华为云

BigDataPro集群搭建指南



华为技术有限公司

目录

[1 准备华为云环境 3](#_Toc174952082)

[1.1 实验介绍 3](#_Toc174952083)

[1.1.1 关于本实验 3](#_Toc174952084)

[1.1.2 实验目的 3](#_Toc174952085)

[1.2 购买华为云ECS 3](#_Toc174952086)

[1.2.1 登录华为云 3](#_Toc174952087)

[1.2.2 购买华为云ECS 4](#_Toc174952088)

[1.3 购买OBS 8](#_Toc174952089)

[1.3.1 选取对象存储服务OBS 8](#_Toc174952090)

[2 搭建Hadoop集群 13](#_Toc174952091)

[2.1 实验介绍 13](#_Toc174952092)

[2.1.1 关于本实验 13](#_Toc174952093)

[2.1.2 实验目的 13](#_Toc174952094)

[2.2 Hadoop集群搭建 14](#_Toc174952095)

[2.2.1 配置ECS 14](#_Toc174952096)

[2.2.2 确认JDK是否安装 18](#_Toc174952097)

[2.3 搭建Hadoop集群 18](#_Toc174952098)

[2.3.1 搭建Hadoop集群 18](#_Toc174952099)

[2.3.2 测试与OBS互联 26](#_Toc174952100)

[3 Spark集群搭建 29](#_Toc174952101)

[3.1 实验介绍 29](#_Toc174952102)

[3.1.1 关于本实验 29](#_Toc174952103)

[3.1.2 实验目的 29](#_Toc174952104)

[3.2 Spark集群存算分离 29](#_Toc174952105)

[3.2.1 搭建Spark集群 29](#_Toc174952106)

[3.2.2 验证存算分离 30](#_Toc174952107)

[4 释放实验资源 32](#_Toc174952108)

[4.1 删除ECS弹性云服务器 32](#_Toc174952109)

[4.2 删除OBS对象存储 33](#_Toc174952110)

[4.3 资源检查 33](#_Toc174952111)

[5 附录 安装OpenJDK 33](#_Toc174952112)

[6 附录 Linux常用命令 34](#_Toc174952113)

[6.1 ls命令 34](#_Toc174952114)

[6.2 cd命令 34](#_Toc174952115)

[6.3 pwd命令 34](#_Toc174952116)

[6.4 mkdir命令 35](#_Toc174952117)

[6.5 rm命令 35](#_Toc174952118)

[6.6 rmdir命令 35](#_Toc174952119)

[6.7 mv命令 36](#_Toc174952120)

[6.8 cp命令 36](#_Toc174952121)

[6.9 cat 命令 36](#_Toc174952122)

[6.10 more命令 37](#_Toc174952123)

[6.11 less命令 38](#_Toc174952124)

[6.12 head命令 39](#_Toc174952125)

[6.13 tail命令 39](#_Toc174952126)

[6.14 vim命令 39](#_Toc174952127)

[6.14.1 启动vim 39](#_Toc174952128)

[6.14.2 vim的模式 40](#_Toc174952129)

[6.14.3 插入命令 40](#_Toc174952130)

[6.14.4 查找命令 40](#_Toc174952131)

[6.14.5 替换命令 40](#_Toc174952132)

[6.14.6 撤销和重做 40](#_Toc174952133)

[6.14.7 删除命令 40](#_Toc174952134)

[6.14.8 拷贝和粘贴 41](#_Toc174952135)

[6.14.9 退出命令 41](#_Toc174952136)

[6.14.10 帮助命令 41](#_Toc174952137)

[7 附录 实验数据Playerinfo 42](#_Toc174952138)

# 准备华为云环境

## 实验介绍

### 关于本实验

本实验通过购买华为云OBS和华为云ECS服务，提供后续集群的搭建基础。本实验的基本步骤包含：购买并配置ECS；购买OBS并获取AK、SK信息；搭建Hadoop集群；搭建Spark集群并验证存算分离。

学生可根据实验手册进行逐步的练习，完成各实验内容。

### 实验目的

掌握华为云OBS的购买。

掌握华为云ECS的购买。

## 购买华为云ECS

### 登录华为云

浏览器登录华为云

华为云地址：https://www.huaweicloud.com/

输入用户名、密码，登录。（如没有账号需点击注册按钮进行账号注册）

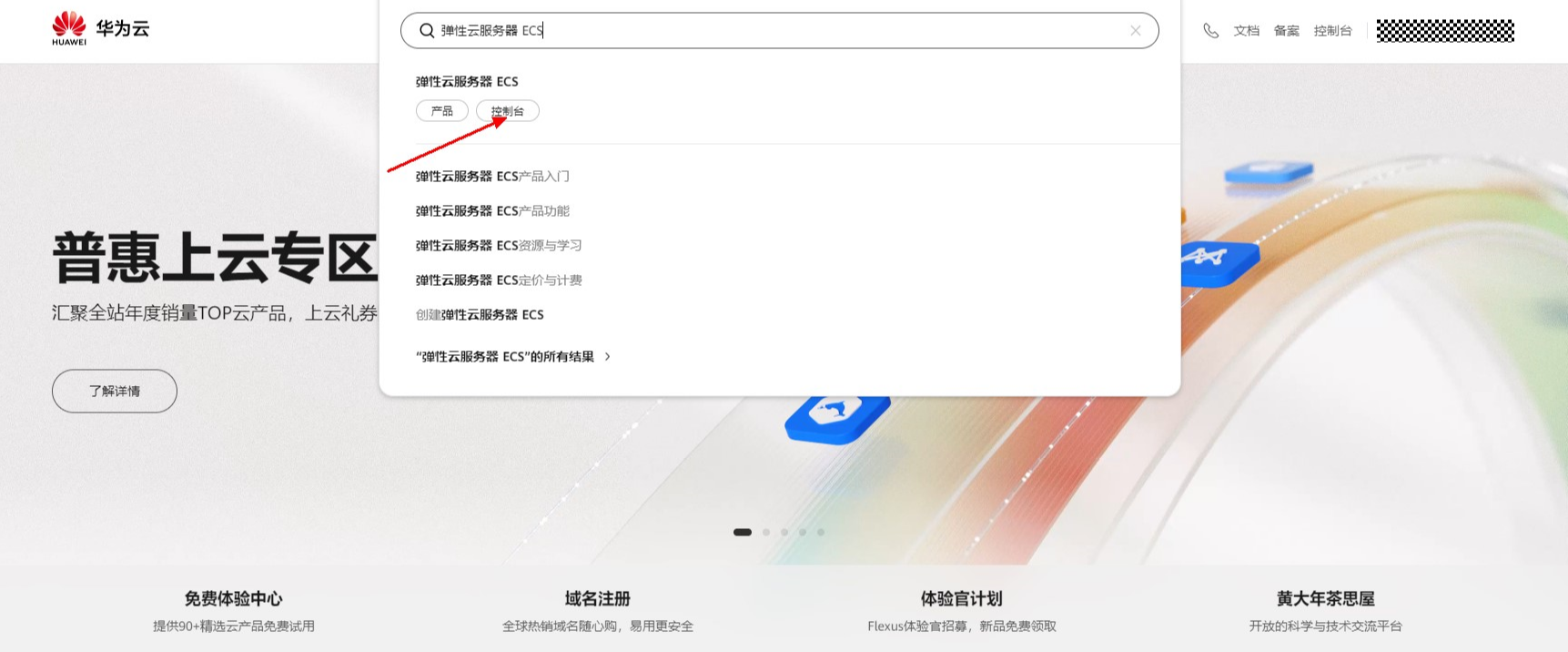




### 购买华为云ECS

搜索框中输入“弹性云服务器 ECS”，点击“控制台”跳转到弹性云服务器ECS

控制台



选择“购买弹性云服务器”



