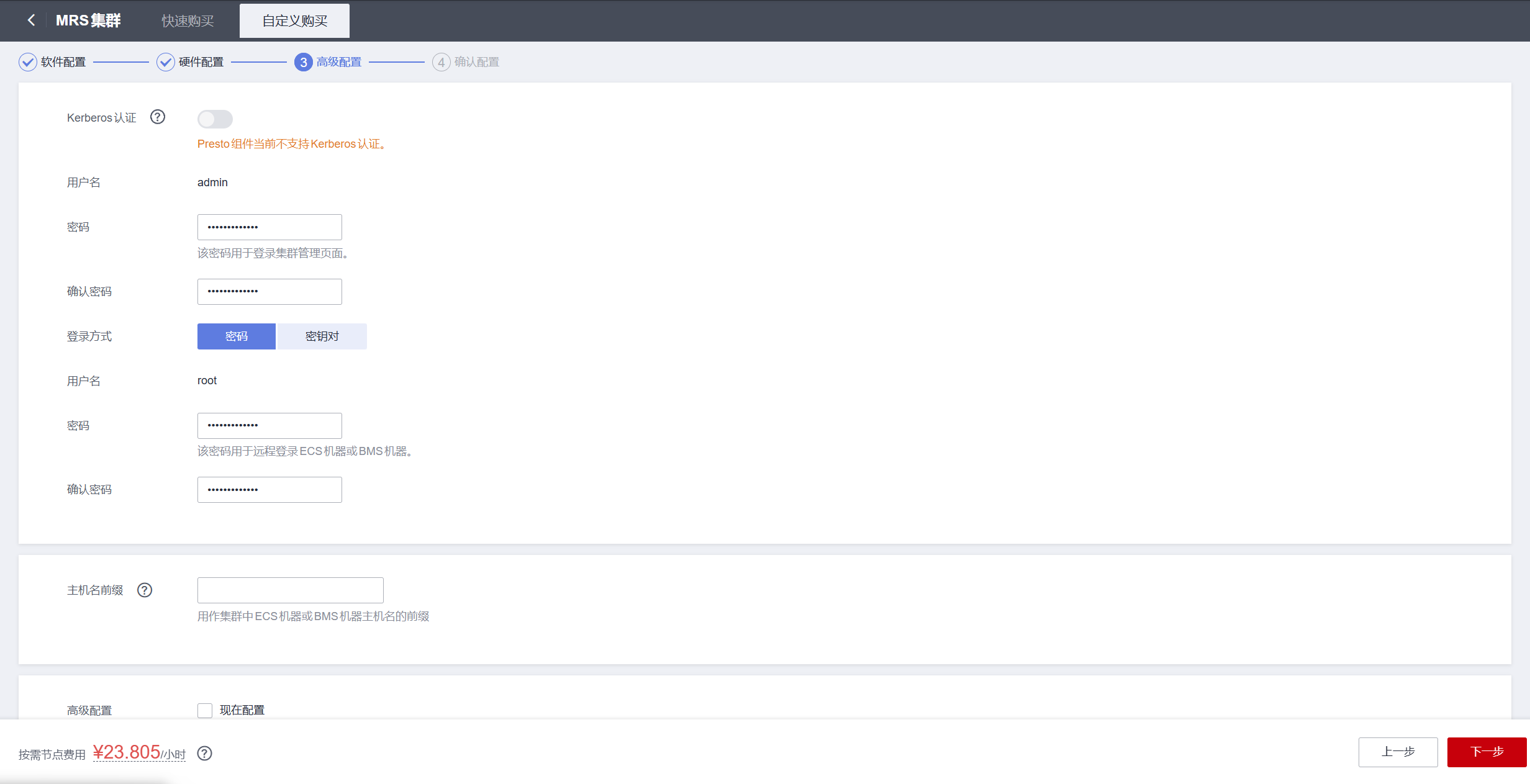
集群名: mrs\_zT57





设置密码:  
1668200060Aa!

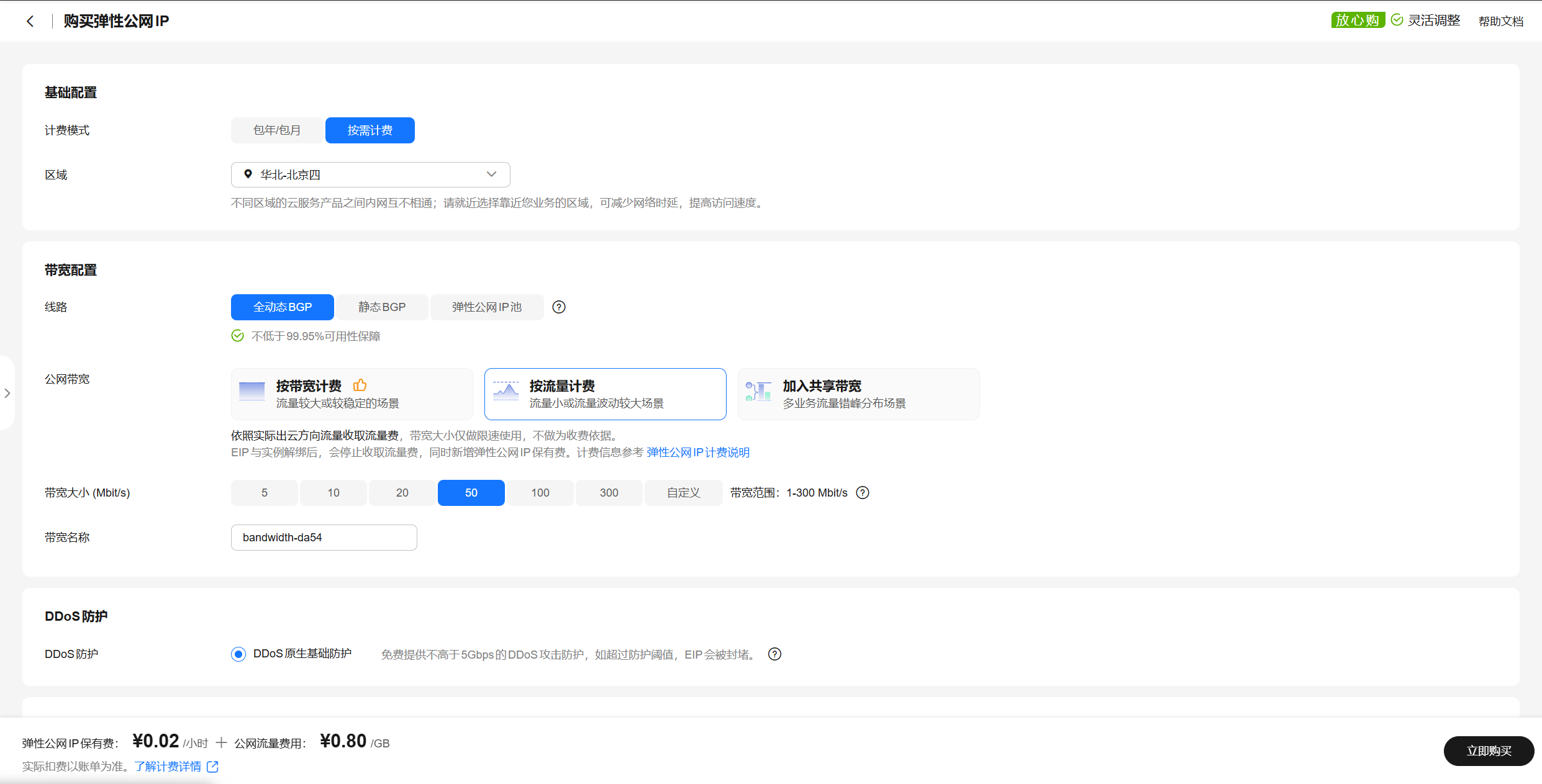
Root密码:

