

**大数据处理实验报告**

实验一：HDFS的基本操作

姓 名：刘日星 （交换）

学 号：X2020I1007

院 系：计算机科学与技术

专 业：计科

年 级：交换

指导教师：石宣化老师

2021年 11 月 30 日

一：实验目的

1、了解HDFS的用途

2、掌握HDFS的基本命令

二：实验工具

1、华为云

2、Hadoop

3、HDFS

三：实验环境配置

## 3.1服务购买

### 3.1.1登录控制台

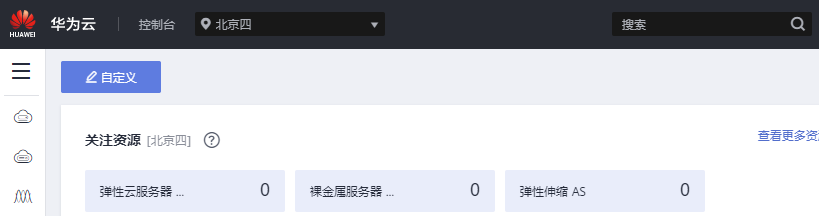
打开华为云官网首页（https://www.huaweicloud.com/），点击“登录”按钮后输入账号信息进行登录。



登录成功后点击“控制台”。



进入控制台后，选择区域为“北京四”。



### 3.1.2 购买MRS服务

在服务列表中点击“EI企业智能”分类下的“MapReduce服务”。



在现有集群界面点击“购买集群”。



选择“自定义购买”，区域选择“华北-北京四”。



集群名称“mrs\_csbd”（可自定义），版本“1.9.2”，类型为“分析集群”，组件默认勾选Hadoop即可。



点击“下一步”进入硬件配置。



计费模式“按需计费”，可用区、虚拟私有云、子网默认，安全组“自动创建”，弹性公网IP“暂不绑定”。