本次ECS购买需遵循以下规格：

计费模式：按需计费

可用区：可用区一

CPU架构：鲲鹏计算

规格：kc1.2xlarge.4 8vcpus|32GB

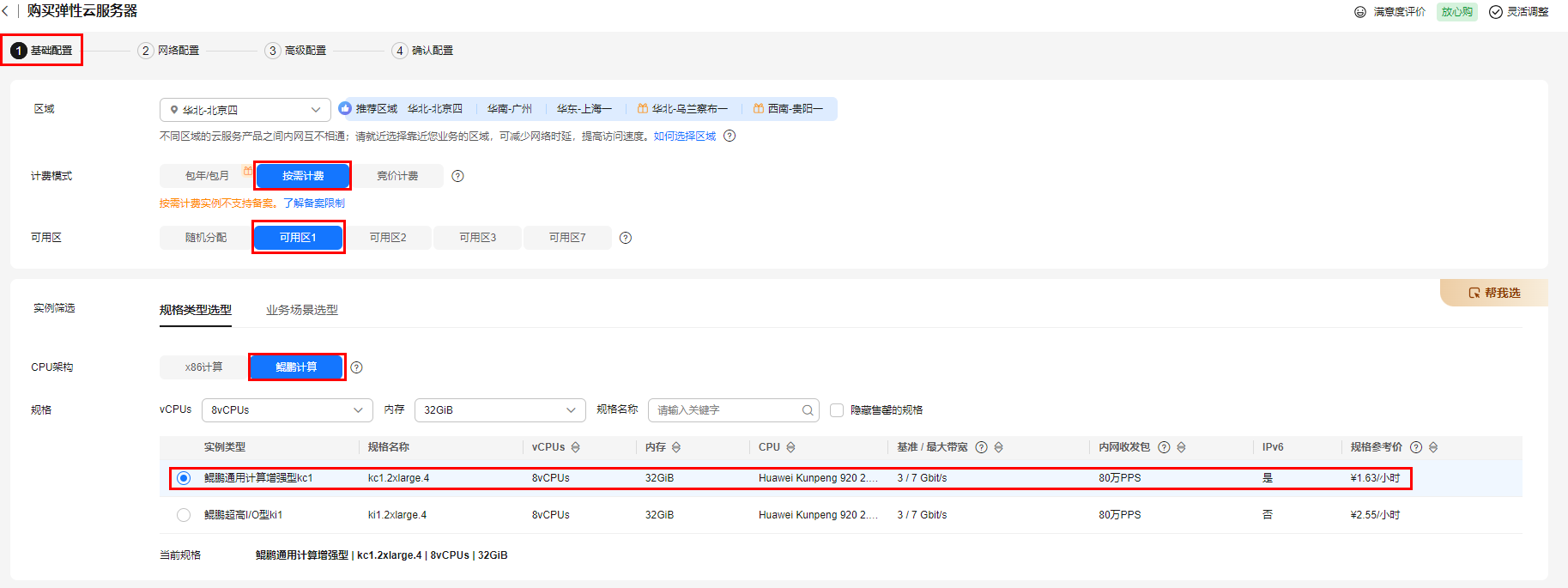
系统：CentOS7.6

系统盘：超高IO、200GB。

网络：按流量计费

其他配置：默认

基础配置



系统选择CentOS7.6

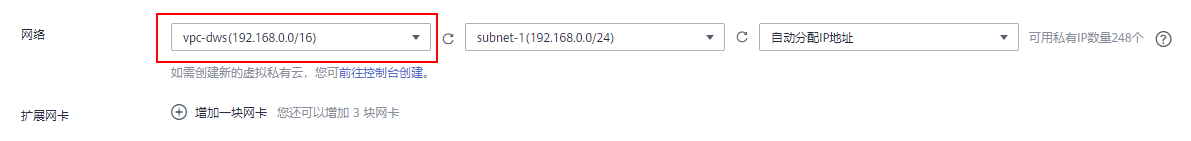
系统盘建议配置200GB，不建议低于100GB。

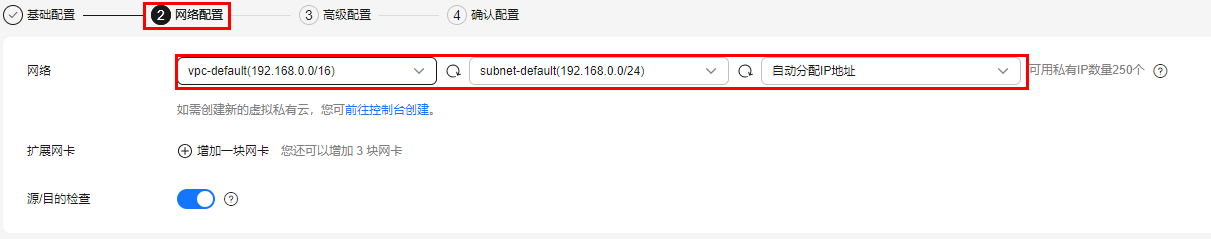


点击“下一步 网络配置”

网络配置

“网络”下拉选择VPC





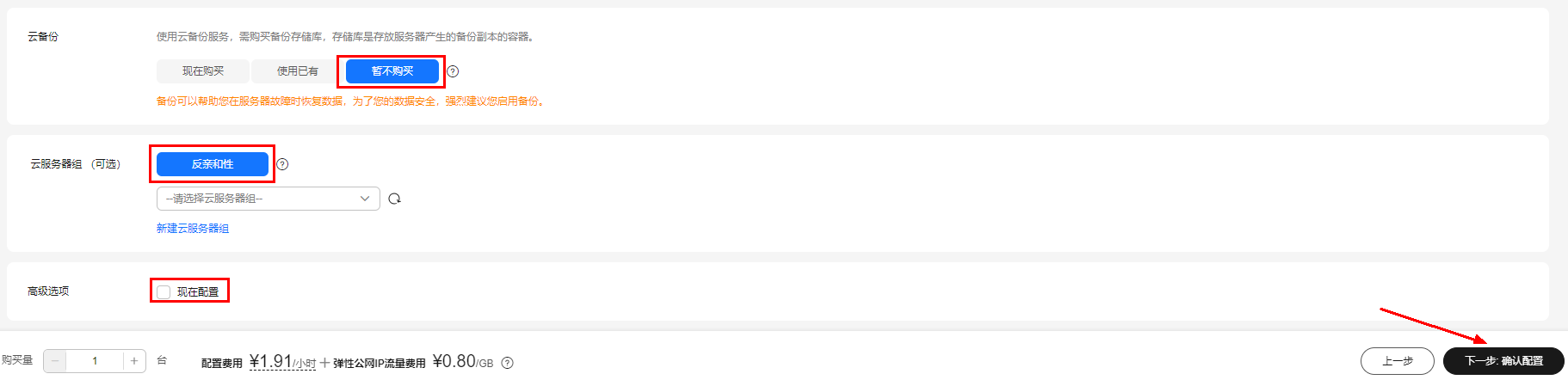
“弹性公网IP”选择“现在购买”，并选择“按流量计费”。



点击“下一步 高级配置”

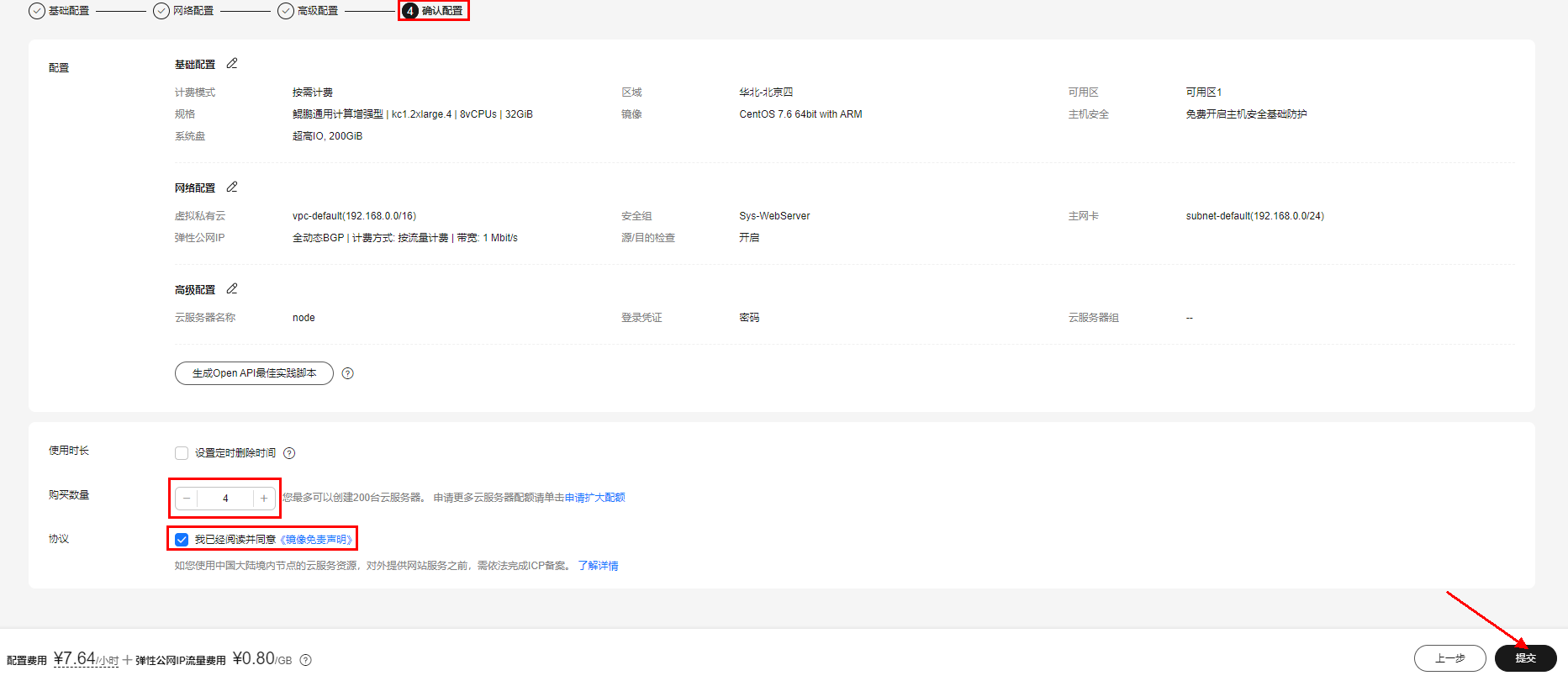
高级配置





ECS名称建议与手册相同，密码自行设定。其他配置默认即可，点击“下一步 确认配置”。

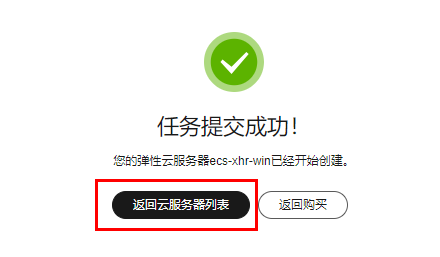
检查配置，点击框线处“我已经阅读并同意[《镜像免责声明》](http://www.huaweicloud.com/declaration/tsa_ims.html?locale=zh-cn)”, 点击“提交”立即购买ECS