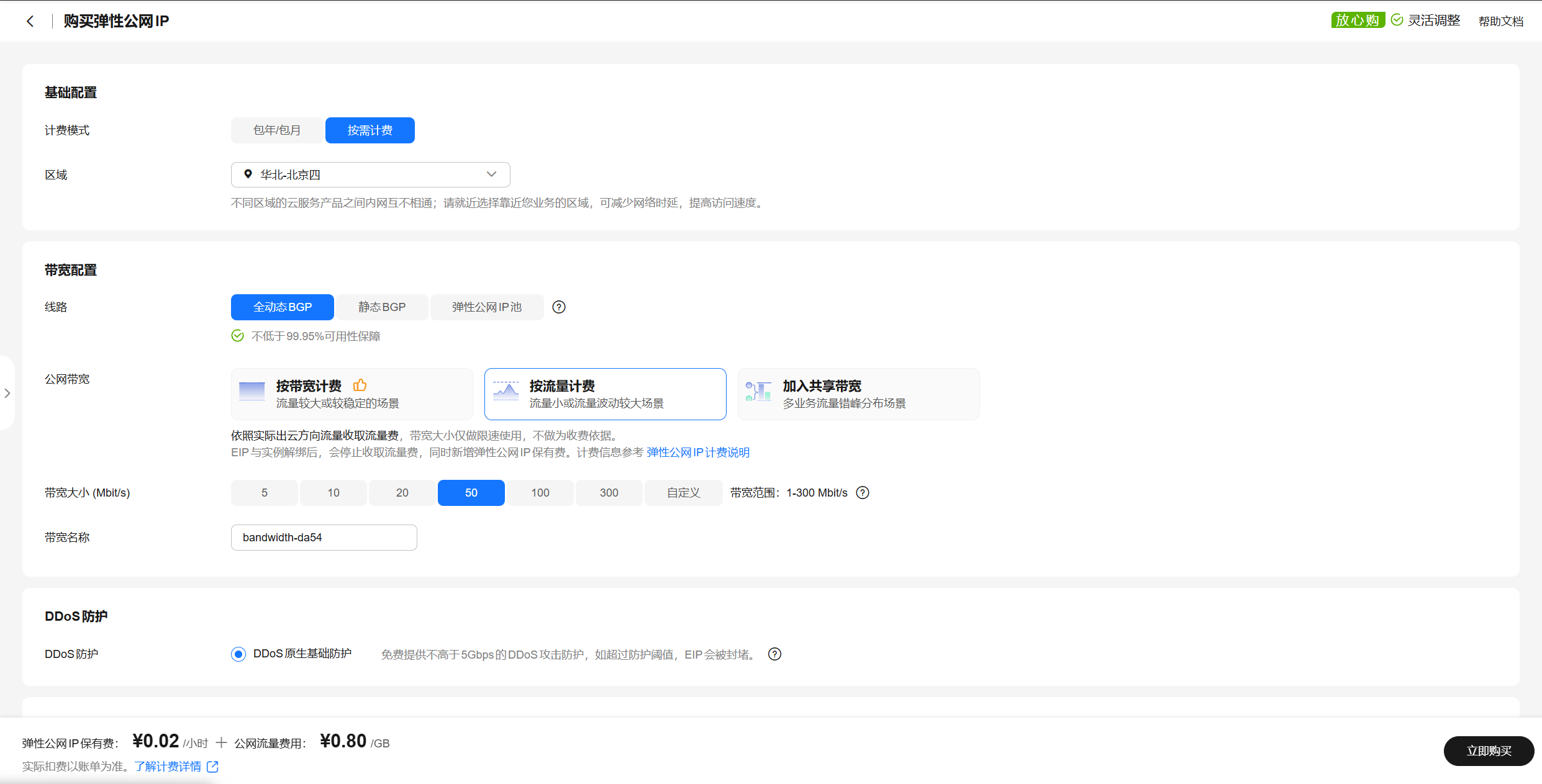
1668200060Aa!  


* 购买弹性公网IP

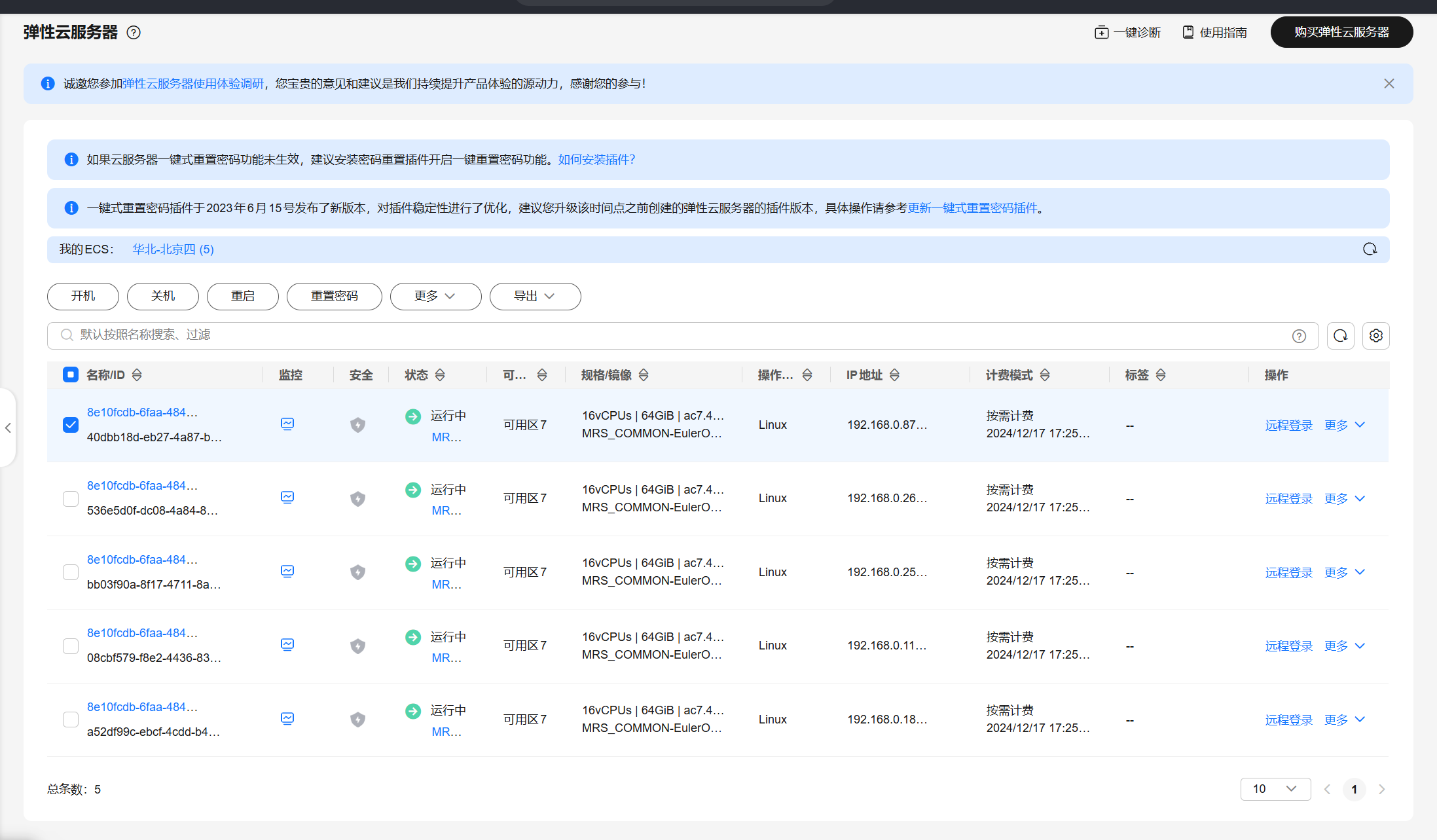
在服务列表中点击“网络”分类下的“弹性公网IP EIP”。



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”  