注：若无虚拟私有云，则点击后面的“查看虚拟私有云”进行创建。



CPU架构选择“鲲鹏计算”，集群节点默认。



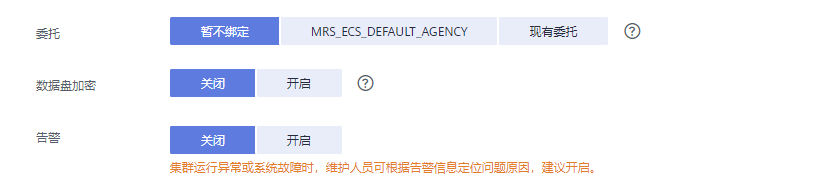
点击“下一步”进入高级配置。



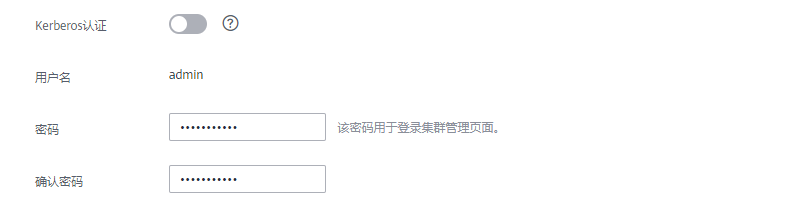
标签、弹性伸缩、引导操作默认。



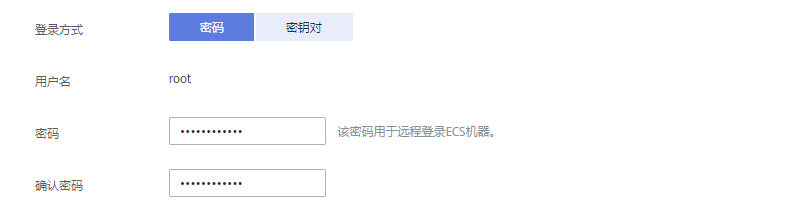
委托、数据盘加密默认，告警“关闭”。



关闭Kerberos认证（“Kerberos认证”关闭时，普通用户可使用MRS集群的所有功能，建议单用户场景下使用），输入密码（该密码用于登录集群管理页面）。



登录方式“密码”，输入密码（该密码用于远程登录集群节点的ECS机器）。



勾选“确认授权”。



点击“立即购买”。



点击“返回集群列表”。



等待集群创建。



集群状态变为“运行中”，创建成功。



### 3.1.3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”。



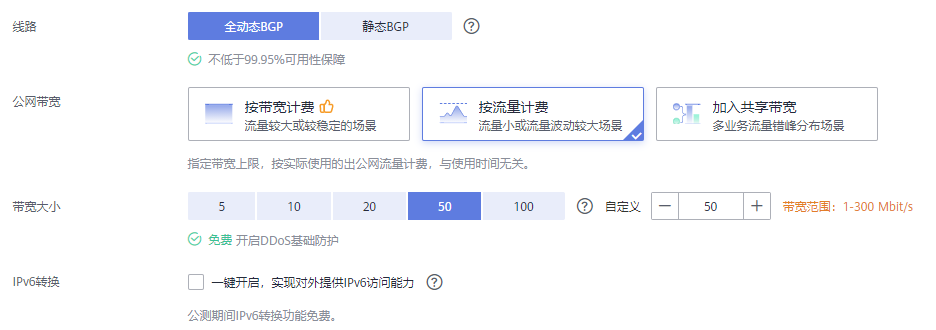
点击“购买弹性公网IP”。