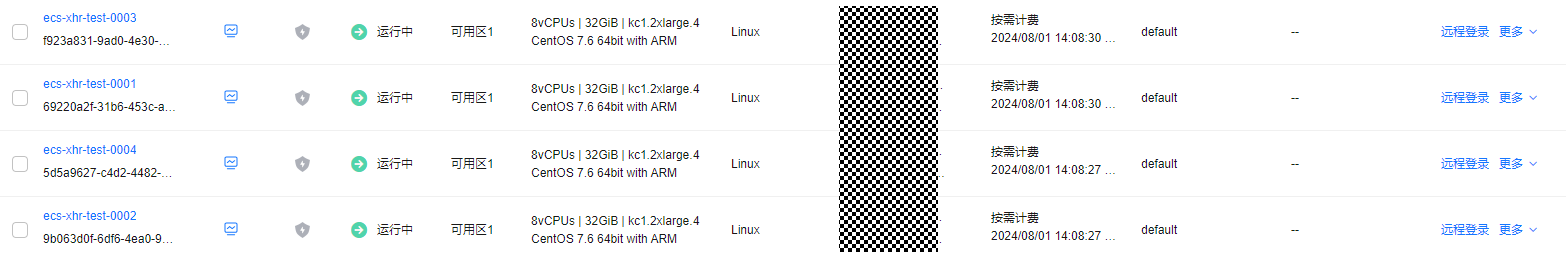
注：本次实验也可使用3台ECS实验。

查看已购买ECS

“步骤6”点击“提交”后，弹出如下界面。



点击“返回云服务器列表”，查看已购买ECS。



## 购买OBS

### 选取对象存储服务OBS

在首页搜索框中输入“对象存储服务 OBS”，点击“控制台”跳转到对象存储服务 OBS

控制台



选择创建桶

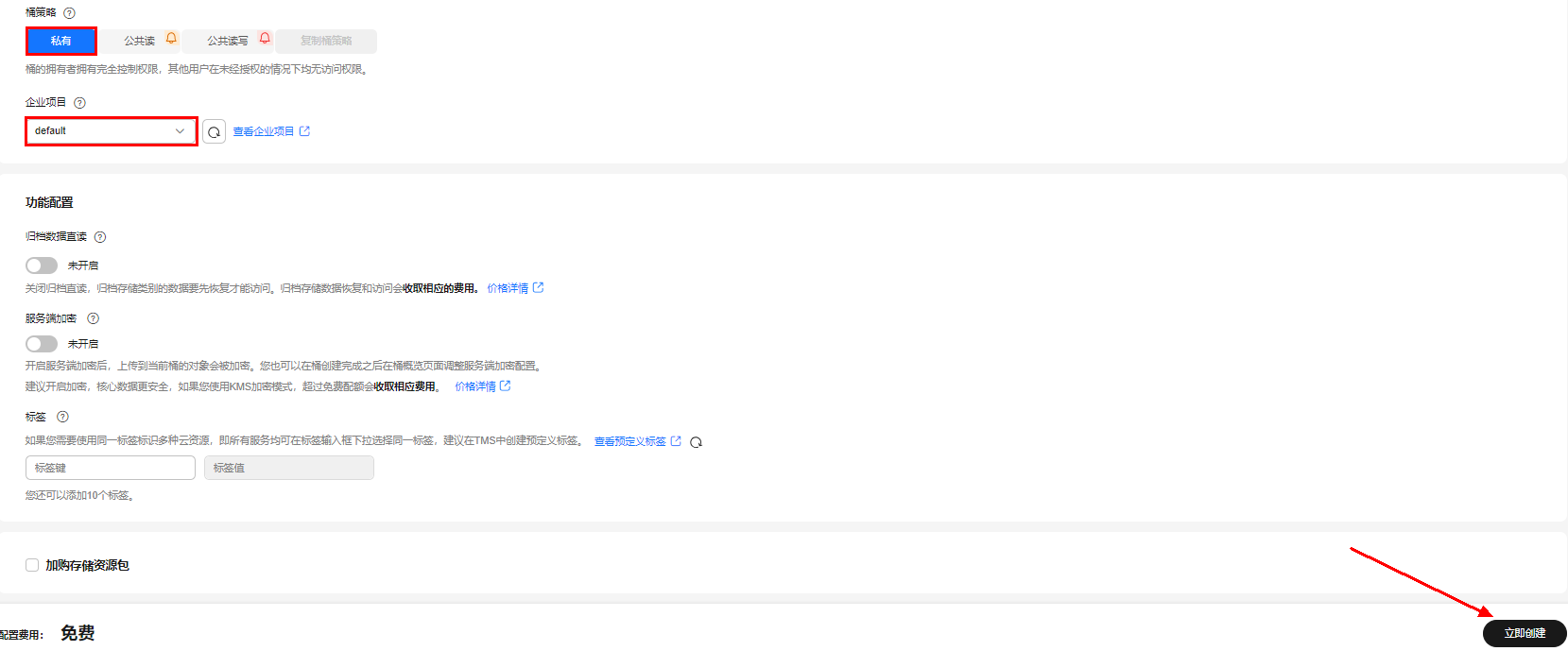
进入OBS服务后，点击“创建桶”

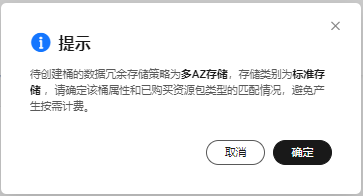


参照下图购买OBS





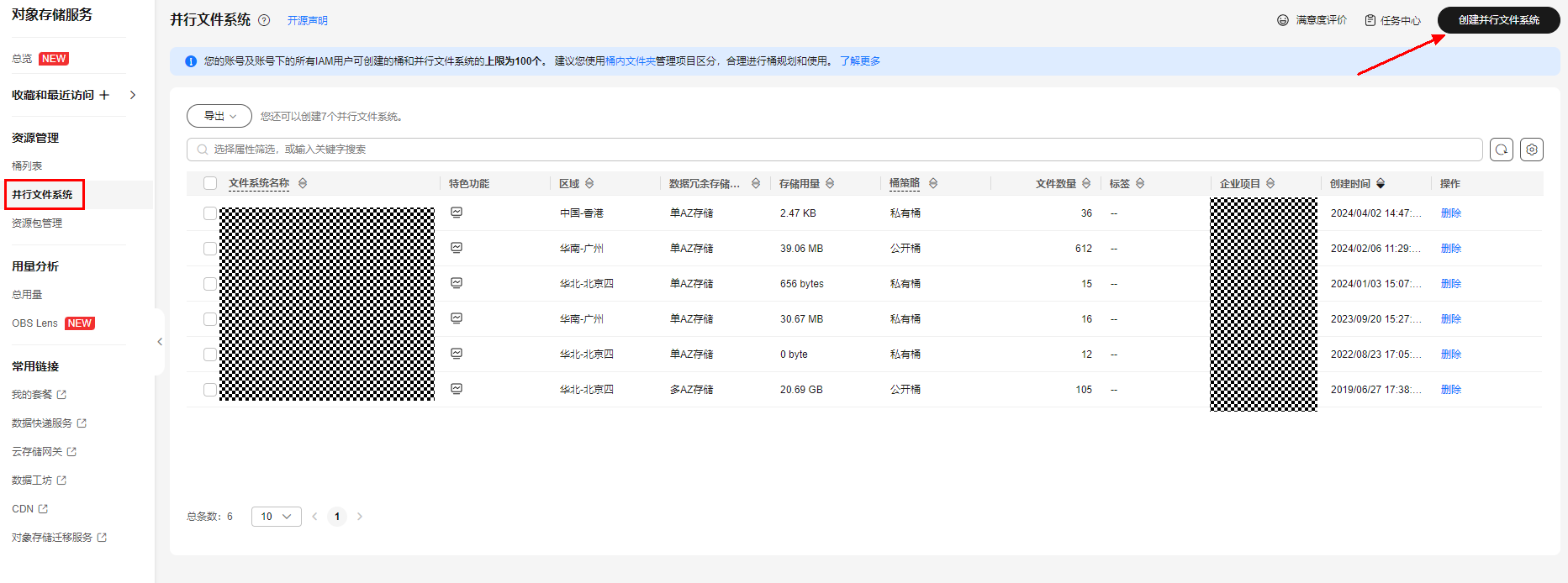




注：区域选择与ECS相同。其他配置默认。点击“立即创建”，然后点击“确认”。

创建并行文件系统

选择“并行文件系统”



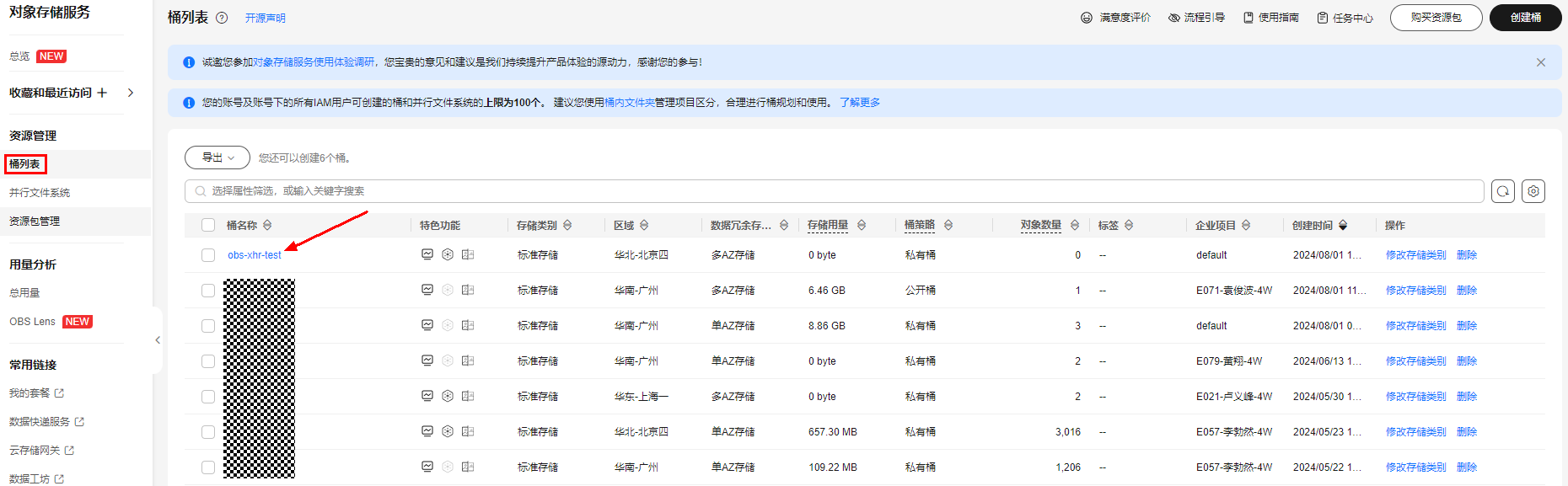
点击右上角“创建并行文件系统”

