华为云实践

实验总结报告

班级：软件2002

学号：20301036

姓名：韩熔

目录

[1 实验概览 4](#_Toc107162027)

[1.1 实验介绍 4](#_Toc107162028)

[1.2 实验流程 5](#_Toc107162029)

[2 华为云环境准备 6](#_Toc107162030)

[2.1 实验介绍 6](#_Toc107162031)

[2.2 购买华为云ECS 6](#_Toc107162032)

[2.2.1 进入华为云控制台 6](#_Toc107162033)

[2.2.2 购买华为云ECS 6](#_Toc107162034)

[2.2.3 配置安全组规则 7](#_Toc107162035)

[2.3 准备OBS服务 8](#_Toc107162036)

[2.3.1 购买OBS 8](#_Toc107162037)

[2.3.2 获取访问密钥AK/SK 9](#_Toc107162038)

[3 搭建Hadoop集群 10](#_Toc107162039)

[3.1 实验介绍 10](#_Toc107162040)

[3.1.1 关于本实验 10](#_Toc107162041)

[3.2 Hadoop集群搭建 10](#_Toc107162042)

[3.2.1 配置ECS 10](#_Toc107162043)

[3.2.2 获取JDK的安装路径 14](#_Toc107162044)

[3.3 搭建Hadoop伪分布式集群 15](#_Toc107162045)

[3.3.1 Hadoop安装 15](#_Toc107162046)

[3.3.2 伪分布式配置 18](#_Toc107162047)

[3.3.3 Hadoop与OBS互联 26](#_Toc107162048)

[4 Spark集群搭建 31](#_Toc107162049)

[4.1 实验介绍 31](#_Toc107162050)

[4.1.1 关于本实验 31](#_Toc107162051)

[4.2 Spark集群存算分离 31](#_Toc107162052)

[4.2.1 搭建Spark集群 31](#_Toc107162053)

[4.2.2 验证存算分离 37](#_Toc107162054)

[5 释放华为云服务 40](#_Toc107162055)

[5.1 释放弹性云服务器ECS 40](#_Toc107162056)

[5.2 删除OBS桶 40](#_Toc107162057)

# 实验概览

## 实验介绍

本实验基于华为云OBS和华为云ECS服务构建一个存算分离的基本架构，并通过运行一个计算程序来完成存算分离架构的验证。本实验的实验数据存储在OBS中，通过在ECS上部署开源组件（Hadoop和Spark）构成计算环境，最后编写Spark程序访问存储在OBS上的数据进行计算（单词出现次数统计）并输出结果。