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”。



购买数量“2”，其他默认。



点击“立即购买”，进入确认页面点击“提交”。

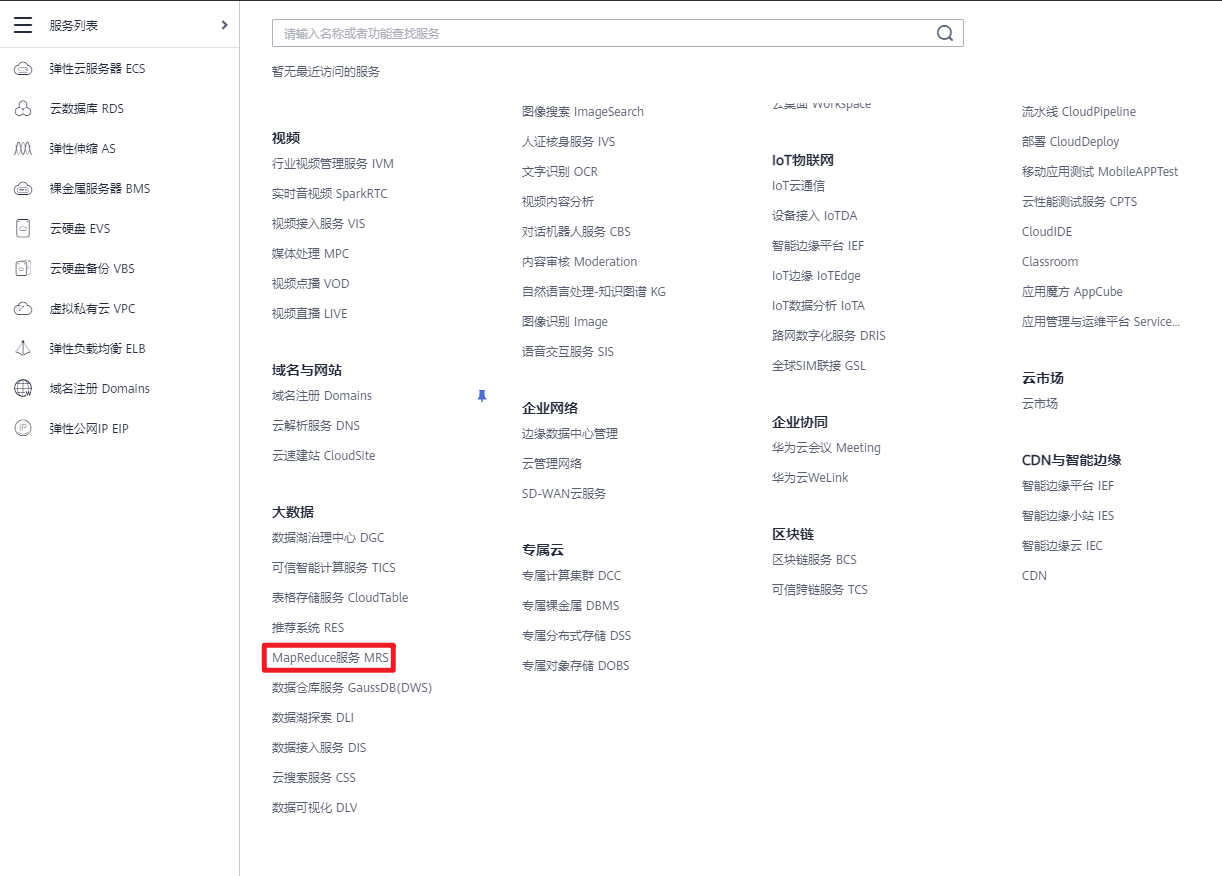


回到弹性公网IP页面后点击刷新按钮，可以看到已经购买的弹性公网IP。



### 3.1.4绑定公网ip

从服务列表切换到MRS。



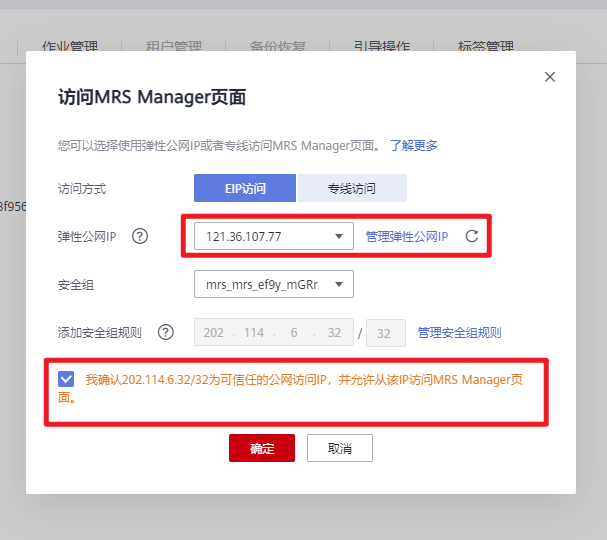
点击新建的集群。



点击前往Manager



在弹出的界面中选择公网ip，并勾选信任，之后点击确定。



在现有集群列表页面中点击MRS集群名称mrs\_batch进入集群页，切换到“节点管理”，展开master节点组并点击其中的节点名称，进入master节点的弹性云服务器控制台。 

点击“弹性公网IP”选项，再点击“绑定弹性公网IP”，会自动跳转到弹性公网IP控制台。



网卡默认，勾选一个IP，点击“确定”。



刷新后可以看到服务器已成功绑定一个弹性公网IP。

### 3.1.5修改安全组

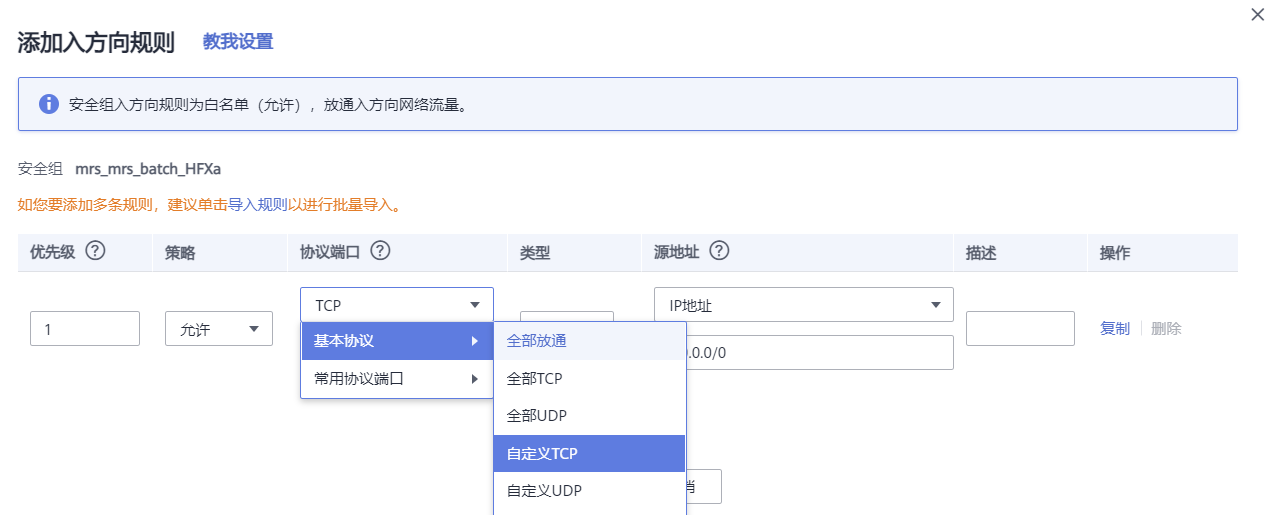
配置好IP后需要修改网络安全组，否则无法登陆到master服务器。点击“安全组”，选择“更改安全组规则”。