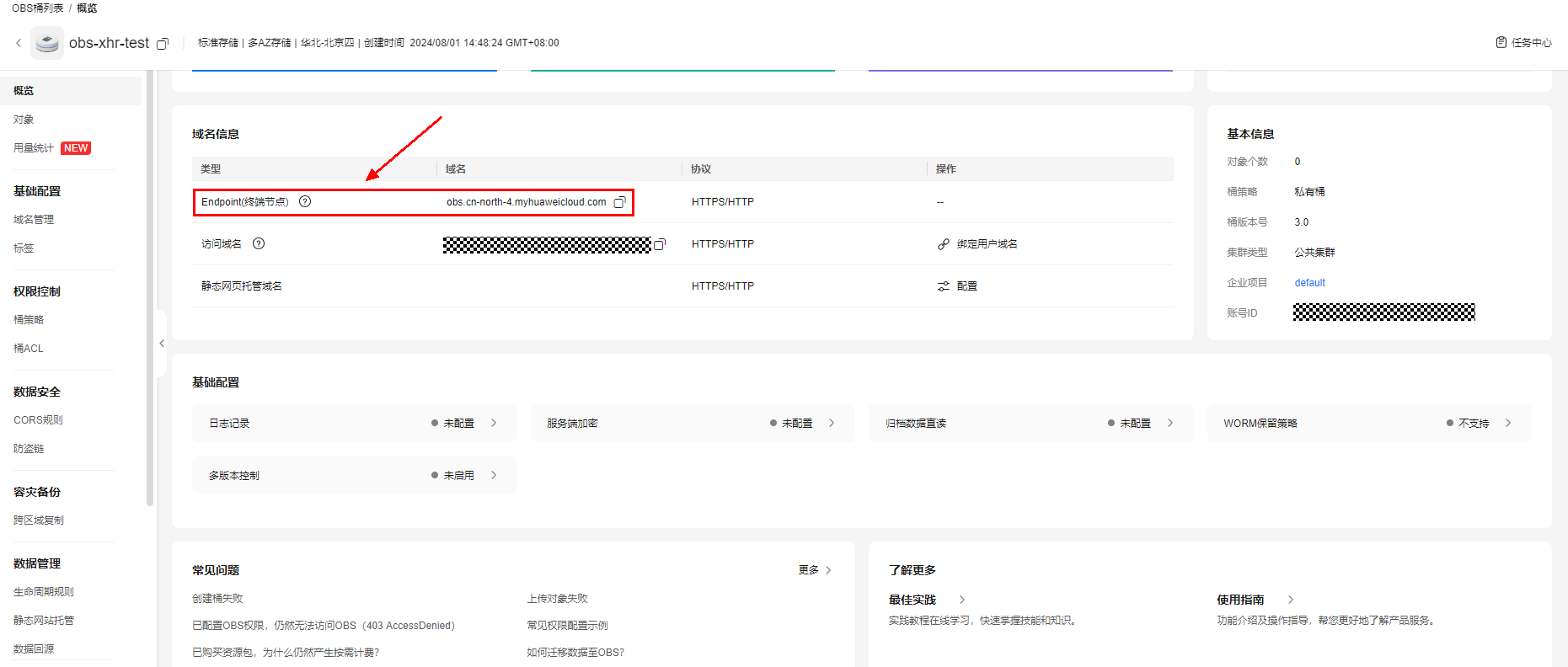




注：区域选择与ECS相同。其他配置默认，点击“立即创建”。

进入创建的OBS桶

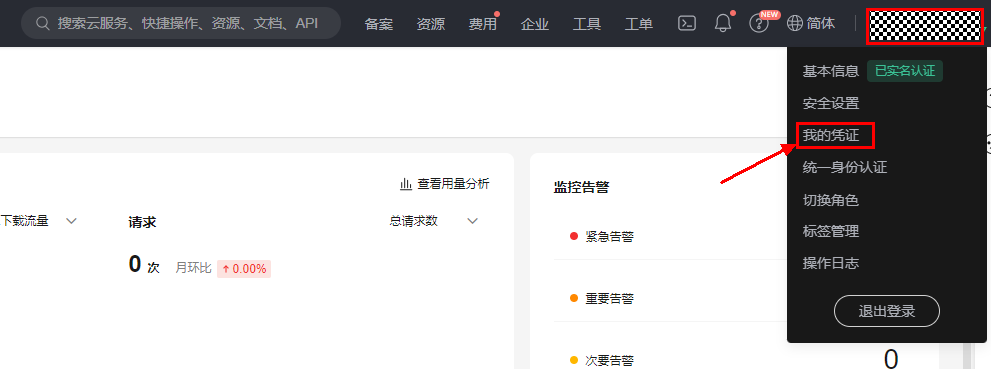




复制该参数，保存到本地文档。

获取AK/SK

点击华为云页面右上角“用户名”，下拉选择“我的凭证”，



点击“访问秘钥”。



点击新增访问秘钥，创建秘钥，然后点击“立即下载”。



得到文件“credentials.csv”，打开即可得到AK/SK。



# 搭建Hadoop集群

## 实验介绍

### 关于本实验

本实验通过搭建开源Hadoop集群帮助学员掌握Hadoop搭建方法。并且使开源Hadoop与华为云OBS服务互联，使Hadoop集群可读取OBS数据。

### 实验目的

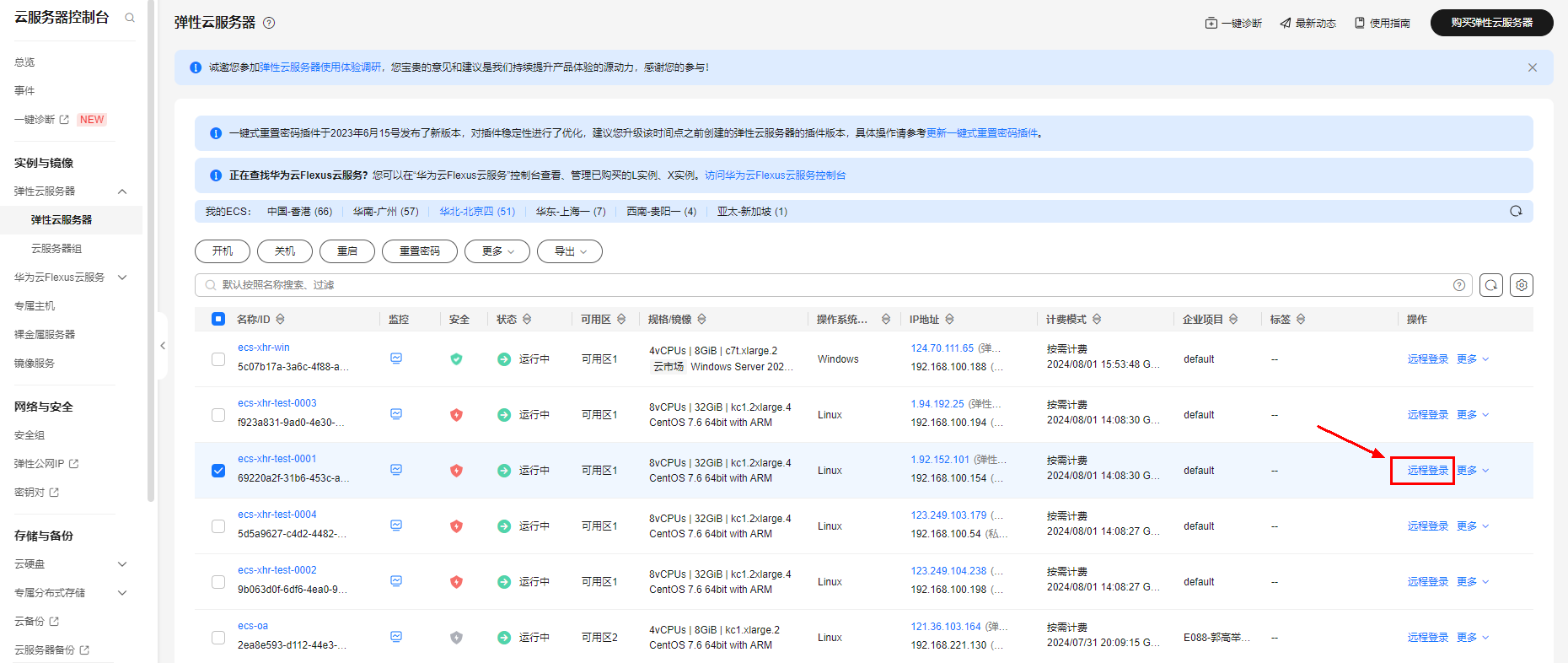
掌握Hadoop集群搭建方法

掌握Hadoop集群与华为云OBS互联方法

## Hadoop集群搭建

### 配置ECS

在ECS页面，点击“远程登录”，弹出框中下拉找到控制台的“立即登录”，登录ECS



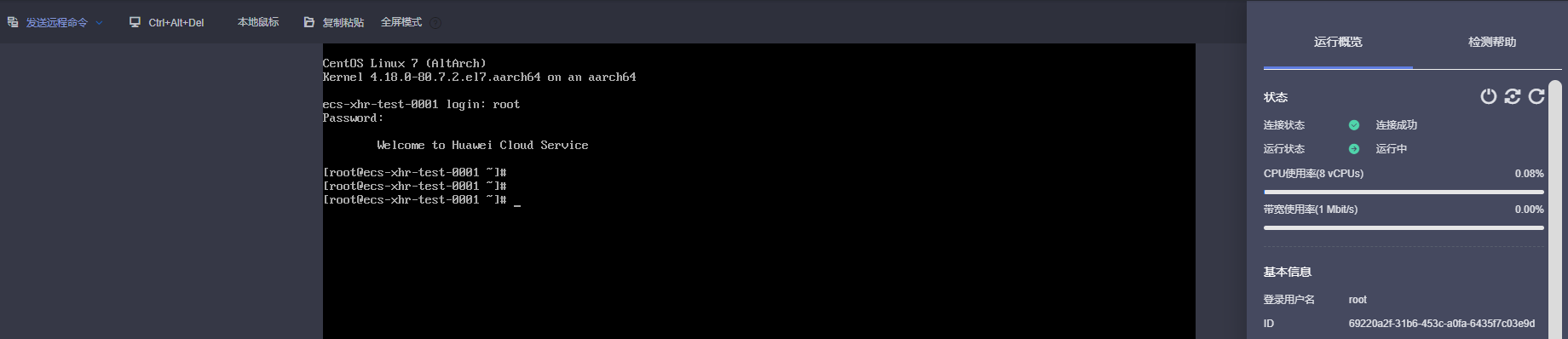
这里推荐使用华为提供的**CodeArts IDE for RemoteShel**来远程连接ecs虚拟机。支持文件传输以及命令复制



其他方式则可以直接在页面登录，但是不支持文件上传和命令复制等功能，建议环境部署完毕后查看系统信息时直接页面登陆查看。



通过root账号登录虚拟机（密码在ecs初始化时设置，如有忘记可在右侧的“重置密码”功能重置）



下载安装包

（1）下载hadoop安装包

cd /opt

wget https://archive.apache.org/dist/hadoop/common/hadoop-3.1.1/hadoop-3.1.1.tar.gz

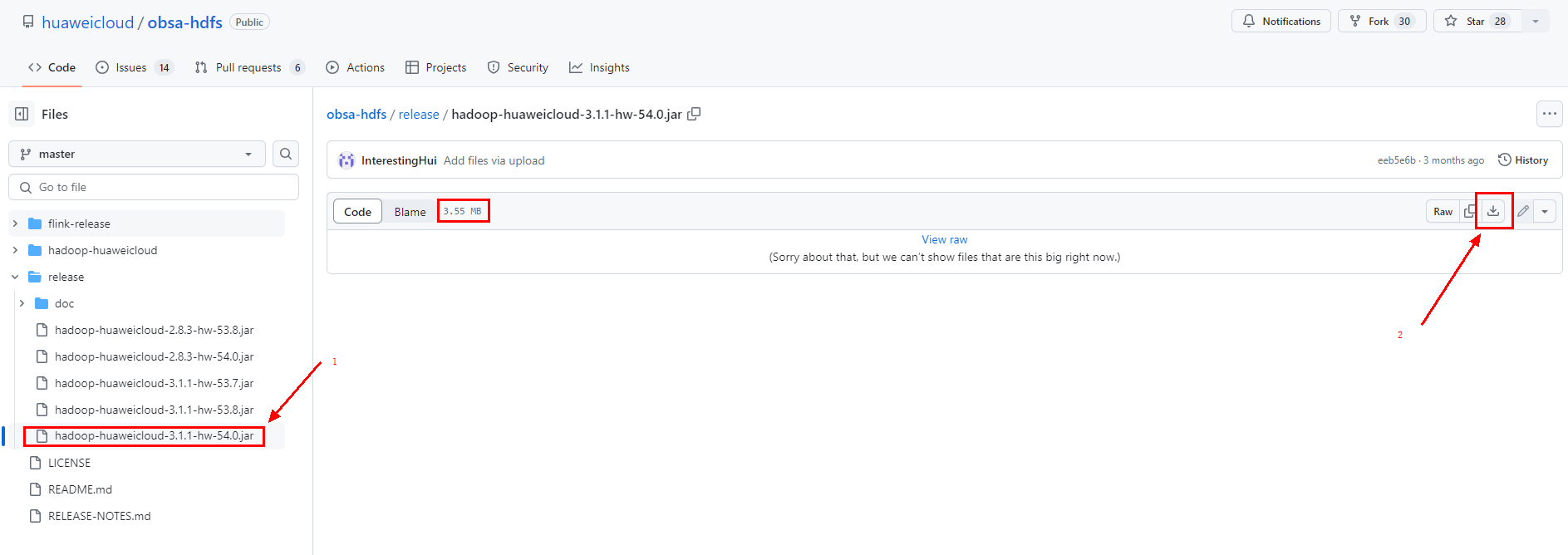
（2）下载OBSFileSystem相关jar包

执行命令，下载hadoop-huaweicloud-3.1.1-hw-54.0.jar

cd /opt

wget <https://github.com/huaweicloud/obsa-hdfs/blob/master/release/hadoop-huaweicloud-3.1.1-hw-54.0.jar>