本实验的基本步骤包含：

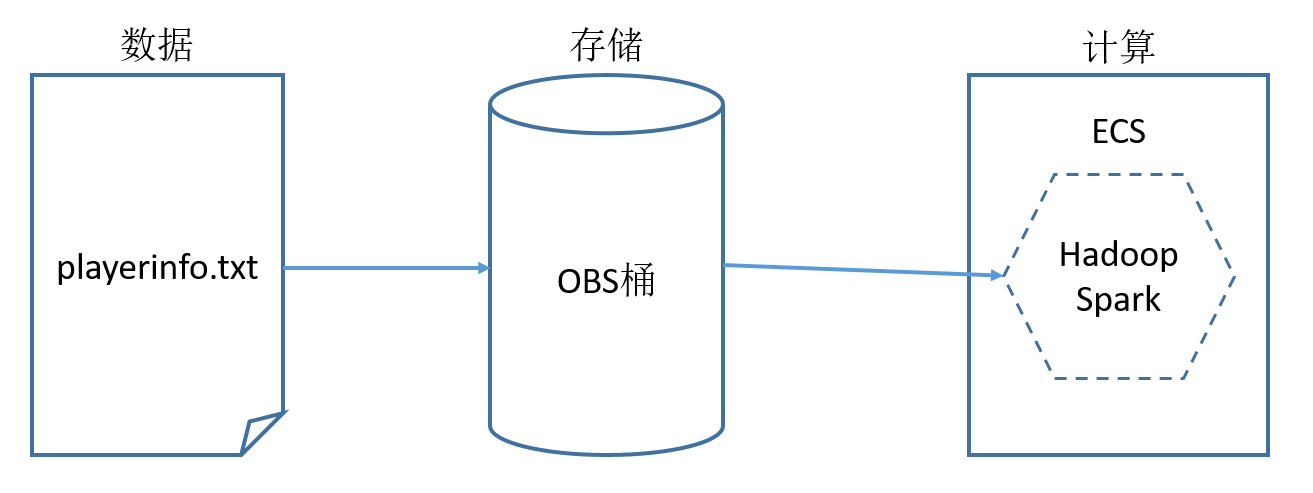
购买并配置ECS；

购买OBS并获取访问密钥AK/SK信息；

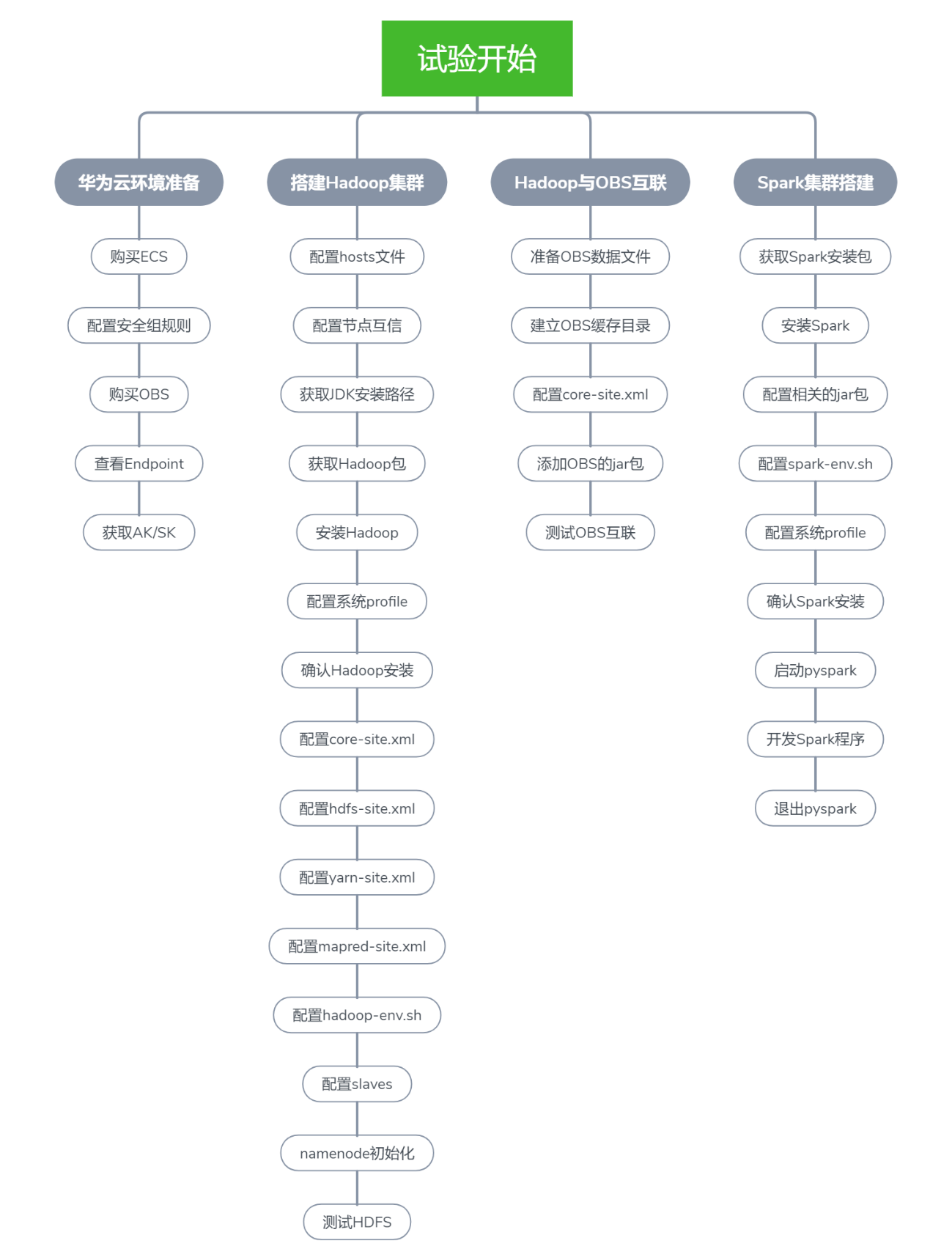
搭建Hadoop集群；

搭建Spark集群；

编写Spark程序验证存算分离。



## 实验流程



验证存算分离

搭建Spark集群

伪分布式配置

安装Hadoop

配置ECS

# 华为云环境准备

## 实验介绍

本部分实验主要是通过购买华为云的相关服务构建起整个实验的云环境，提供后续集群搭建及实验的基础。主要包括购买华为云ECS和OBS，下载访问密钥文件获取AK/SK。

## 购买华为云ECS

### 进入华为云控制台

登录华为云

进入控制台

登录成功后点击“控制台”。

进入到华为云控制台，选择区域为“北京四”。

### 购买华为云ECS

服务器名：ecs-bigdatapro

### 配置安全组规则

进入安全组

在云服务器控制台界面点击ECS名称。

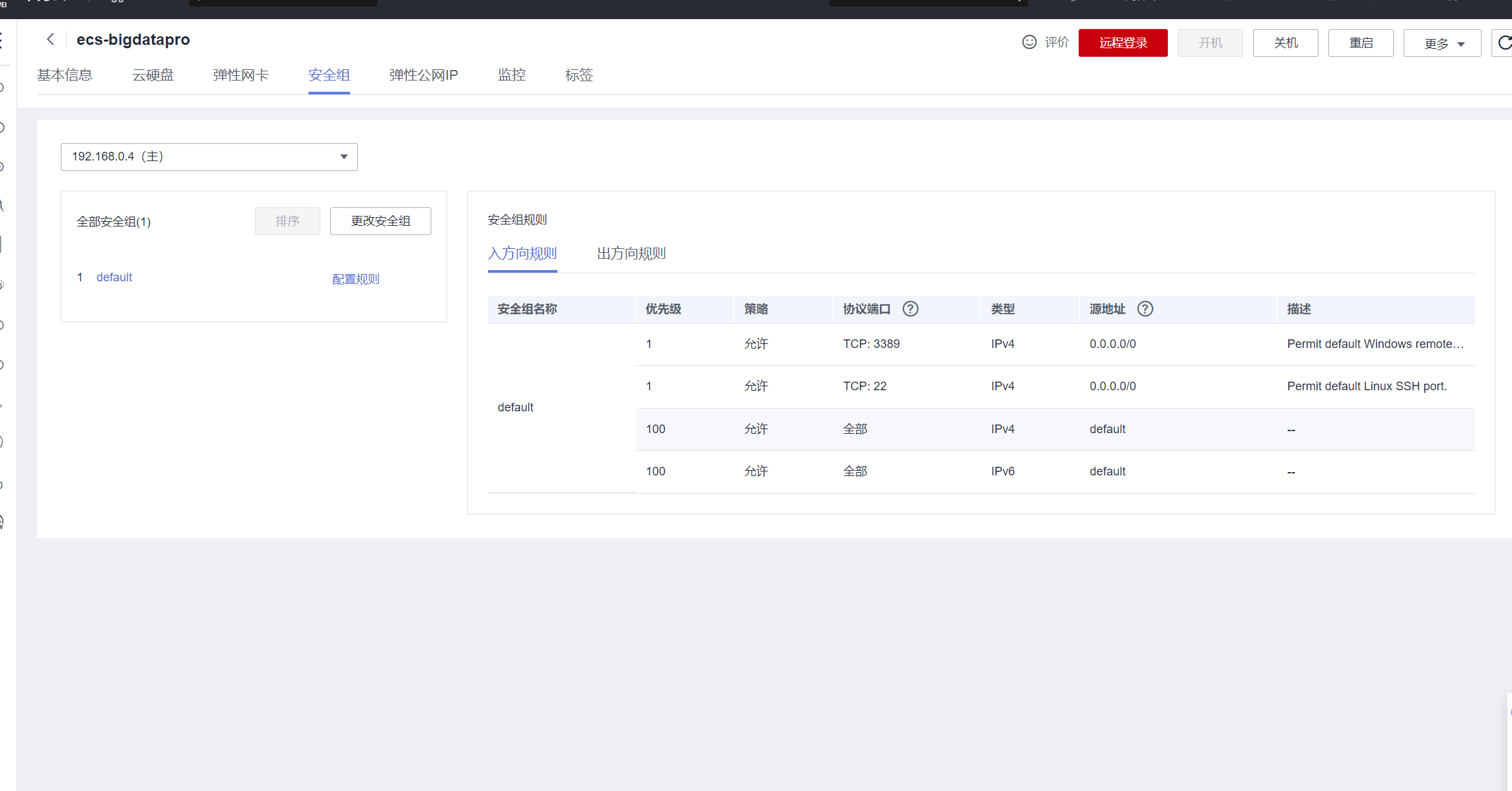
进入ECS后选择“安全组”。

修改规则

找到default安全组，点击“配置规则”。



选择“入方向规则”，点击“添加规则”。



点击下拉框，选择“基本协议”中的“全部协议”，然后点击“确定”。



配置好安全组规则后，后面就可以通过远程工具进行连接了。

## 准备OBS服务

### 购买OBS

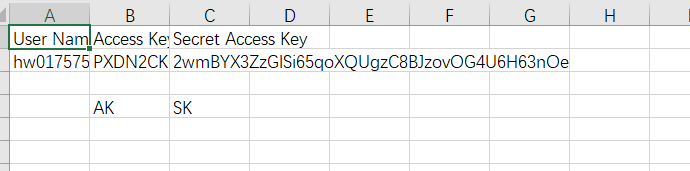


点击刚创建的桶名称，进入概览后下拉滚动条可以看到桶的基本信息，记录其中的Endpoint内容（IPv4地址），后面配置连接OBS时会使用。



Endpoint（IPv4地址）：obs.cn-north-4.myhuaweicloud.com

获取访问密钥AK/SK



Access Key Id：PXDN2CKOMLGBFZLQM6KL

Secret Access Key：2wmBYX3ZzGlSi65qoXQUgzC8BJzovOG4U6H63nOe

# 搭建Hadoop集群

## 实验介绍

### 关于本实验

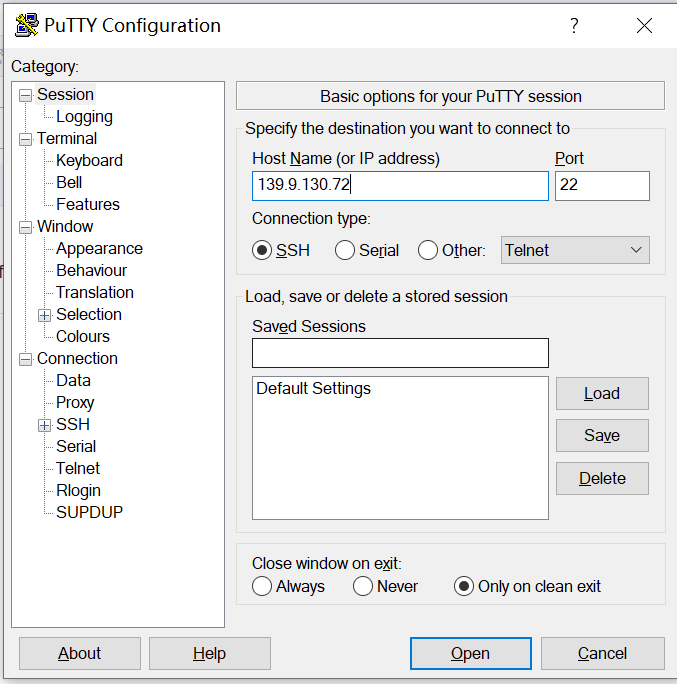
本部分实验需要在已经购买的ECS上搭建Hadoop集群，并且通过配置与华为云OBS服务互联，使Hadoop集群可读取OBS数据。

## Hadoop集群搭建

### 配置ECS

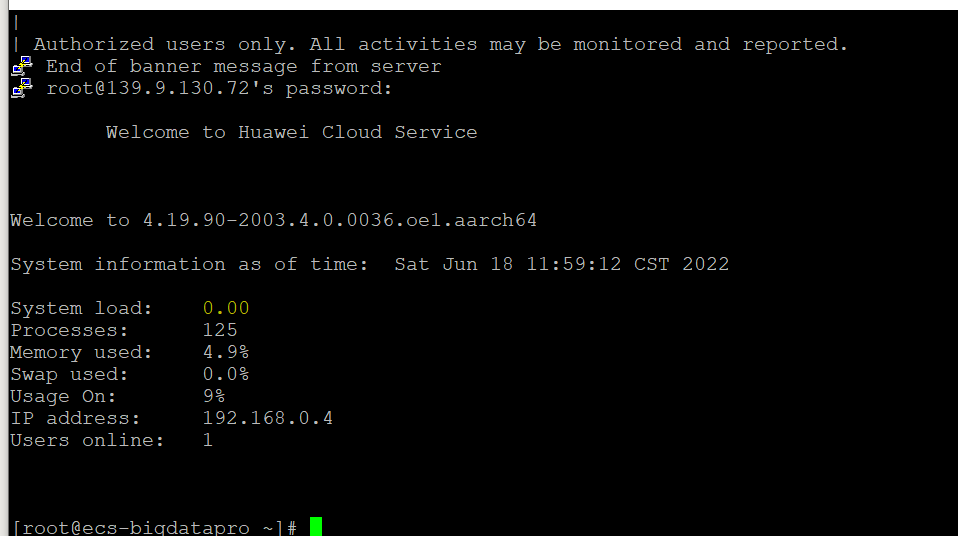
使用PuTTY远程登录ECS

下载并启动PuTTY，在Host Name处输入前面申请的ECS的公网IP地址139.9.130.72，选择SSH，点击“Open”。



弹出的对话框选择“Apply”（这里只会在第一次打开时弹出，第二次及之后打开没弹出，所以没截到图）

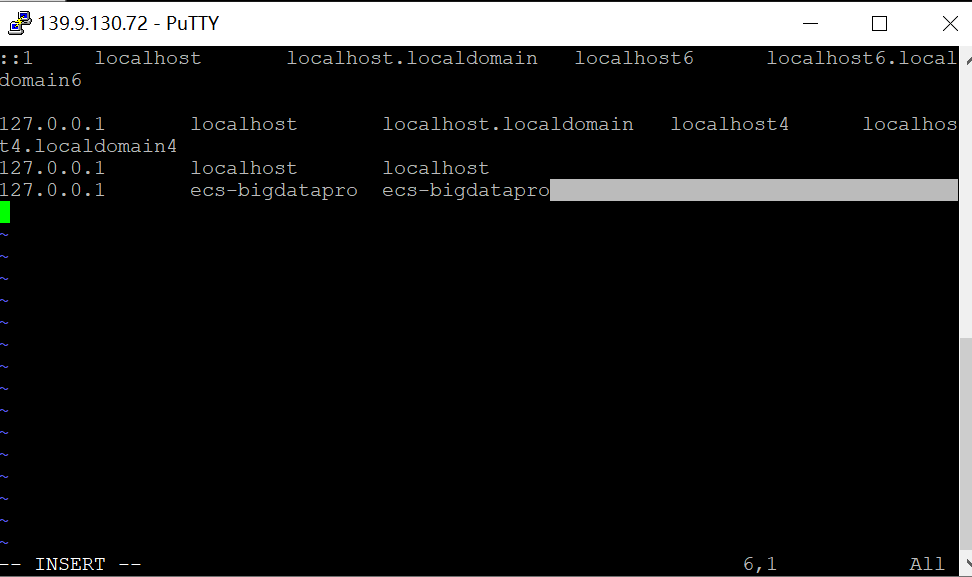
输入用户名root和密码（此处输入密码不返显字符，直接输入即可）进行登录。



配置/etc/hosts文件

执行vim /etc/hosts命令使用vim编辑器编辑hosts文件。

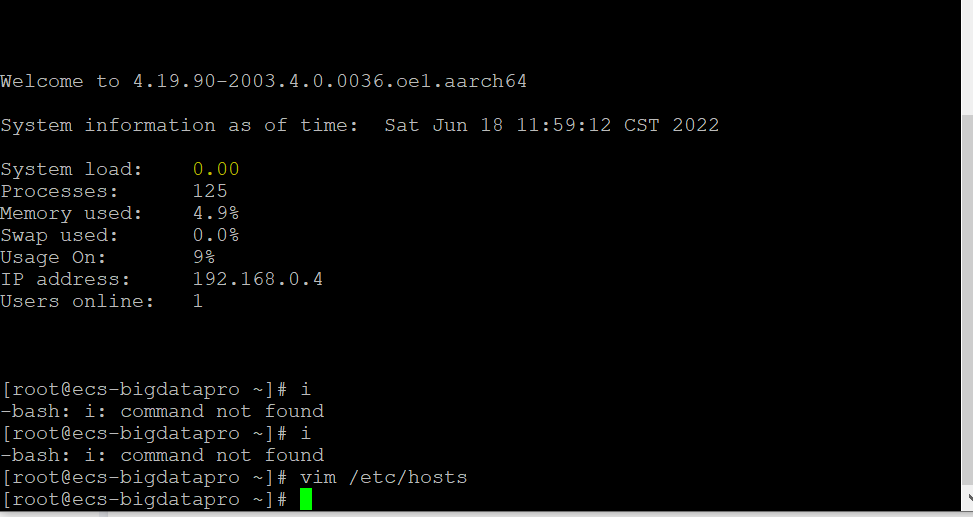
按下i键进入编辑模式，添加ECS的私有IP和主机名esc-bigdatapro（IP和主机名之间留有一个空格）。