选择入方向规则，点击“添加规则”。



优先级“1”，策略“允许”，协议端口点击下拉框选择“基本协议”中的“全部放通”，然后点击确定。



至此，MRS服务配置完成。

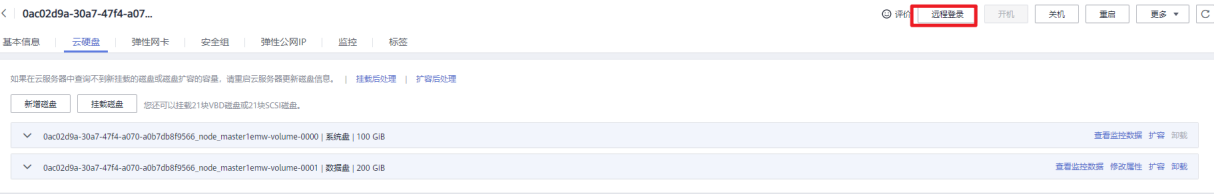
### 3.1.6登录服务器

以下三种方法三选一即可：

1在MRS服务中选择master节点



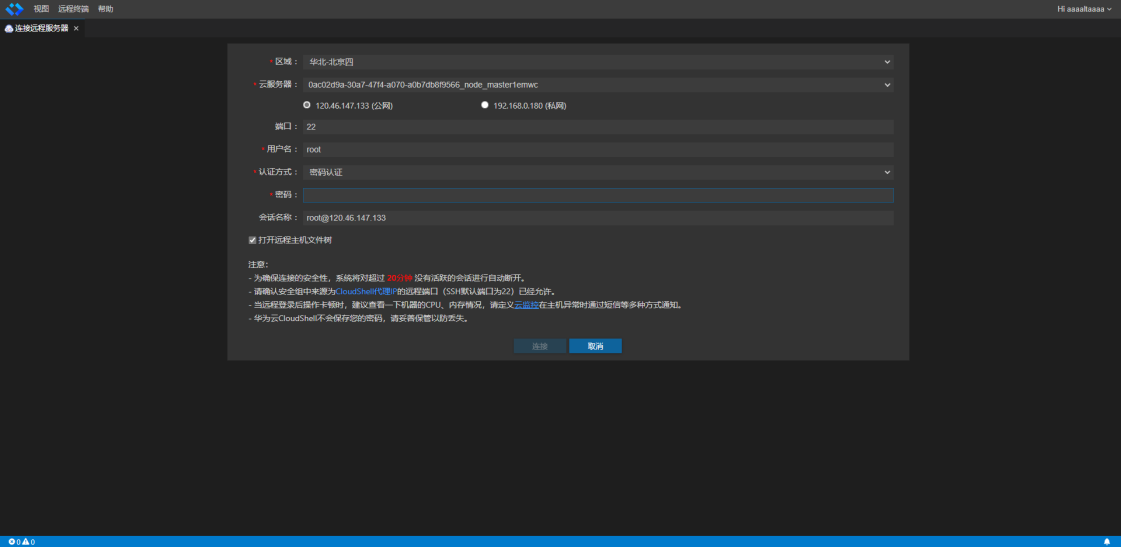
点击右上角远程登陆



点击CloudShell登录



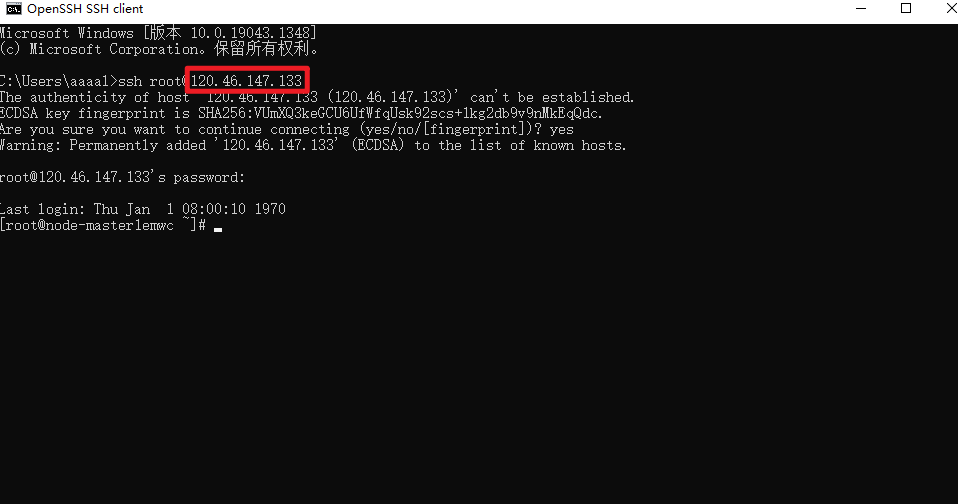
在弹出的界面中输入密码登录



