

《云综合实训》报告一

题目：云主机实现大数据

学 院 计算机科学与技术学院

班 级 计科20-4班

姓 名 刘宸

学 号 12205006

中国矿业大学

2023年06月

课程设计报告格式要求：

1.报告中应包含一、二级标题；

2.正文宋体小4号，1.5倍行距；

3.报告正文不少于1000字；

4.上交报告电子版，各班学委统一收齐后打包发邮件到wjyang@cumt.edu.cn。

课程设计报告评分要点：

1. 报告中云计算技术的应用情况（20%）
2. 报告中功能有效及完整性（40%）
3. 报告中的个人总结及体会情况（20%）
4. 报告格式规范程度（20%）

评语：

评分： 日期：

**目 录**

**[一.背景及](#_Toc21901)[功能概要 3](#_Toc2601)**

**[二、准备华为云环境](#_Toc80)** [3](#_Toc80)

[1.实验内容 3](#_Toc2904)

[2.实验目的 3](#_Toc25976)

[1.1购买华为云ECS 3](#_Toc23134)

[1.2 购买OBS 5](#_Toc21256)

**[三、搭建hadoop集群](#_Toc32613)** [6](#_Toc32613)

[1、实验内容 6](#_Toc8687)

[2.实验目的 6](#_Toc17252)

[1.1配置ECS 6](#_Toc1021)

[1.2安装OpenJDK 11](#_Toc21174)

[1.3搭建Hadoop集群 13](#_Toc30380)

[1实验内容 22](#_Toc21844)

[2.实验目的 22](#_Toc10479)

**[四.MapReduce存算分离](#_Toc12761)** [22](#_Toc12761)

[3.1 测试Hadoop集群功能 22](#_Toc21661)

[步骤 1 node1节点启动YARN执行命令 22](#_Toc6590)

[步骤 2 测试文件 22](#_Toc20544)

1. **背景及功能概要**

**1.背景**

使用hadoop主要是为了解决海量数据的存储和处理问题，特别是非结构化或半结构化数据的分析，利用廉价的硬件和分布式算法实现高容错性、高可扩展性和高效率的数据平台。

**2.功能概要**

hadoop是一个开源的分布式计算框架，可以在大量廉价的机器上存储和处理海量的数据。hadoop主要包含两个核心组件：HDFS和MapReduce。

HDFS是hadoop分布式文件系统，负责将数据切分成多个块，并分布式地存储在不同的节点上，提供高容错性和高吞吐量的数据访问。

MapReduce是hadoop分布式计算框架，负责将用户编写的业务逻辑代码和自带默认组件整合成一个完整的分布式运算程序，并发运行在一个hadoop集群上。MapReduce程序通常包含两个阶段：Map阶段和Reduce阶段，分别负责对输入数据进行并行处理和汇总结果。

### **二、准备华为云环境**

**1.实验内容**

购买华为云ECS并在华为云上选取对象存储服务OBS服务，为后续实验提供环境基础。

### **2.实验目的**

掌握华为云上购买ECS步骤。掌握华为云上选取对象存储服务OBS。

## **1.1购买华为云ECS**

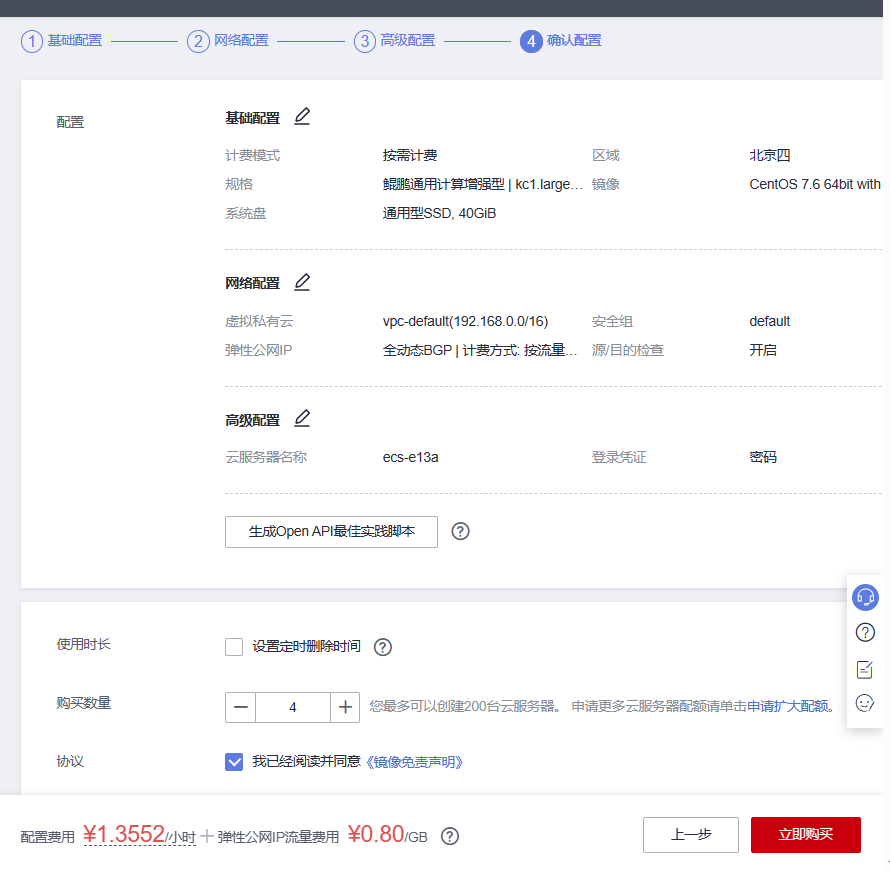
选择“买弹性云服务器ECS”：

配置操作系统和磁盘

选择“公共镜像”，CentOS7.6，系统盘建议配置40GB，购买数量4台，点击“网络配置”

