课程实验一：云主机实现大数据

实验时间：2021年03月25日

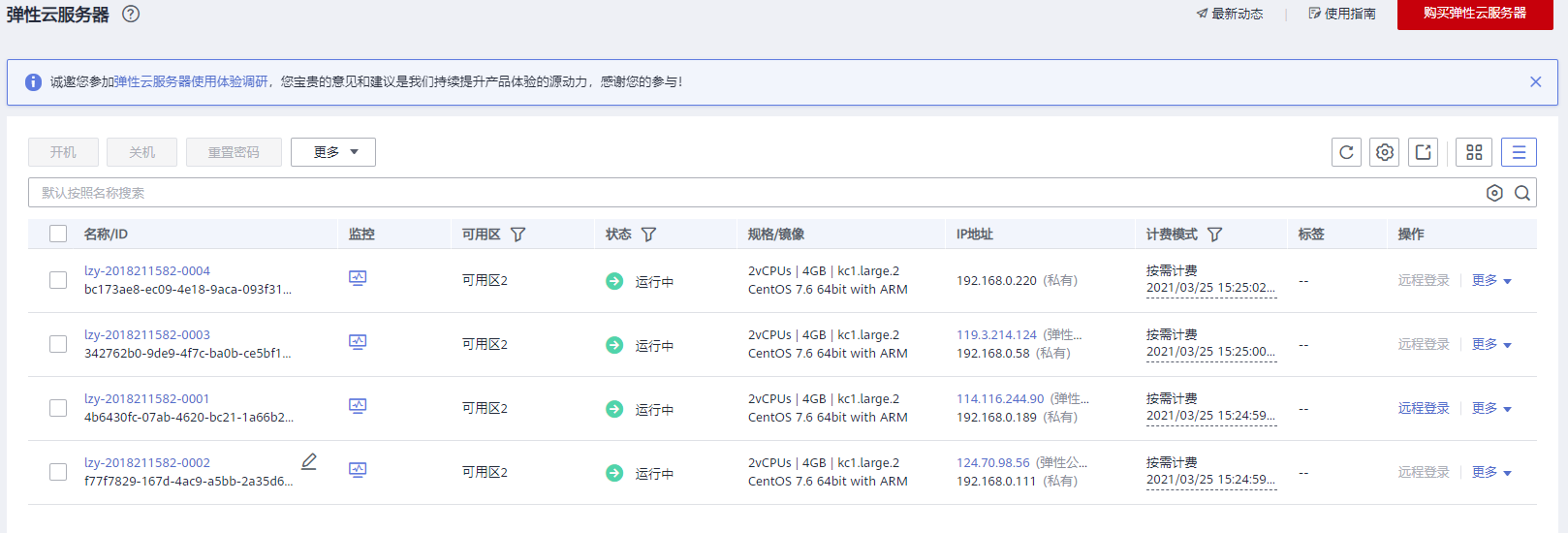
实验学生姓名：李志毅

学生班号、学号：2018211314班 2018211582

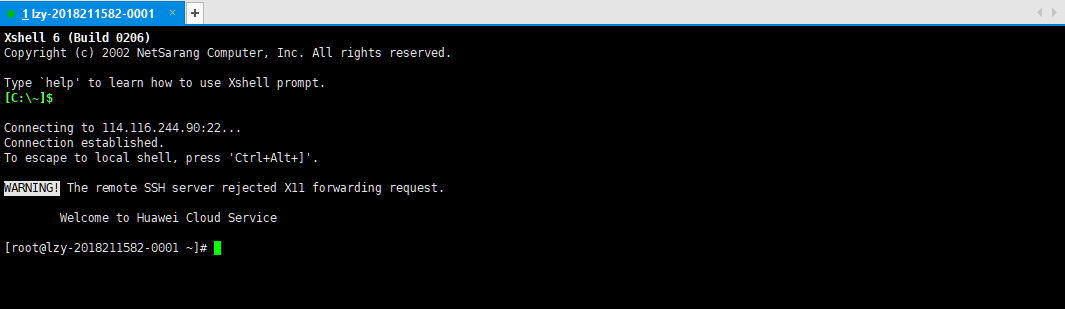
**一、实验步骤**

请大家参照《云主机实现大数据实验指导书》完成本次实验，并将实验中的结果截图，完成本次实验报告。

1. “1.2 购买华为云ECS”中的步骤6，要求自定义云服务器名称为“姓名+学号+节点序号（1234）”。

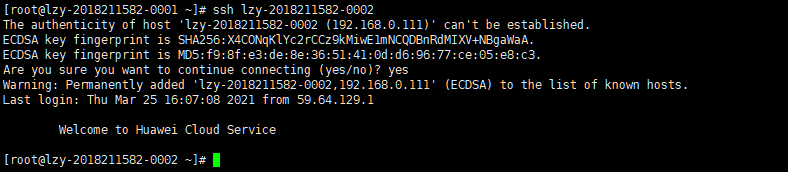


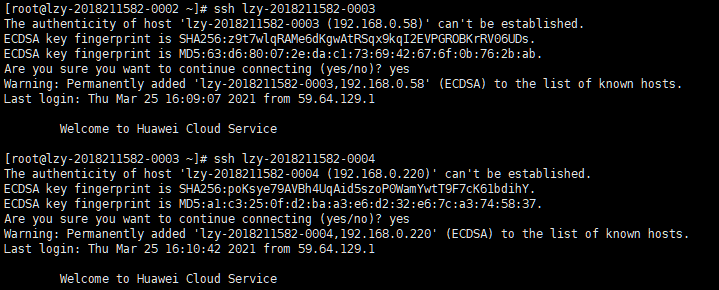
2. “2.2.1 配置ECS”中的步骤1，使用putty连接成功后，在这里贴node1主机登录成功的命令行界面，必须要体现出主机名和IP地址



**个人理解：**我这里选用的是xshell登录我的node1服务器，使用xshell可以和xftp搭配能够上传文件。我的node1服务器名字为‘lzy-2018211582-0001’，图示为登录成功后显示内容，公网ip 114.116.244.90

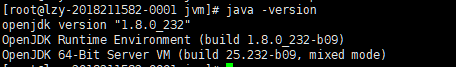
3. “2.2.1 配置ECS”中的步骤6配置节点互信，在这里贴任意一个节点执行ssh命令跳转成功的截图，要体现出执行的命令和运行结果



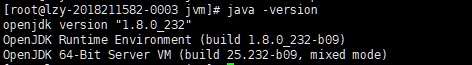


**个人理解：**配置ssh互信，使得四个节点之间可以无需密码即可ssh命令登录，为之后的MapReduce项目铺垫，这里展示了配置完成后的截图，如图所示，node1节点**‘lzy-2018211582-0001’**服务器执行命令**ssh lzy-2018211582-0002**后可直接登录到‘**lzy-2018211582-0002’**，在node2节点上执行**ssh lzy-2018211582-0003**可直接登陆到node3节点，node3节点执行命令可直接登录到node4节点。