需要注意的是，下载下来的hadoop-huaweicloud-3.1.1-hw-54.0.jar 包，需要确认大小是否正常（3.5MB）,如果直接wget下载的包大小异常，请登录<https://github.com/huaweicloud/obsa-hdfs/tree/master/release> 页面进行页面下载然后再传到ecs机器上。



配置/etc/hosts文件

查看ECS列表各节点IP



各节点执行

vim /etc/hosts

备注：vim编辑器的使用

（1） 点击键盘字母i，输入数据；

（2） 数据输入完成后，点击键盘esc，输入:wq即完成数据保存和退出编辑界面。

添加内容：

192.168.100.154 ecs-xhr-test-0001

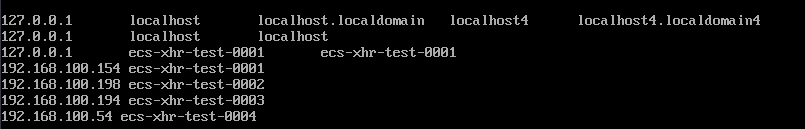
192.168.100.198 ecs-xhr-test-0002

192.168.100.194 ecs-xhr-test-0003

192.168.100.54 ecs-xhr-test-0004

（注意：ecs-xhr-test-0001为申请的ecs机器名）

参考如下：



保存退出。

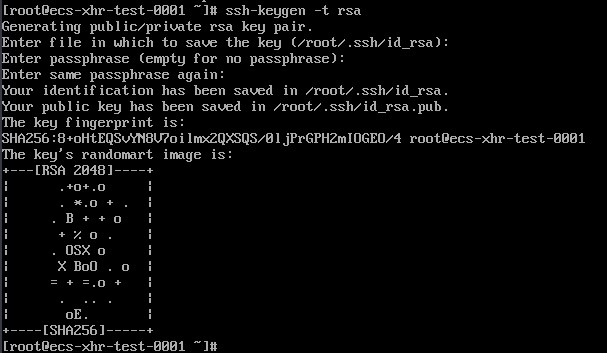
注：此处需要根据ECS的实际内网IP配置。

配置节点互信

各节点执行

ssh-keygen -t rsa

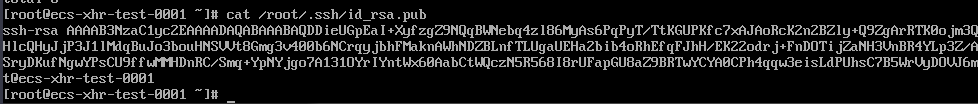
连续回车生成/root/.ssh/id\_rsa.pub文件；



各节点执行：

cat /root/.ssh/id\_rsa.pub

输出如下内容：



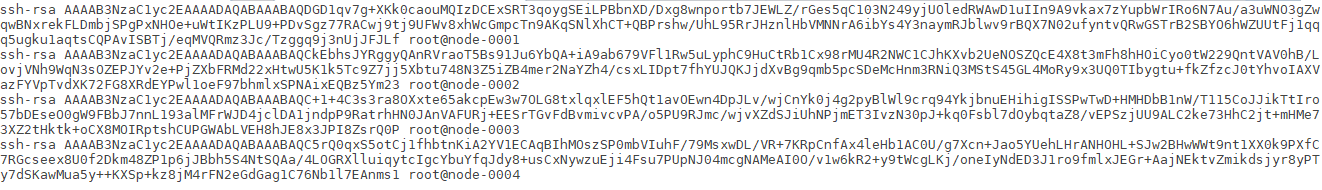
复制该命令在各节点的输出内容。

各节点执行：

vim /root/.ssh/authorized\_keys

输入各节点的复制内容，保存退出。

authorized\_keys内容参考如下：



各节点执行：

ssh ecs-xhr-test-0001

ssh ecs-xhr-test-0002

ssh ecs-xhr-test-0003

ssh ecs-xhr-test-0004，选择yes后，确保能够无密码跳转到目标节点。

注意：ssh链接其他服务器之后，需要输入exit退出连接然后再进行其他机器的连接。



### 确认JDK是否安装

确认是否已安装JDK

各节点执行命令：

java -version

若输出如下结果，则ECS已安装JDK，无需执行OpenJDK安装操作。

openjdk version "1.8.0\_232"

OpenJDK Runtime Environment (build 1.8.0\_232-b09)

OpenJDK 64-Bit Server VM (build 25.232-b09, mixed mode)

若无则参考附录5[安装OpenJDK](#_附录_安装OpenJDK)。

## 搭建Hadoop集群

### 搭建Hadoop集群

创建目录

各节点执行：

mkdir -p /opt/data/buf

mkdir -p /opt/nm/localdir

登录ecs-xhr-test-0001节点，解压hadoop安装包

cd /opt

tar -zxvf hadoop-3.1.1.tar.gz

配置hadoop core-site.xml配置文件

ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/core-site.xml

参数配置如下：

<configuration>

<property>

<name>fs.obs.readahead.inputstream.enabled</name>

<value>true</value>

</property>

<property>

<name>fs.obs.buffer.max.range</name>

<value>6291456</value>

</property>

<property>

<name>fs.obs.buffer.part.size</name>

<value>2097152</value>

</property>

<property>

<name>fs.obs.threads.read.core</name>

<value>500</value>

</property>

<property>

<name>fs.obs.threads.read.max</name>

<value>1000</value>

</property>

<property>

<name>fs.obs.write.buffer.size</name>

<value>8192</value>

</property>

<property>

<name>fs.obs.read.buffer.size</name>

<value>8192</value>

</property>

<property>

<name>fs.obs.connection.maximum</name>

<value>1000</value>

</property>

<property>

<name>fs.defaultFS</name>

<value>hdfs://ecs-xhr-test-0001:8020</value>

</property>

<property>

<name>hadoop.tmp.dir</name>

<value>/opt/hadoop-3.1.1/tmp</value>

</property>

<property>

<name>fs.obs.access.key</name>

<value>U6YSHXXWCOHJGCBCMOQG</value>

</property>

<property>

<name>fs.obs.secret.key</name>

<value>kjHdB4s1ldylm1sv0CwAC73UK075LCJ7E7qaBuAF</value>

</property>

<property>

<name>fs.obs.endpoint</name>

<value>obs.cn-north-4.myhuaweicloud.com:5080</value>

</property>

<property>

<name>fs.obs.buffer.dir</name>

<value>/opt/data/buf</value>

</property>

<property>

<name>fs.obs.impl</name>

<value>org.apache.hadoop.fs.obs.OBSFileSystem</value>

</property>

<property>

<name>fs.obs.connection.ssl.enabled</name>

<value>false</value>

</property>

<property>

<name>fs.obs.fast.upload</name>

<value>true</value>

</property>

<property>

<name>fs.obs.socket.send.buffer</name>

<value>65536</value>

</property>

<property>

<name>fs.obs.socket.recv.buffer</name>

<value>65536</value>

</property>

<property>

<name>fs.obs.max.total.tasks</name>

<value>20</value>

</property>

<property>

<name>fs.obs.threads.max</name>

<value>20</value>

</property>

</configuration>

注：fs.obs.access.key、fs.obs.secret.key、fs.obs.endpoint、fs.defaultFS需根据实际情况修改。

fs.obs.access.key：[credentials.csv（步骤6）](#_选取对象存储服务OBS) 文件中的Access Key Id

fs.obs.secret.key：[credentials.csv（步骤6）](#_选取对象存储服务OBS) 文件中的Secret Access Key

fs.obs.endpoint：[credentials.csv（步骤5）](#_选取对象存储服务OBS) 文件中的Endpoint(终端节点)

fs.defaultFS：ecs-xhr-test-0001 替换为自己机器名称

配置hdfs-site.xml

ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/hdfs-site.xml

参数配置如下：

<configuration>

<property>

<name>dfs.replication</name>

<value>3</value>

</property>

<property>

<name>dfs.namenode.secondary.http-address</name>

<value>ecs-xhr-test-0001:50090</value>

</property>

<property>

<name>dfs.namenode.secondary.https-address</name>

<value>ecs-xhr-test-0001:50091</value>

</property>

</configuration>

注：主机名ecs-xhr-test-0001需要根据实际替换为自己的ecs机器名。

配置yarn-site.xml

ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/yarn-site.xml

参数配置如下：

<configuration>

<property>

<name>yarn.nodemanager.local-dirs</name>

<value>/home/nm/localdir</value>

</property>

<property>

<name>yarn.nodemanager.resource.memory-mb</name>

<value>28672</value>

</property>

<property>

<name>yarn.scheduler.minimum-allocation-mb</name>

<value>3072</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-mb</name>

<value>28672</value>

</property>

<property>

<name>yarn.nodemanager.resource.cpu-vcores</name>

<value>38</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-vcores</name>

<value>38</value>

</property>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.resourcemanager.hostname</name>

<value>ecs-xhr-test-0001</value>

</property>

<property>

<name>yarn.log-aggregation-enable</name>

<value>true</value>

</property>

<property>

<name>yarn.log-aggregation.retain-seconds</name>

<value>106800</value>

</property>

<property>

<name>yarn.nodemanager.vmem-check-enabled</name>

<value>false</value>

<description>Whether virtual memory limits will be enforced for containers</description>

</property>

<property>

<name>yarn.nodemanager.vmem-pmem-ratio</name>

<value>4</value>

<description>Ratio between virtual memory to physical memory when setting memory limits for containers</description>

</property>

<property>

<name>yarn.resourcemanager.scheduler.class</name>

<value>org.apache.hadoop.yarn.server.resourcemanager.scheduler.fair.FairScheduler</value>