绑定



## 云服务实现大数据

### 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。

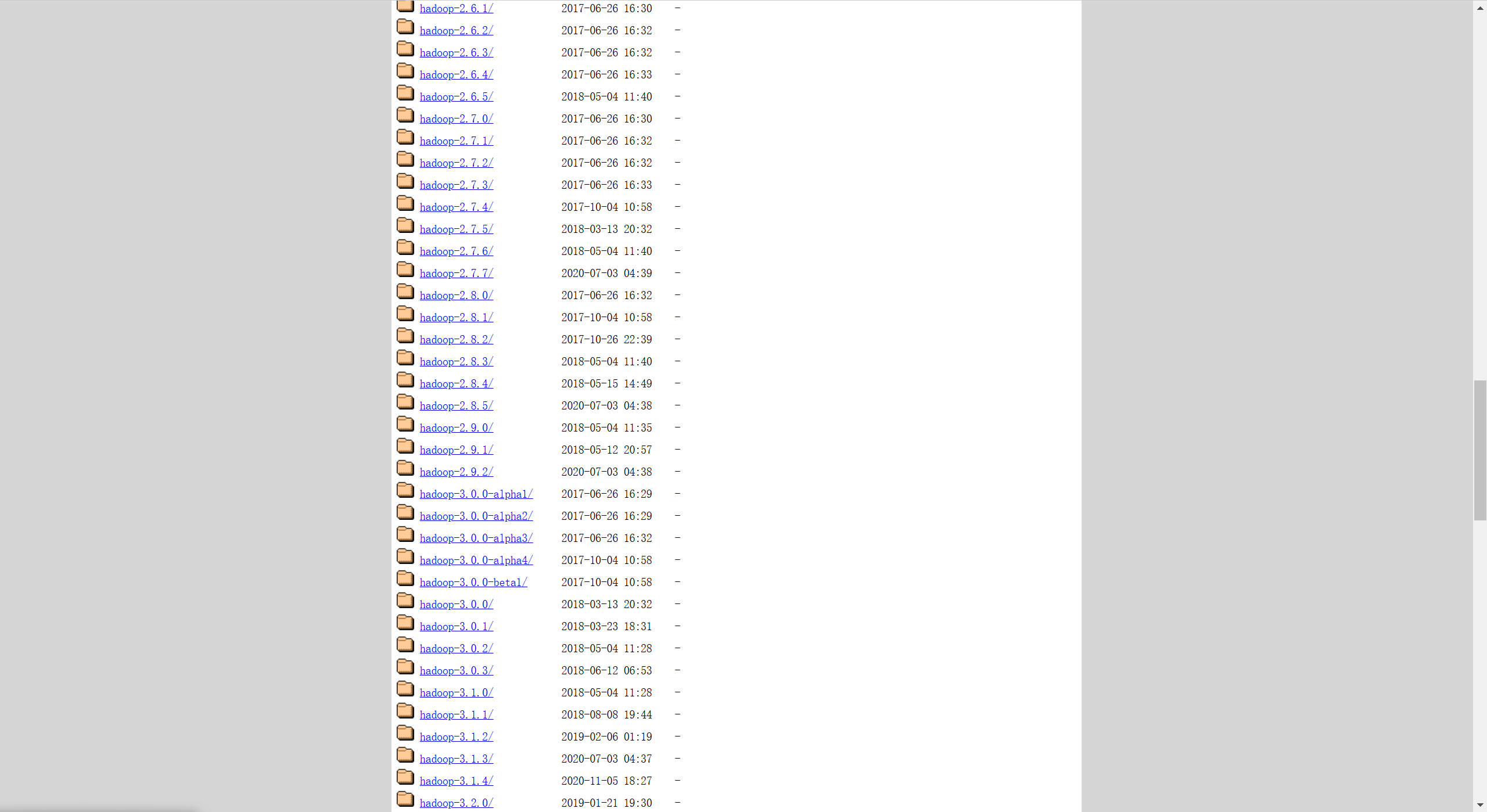
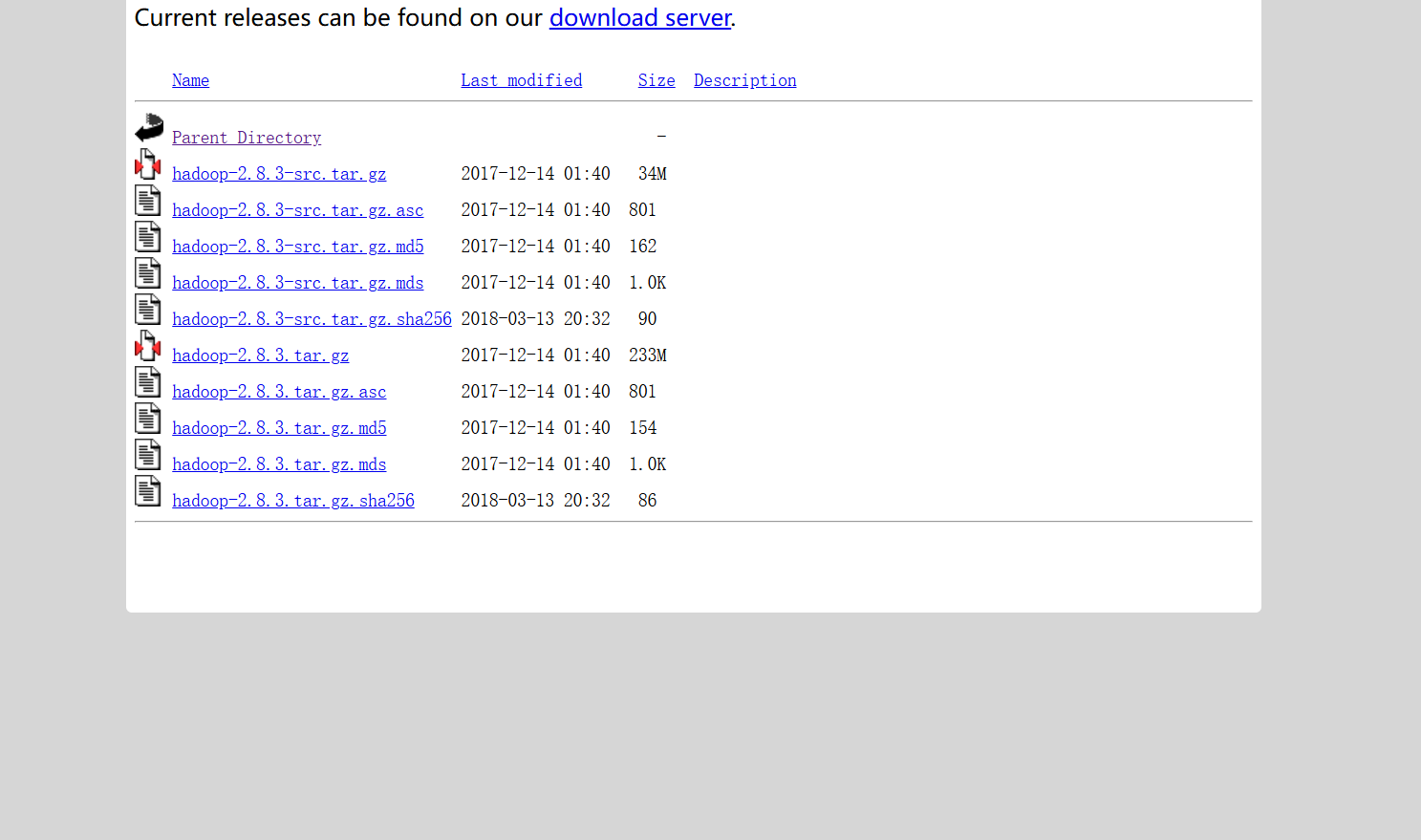
### 实验流程