2 通过ssh

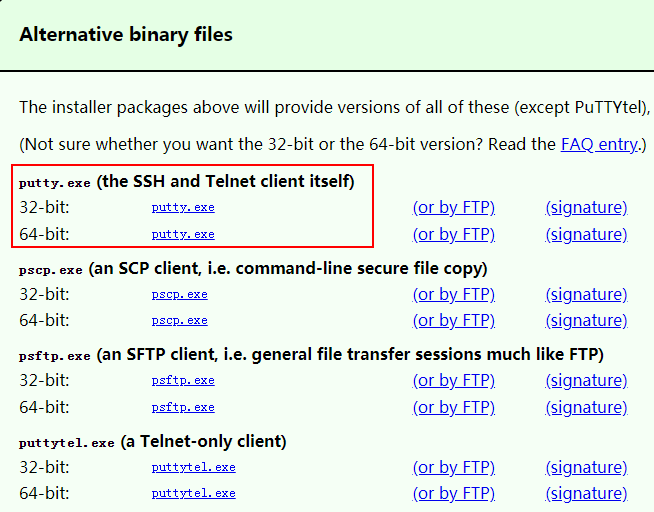
将ip替换为master节点的ip

使用之前配置的密码登录

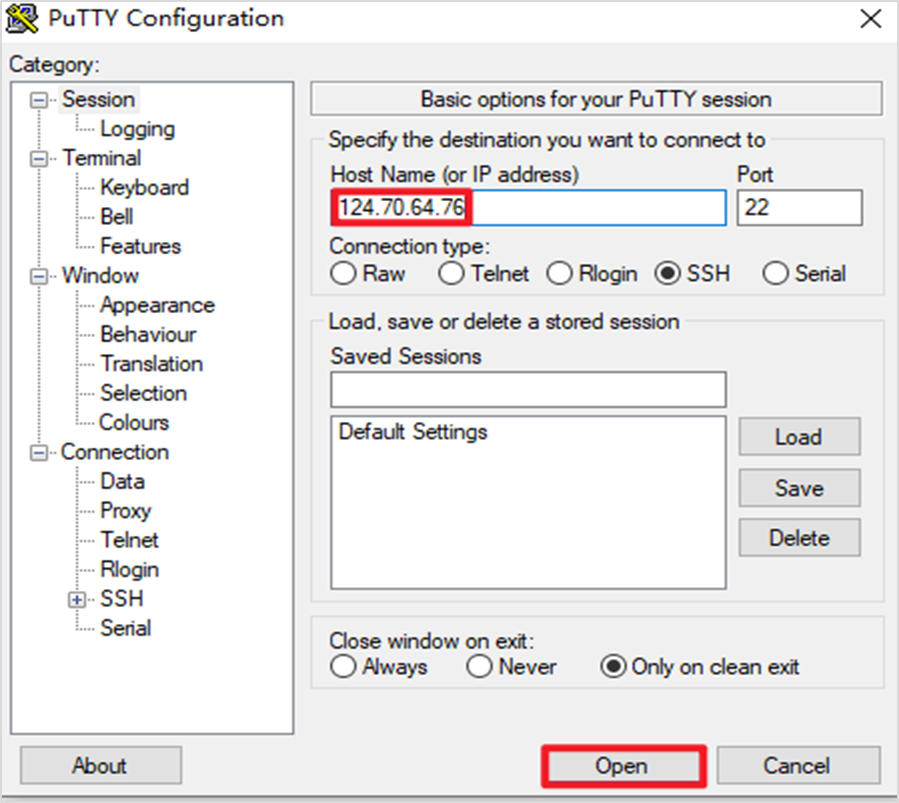


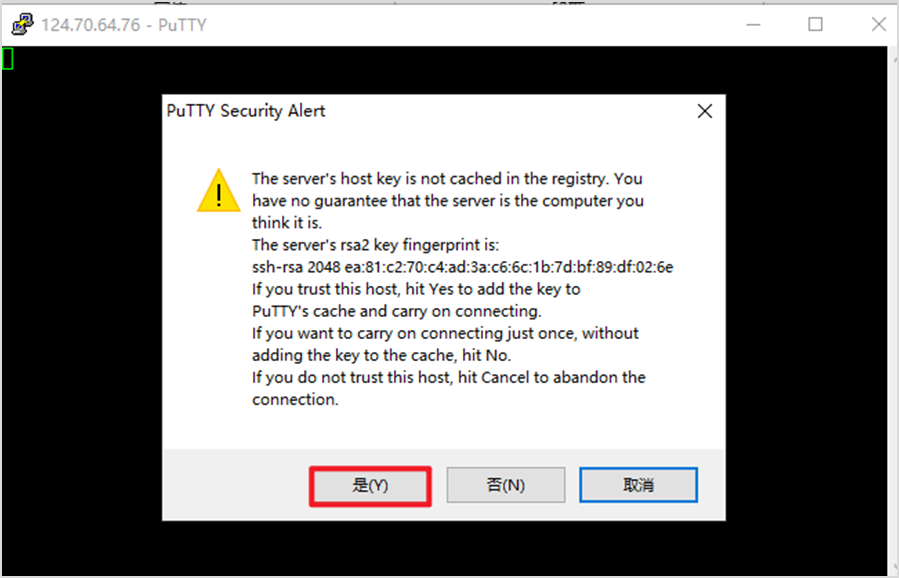
3 通过putty

https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html，选择putty.exe



网址根据实际node节点IP地址进行填写。