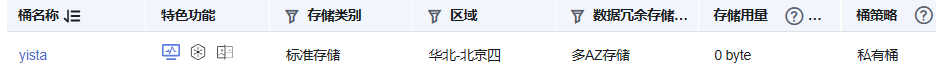
配置网络，网络选择“vpc-default”，安全组选择“Sys-default”，“现在购买”，选择“全动态BGP”，“按流量计算”，“5M”，点击“高级配置”

配置密码，自定义云服务器名称，自行设置root登录密码，云备份选择“暂不购买”，点击“确认配置”，点中“我已经阅读并同意”，点击“立即购买”，如下图：

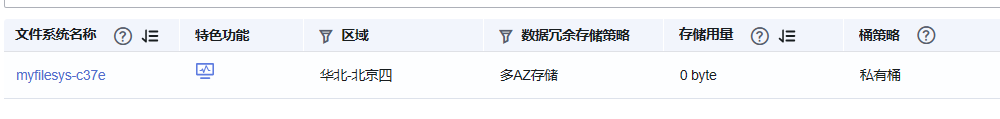


## **1.2 购买OBS**

进入控制台，选择“对象存储服务”，选择“创建桶”，自定义桶名称，选择“标准存储”，点击“立即创建”：

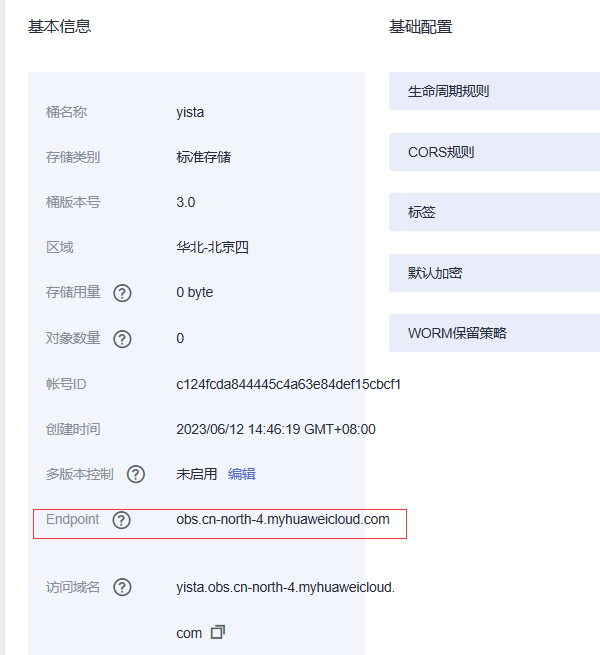


选择“创建并行文件系统”，自定义文件系统名称，点击“自定义立即创建”，如下图：



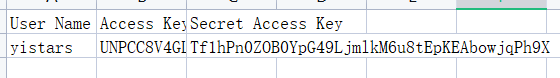
进入创建的OBS桶，复制该参数，保存到本地：

obs.cn-north-4.myhuaweicloud.com



获取AK/SK，点击“我的凭证”，选择“访问秘钥”，点击新增访问秘钥，根据提示进行操作。通过手机号码，接收短信验证码，点击“确定”，操作完成后，得到文件，打开即可得到AK/SK。