4. “2.2.2 安装JDK”中的步骤5，在这里贴执行“java -version”命令后的结果，显示Java版本即安装成功



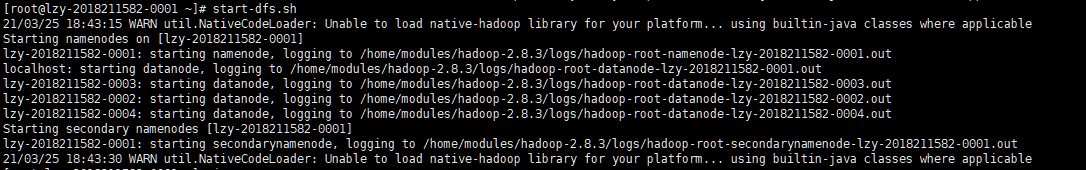


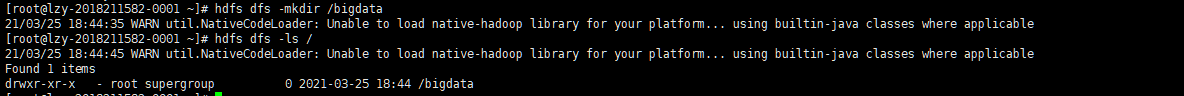




**个人理解：**解压jdk文件，配置环境变量java\_home，执行source命令使其生效后，四个节点执行**java -vesion**后都显示**openjdk** **version “1.8.0\_232”**

5. “2.3.1 搭建Hadoop集群”中的步骤12、13，在这里贴执行启动hdfs与执行hdfs命令的结果，要体现出执行的命令和运行结果

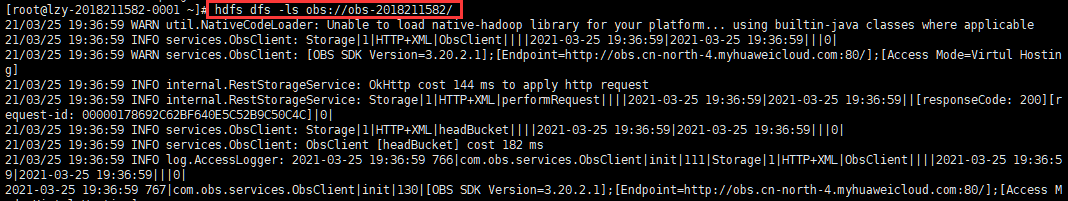




**个人理解：**启动hdfs，node1启动namenode，node2、3、4启动datanode，hdfs启动后，node1执行jps可以看到Jps、SecondaryNameNode、ResourceManager、NameNode，其余三个节点执行jps可以看到NodeManager、Jps、DataNode

执行hdfs命令**hdfs dfs -mkdir/bigdata**，在hdfs上创建bigdata文件夹，执行**hdfs dfs -ls**可以查看到该文件夹已创建成功

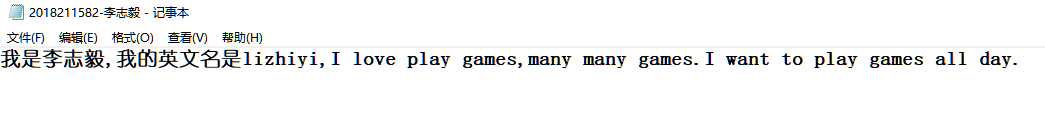
6. “2.3.2 测试与OBS互联”中，要求上传的文件以自己的学号+姓名命名，在这里贴上传成功后执行hdfs命令查看OBS文件的结果

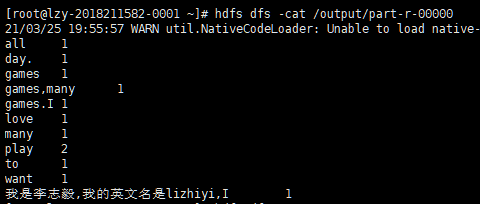




**个人理解：**我的obs桶名为“obs-2018211582”，上传文件名为“2018211582-李志毅.txt”，上传完成执行hdfs命令查看桶内文件可以看到该文件已经上传成功，在OBS中展示。

7. “3.3.1 测试Hadoop集群功能”中，步骤2的测试文件请同学们自定义文件内容，要求包含自己的姓名中英文，且至少有一个单词的数量大于等于2，在这里贴wordcount的结果





**个人理解：**事先编写好txt文档内容，包含我的中文名和文名，以及出现两次的单词**“play”**，执行wordcount命令后可以看到结果如上，其统计出play出现两次，其余单词一次。

8. 请同学们实验后一定按照“4 释放云服务器资源”中的说明完成ECS资源和OBS桶的释放，否则会继续计费





**二、结果分析**

1. hdfs-site.xml中参数dfs.replication的含义是什么？为什么要设置为3？

**解：**hdfs-site.xml中的参数dfs.replication代表备份系数，即缺省的块复制数量，代指DataNode存储block的副本数量，默认是3，此数不能大于集群的机器数，理论上replication值越大跑数速度越快，但是需要的存储空间也更多，默认选择3是因为HDFS采用一种称为机架感知的策略来改进数据的可靠性、可用性和网络带宽的利用率。

在大多数情况下，HDFS的副本系数是3，HDFS的存放策略是一个副本存放在本地机架节点上，另一个副本存放在同一机架的另一个节点上，第三个副本存放在在不同机架的节点上。这种策略减少了机架间的数据传输，提高了写操作的效率。机架错误的概率远比节点错误的概率小，所以这种策略不会对数据的可靠性和可用性造成影响。与此同时，因为数据只存在两个机架上，这种策略减少了读数据时需要的网络传输带宽。在这种策略下，副本并不是均匀地分布在机架上。这种策略在不损坏可靠性和读取性能的情况下，改善了写的性能。