139.9.130.72 -esc-bigdatapro

按ECS退出编辑模式，输入:wq保存并退出编辑器。

配置节点互信

执行命令ssh-keygen -t rsa生成密钥文件。（第一次弄时没截图，这是第二次的截图，所以显示already exists）

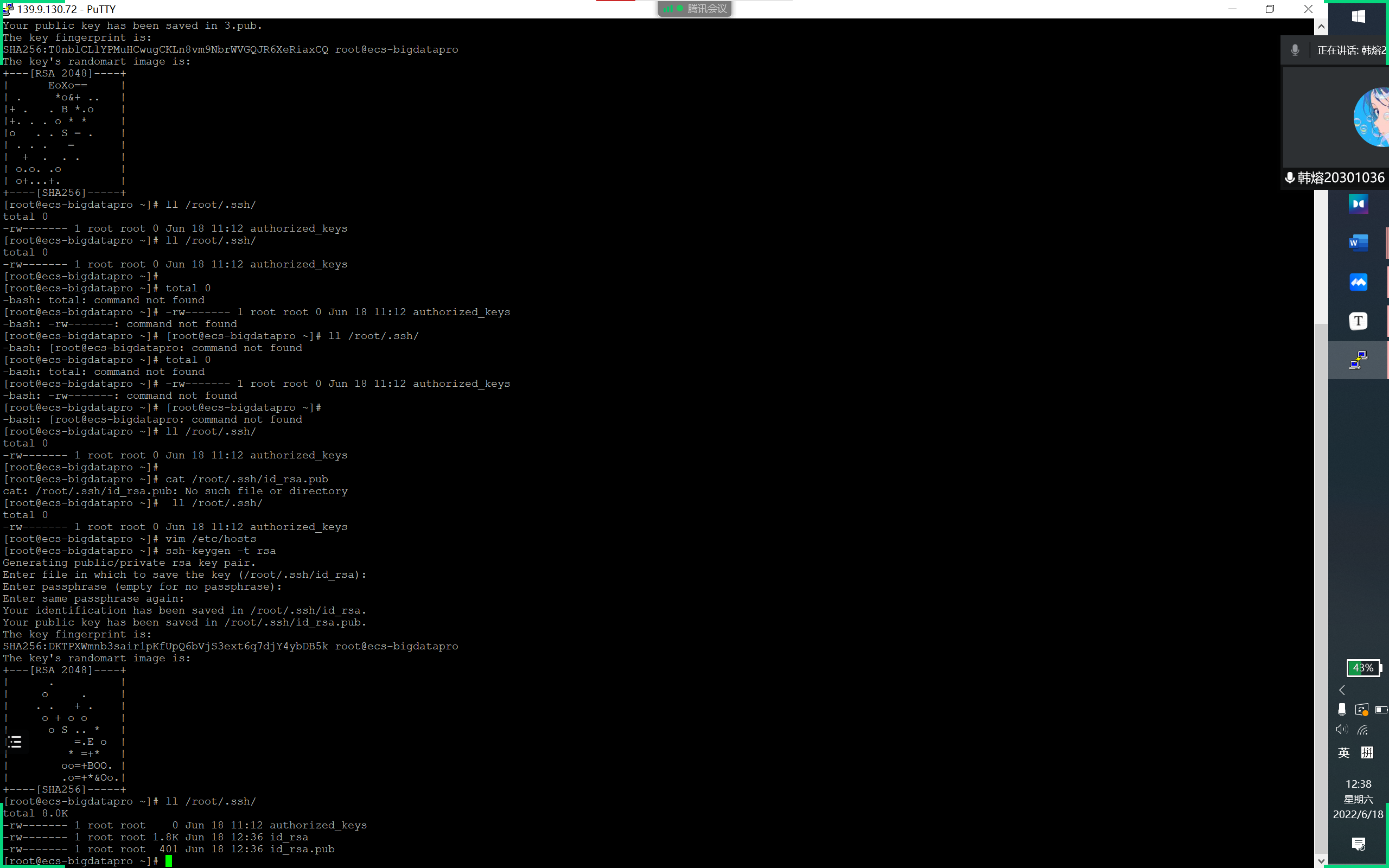


可以在/root/.ssh目录中看到生成的文件。

ll /root/.ssh/

使用cat命令查看文件内容。

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



复制上面命令输出的内容。

**ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABAQDVO8elxzwk1lSbq0Ev7sHP7QCOvxY2LJ7z88yZtXdDU8QcJ553kTIjQ0DupuH+6uCFzWNICvV+b51pDlDzQ01OOvyCe7E21EWD7WBgXLhc9Y2CBU0se6QRWmRiJLXqpXdRAxQXRqIIlTFT15G1uBhs7GjBeNCpmH9sykIMu6OoZoQxI5stJHaZu6hMiIT5flE5XhJwrpN0Tqtr8JlUW3PupoE0hdPEs+n7t5rvjDhmXA8HVqBcggRTe4ivo+pum7FVPvJs0pmyePlNIRl459o0XONhXdU9GC8oKEHVKP3UPH+BKotkP9t9VUULLuFkA0j84UN9t9NA8HlFTsF6YUq5 root@ecs-bigdatapro**

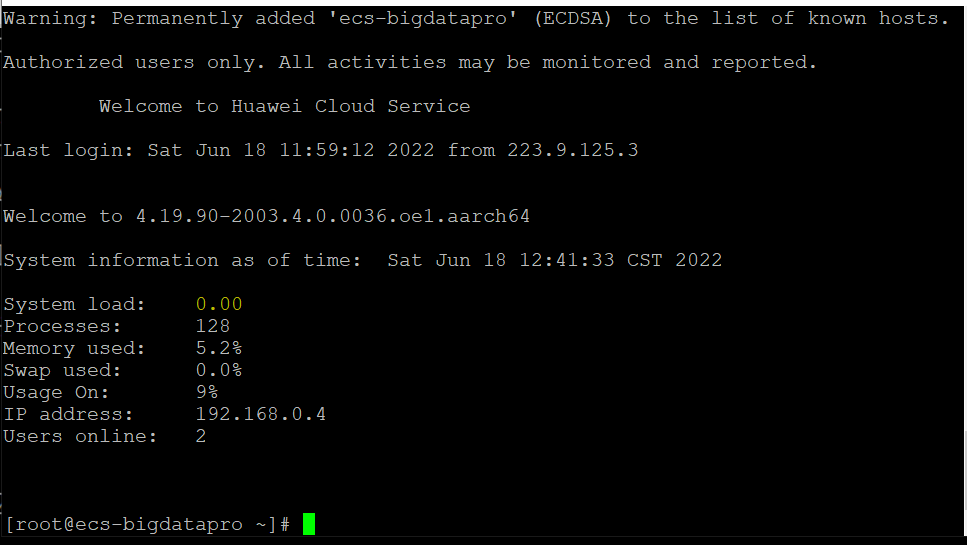
使用vim命令打开authorized\_keys文件，并把复制的内容粘贴到文件中。