## **三、搭建hadoop集群**

**1、实验内容**

搭建开源hadoop集群，使开源hadoop与华为云OBS服务互联，使hadoop集群可读取OBS数据。

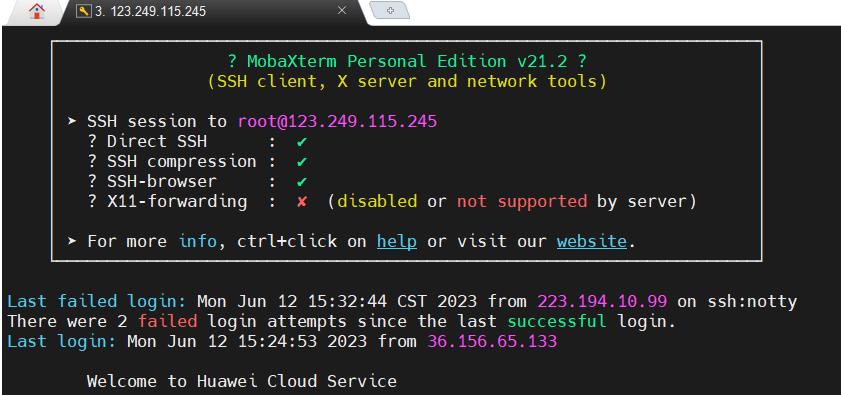
**2.实验目的**

掌握hadoop集群搭建方法。掌握hadoop集群与华为云OBS互联方法

### **1.1配置ECS**

使用mobaXterm远程登录ECS

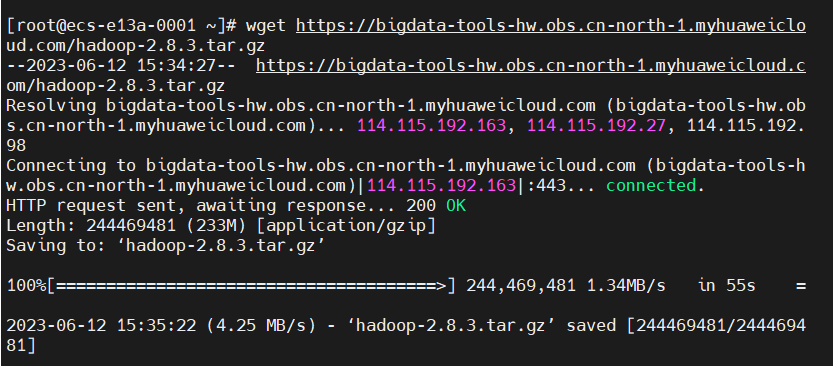
网址根据实际node节点IP地址进行填写。



node1节点登录成功后执行命令在线下载安装包

wget

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



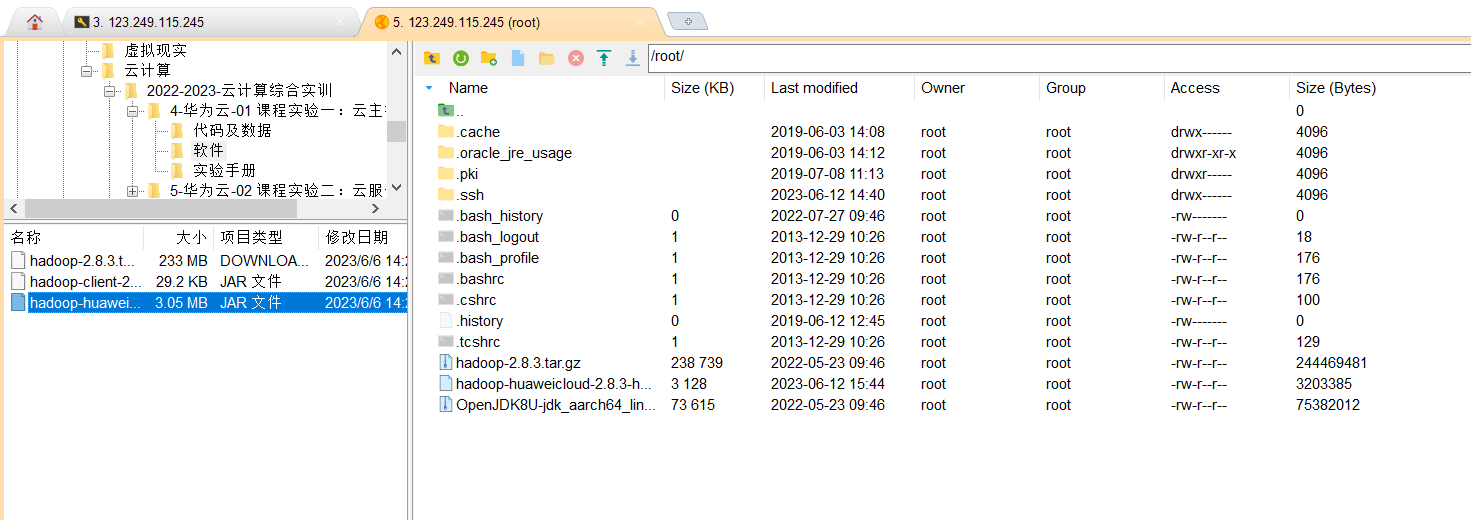
下载OBSFileSystem

登录

https://github.com/huaweicloud/obsa-hdfs/tree/master/release

下载hadoop-huaweicloud-2.8.3-hw-39.jar

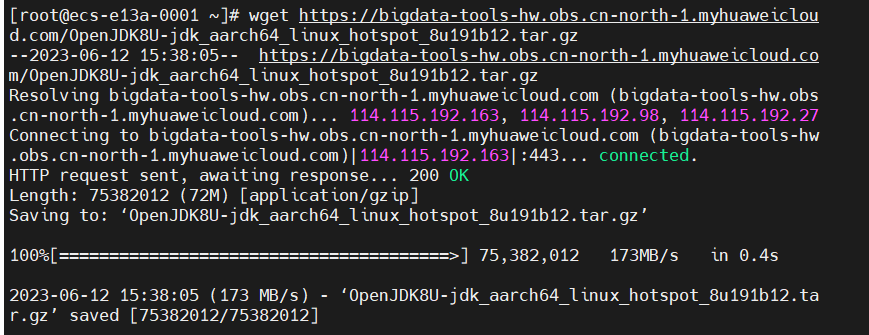
下载后通过SFTP连接，从本地将文件传给服务器



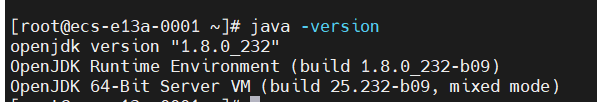
下载OpenJDK

执行命令

wget https://bigdata-tools-hw.obs.cn-north-1.myhuaweicloud.com/OpenJDK8U-jdk\_aarch64\_linux\_hotspot\_8u191b12.tar.gz