* 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

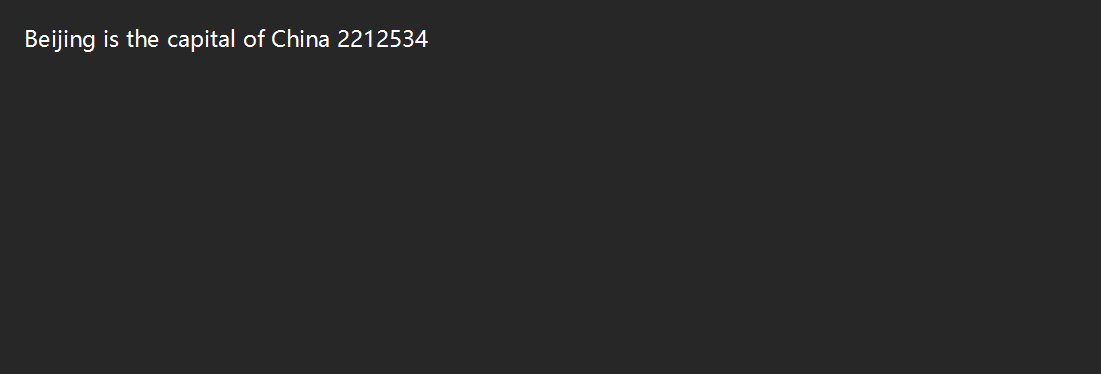
* 程序及数据准备

1. 准备示例程序
2. 下载hadoop安装包
3. # 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