vim /root/.ssh/authorized\_keys

输入:wq保存并退出。

验证互信

输入ssh ecs-32df，确保能够无密码登录即为成功。



输入exit可以关闭ssh连接。

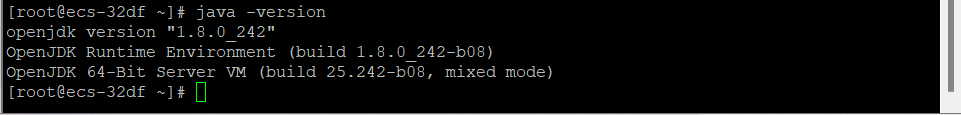
### 获取JDK的安装路径

确认是否已安装JDK

执行命令查询Java版本。

java -version

输出如下结果，说明ECS已安装JDK。

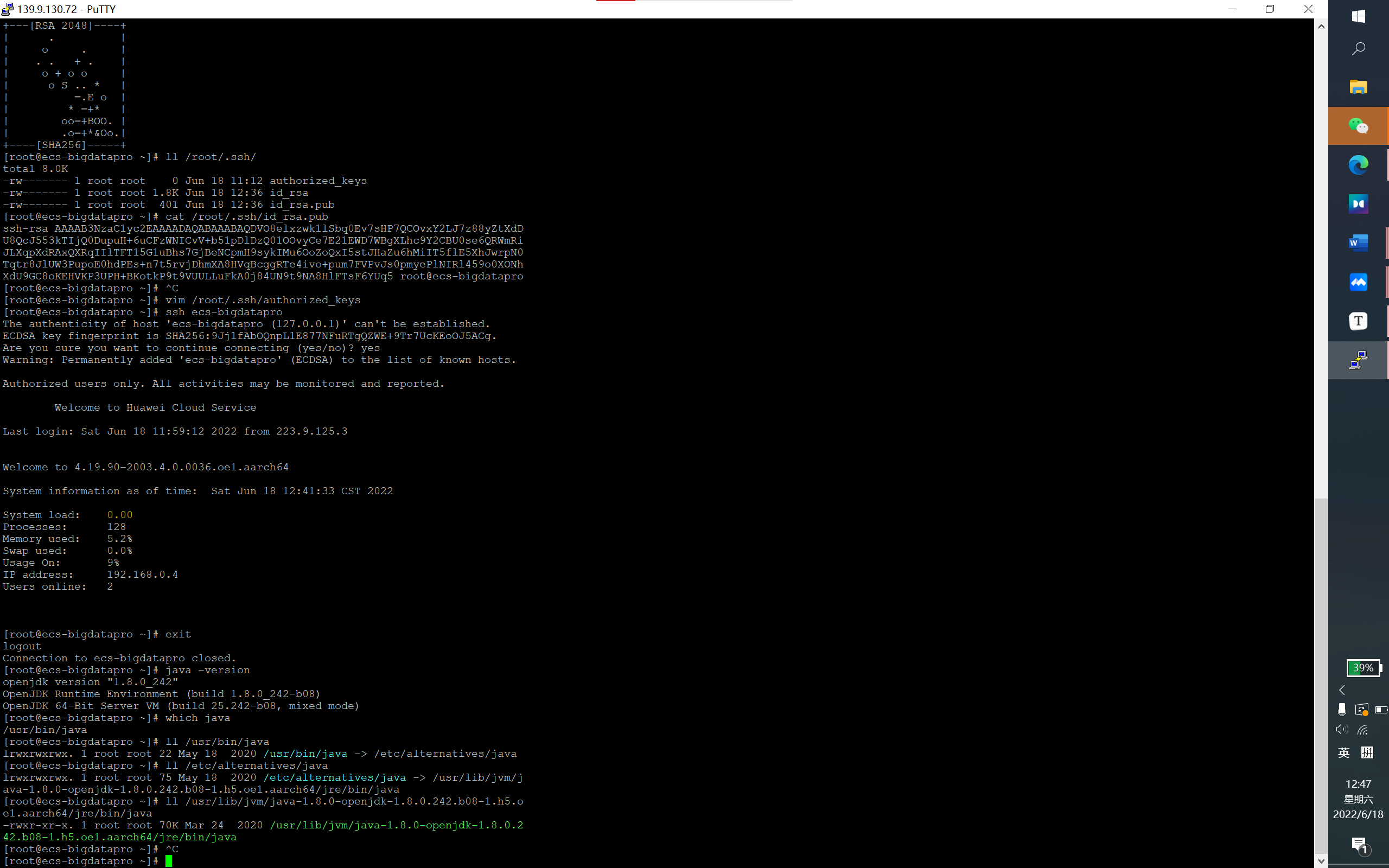


获取JDK的实际安装目录（为后面配置JAVA\_HOME所需要的值）

输入命令which获取JDK的位置。

which java

查看目录信息。（可知目录已经是一个普通文件）。

ll /usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64/bin/java

则bin之前的部分为JDK的实际安装目录。

/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64

记录下此内容，作为后面配置**JAVA\_HOME**的值。

## 搭建Hadoop伪分布式集群

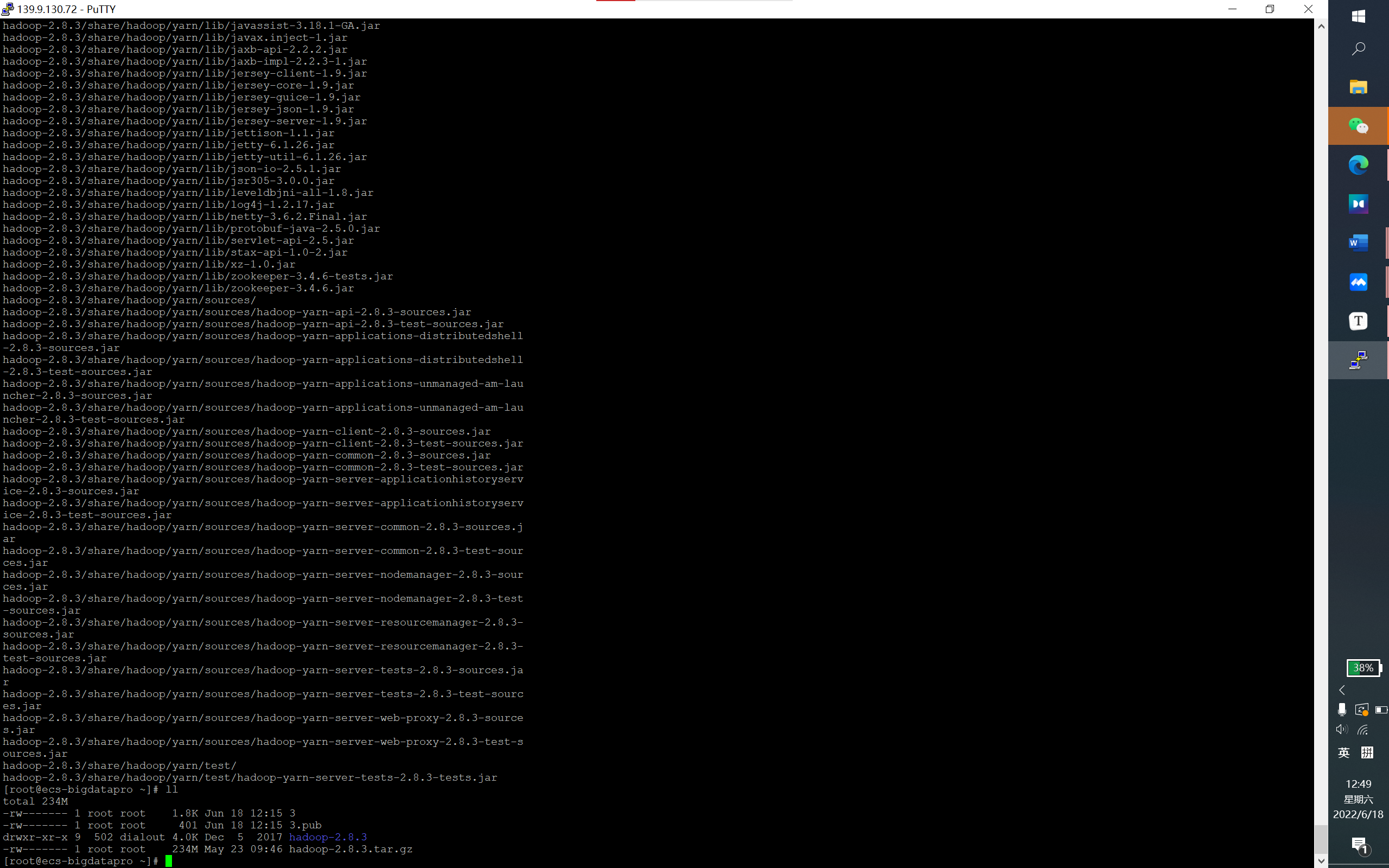
### Hadoop安装

下载hadoop安装包

进入root目录，使用wget命令下载hadoop安装包。

cd /root

wget <https://bigdata-tools-hw.obs.cn-north-1.myhuaweicloud.com/hadoop-2.8.3.tar.gz>



创建hadoop所需的目录

mkdir -p /home/modules

mkdir -p /home/nm/localdir

**/home/modules**作为hadoop的安装目录；

**/home/nm/localdir**作为NodeManager的数据目录，存放执行Container所需的数据和运行过程中产生的临时数据，后面在配置文件yarn-site.xml中会使用yarnyarn.nodemanager.local-dirs属性指定。

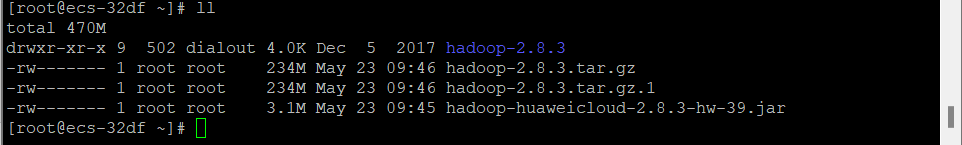
安装hadoop

解压hadoop安装包。

cd /root

tar -zxvf hadoop-2.8.3.tar.gz

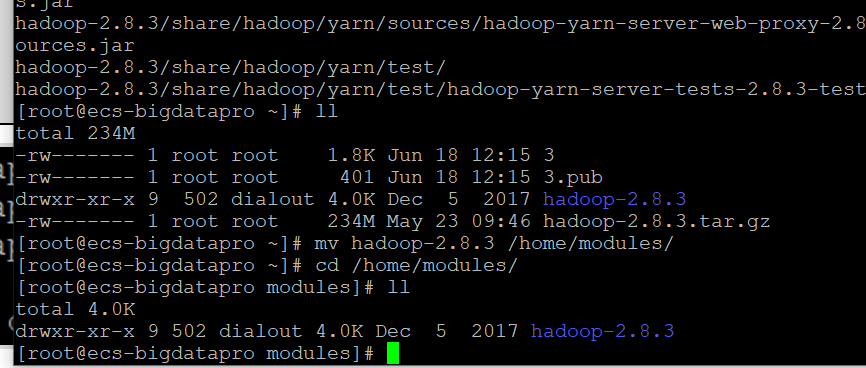
ll



把解压后的目录移动到安装目录下。

mv hadoop-2.8.3 /home/modules/

cd /home/modules/

ll

配置系统环境变量

使用vim打开系统配置文件。

vim /etc/profile

按i键进入编辑模式，在文件尾部添加如下内容：

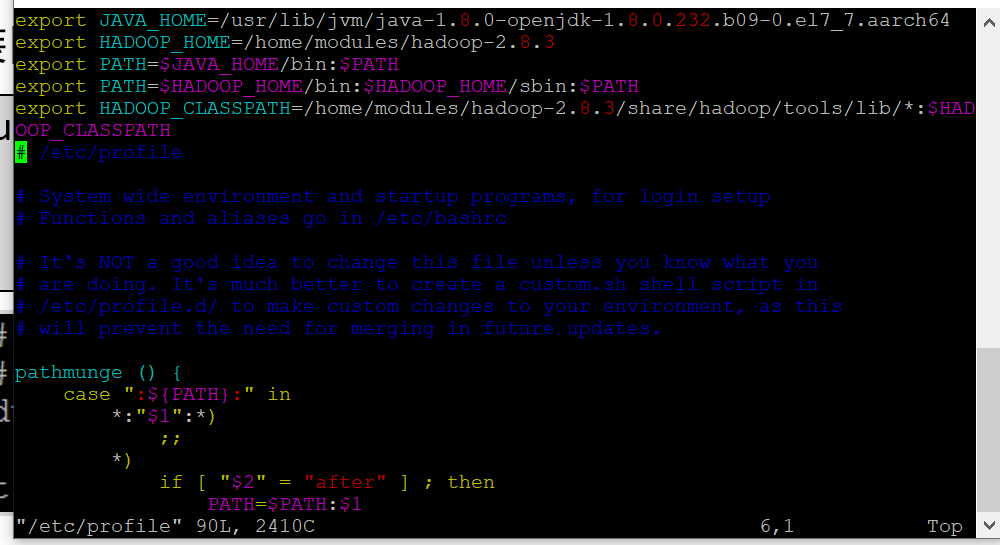
export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64

export HADOOP\_HOME=/home/modules/hadoop-2.8.3

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

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

export HADOOP\_CLASSPATH=/home/modules/hadoop-2.8.3/share/hadoop/tools/lib/\*:$HADOOP\_CLASSPATH