***执行java –version，查看到java安装成功***



关闭防火墙

systemctl stop firewalld

systemctl disable firewalld

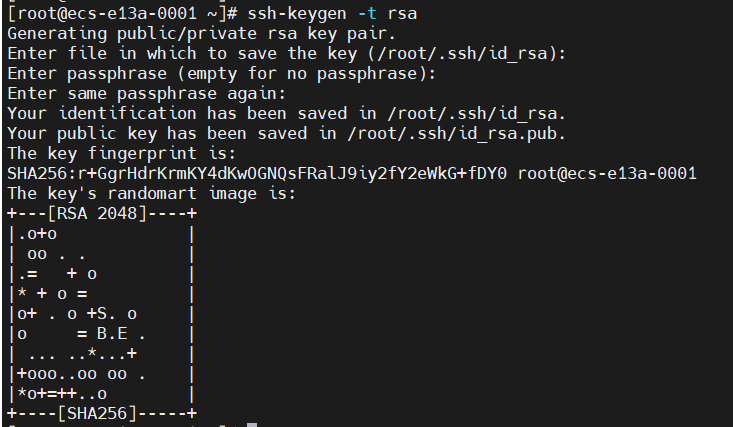


配置节点互信

node1~node4节点分别执行如下命令：

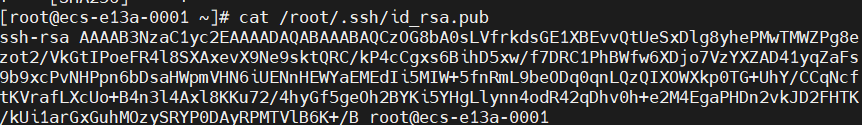
#ssh-keygen -t rsa

提问框按默认连续回车即可，生成/root/.ssh/id\_rsa.pub文件。



node1~node4节点分别执行命令cat /root/.ssh/id\_rsa.pub命令。

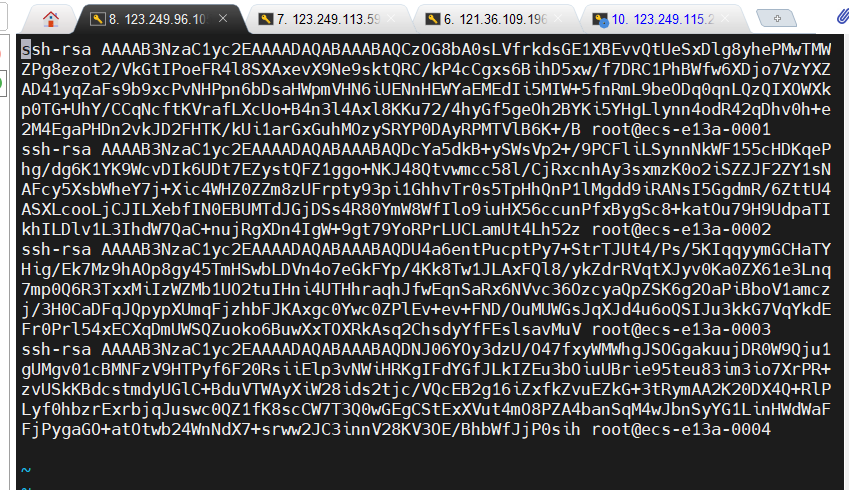
cat /root/.ssh/id\_rsa.pub



将4个节点的内容拷贝汇总到一个文本中，再将该文本内容拷贝到node1、node2、node3、node4的/root/.ssh/authorized\_keys中。

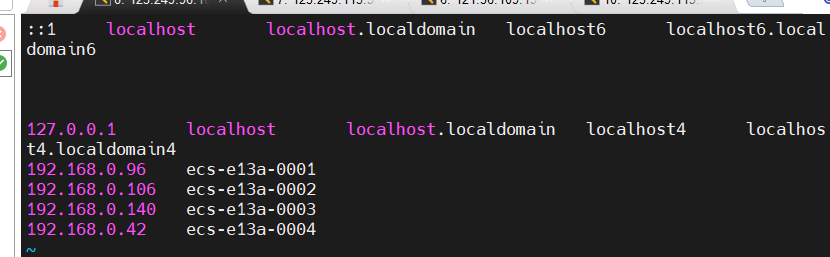
vim /root/.ssh/authorized\_keys





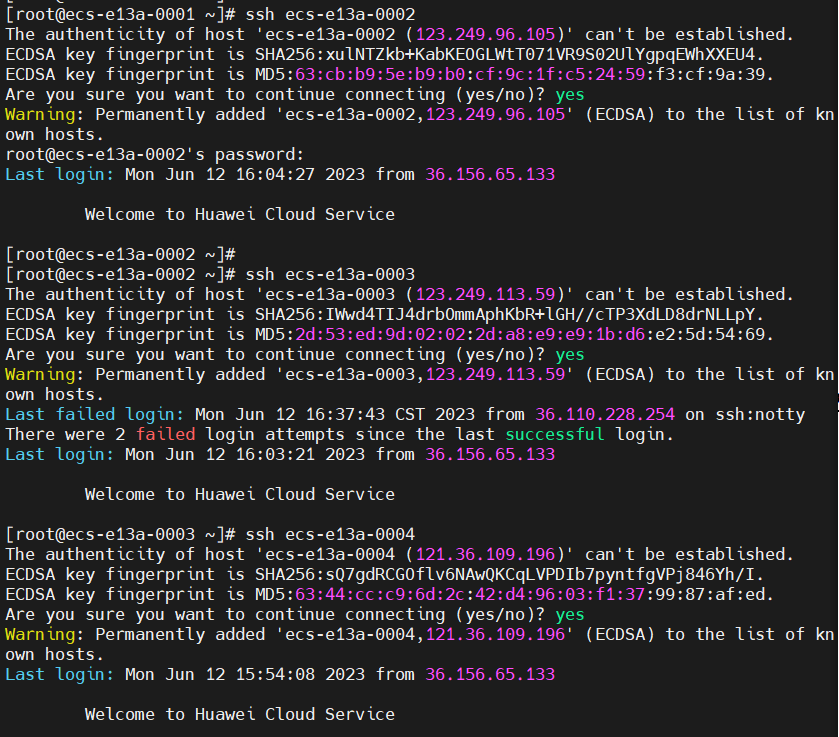
node1~node4节点分别执行命令vim /etc/hosts，加入node1~node4对应IP及node节点名。