</property>

<property>

<name>yarn.log.server.url</name>

<value>http://ecs-xhr-test-0001:19888/jobhistory/logs</value>

</property>

</configuration>

注：主机名ecs-xhr-test-0001需要根据实际替换为自己的ecs机器名。

配置mapred-site.xml

ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/mapred-site.xml

参数配置如下：

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

<property>

<name>mapreduce.jobhistory.address</name>

<value>ecs-xhr-test-0001:10020</value>

</property>

<property>

<name>mapreduce.jobhistory.webapp.address</name>

<value>ecs-xhr-test-0001:19888</value>

</property>

<property>

<name>mapred.task.timeout</name>

<value>1800000</value>

</property>

</configuration>

注：主机名ecs-xhr-test-0001需要根据实际替换为自己的ecs机器名。

配置slaves

ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/slaves

添加内容如下：

ecs-xhr-test-0002

ecs-xhr-test-0003

ecs-xhr-test-0004

配置hadoop环境变量

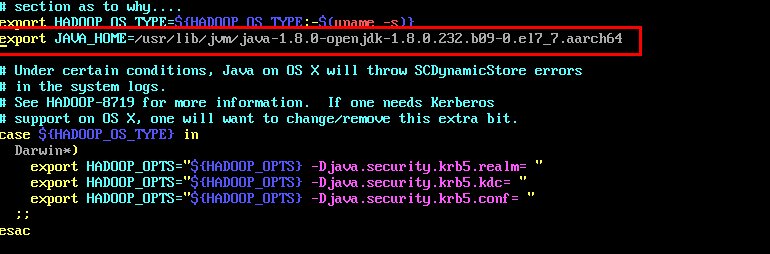
ecs-xhr-test-0001节点执行下列命令：

vim /opt/hadoop-3.1.1/etc/hadoop/hadoop-env.sh

添加如下内容：

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.232.b09-0.el7\_7.aarch64

参考如下：



配置jar包

ecs-xhr-test-0001节点执行下列命令：

cp /opt/hadoop-huaweicloud-3.1.1-hw-54.0.jar /opt/hadoop-3.1.1/share/hadoop/tools/lib/

cp /opt/hadoop-huaweicloud-3.1.1-hw-54.0.jar /opt/hadoop-3.1.1/share/hadoop/common/lib/

cp /opt/hadoop-huaweicloud-3.1.1-hw-54.0.jar /opt/hadoop-3.1.1/share/hadoop/hdfs/lib/

分发hadoop包到各节点

ecs-xhr-test-0001下执行下列命令：

分发hadoop包到node2

scp -r /opt/hadoop-3.1.1/ root@ecs-xhr-test-0002:/opt/

分发hadoop包到node3

scp -r /opt/hadoop-3.1.1/ root@ecs-xhr-test-0003:/opt/

分发hadoop包到node4

scp -r /opt/hadoop-3.1.1/ root@ecs-xhr-test-0004:/opt/

配置环境变量

各节点执行：

vim /etc/profile

在文件尾部添加如下内容，注意“=”两边没有空格：

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.232.b09-0.el7\_7.aarch64

export HADOOP\_HOME=/opt/hadoop-3.1.1

export PATH=$JAVA\_HOME/bin:$PATH

export PATH=$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin:$PATH

export HADOOP\_CLASSPATH=/opt/hadoop-3.1.1/share/hadoop/tools/lib/\*:$HADOOP\_CLASSPATH