### 3.1.7释放资源

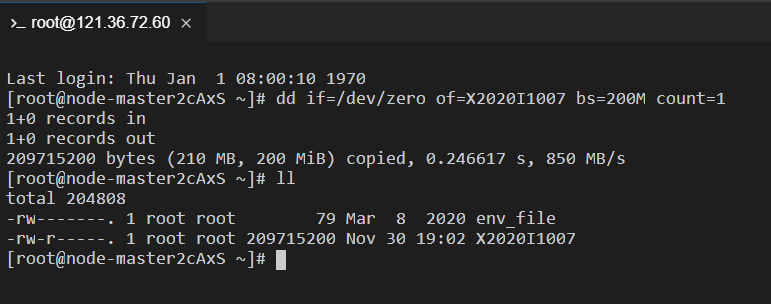
点开之前的MRS和弹性公网IP，分别点击删除按钮或者相关释放按钮，释放相关资源，以免产生过多费用。

四：实验内容及步骤、实验的详细记录、实验结果分析

### 4.1文件准备(20’)（附上实验过程截图以及必要的文字分析）

4.1.1创建文件，文件名X2020I1007是个人学号。(10’)

dd if=/dev/zero of=X2020I1007 bs=200M count=1

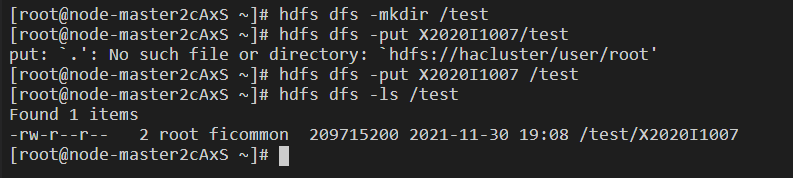


4.1.2在HDFS中创建文件夹，将文件移动到hdfs并显示。(10’)