vim /etc/hosts



***注意：此处要删除127.0.0.1 ecs-xxx-0001 ecs-xxx-0001一行。hosts配置完成后，执行chattr +i /etc/hosts命令***

node1~node4节点分别执行命令ssh node1~node4，选择yes后跳转到目的节点。node1节点无密码跳转到node2节点，其余同理。



### **1.2安装OpenJDK**

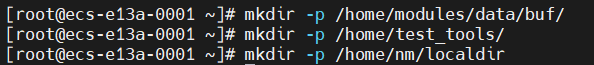
创建目录

node1~node4四个节点分别执行下列命令。

mkdir -p /home/modules/data/buf/

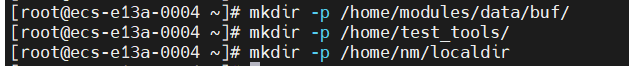
mkdir -p /home/test\_tools/

mkdir -p /home/nm/localdir









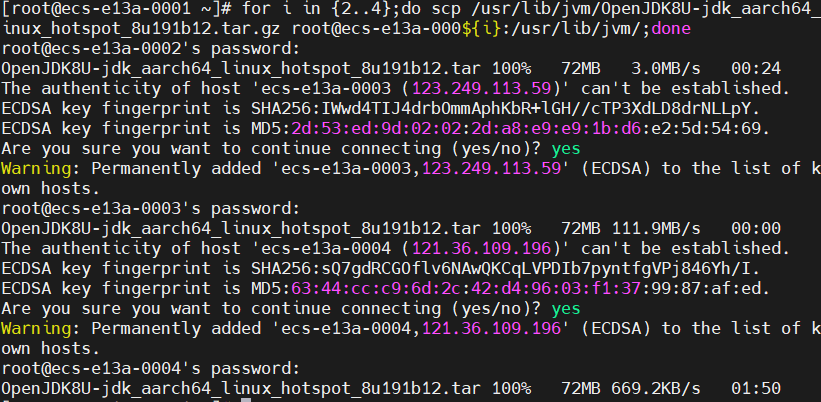
安装OpenJDK

在node1节点，执行如下命令，将jdk安装包拷贝到/usr/lib/jvm目录下。

cp OpenJDK8U-jdk\_aarch64\_linux\_hotspot\_8u191b12.tar.gz /usr/lib/jvm/

在node1节点，执行如下命令，将jdk安装包拷贝到其他几个节点，

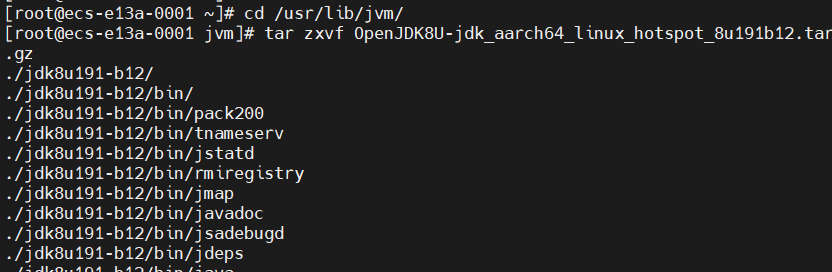
For in {2..4};do scp /usr/lib/jvm/OpenJDK8U-jdk\_aarch64\_linux\_hotspot\_8u191b12.tar.gz root@ecs-e13a-000${i}:/usr/lib/jvm/;done



在node1~node4四个节点分别执行命令

cd /usr/lib/jvm/

tar zxvf OpenJDK8U-jdk\_aarch64\_linux\_hotspot\_8u191b12.tar.gz

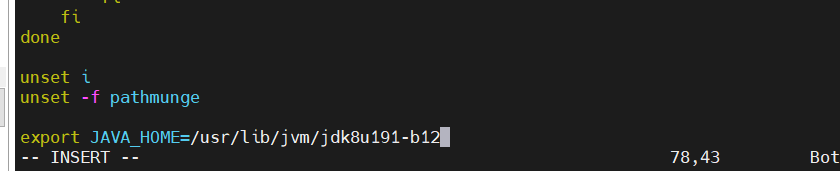


在node1~node4四个节点上编辑/etc/profile增加如下的配置

vim /etc/profile



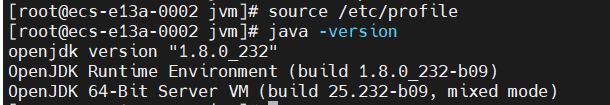
添加export JAVA\_HOME=/usr/lib/jvm/jdk8u191-b12