各节点执行：

source /etc/profile

再在各节点执行

vim /opt/hadoop-3.1.1/sbin/start-dfs.sh

vim /opt/hadoop-3.1.1/sbin/stop-dfs.sh

在start-dfs.sh和stop-dfs.sh的头部添加以下root执行用户

HDFS\_ZKFC\_USER=root

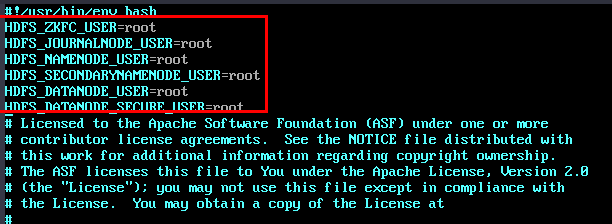
HDFS\_JOURNALNODE\_USER=root

HDFS\_NAMENODE\_USER=root

HDFS\_SECONDARYNAMENODE\_USER=root

HDFS\_DATANODE\_USER=root

HDFS\_DATANODE\_SECURE\_USER=root



相同的执行一下命令：

vim /opt/hadoop-3.1.1/sbin/start-yarn.sh

vim /opt/hadoop-3.1.1/sbin/stop-yarn.sh

在start-yarn.sh和stop-yarn.sh的头部添加以下root执行用户

#HADOOP\_SECURE\_DN\_USER=root

HDFS\_DATANODE\_SECURE\_USER=root

YARN\_NODEMANAGER\_USER=root

YARN\_RESOURCEMANAGER\_USER=root

namenode初始化

ecs-xhr-test-0001节点执行namenode初始化

执行下列命令：

hdfs namenode -format

初始化成功后，启动hdfs。

启动HDFS

ecs-xhr-test-0001节点执行

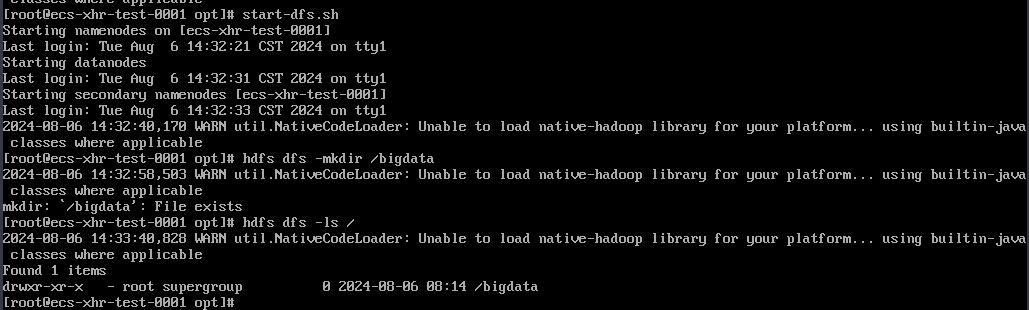
start-dfs.sh

执行hdfs命令

hdfs dfs -mkdir /bigdata

hdfs dfs -ls /

效果如下。即完成了Hadoop集群的搭建。

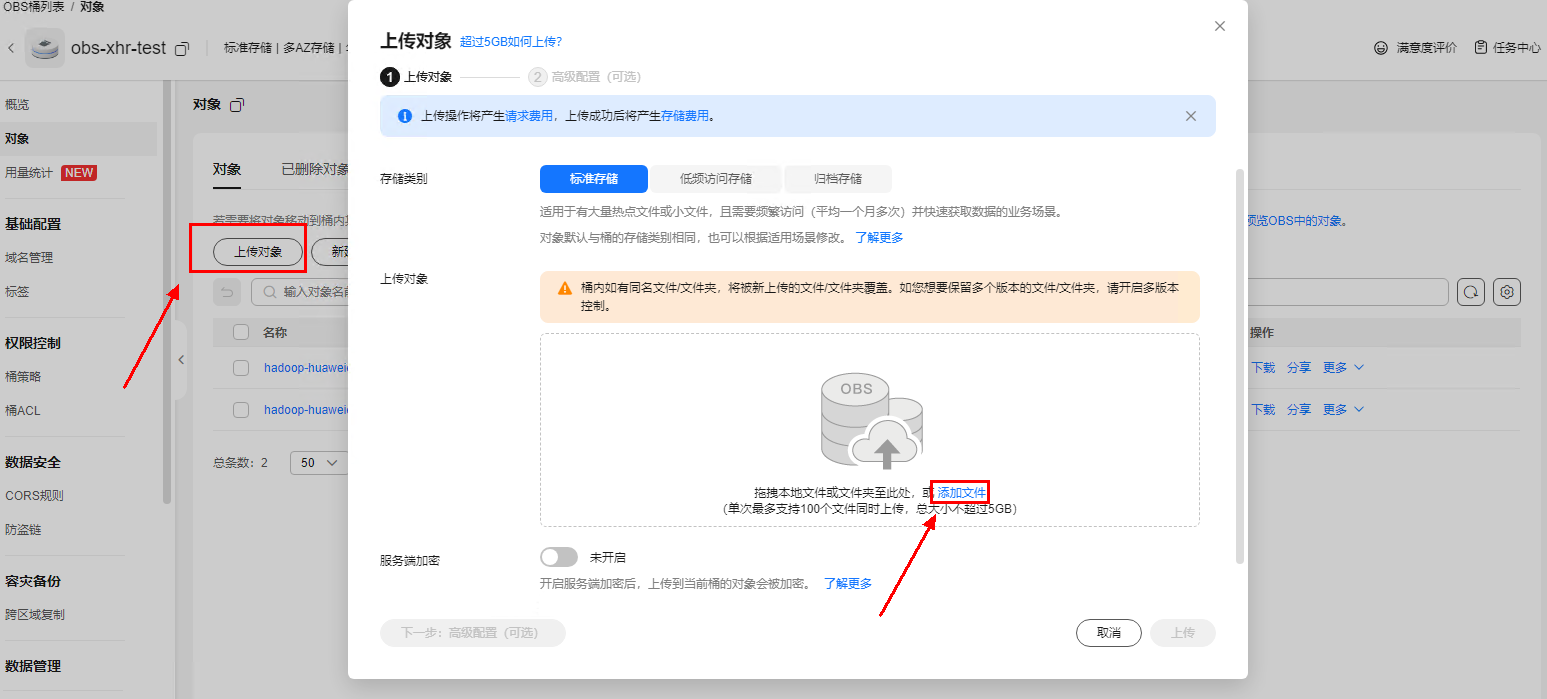


### 测试与OBS互联

在OBS上传文件

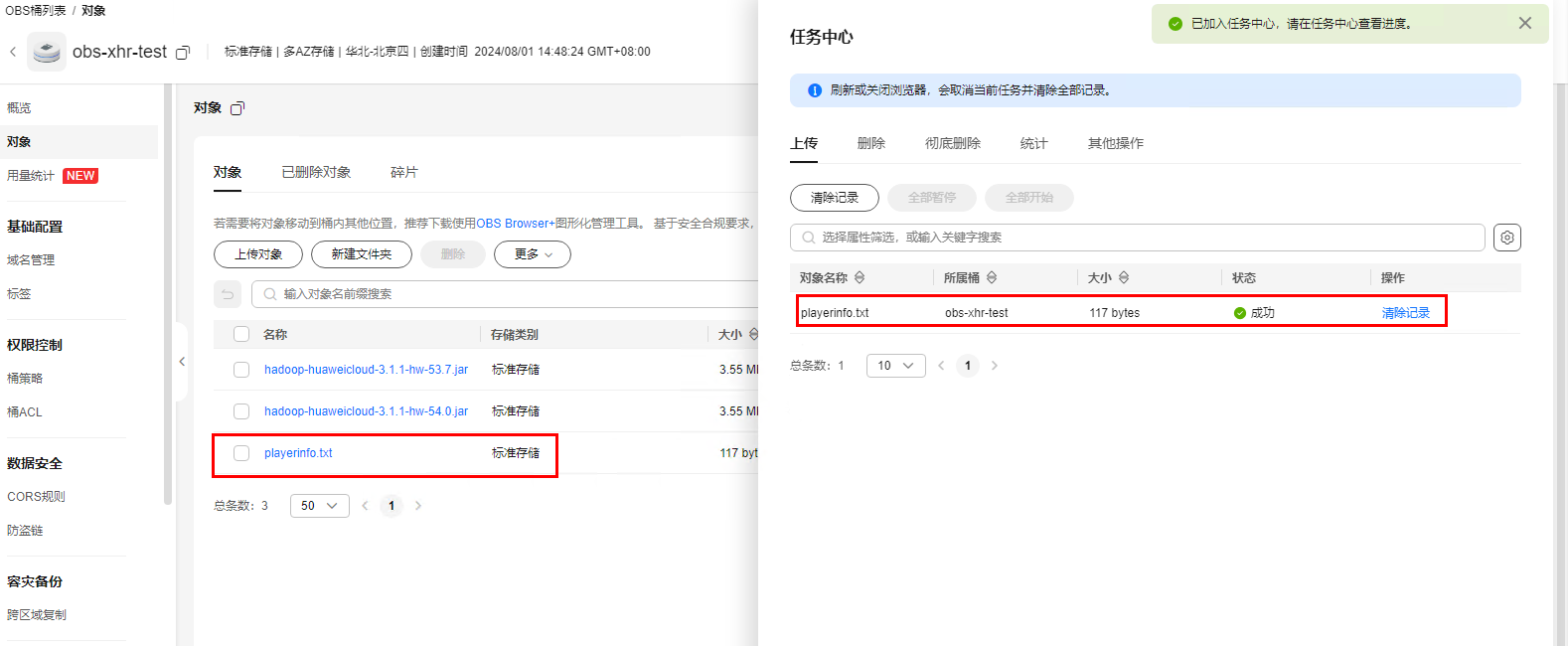
进入OBS桶，选择“对象”上传文件





上传我们实验附件中的“[playerinfo.txt](#_附录_实验数据Playerinfo)“文件到OBS桶中



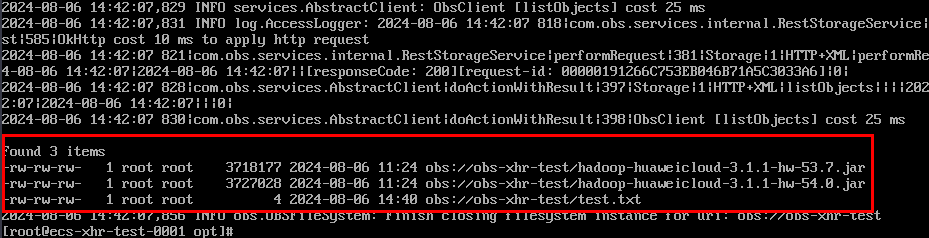


OBS文件上传成功

执行hdfs命令查看OBS文件

登录到ecs机器ecs-xhr-test-0001执行如下命令：

hdfs dfs -ls obs://obs-xhr-test/



如上图出现如上日志即表示Hadoop集群与OBS互联成功。

注：命令参数“bigdatapro-obs”需要根据实际情况修改，该参数为下图OBS桶名。



# Spark集群搭建

## 实验介绍

### 关于本实验

本实验介绍安装Spark集群，并使Spark能够读取OBS数据，通过该实验使Spark集群能够实现存算分离，提高计算性能。

### 实验目的

掌握Spark集群搭建

掌握Spark集群与OBS互联，实现存算分离

## Spark集群存算分离

### 搭建Spark集群

获取spark安装包

ecs-xhr-test-0001节点下载Spark安装包

cd /opt

wget https://archive.apache.org/dist/spark/spark-2.3.3/spark-2.3.3-bin-without-hadoop.tgz

解压spark安装包

ecs-xhr-test-0001节点执行下列命令：

到/opt目录下

cd /opt

解压安装包

tar -zxvf spark-2.3.3-bin-without-hadoop.tgz

mv spark-2.3.3-bin-without-hadoop spark-2.3.3

配置spark jar包

在ecs-xhr-test-0001节点，复制jar包到spark/jar下

cp /opt/hadoop-huaweicloud-3.1.1-hw-54.0.jar /opt/spark-2.3.3/jars/

cp /opt/hadoop-3.1.1/share/hadoop/common/lib/snappy-java-1.0.5.jar /opt/spark-2.3.3/jars/

配置spark配置文件

ecs-xhr-test-0001节点执行下列命令：

mv /opt/spark-2.3.3/conf/spark-env.sh.template /opt/spark-2.3.3/conf/spark-env.sh

vim /opt/spark-2.3.3/conf/spark-env.sh

文件末尾添加如下内容：

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.232.b09-0.el7\_7.aarch64

export SCALA\_HOME=/opt/spark-2.3.3/examples/src/main/scala

export HADOOP\_HOME=/opt/hadoop-3.1.1

export HADOOP\_CONF\_DIR=/opt/hadoop-3.1.1/etc/hadoop

export SPARK\_HOME=/opt/spark-2.3.3

export SPARK\_DIST\_CLASSPATH=$(/opt/hadoop-3.1.1/bin/hadoop classpath)

分发Spark

ecs-xhr-test-0001节点执行下列命令：

scp -r /opt/spark-2.3.3/ root@ecs-xhr-test-0002:/opt/

scp -r /opt/spark-2.3.3/ root@ecs-xhr-test-0003:/opt/

scp -r /opt/spark-2.3.3/ root@ecs-xhr-test-0004:/opt/

配置环境变量

各节点执行：

vim /etc/profile

添加如下内容：

export SPARK\_HOME=/opt/spark-2.3.3

export PATH=${SPARK\_HOME}/bin:${SPARK\_HOME}/sbin:$PATH

保存退出

各节点执行如下命令，使环境变量生效：

source /etc/profile

### 验证存算分离

查看要计算的文件

本次实验验证Spark与OBS实现存算分离，数据使用实验2.3.2中上传的playerinfo.txt文件。

数据如下：

Alex,James,Lax,Genu

Kerry,Mary,Olivia,William

Hale,Edith,Vera,Robert

Mary,Olivia,James,Lax

Edith,Vera,Robertm,Genu

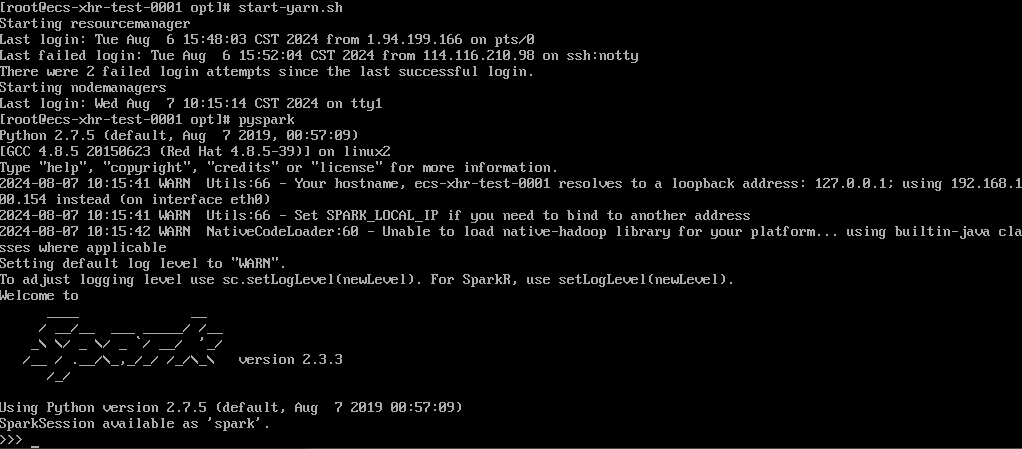
计算上述数据的wordcount

启动yarn

start-yarn.sh

启动pyspark

[root@ecs-xhr-test-0001 conf]# pyspark



输入下列代码：

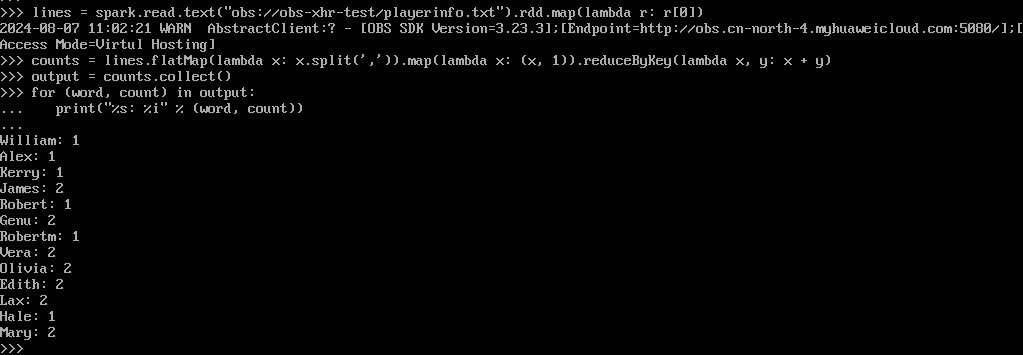
lines = spark.read.text("obs://obs-xhr-test/playerinfo.txt").rdd.map(lambda r: r[0])

counts = lines.flatMap(lambda x: x.split(',')).map(lambda x: (x, 1)).reduceByKey(lambda x, y: x + y)

output = counts.collect()

for (word, count) in output:

print("%s: %i" % (word, count))



出现如上图所示的结果即表示存算分离计算成功！

需要注意的是，obs://obs-xhr-test/playerinfo.txt是自己的obs文件路径



# 释放实验资源

## 删除ECS弹性云服务器

进入ECS列表，点击全选按钮，点击“更多 -> 删除”。

在对话框中选择“释放云服务器绑定的公网IP地址”和“删除云服务器挂载的数据盘”，点击“是”。

## 删除OBS对象存储

进入OBS服务的并行文件系统列表，点击创建的文件系统。

选择左侧“文件”标签，在右侧全选所有文件，点击“删除”并确认。

返回并行文件系统列表，在对应的文件系统右侧点击“删除”并确认。

## 资源检查

在控制台点击“更多 | 资源 | 我的资源”菜单项，检查资源是否全部删除。

注意：(1) 虚拟私有云VPC和安全组可以不删除，以留下次使用。(2) 若在其他区域购买了ECS和EIP等资源，请切换到那个区域查看。

# 附录 安装OpenJDK

安装OpenJDK, 在各个节点上，执行下列命令

在节点，执行如下命令，将jdk安装到各个节点

yum install -y java-1.8.0-openjdk

确认java版本

java -version

[root@ecs-xhr-test-0001 ~]# java -version

openjdk version "1.8.0\_191"

OpenJDK Runtime Environment (AdoptOpenJDK)(build 1.8.0\_191-b12)

# 附录 Linux常用命令

## ls命令

ls 命令不仅可以查看 linux 文件夹包含的文件而且可以查看文件权限(包括目录、文件夹、文件权限)查看目录信息等等。

命令格式：ls [选项][目录名]