hdfs dfs -mkdir /test

hdfs dfs -put X2020I1007 /test

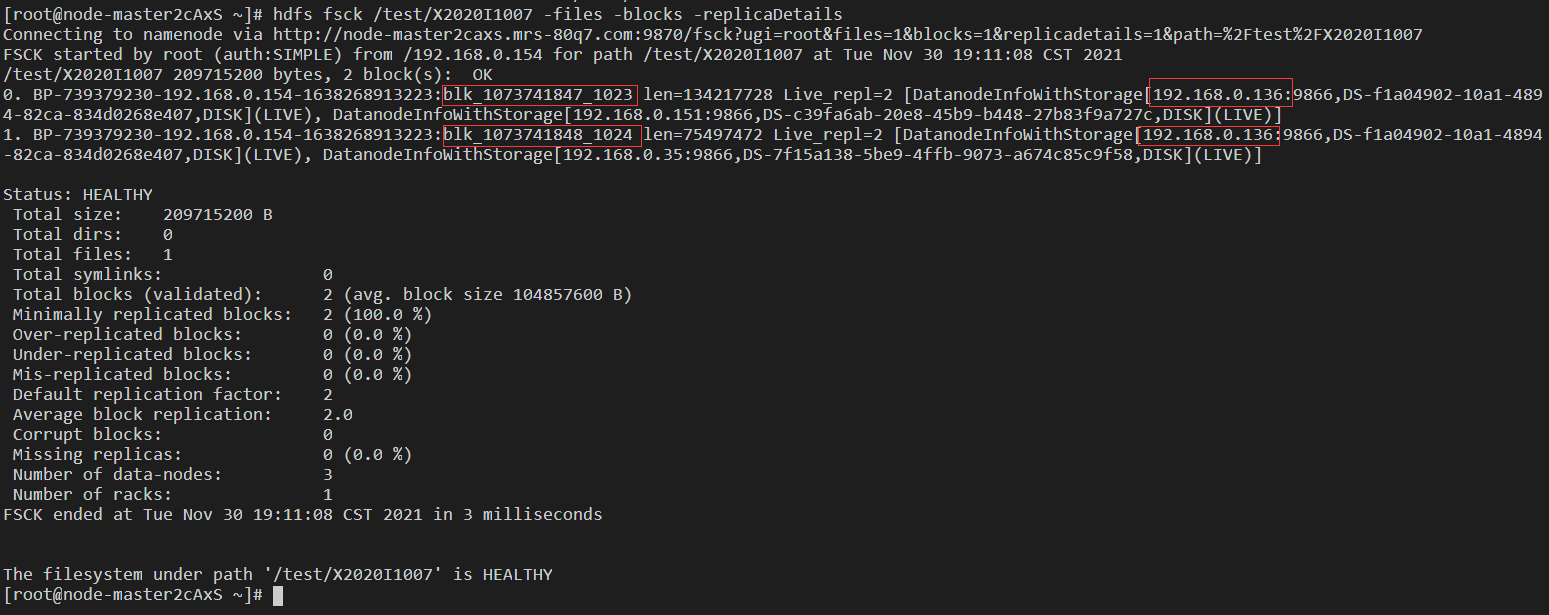
hdfs dfs -ls /test



### 4.2元数据及副本查看(30’)（附上实验过程截图以及必要的文字分析）

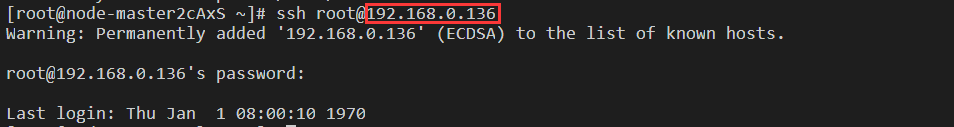
4.2.1查看hdfs文件信息，试解释各字段含义，记录0号块所在的namenode ip和块ID。(10’)

hdfs fsck /test/X2020I1007 -files -blocks -replicaDetails



4.2.2通过ssh进入0号块第一个副本所在的数据节点。(5’)