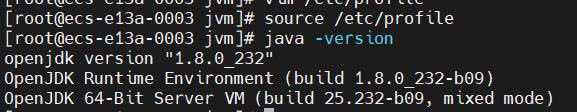


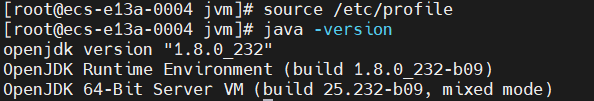
确认java版本

source /etc/profile

java -version







## **1.3搭建Hadoop集群**

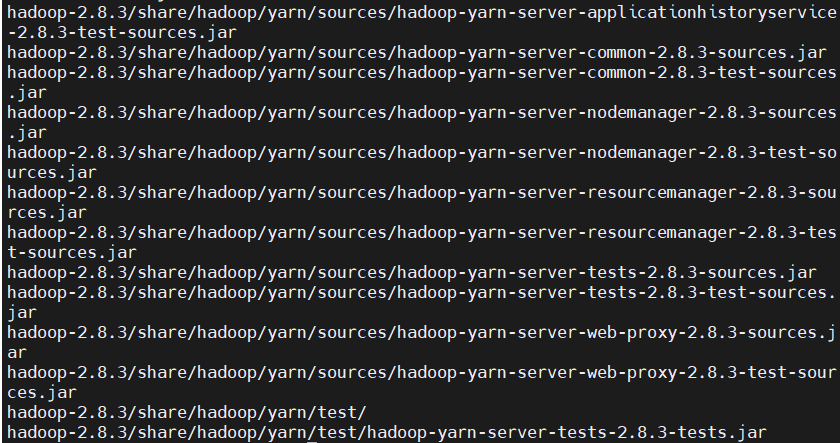
### **1.3.1搭建Hadoop集群**

登录node1节点，解压hadoop安装包

cp hadoop-2.8.3.tar.gz /home/modules/

cd /home/modules/

tar zxvf hadoop-2.8.3.tar.gz

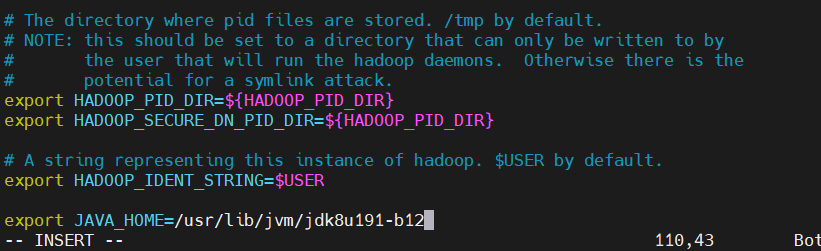


配置hadoop环境变量

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

编辑加入

export JAVA\_HOME=/usr/lib/jvm/jdk8u191-b12

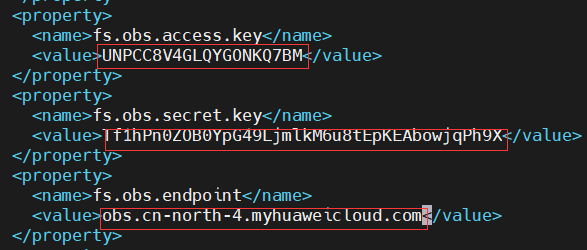


配置hadoop core-site.xml配置文件

执行命令：

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

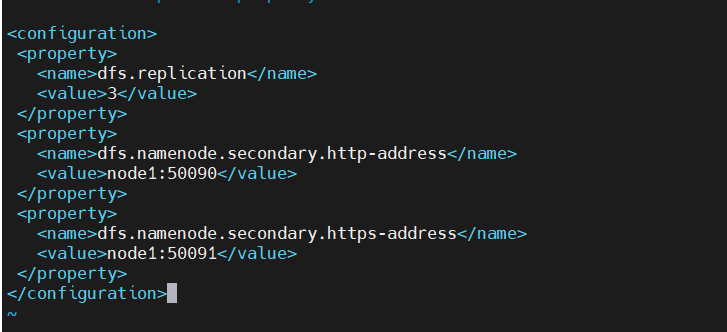




配置hdfs-site.xml

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

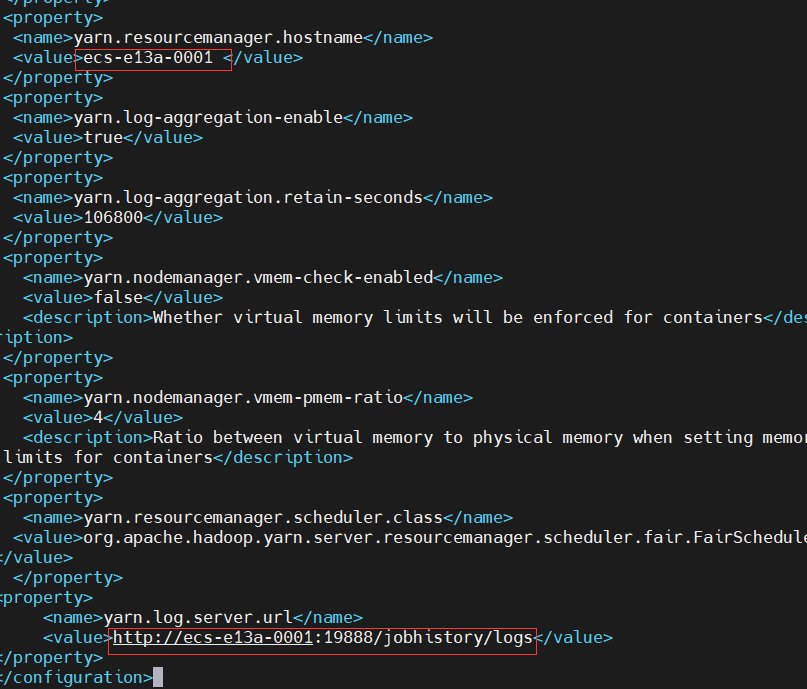




配置yarn-site.xml

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



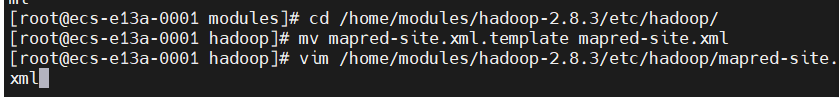
配置mapred-sit.xml

执行下列命令

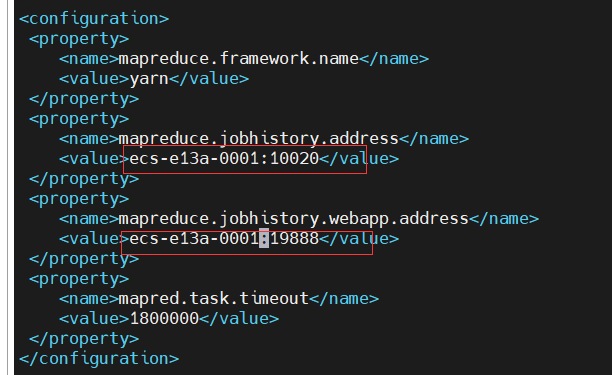
cd /home/modules/hadoop-2.8.3/etc/hadoop/