常用参数

-l ：列出长数据串，包含文件的属性与权限数据等

-a ：列出全部的文件，连同隐藏文件（开头为.的文件）一起列出来（常用）

-d ：仅列出目录本身，而不是列出目录的文件数据

-h ：将文件容量以较易读的方式（GB，kB等）列出来

-R ：连同子目录的内容一起列出（递归列出），等于该目录下的所有文件都会显示出来

## cd命令

最基本的命令语句，其他的命令语句要进行操作，都是建立在使用 cd 命令上的。用于切换当前目录至dirName。

命令格式：cd [目录名]

## pwd命令

查看"当前工作目录"的完整路径。

命令格式：pwd [选项]

常用参数：

-P :显示实际物理路径，而非使用连接（link）路径

-L :当目录为连接路径时，显示连接路径

## mkdir命令

用来创建指定的名称的目录，要求创建目录的用户在当前目录中具有写权限，并且指定的目录名不能是当前目录中已有的目录。

命令格式：mkdir [选项] 目录

常用参数

-m, --mode=模式，设定权限<模式> (类似 chmod)，而不是 rwxrwxrwx 减 umask

-p, --parents 可以是一个路径名称。此时若路径中的某些目录尚不存在,加上此选项后,系统将自动建立好那些尚不存在的目录,即一次可以建立多个目录;

-v, --verbose 每次创建新目录都显示信息

--help 显示此帮助信息并退出

--version 输出版本信息并退出

## rm命令

删除一个目录中的一个或多个文件或目录，如果没有使用- r选项，则rm不会删除目录。如果使用 rm 来删除文件，通常仍可以将该文件恢复原状。

命令格式：rm [选项] 文件

常用参数

-f, --force 忽略不存在的文件，从不给出提示。

-i, --interactive 进行交互式删除

-r, -R, --recursive 指示rm将参数中列出的全部目录和子目录均递归地删除。

-v, --verbose 详细显示进行的步骤

--help 显示此帮助信息并退出

--version 输出版本信息并退出

## rmdir命令

该命令从一个目录中删除一个或多个子目录项，删除某目录时也必须具有对父目录的写权限。

命令格式：rmdir [选项] 目录

常用参数

-p 递归删除目录dirname，当子目录删除后其父目录为空时，也一同被删除。如果整个路径被删除或者由于某种原因保留部分路径，则系统在标准输出上显示相应的信息。

-v, --verbose 显示指令执行过程

## mv命令

可以用来移动文件或者将文件改名（move (rename) files）。当第二个参数类型是文件时，mv命令完成文件重命名。当第二个参数是已存在的目录名称时，源文件或目录参数可以有多个，mv命令将各参数指定的源文件均移至目标目录中。

命令格式：mv [选项] 源文件或目录 目标文件或目录

常用参数

-b ：若需覆盖文件，则覆盖前先行备份

-f ：force 强制的意思，如果目标文件已经存在，不会询问而直接覆盖

-i ：若目标文件 (destination) 已经存在时，就会询问是否覆盖

-u ：若目标文件已经存在，且 source 比较新，才会更新(update)

-t ：--target-directory=DIRECTORY move all SOURCE arguments into DIRECTORY，即指定mv的目标目录，该选项适用于移动多个源文件到一个目录的情况，此时目标目录在前，源文件在后

## cp命令

将源文件复制至目标文件，或将多个源文件复制至目标目录。

命令格式：cp [选项] 源文件 目录 或 cp [选项] -t 目录 源文件

常用参数

-t --target-directory 指定目标目录

-i --interactive 覆盖前询问（使前面的 -n 选项失效）

-n --no-clobber 不要覆盖已存在的文件（使前面的 -i 选项失效）

-f --force 强行复制文件或目录，不论目的文件或目录是否已经存在

-u --update 使用这项参数之后，只会在源文件的修改时间较目的文件更新时，或是对应的目的文件并不存在，才复制文件

## cat 命令

用来显示文件内容，或者将几个文件连接起来显示，或者从标准输入读取内容并显示，它常与重定向符号配合使用。

命令格式：cat [选项] [文件]

常用参数

-A, --show-all 等价于 -vET

-b, --number-nonblank 对非空输出行编号

-e 等价于 -vE

-E, --show-ends 在每行结束处显示 $

-n, --number 对输出的所有行编号,由1开始对所有输出的行数编号

-s, --squeeze-blank 有连续两行以上的空白行，就代换为一行的空白行

-t 与 -vT 等价

-T, --show-tabs 将跳格字符显示为 ^I

-v, --show-nonprinting 使用 ^ 和 M- 引用，除了 LFD 和 TAB 之外

## more命令

more 命令和 cat 的功能一样都是查看文件里的内容，但有所不同的是more可以按页来查看文件的内容，还支持直接跳转行等功能。

命令格式：more [-dlfpcsu ] [-num ] [+/ pattern] [+ linenum] [file ... ]

常用参数

+n 从笫n行开始显示

-n 定义屏幕大小为n行

+/pattern 在每个档案显示前搜寻该字串（pattern），然后从该字串前两行之后开始显示

-c 从顶部清屏，然后显示

-d 提示“Press space to continue，’q’ to quit（按空格键继续，按q键退出）”，禁用响铃功能

-l 忽略Ctrl+l（换页）字符

-p 通过清除窗口而不是滚屏来对文件进行换页，与-c选项相似

-s 把连续的多个空行显示为一行

-u 把文件内容中的下画线去掉

操作指令

Enter：向下n行，需要定义。默认为1行

Ctrl+F：向下滚动一屏

空格键：向下滚动一屏

Ctrl+B：返回上一屏

= ：输出当前行的行号

：f ：输出文件名和当前行的行号

V ：调用vi编辑器

!命令 ：调用Shell，并执行命令

q ：退出more

## less命令

less 与 more 类似，但使用 less 可以随意浏览文件，而 more 仅能向前移动，却不能向后移动，而且 less 在查看之前不会加载整个文件。

命令格式：less [参数] 文件

常用参数

-b <缓冲区大小> 设置缓冲区的大小

-e 当文件显示结束后，自动离开

-f 强迫打开特殊文件，例如外围设备代号、目录和二进制文件

-g 只标志最后搜索的关键词

-i 忽略搜索时的大小写

-m 显示类似more命令的百分比

-N 显示每行的行号

-o <文件名> 将less 输出的内容在指定文件中保存起来

-Q 不使用警告音

-s 显示连续空行为一行

-S 行过长时间将超出部分舍弃

-x <数字> 将“tab”键显示为规定的数字空格

操作命令

/字符串：向下搜索“字符串”的功能

?字符串：向上搜索“字符串”的功能

n：重复前一个搜索（与 / 或 ? 有关）

N：反向重复前一个搜索（与 / 或 ? 有关）

b 向后翻一页

d 向后翻半页

h 显示帮助界面

Q 退出less 命令

u 向前滚动半页

y 向前滚动一行

空格键 滚动一行

回车键 滚动一页

[pagedown]：向下翻动一页

[pageup]：向上翻动一页

## head命令

head 用来显示档案的开头至标准输出中，默认 head 命令打印其相应文件的开头 10 行。

命令格式：head [参数] [文件]

常用参数

-q 隐藏文件名

-v 显示文件名

-c<字节> 显示字节数

-n<行数> 显示的行数

## tail命令

显示指定文件末尾内容，不指定文件时，作为输入信息进行处理。常用查看日志文件。

命令格式：tail [必要参数] [选择参数] [文件]

常用参数

-f 循环读取

-q 不显示处理信息

-v 显示详细的处理信息

-c<数目> 显示的字节数

-n<行数> 显示行数

--pid=PID 与-f合用,表示在进程ID,PID死掉之后结束.

-q, --quiet, --silent 从不输出给出文件名的首部

-s, --sleep-interval=S 与-f合用,表示在每次反复的间隔休眠S秒

## vim命令

vim编辑器是所有Unix及Linux系统下标准的编辑器，它的强大不逊色于任何最新的文本编辑器;vi也是Linux中最基本的文本编辑器,vim就是vi的升级版。

### 启动vim

在命令行窗口中输入以下命令即可

vim

直接启动vim

vim filename 打开vim并创建名为filename的文件

### vim的模式

正常模式（按Esc或Ctrl+[进入） 左下角显示文件名或为空

插入模式（按i键进入） 左下角显示--INSERT--

可视模式（不知道如何进入） 左下角显示--VISUAL--

导航命令

% 括号匹配

### 插入命令

i 在当前位置生前插入

I 在当前行首插入

a 在当前位置后插入

### 查找命令

/text　查找text，按n健查找下一个，按N健查找前一个。

?text　查找text，反向查找，按n健查找下一个，按N健查找前一个。

:set hlsearch　高亮搜索结果，所有结果都高亮显示，而不是只显示一个匹配。

:set nohlsearch　关闭高亮搜索显示

:nohlsearch　关闭当前的高亮显示，如果再次搜索或者按下n或N键，则会再次高亮。

:set incsearch　逐步搜索模式，对当前键入的字符进行搜索而不必等待键入完成。

:set wrapscan　重新搜索，在搜索到文件头或尾时，返回继续搜索，默认开启。

:set nu 显示行号

### 替换命令

ra 将当前字符替换为a，当期字符即光标所在字符。

s/old/new/ 用old替换new，替换当前行的第一个匹配

### 撤销和重做

u 撤销（Undo）

U 撤销对整行的操作

Ctrl + r 重做（Redo），即撤销的撤销。

### 删除命令

x 删除当前字符

3x 删除当前光标开始向后三个字符

X 删除当前字符的前一个字符。X=dh

dl 删除当前字符， dl=x

dh 删除前一个字符

dd 删除当前行

10d 删除当前行开始的10行。

D 删除当前字符至行尾。D=d$

d$ 删除当前字符之后的所有字符（本行）

### 拷贝和粘贴

yy 拷贝当前行

nyy 拷贝当前后开始的n行，比如2yy拷贝当前行及其下一行。

p 在当前光标后粘贴,如果之前使用了yy命令来复制一行，那么就在当前行的下一行粘贴。

shift+p 在当前行前粘贴

### 退出命令

:wq 保存并退出

ZZ 保存并退出

:q! 强制退出并忽略所有更改

:e! 放弃所有修改，并打开原来文件。

### 帮助命令

:help or F1 显示整个帮助

:help xxx 显示xxx的帮助，比如 :help i, :help CTRL-[（即Ctrl+[的帮助）。

:help 'number' Vim选项的帮助用单引号括起

:help 特殊键的帮助用<>扩起

:help -t Vim启动参数的帮助用-

：help i\_ 插入模式下Esc的帮助，某个模式下的帮助用模式\_主题的模式

帮助文件中位于||之间的内容是超链接，可以用Ctrl+]进入链接，Ctrl+o（Ctrl + t）返回

其他非编辑命令

# 附录 实验数据Playerinfo

Playerinfo内容：

创建一个txt文件，重命名为playerinfo.txt，将下面的值全部复制到playerinfo.txt中，保存

Alex,James,Lax,Genu

Kerry,Mary,Olivia,William

Hale,Edith,Vera,Robert

Mary,Olivia,James,Lax

Edith,Vera,Robertm,Genu