按ECS退出编辑模式，输入:wq保存并退出。

使用source命令使得环境变量生效。

source /etc/profile



验证hadoop安装

进入root目录，查看hadoop版本信息。

cd /root

hadoop version

正确显示hadoop版本信息，安装完成。

### 伪分布式配置

配置hadoop的core-site.xml文件

使用vim命令打开配置文件。

vim /home/modules/hadoop-2.8.3/etc/hadoop/core-site.xml

按i键进入编辑模式，复制下面的配置内容粘贴到<configuration>和</configuration>之间

<property>

<name>fs.defaultFS</name>

<value>hdfs://**esc-bigdatapro**:8020</value>

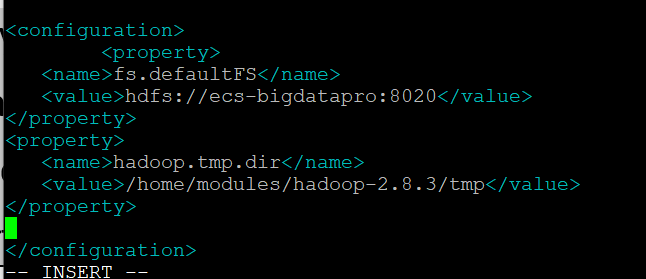
</property>

<property>

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

<value>/home/modules/hadoop-2.8.3/tmp</value>

</property>



按ECS退出编辑模式，输入:wq保存并退出。

配置hadoop的hdfs-site.xml文件

使用vim打开配置文件。

vim /home/modules/hadoop-2.8.3/etc/hadoop/hdfs-site.xml

按i键进入编辑模式，复制下面的配置内容粘贴到<configuration>和</configuration>之间：

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

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

<value>**esc-bigdatapro**:50090</value>

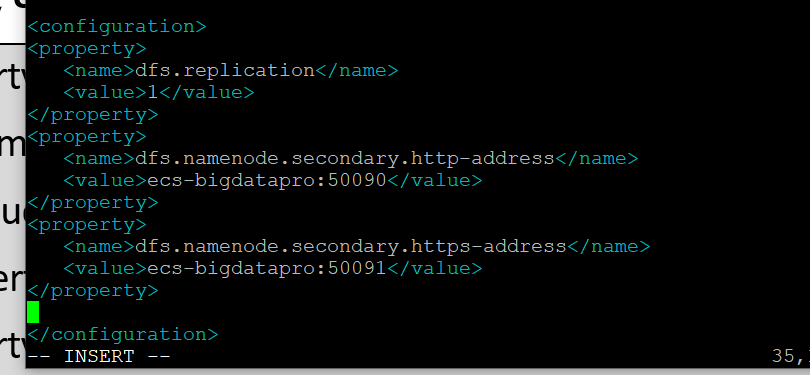
</property>

<property>

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

<value>**esc-bigdatapro**:50091</value>

</property>



按ECS退出编辑模式，输入:wq保存并退出。

配置hadoop的yarn-site.xml文件

使用vim打开配置文件。

vim /home/modules/hadoop-2.8.3/etc/hadoop/yarn-site.xml

按i键进入编辑模式，复制下面的配置内容粘贴到<configuration>和</configuration>之间：

<property>

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

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

</property>

<property>

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

<value>mapreduce\_shuffle</value>

</property>

<property>

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

<value> **esc-bigdatapro** </value>

</property>

<property>

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

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

</property>

<property>

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

<value>true</value>

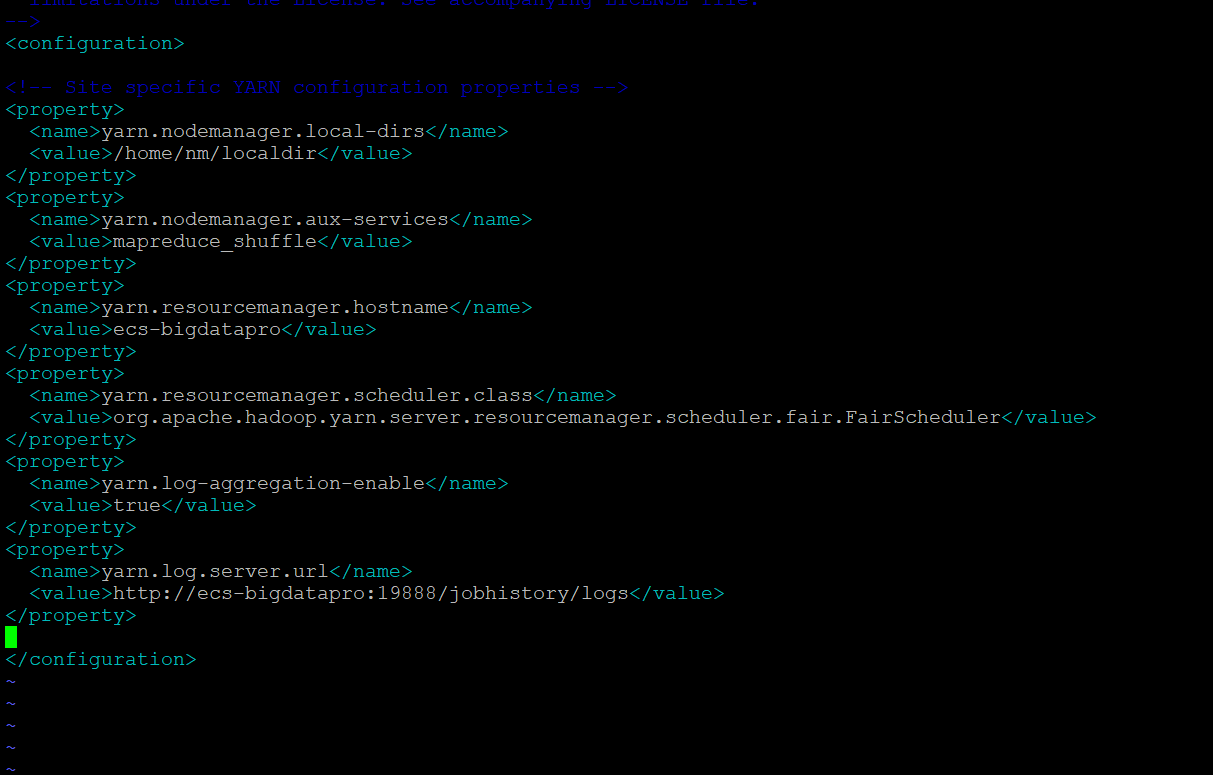
</property>

<property>

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

<value>http://**esc-bigdatapro**:19888/jobhistory/logs</value>

</property>



按ECS退出编辑模式，输入:wq保存并退出。

配置hadoop的mapred-site.xml文件

使用hadoop自带模板文件创建mapred-site.xml配置文件：

cd /home/modules/hadoop-2.8.3/etc/hadoop/

mv mapred-site.xml.template mapred-site.xml

使用vim打开配置文件。