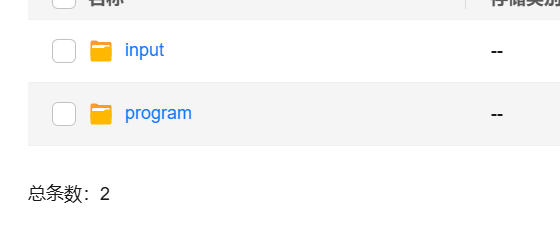
 

1. 准备数据文件

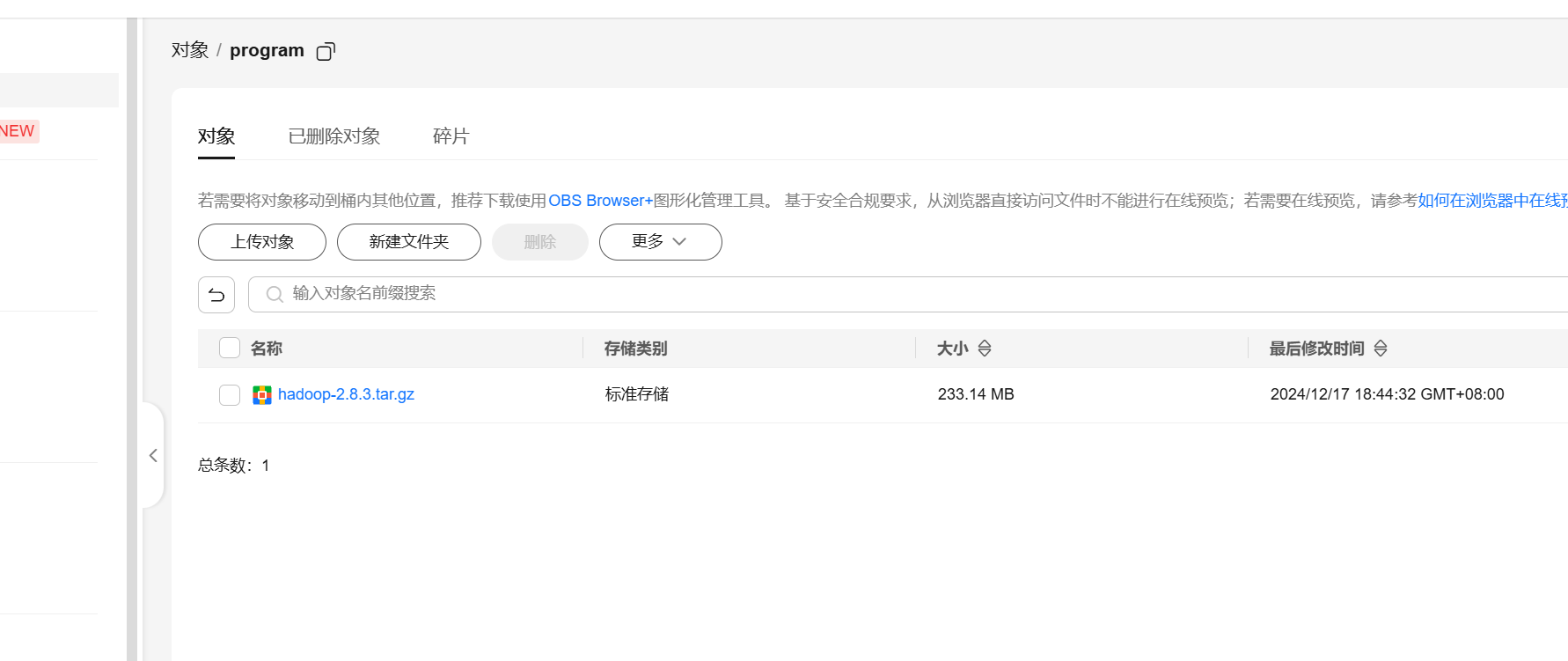




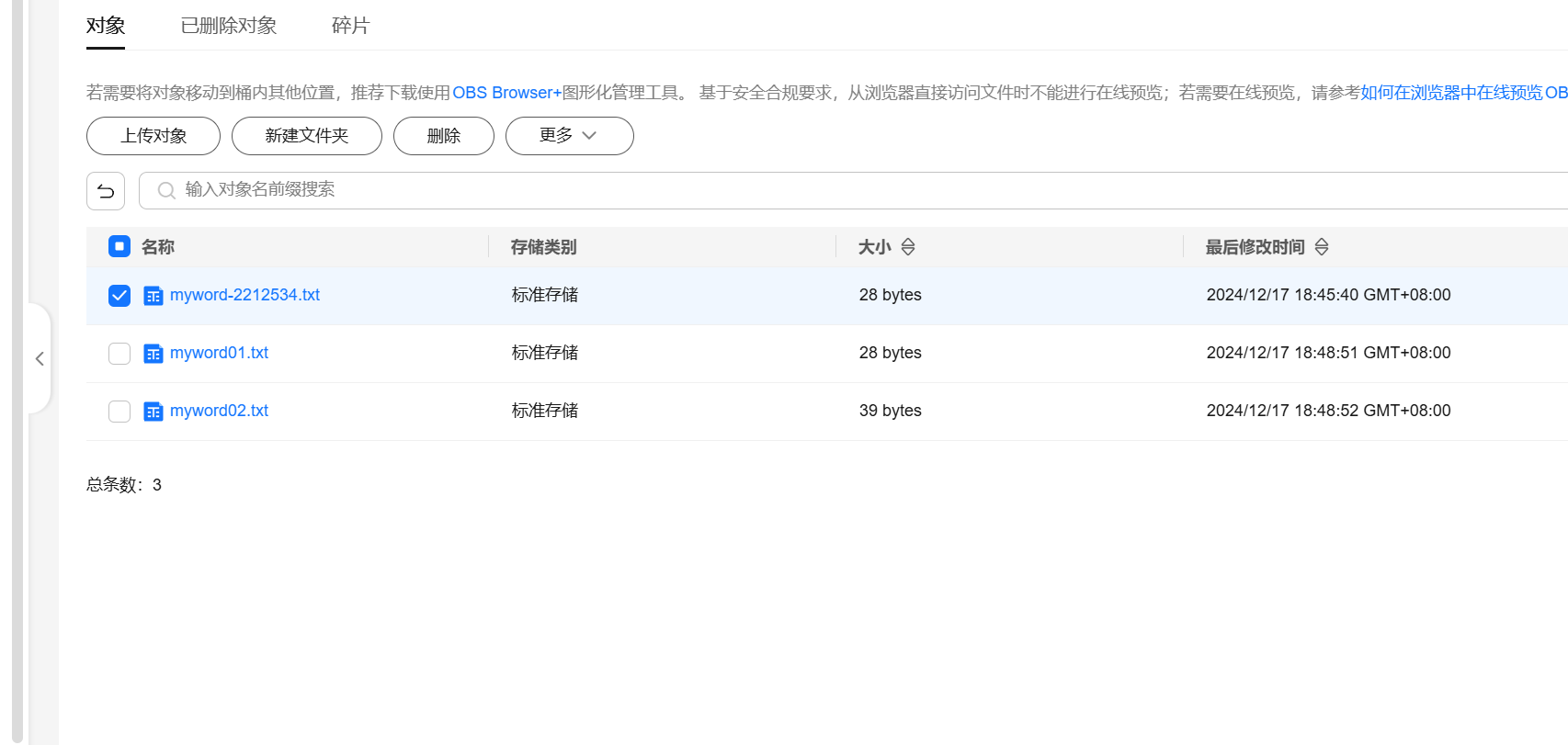
1. 存储到OBS









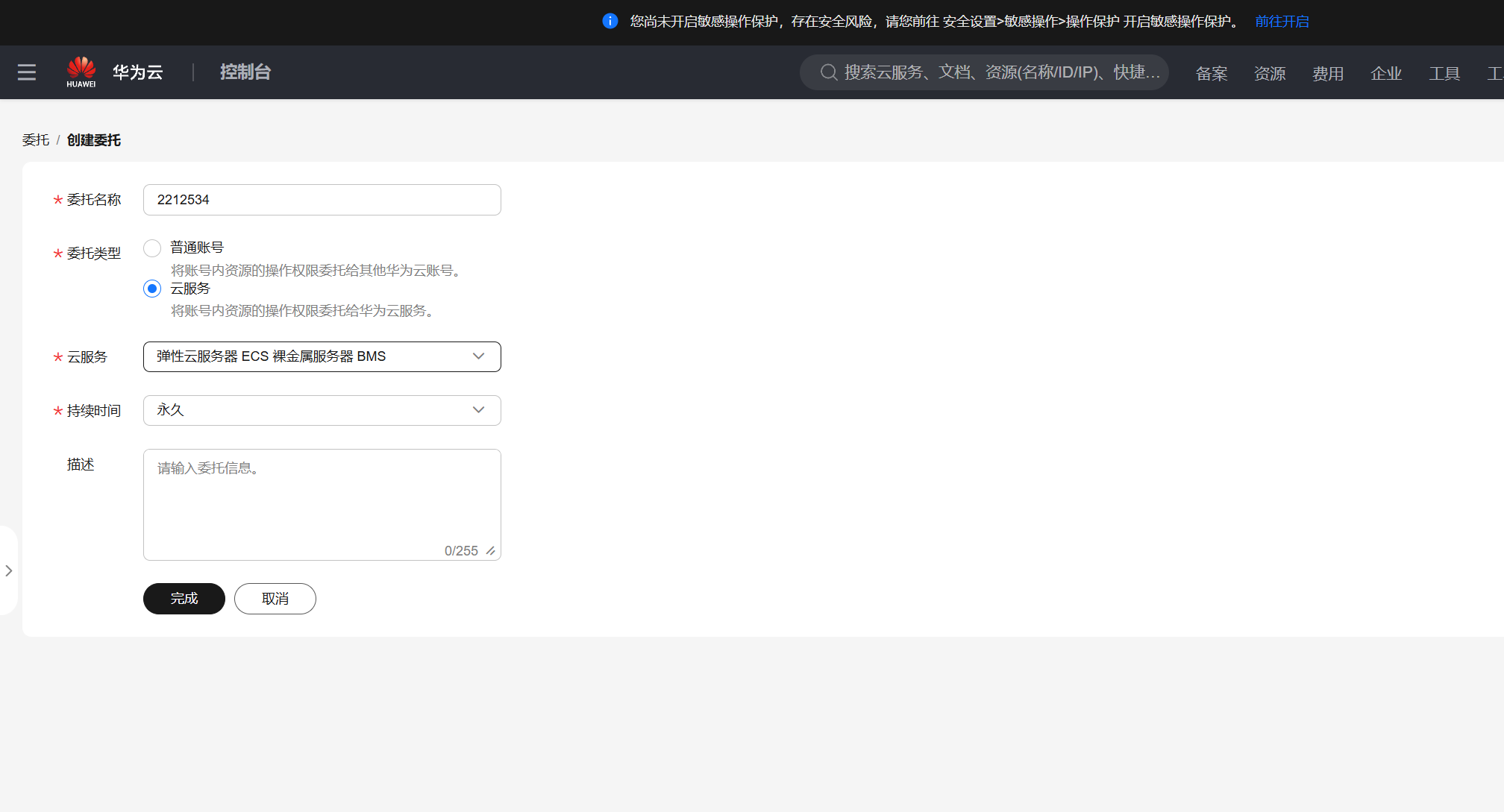