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

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



参数配置如下(ecs-e13a-0001为节点名)：



配置slaves

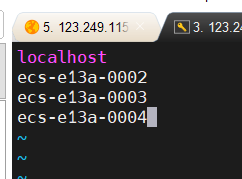
vim /home/modules/hadoop-2.8.3/etc/hadoop/slaves

编辑内容如下：

node2

node3

node4



配置jar包

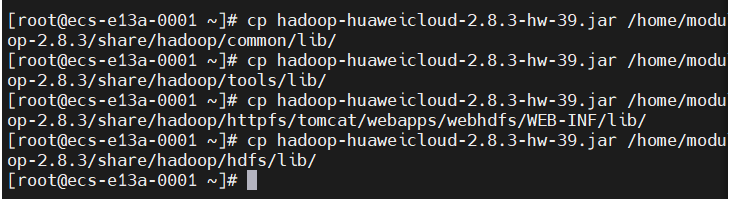
通过mobaXterm将hadoop-huaweicloud-2.8.3-hw-39.jar文件上传至服务器后执行以下命令

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/

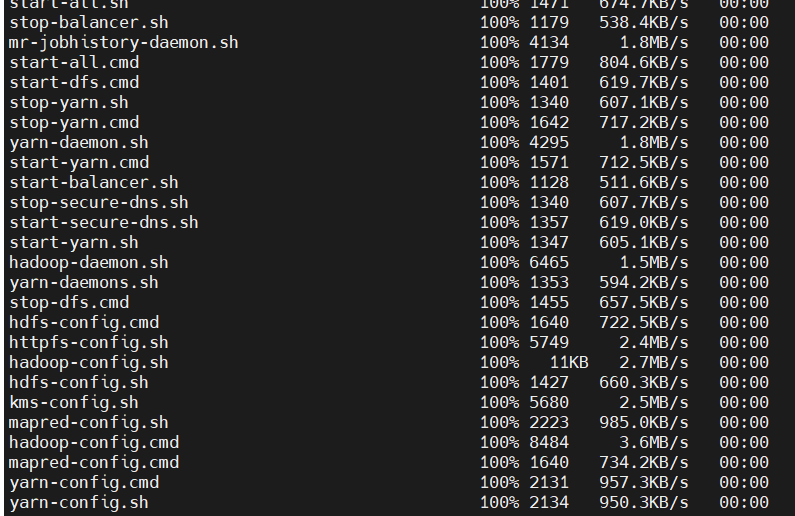


分发hadoop包到各节点

node1~node4四个节点下执行下列命令(node-000为自己节点前面名称，i为节点数，根据自己实际节点名进行编写)

for i in {2..4};do scp -r /home/modules/hadoop-2.8.3 root@node-000${i}:/home/modules/;done





配置环境变量

node1~node4四个节点下执行下列命令：

vim /etc/profile

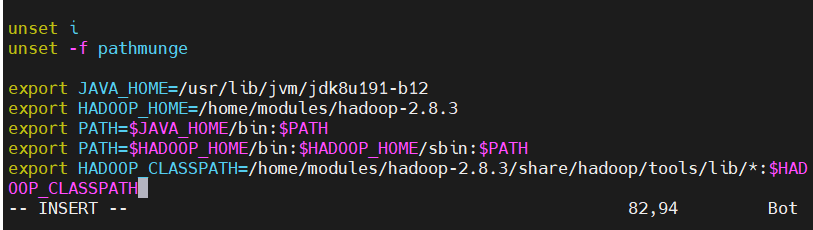
添加如下：

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

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

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

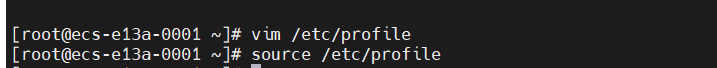
export HADOOP\_CLASSPATH=/home/modules/hadoop-2.8.3/share/hadoop/tools/lib/\*:$HADOOP\_CLASSPATH