vim /home/modules/hadoop-2.8.3/etc/hadoop/mapred-site.xml

按i键进入编辑模式，复制下面的配置内容粘贴到<configuration>和</configuration>之间：

<property>

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

<value>yarn</value>

</property>

<property>

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

<value>**esc-bigdatapro**:10020</value>

</property>

<property>

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

<value>**esc-bigdatapro**:19888</value>

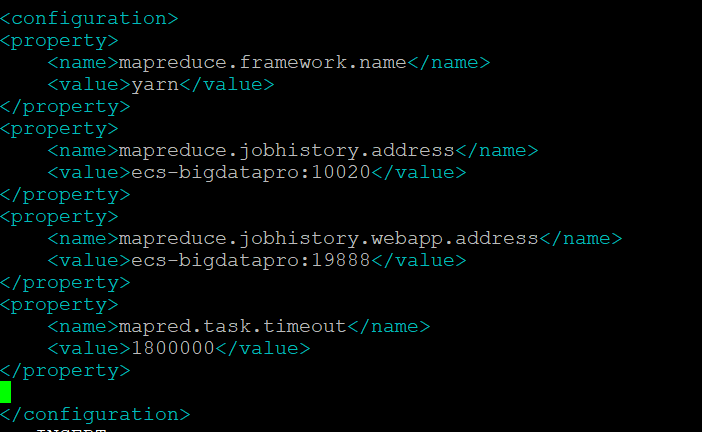
</property>

<property>

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

<value>1800000</value>

</property>



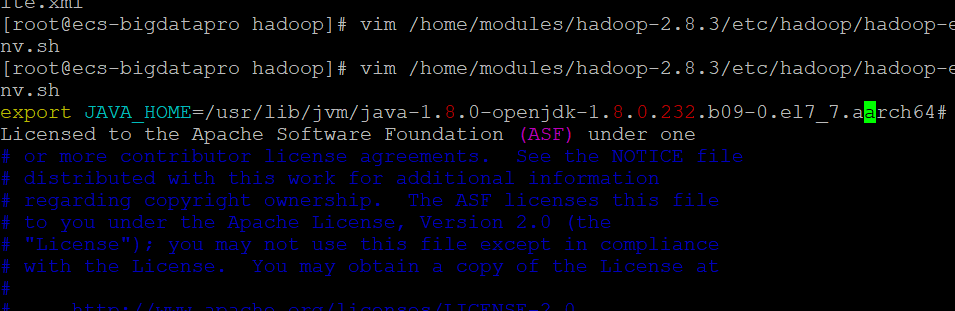
按ECS退出编辑模式，输入:wq保存并退出。

配置hadoop环境变量

使用vim命令打开配置文件。

vim /home/modules/hadoop-2.8.3/etc/hadoop/hadoop-env.sh

按i键进入编辑模式，找到JAVA\_HOME并修改如下：

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64

按ECS退出编辑模式，输入:wq保存并退出。

配置slaves

使用vim命令打开配置文件

vim /home/modules/hadoop-2.8.3/etc/hadoop/slaves

按i键进入编辑模式，删去原有内容，添加以下内容：

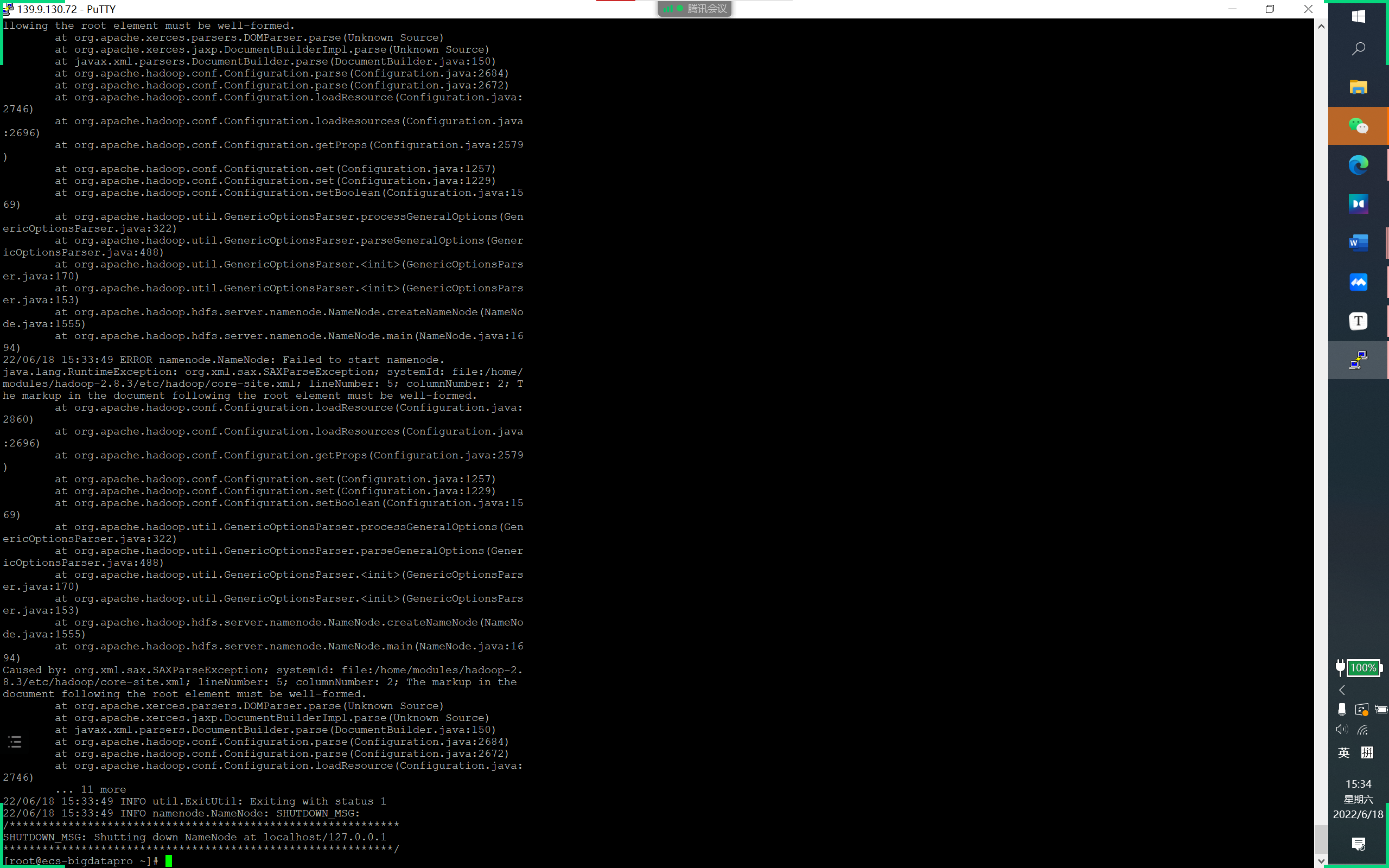
ecs-bigdatapro

按ECS退出编辑模式，输入:wq保存并退出。

namenode初始化

执行下列命令初始化namenode。

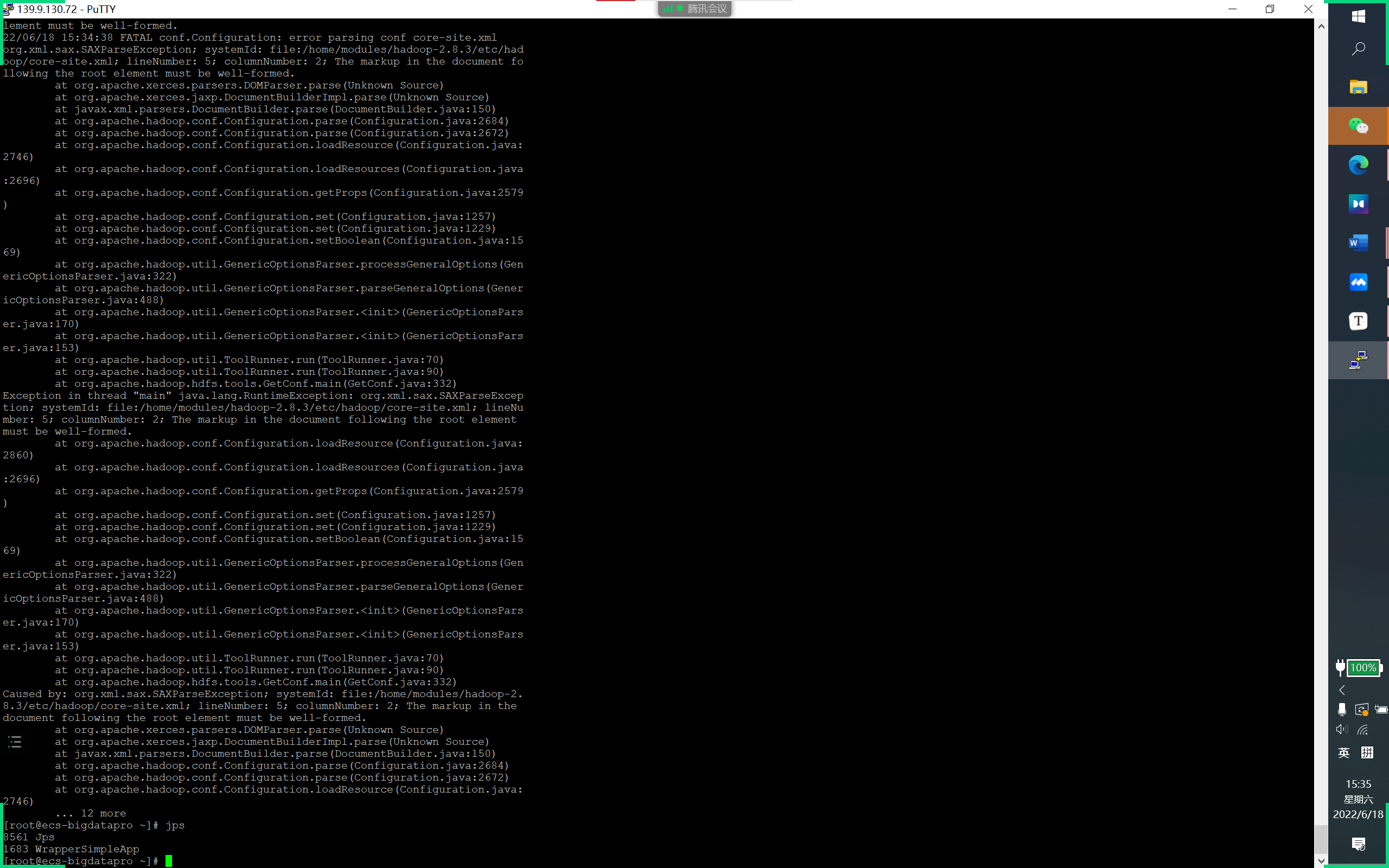
hdfs namenode -format



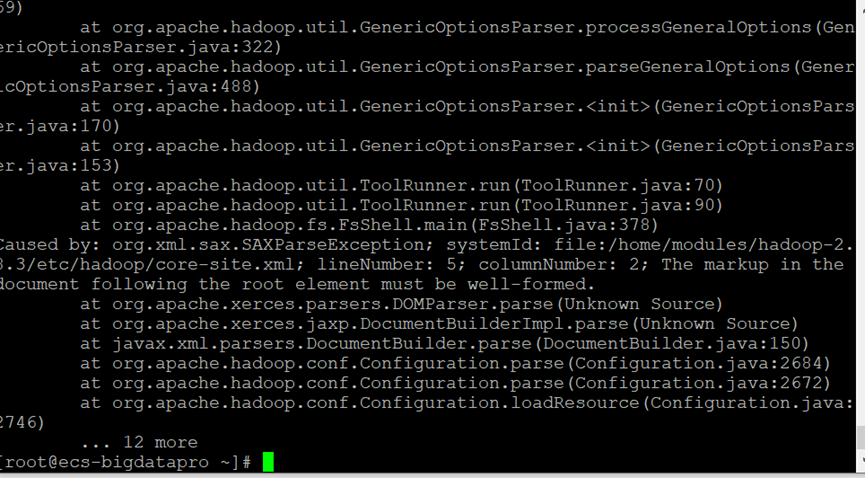
初始化成功。

启动HDFS

执行以下命令启动HDFS。

start-dfs.sh

使用JPS查看启动的进程。

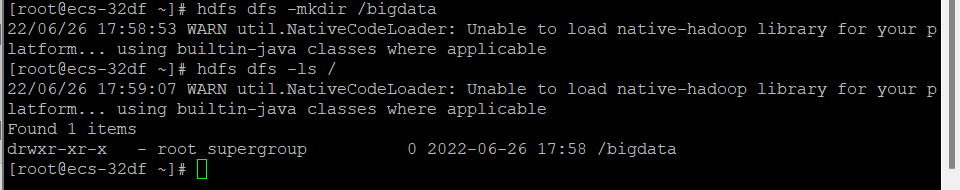


测试HDFS命令 ?

使用mkdir命令建立bigdata目录

hdfs dfs -mkdir /bigdata

hdfs dfs -ls /

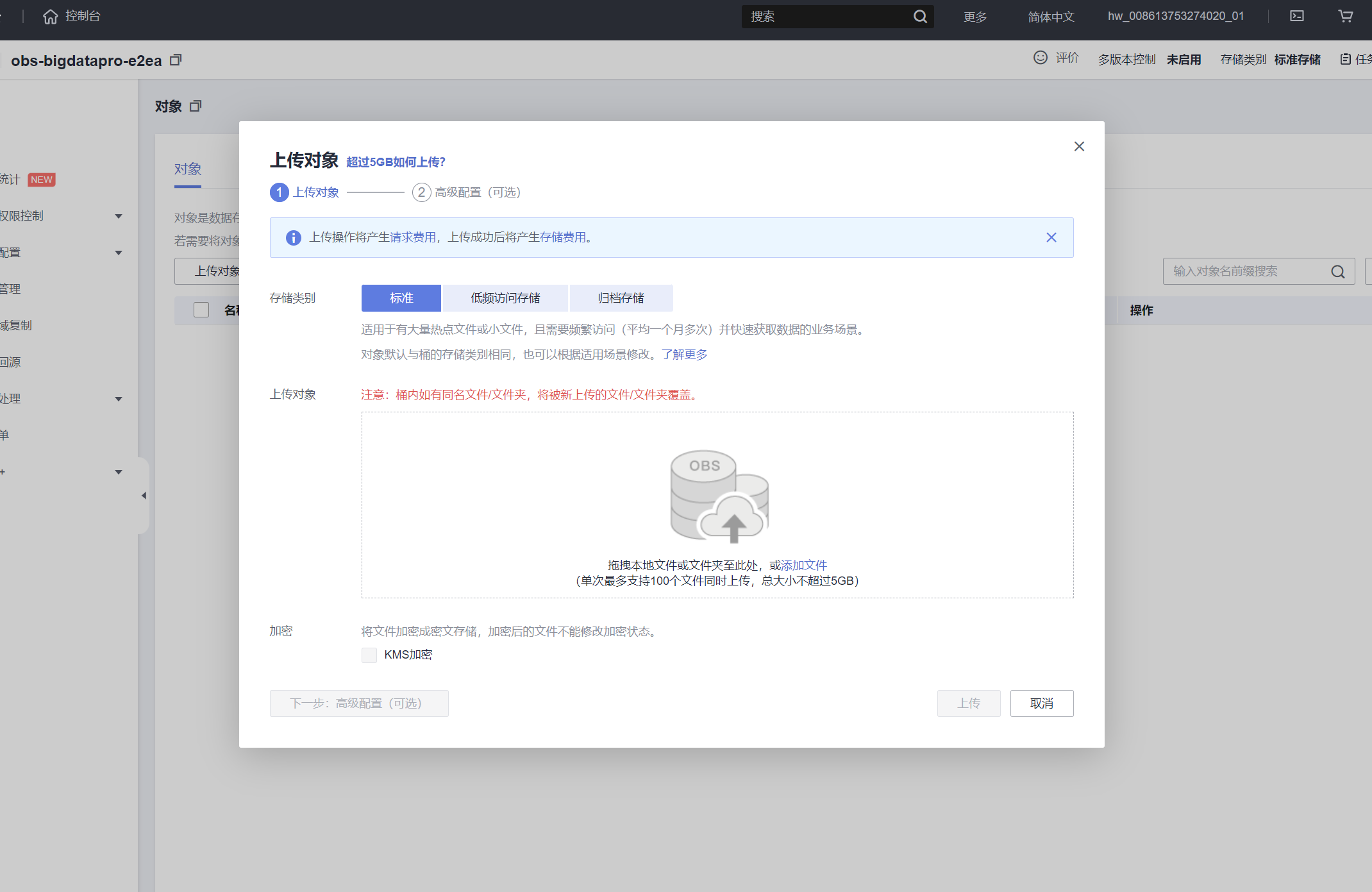


### Hadoop与OBS互联

在OBS上传文件

进入前面建立的OBS桶obs-bigdatapro-e2ea，选择“对象”点击“上传对象”。

在弹出的对话框中点击“添加文件”，选择playerinfo.txt文件。点击“上传”。OBS文件上传成功。（文件在我第一次实验时已上传，所以时间显示为06/12）



建立OBS的缓存目录

使用mkdir建立OBS所需的缓存目录。

mkdir -p /home/modules/data/buf

配置OBS信息

使用vim打开hadoop的配置文件core-site.xml。

vim /home/modules/hadoop-2.8.3/etc/hadoop/core-site.xml

按i键进入编辑模式，在<configuration>和</configuration>之间添加如下内容。

<property>

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

<value>true</value>

</property>

<property>

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

<value>**PXDN2CKOMLGBFZLQM6KL**</value>

</property>

<property>

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

<value>**2wmBYX3ZzGlSi65qoXQUgzC8BJzovOG4U6H63nOe**</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>/home/modules/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>