* 执行程序

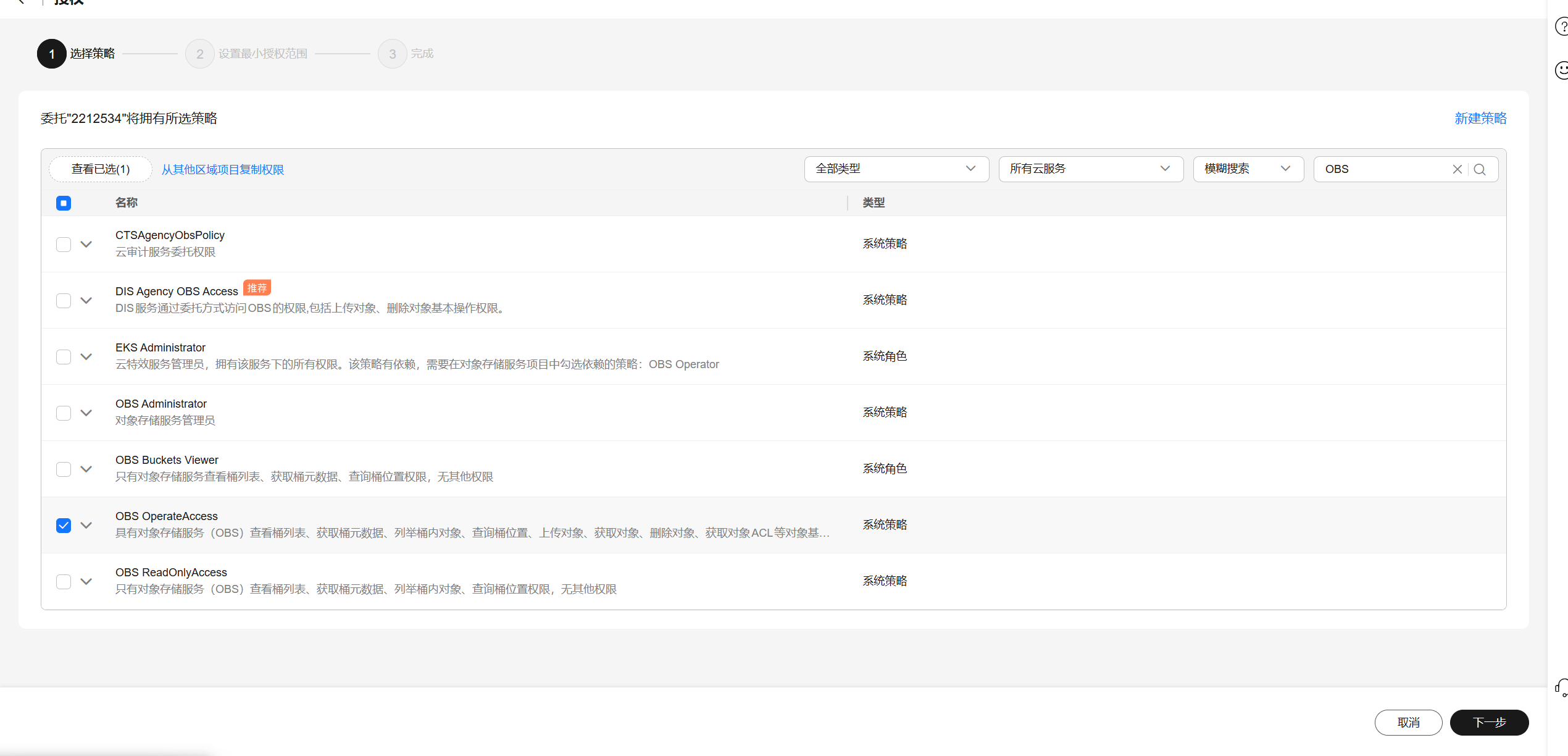
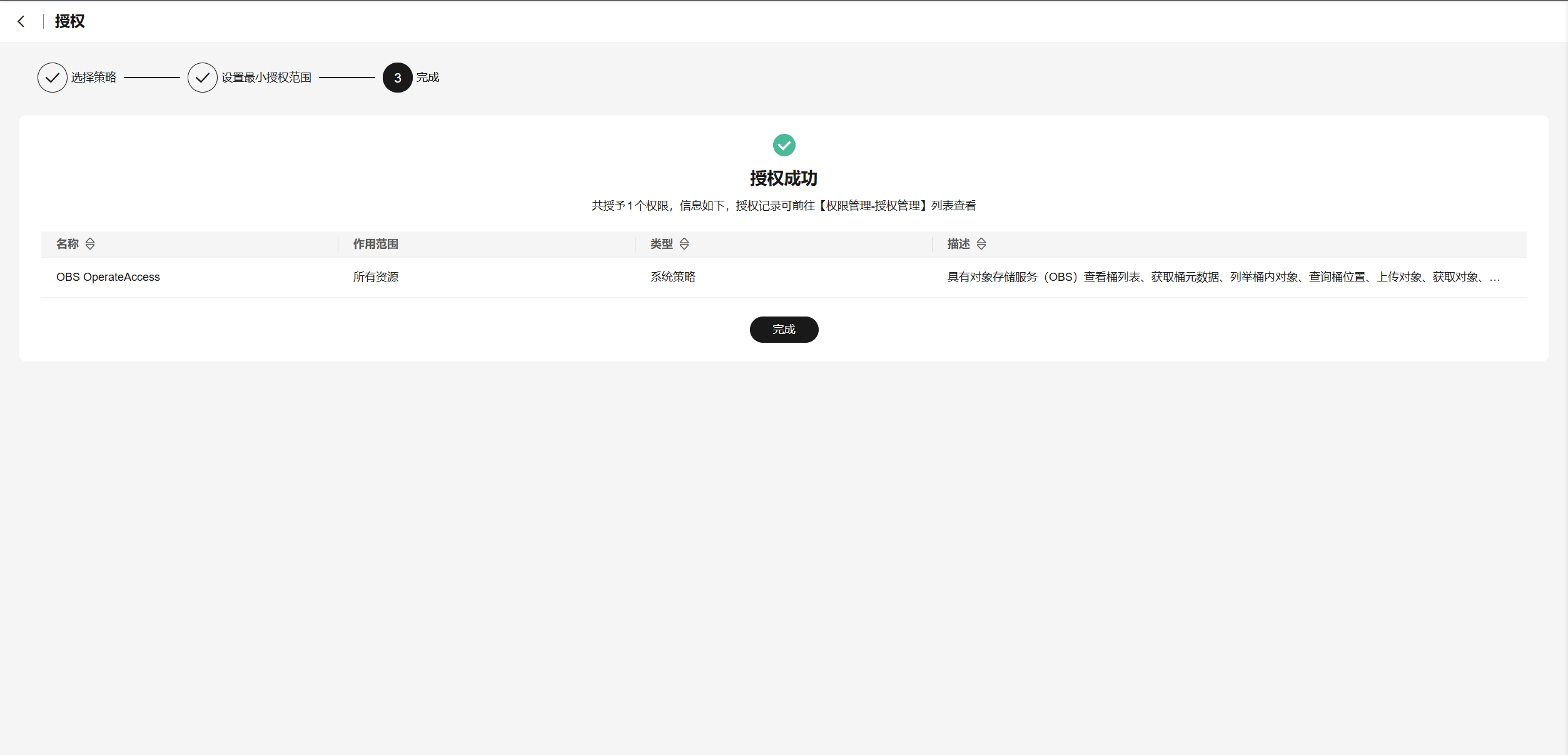
步骤 1 创建并运行作业

# 进入控制台MRS现有集群页面，点击前面创建的集群名称

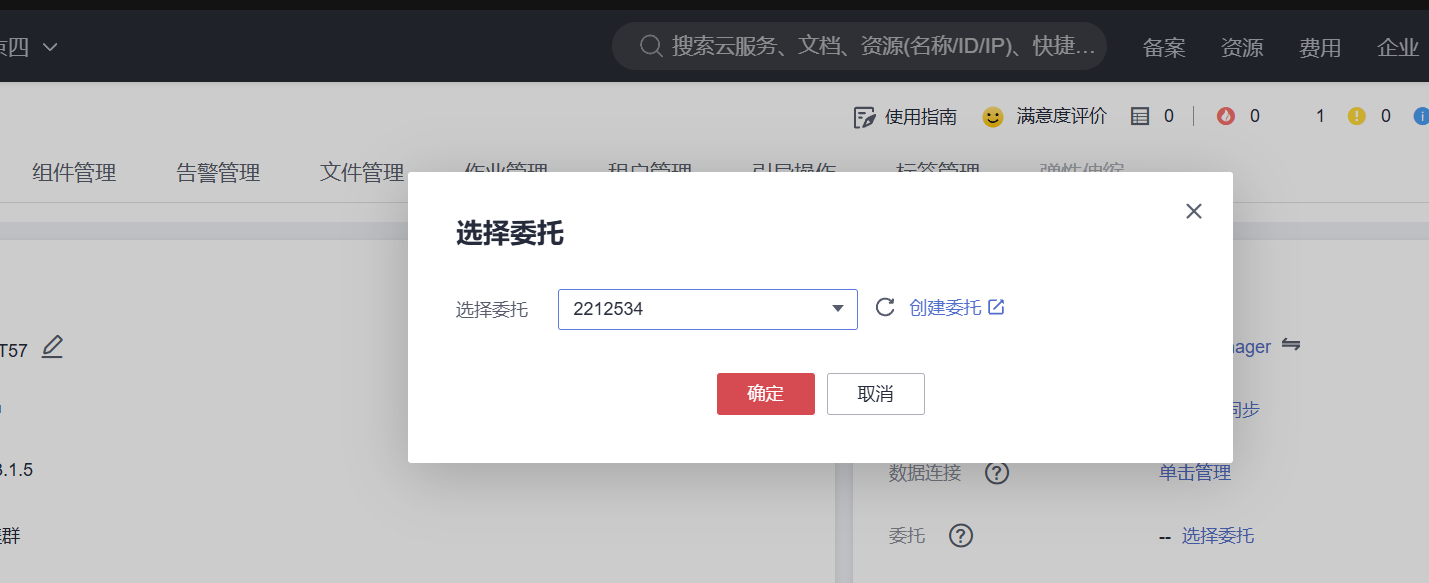


创建委托:

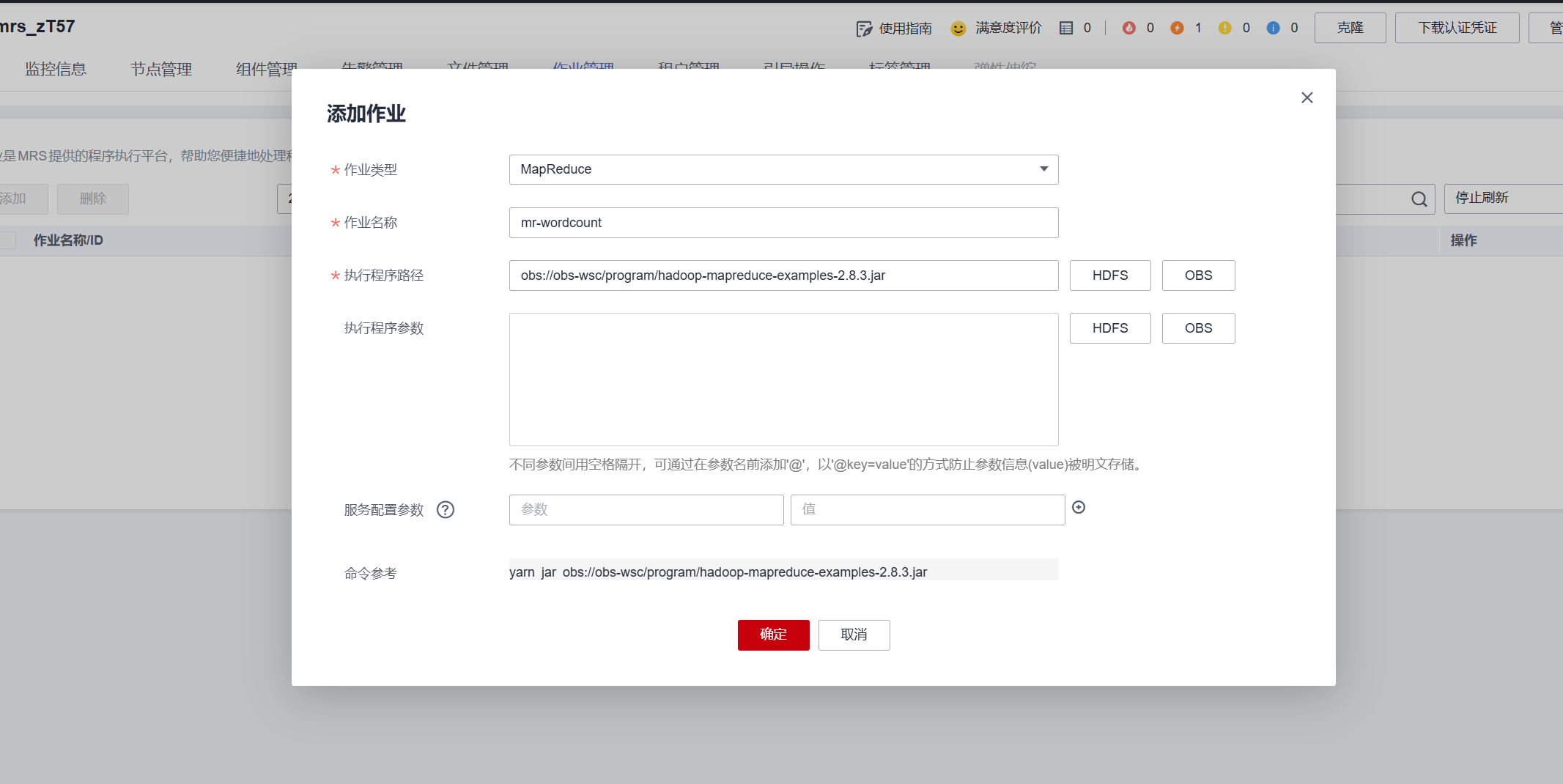


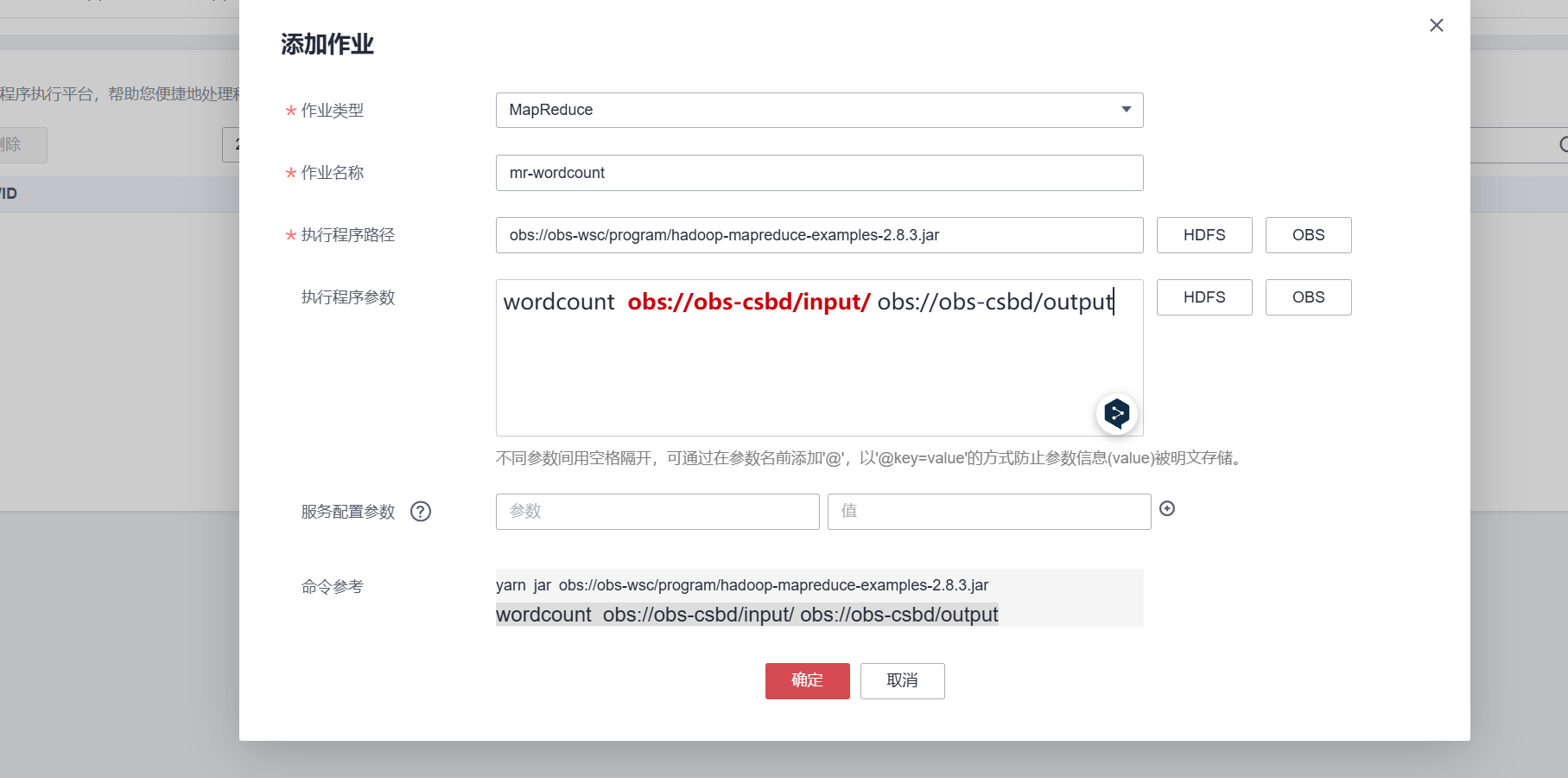
 

绑定:



作业部分:



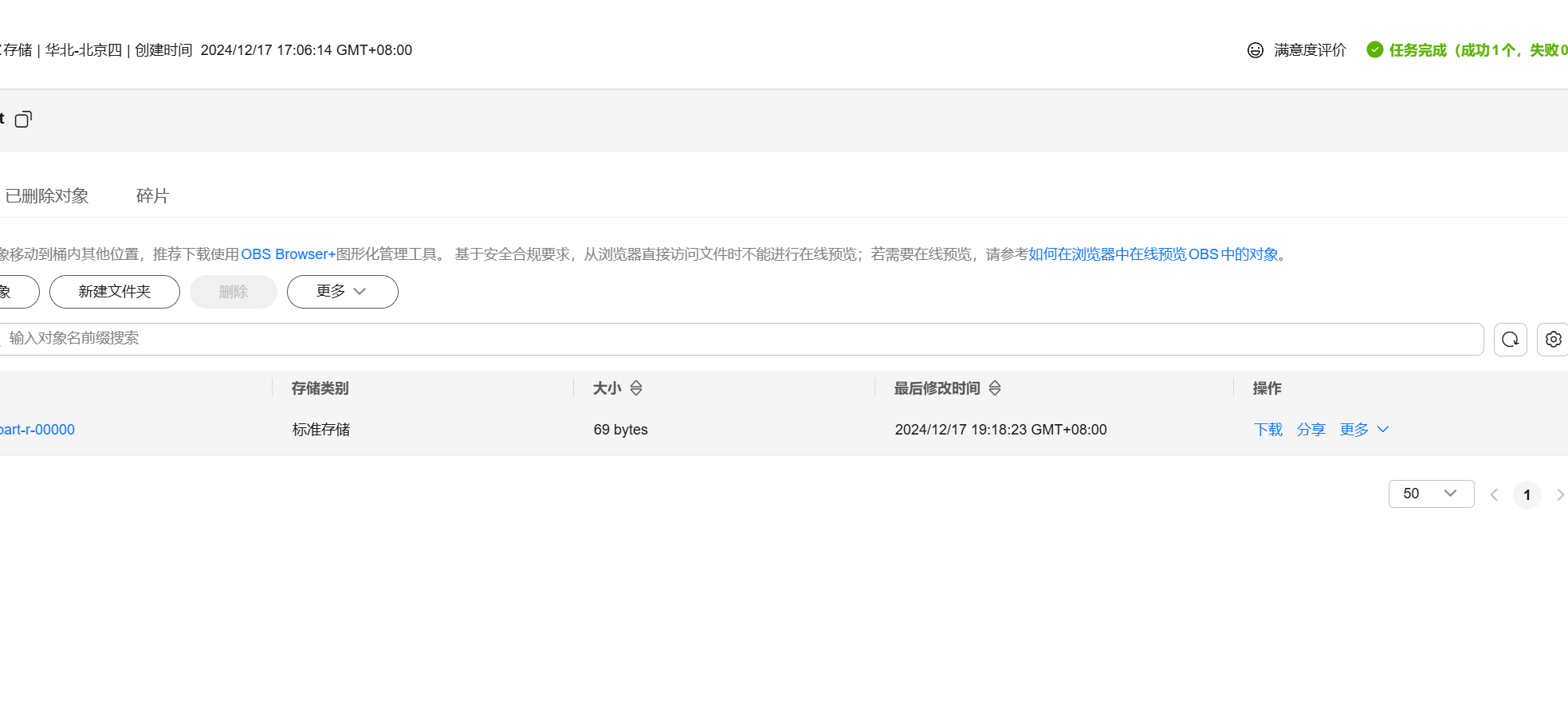


* 查看作业结果

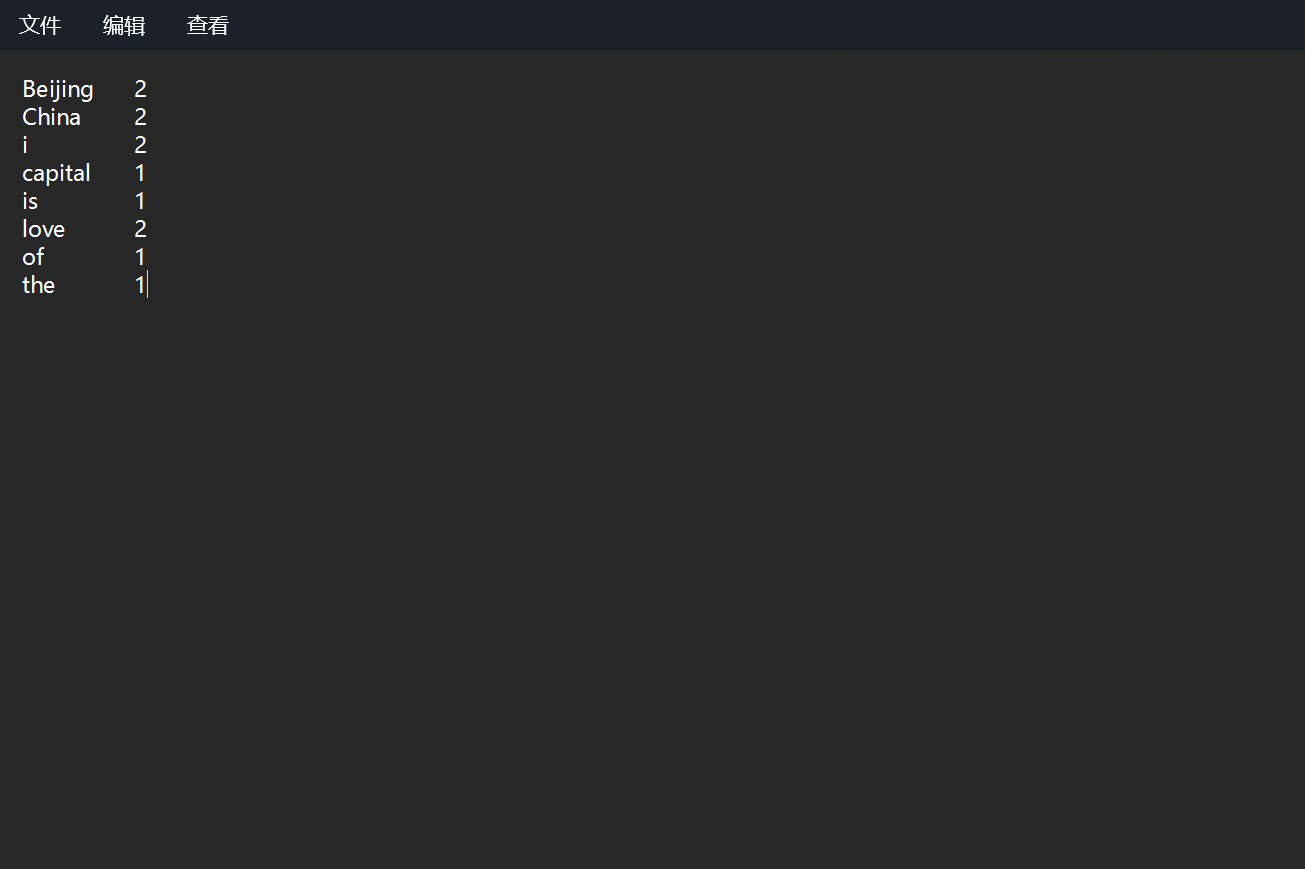




## 下载结果文件查看内容



* 成功运行:



## 思考题

### 我们的现在使用的是MapReduce作业，使用Spark呢?

MapReduce 作业

计算模型：基于批处理的编程模型，由“Map（映射）” 和 “Reduce（归约）”两个阶段组成。

数据处理：将数据分布到多个节点上，通过键值对进行分布式计算。

特点：每一个 MapReduce 任务需要经过中间数据写入磁盘的过程（如 HDFS），从而导致性能较慢。

Spark 作业

计算模型：基于内存的分布式计算框架，核心是RDD（弹性分布式数据集）。

数据处理：数据在内存中进行多次迭代计算，减少磁盘 I/O，提高速度。

特点：支持批处理、流处理、交互式查询等多种计算模式。

具体方法:

选择集群类型：**分析型 Hadoop 集群（包括 Spark 组件）**

编写 Spark 作业代码（例如使用 Python、Scala、Java）。

将输入数据和作业文件上传到 **OBS（对象存储服务）** 或 **HDFS**。

点击“**新建作业**”，选择作业类型为 **Spark**。

后面基本一致

# 实验结果或结论

## 结果

* 实验结果正确,输出了正确的答案



## 结论与收获

* 在此次实现中，我使用了华为云服务的云计算平台，基于Hadoop组件实现了MapReduce算法。这一过程中，我对于Hadoop集群的搭建，包括云上虚拟机创建、节点配置以及运行环境的部署，有了更深刻的了解。同时，了解到MapReduce的工作原理，包括Map阶段的分考和Reduce阶段的强聚总结，通过一些实际代码操作，我技术的熟练度有了大幅度提高。
* 华为云服务提供的平台高效且便捷，在体验过程中，我感受到了云服务的优势，包括开放的API和可接入的工具集，让我很快地完成了虚拟机部署以及Hadoop组件的运行。并且，华为云的安全性和可靠性也使我更方便地进行大数据的分析和处理。