node1节点添加如图所示，其余节点同理。

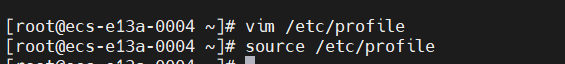
node1~node4四个节点下执行下列命令：

source /etc/profile









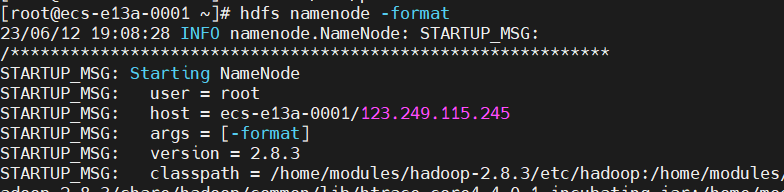
namenode初始化

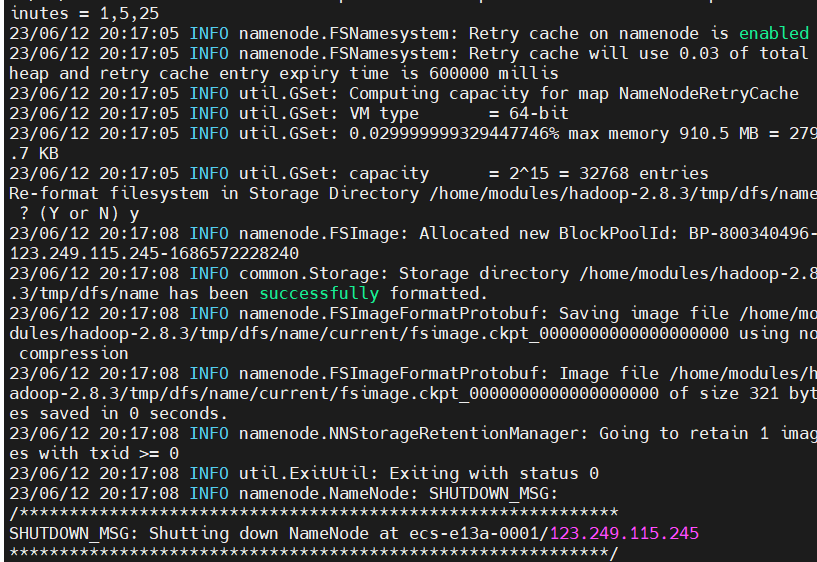
node1节点执行namenode初始化

执行下列命令：

hdfs namenode -format

初始化成功后，启动hdfs如图所示：





启动HDFS

node1节点执行命令：

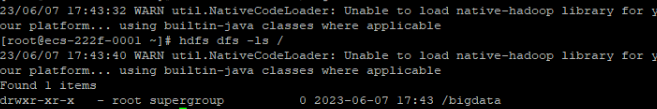
start-dfs.sh



执行hdfs命令

hdfs dfs -mkdir /bigdata

hdfs dfs -ls /



### **1.3.2测试与OBS互联**

在OBS上传文件 选择“对象存储服务OBS”,选择“桶名称”,如下图：



OBS文件上传成功

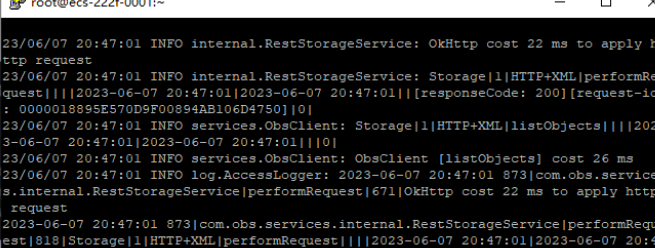


执行hdfs命令查看OBS文件

obs-bxw为实际自己创建的桶名称

# hdfs dfs -ls obs://obs-bxw/

Hadoop集群与OBS互联成功



**四、MapReduce程序实验**

**1实验内容**

本实使用MapReduce验证存算分离，数据存放在OBS上，计算在ECS执行。通过本实验可以掌握大数据存算分离操作。

**2.实验目的**

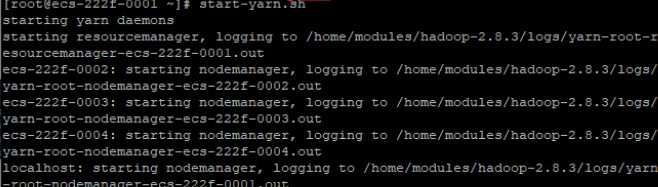
掌握MapReduce程序执行方法。掌握存算分离执行方法

**3.MapReduce存算分离**

**3.1 测试Hadoop集群功能**

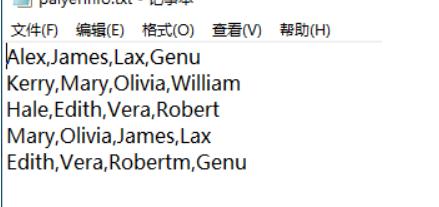
node1节点启动YARN执行命令

start-yarn.sh



测试文件

palyerinfo.txt内容如下：



Alex James Lax Genu

Kerry Mary Olivia William

Hale Edith Vera Robert

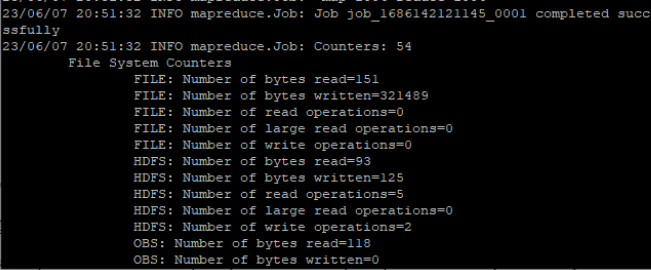
Mary Olivia James Lax

Edith Vera Robertm Genu

执行hadoop wordcount

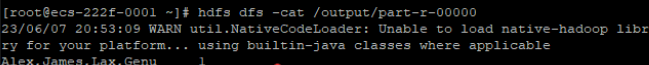
obs-bxw为实际自己创建的桶名

hadoop jar /home/modules/hadoop-2.8.3/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.8.3.jar wordcount obs://obs-bxw/ /output



查看结果

hdfs dfs –cat /output/part-r-00000



至此应用MapReduce实现存算分离计算成功