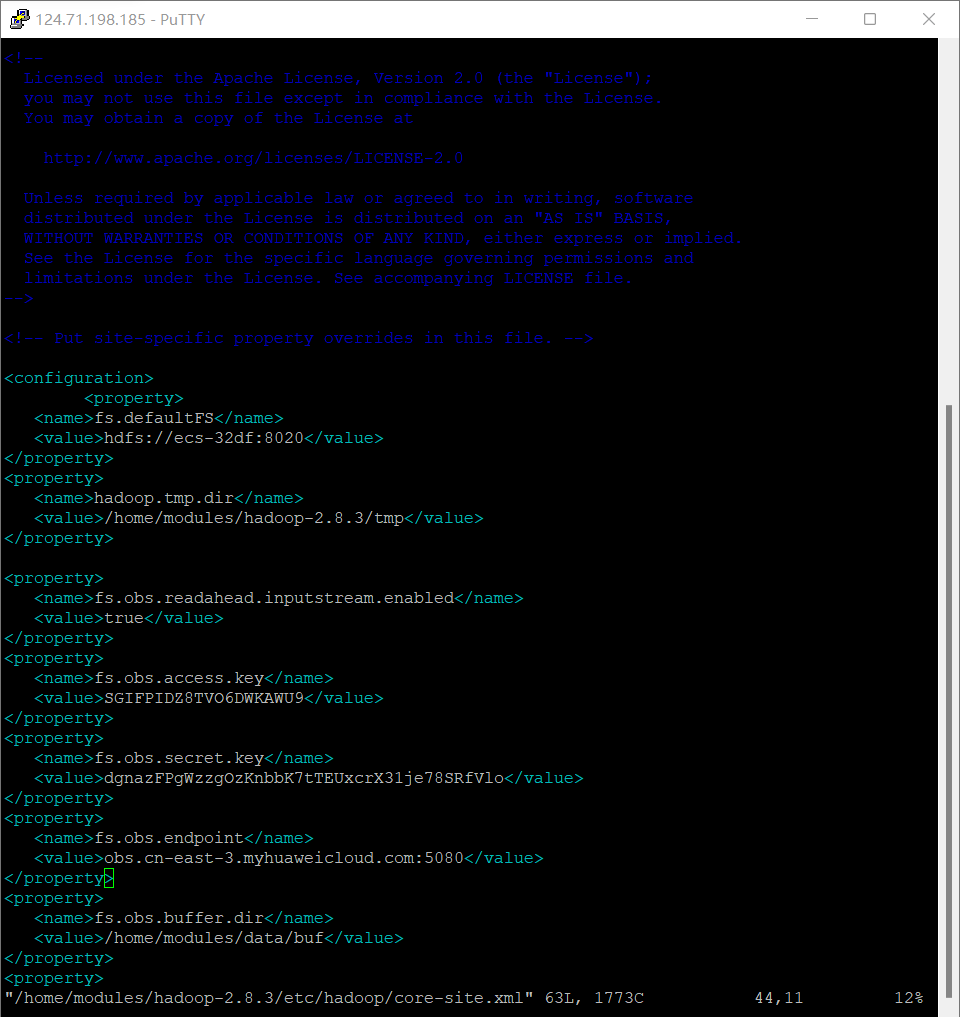
<name>fs.obs.fast.upload</name>

<value>true</value>

</property>

**fs.obs.access.key、fs.obs.secret.key**为2.3.2步骤5中获取的AK和SK的值。

**fs.obs.endpoint**为2.3.1步骤5中记录的OBS的endpoint的值。



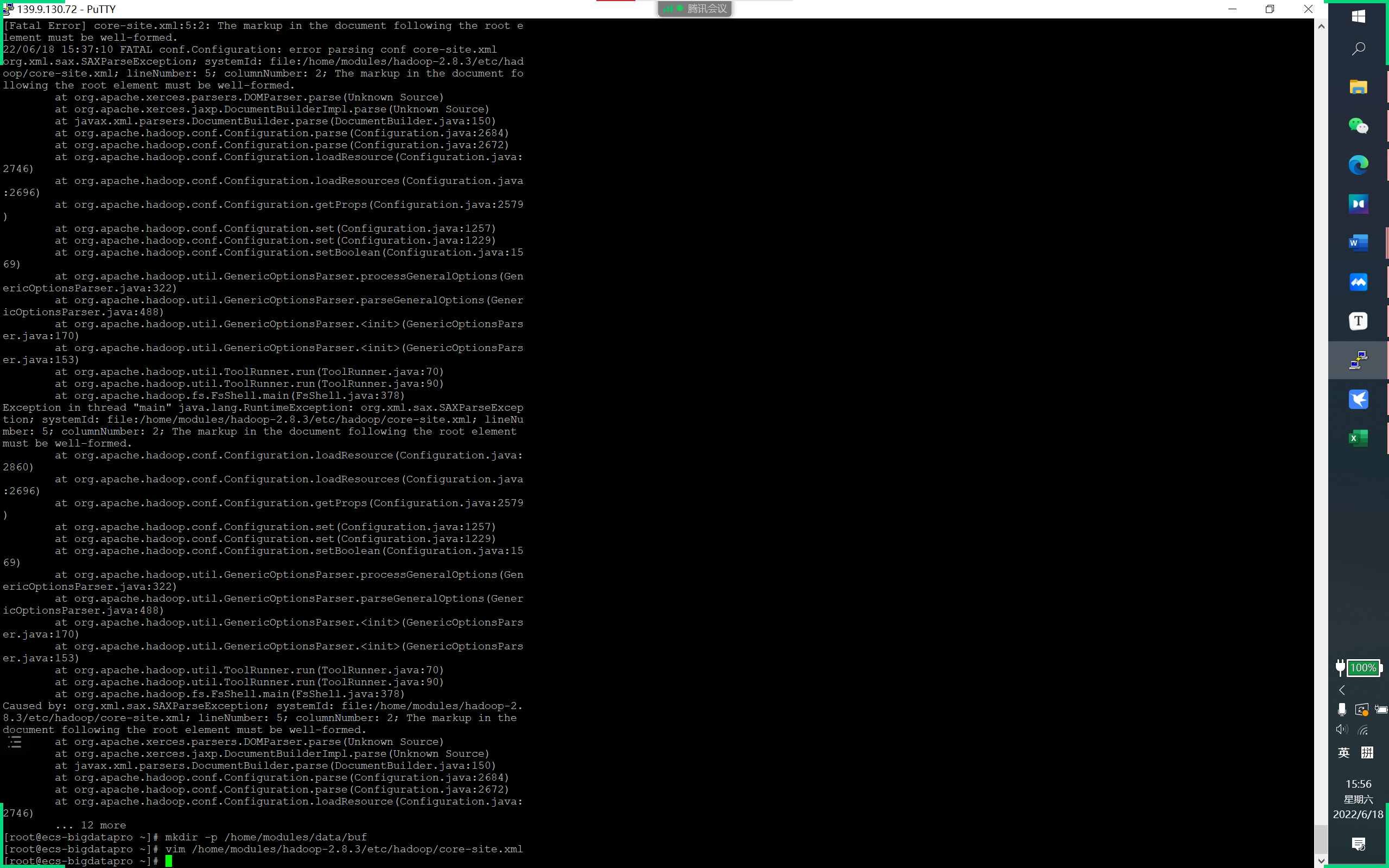
按ECS退出编辑模式，输入:wq保存并退出。

添加所需jar包

进入root目录，使用wget命令下载OBSFileSystem相关jar包：

cd /root

wget https://bigdata-tools-hw.obs.cn-north-1.myhuaweicloud.com/hadoop-huaweicloud-2.8.3-hw-39.jar



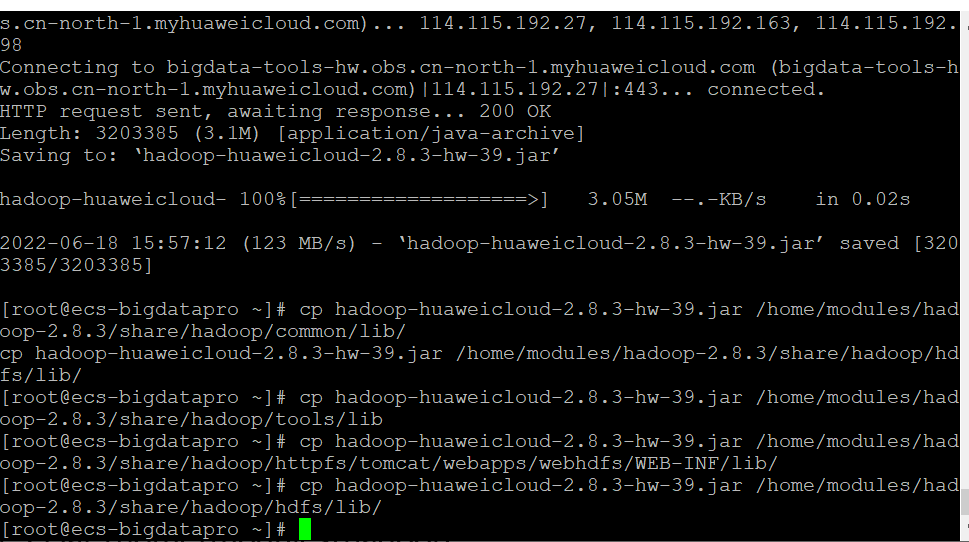
把下载的jar包复制到所需目录中（第一次实验时已复制，所以截图这里询问是否要overwrite）。

cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/common/lib/

cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/tools/lib

cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/httpfs/tomcat/webapps/webhdfs/WEB-INF/lib/

cp hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/hadoop-2.8.3/share/hadoop/hdfs/lib/



测试OBS互联 ??

执行HDFS命令查看OBS文件。

hdfs dfs -ls obs://obs-bigdatapro-e2ea/



Hadoop集群与OBS互联成功。

# Spark集群搭建

## 实验介绍

### 关于本实验

本部分实验介绍安装Spark集群，并使Spark能够读取OBS数据，使用Python编写Spark程序处理OBS中的数据（单词统计）。该实验使用Spark集群+OBS实现存算分离，提高计算性能。

## Spark集群存算分离

### 搭建Spark集群

获取Spark安装包

进入root目录下载Spark安装包。

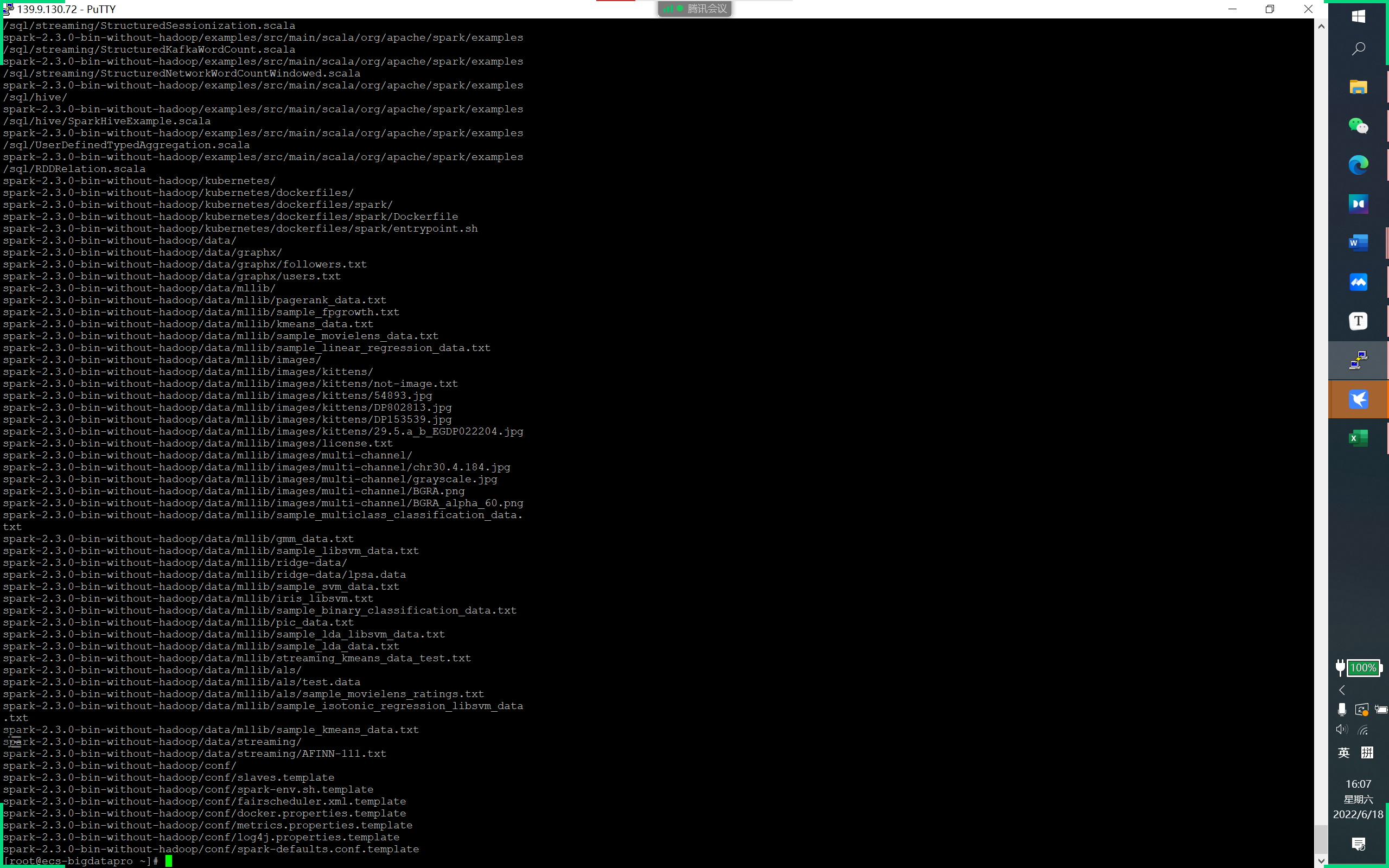
cd /root

wget <https://bigdata-tools-hw.obs.cn-north-1.myhuaweicloud.com/spark-2.3.0-bin-without-hadoop.tgz>

解压Spark安装包

使用tar命令解压Spark的安装包。

tar -zxvf spark-2.3.0-bin-without-hadoop.tgz



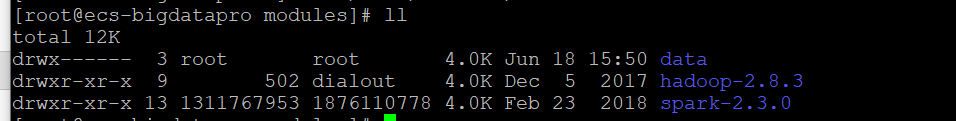
移动Spark到/home/modules目录下并重命名为spark-2.3.0。

mv spark-2.3.0-bin-without-hadoop /home/modules/spark-2.3.0

进入安装目录查看Spark。

cd /home/modules/

ll



配置相关的jar包

复制前面下载的OBSFileSystem相关jar包和hadoop的相关jar包到Spark的jars目录下。

cp /root/hadoop-huaweicloud-2.8.3-hw-39.jar /home/modules/spark-2.3.0/jars/

cp /home/modules/hadoop-2.8.3/share/hadoop/common/lib/snappy-java-1.0.4.1.jar /home/modules/spark-2.3.0/jars/

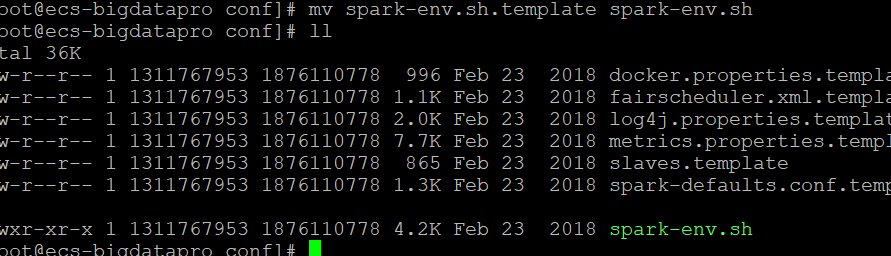
进行Spark的配置

通过mv命令使用Spark自带的配置文件模板创建出所需的配置文件。

cd /home/modules/spark-2.3.0/conf/

mv spark-env.sh.template spark-env.sh

ll



使用vim命令打开配置文件。

vim spark-env.sh

按下i键盘进入编辑模式，在文件末尾添加如下内容。

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.242.b08-1.h5.oe1.aarch64

export SCALA\_HOME=/home/modules/spark-2.3.0/examples/src/main/scala

export HADOOP\_HOME=/home/modules/hadoop-2.8.3

export HADOOP\_CONF\_DIR=/home/modules/hadoop-2.8.3/etc/hadoop

export SPARK\_HOME=/home/modules/spark-2.3.0

export SPARK\_DIST\_CLASSPATH=$(/home/modules/hadoop-2.8.3/bin/hadoop classpath)

按ESC退出编辑模式，输入:wq保存并退出。

配置系统环境变量

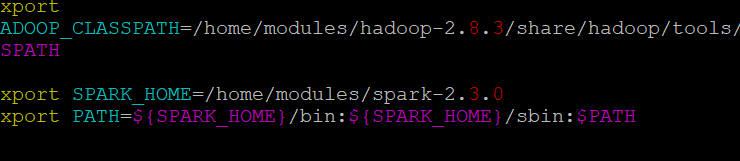
使用vim命令打开系统配置文件profile。

vim /etc/profile

按i键进入编辑模式，在文件尾添加如下内容。

export SPARK\_HOME=/home/modules/spark-2.3.0

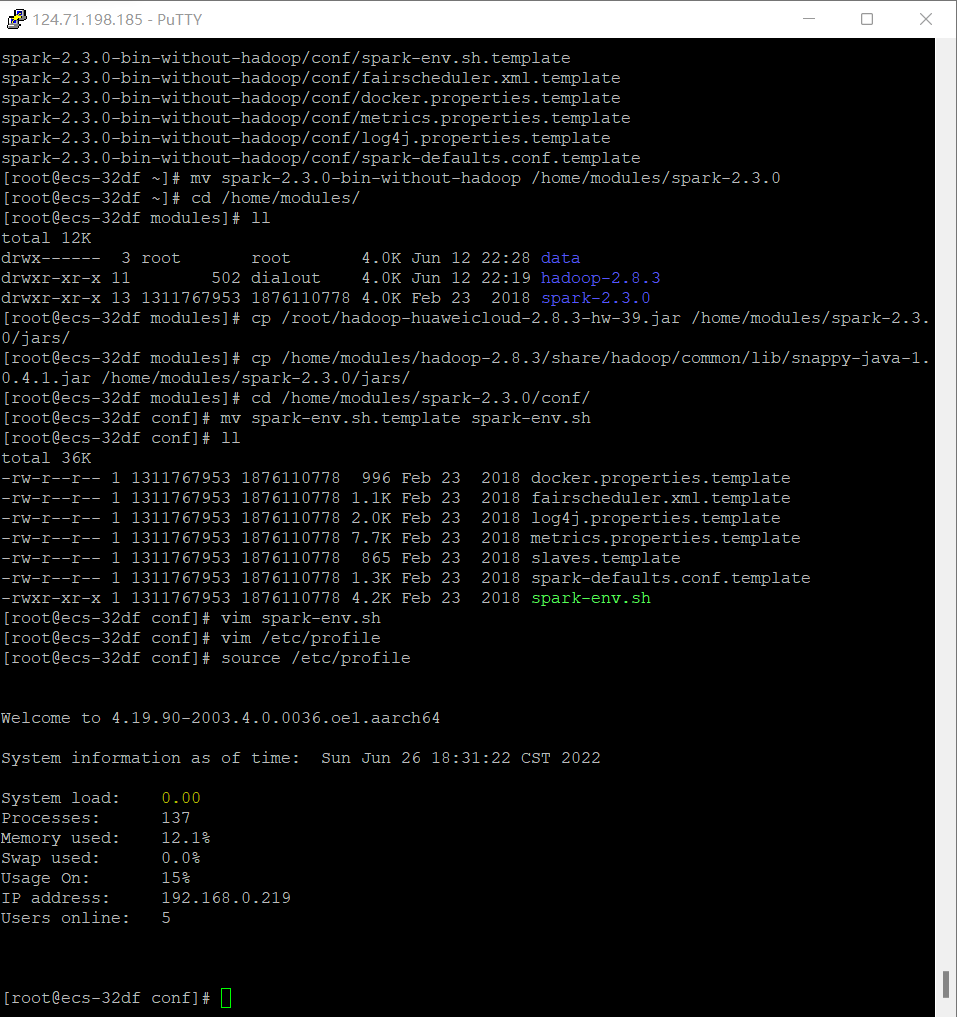
export PATH=${SPARK\_HOME}/bin:${SPARK\_HOME}/sbin:$PATH



按ECS退出编辑，输入:wq保存并退出。

使用source命令使用环境变量生效。

source /etc/profile



验证Spark安装

进入root目录，输入spark-shell命令。

spark-shell



显示版本号并进入命令提示符，Spark安装完成。

输入:quit退出spark-shell。

:quit



### 验证存算分离

查看要计算的文件

本次实验数据使用实验3.3.2中上传的playerinfo.txt文件。

数据内容如下：

Alex James Lax Genu

Kerry Mary Olivia William

Hale Edith Vera Robert

Mary Olivia James Lax

Edith Vera Robertm Genu

启动pyspark

输入**pyspark**回车启动。

pyspark



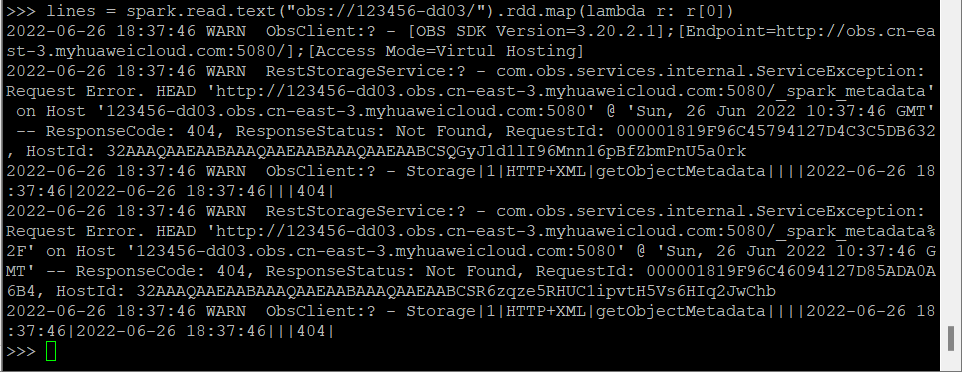
进入提示符“>>>”表示启动成功。

验证存算分离

编写Spark程序，读取OBS中存储的数据，然后计算单词出现次数。在pyspark中依次输入以下代码。

**读取OBS内容：**

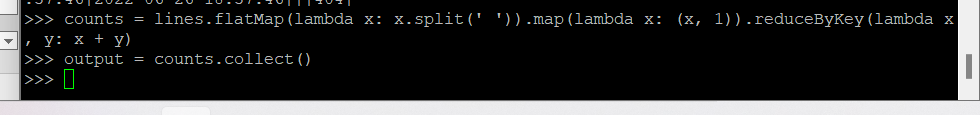
lines = spark.read.text("obs://**obs-bigdatapro-e2ea**/").rdd.map(lambda r: r[0])



**进行统计：**

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

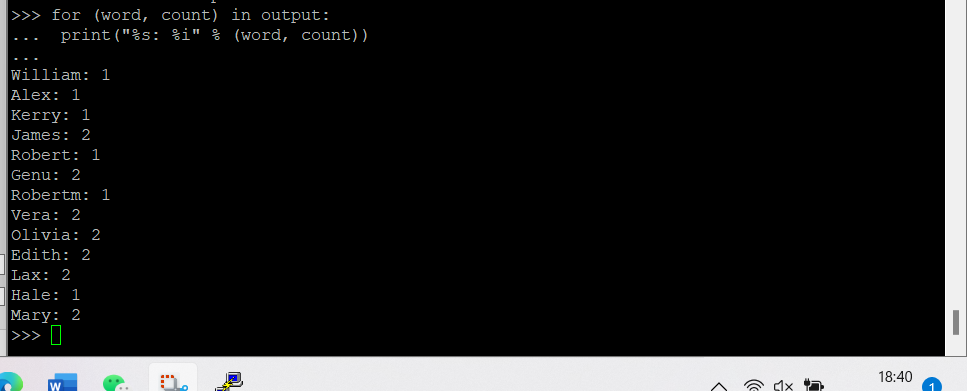
output = counts.collect()



**输出结果：**

for (word, count) in output:

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



存算分离验证成功！

退出pyspark

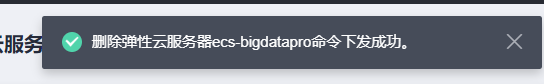
输入exit()退出pyspark。

# 释放华为云服务

## 释放弹性云服务器ECS

在控制台进入弹性云服务管理页面，选中要删除的实例对应的复选框，然后点击上面的“更多”在弹出的菜单中点击“删除”。

在弹出的对话框中选中两个复选框，然后点击“是”即可删除ECS和相关资源。



## 删除OBS桶

在控制台进入OBS服务，选择“对象存储”，然后点击对应桶后面的“删除”（如果桶非空需要先点击桶名进入桶内删除所有的对象文件）。

这里先删除桶中的对象文件playerinfo.txt；



再删除桶obs-bigdatapro-e2ea；

弹出的对话框选择“是”即可删除桶。