ssh root@192.168.0.136

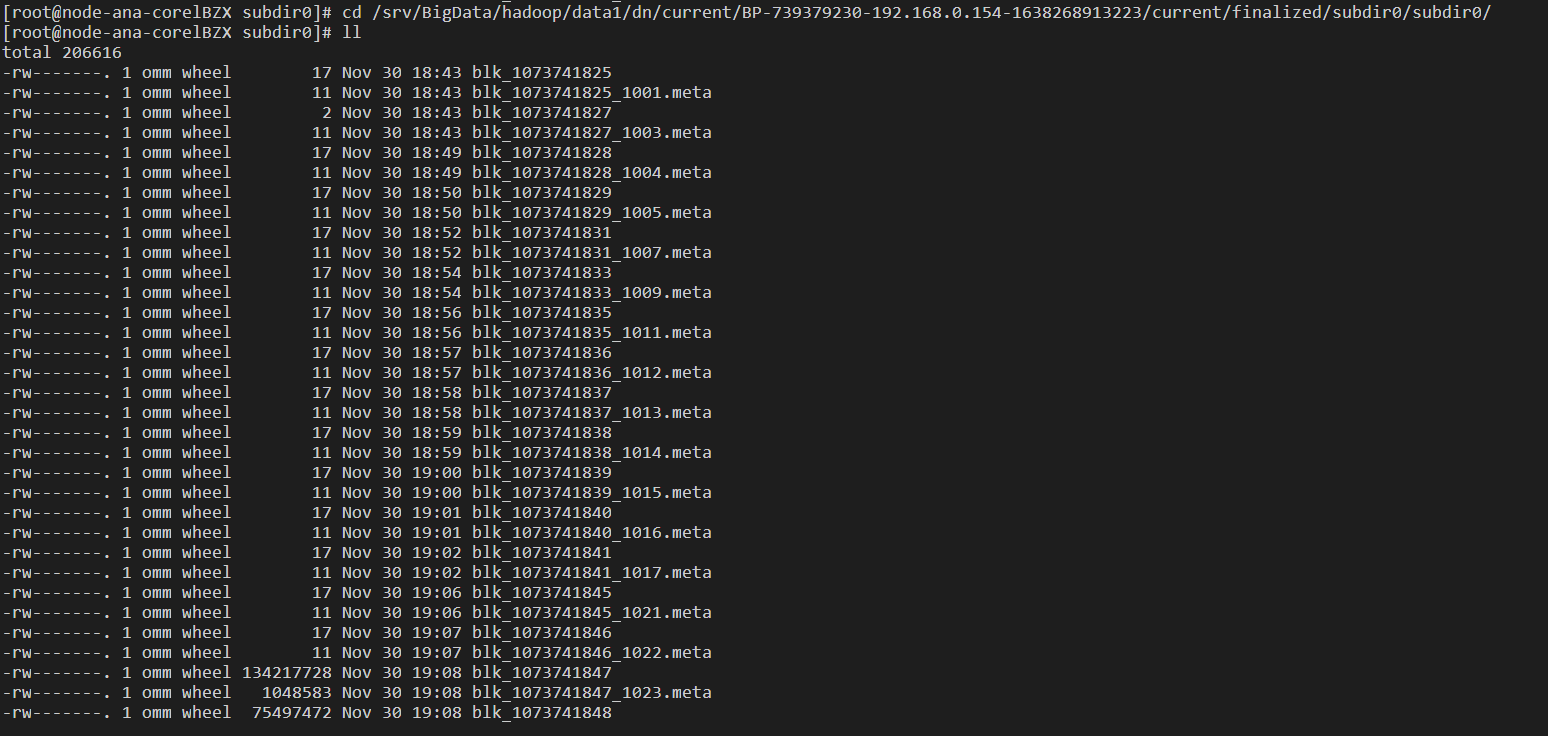


4.2.3查找该块，文件名为块ID，后缀为.meta。(5’)

find /srv -name blk\_1073741847\_1023.meta

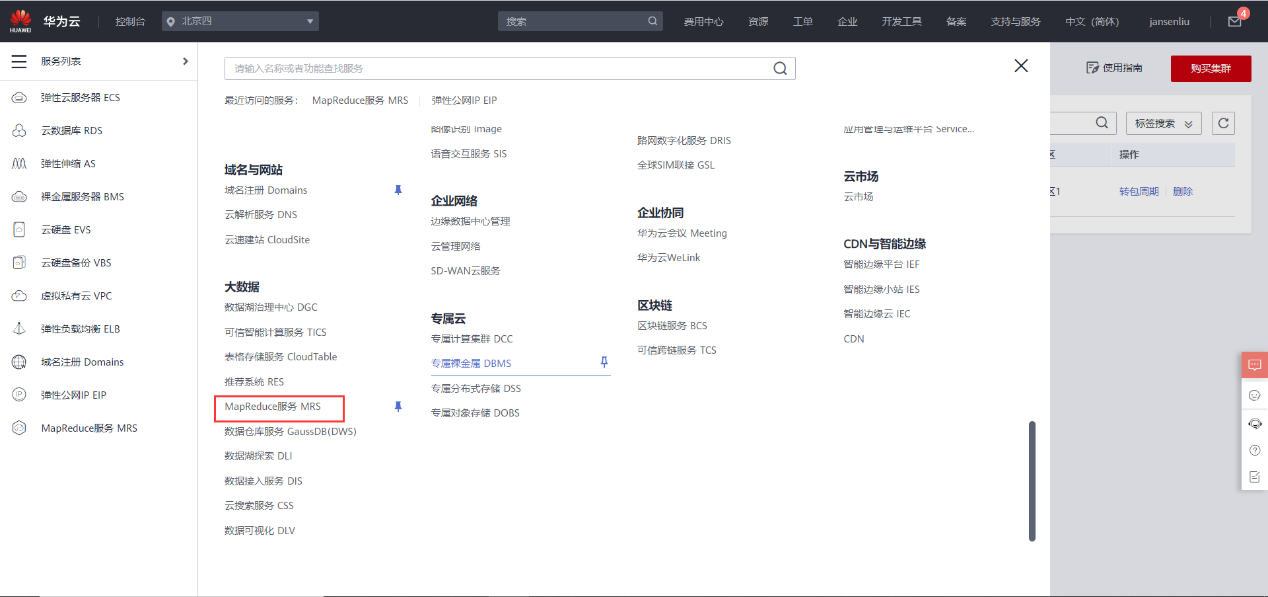


4.2.4进入该文件的上层目录，查看该目录下的的块文件。(10’)

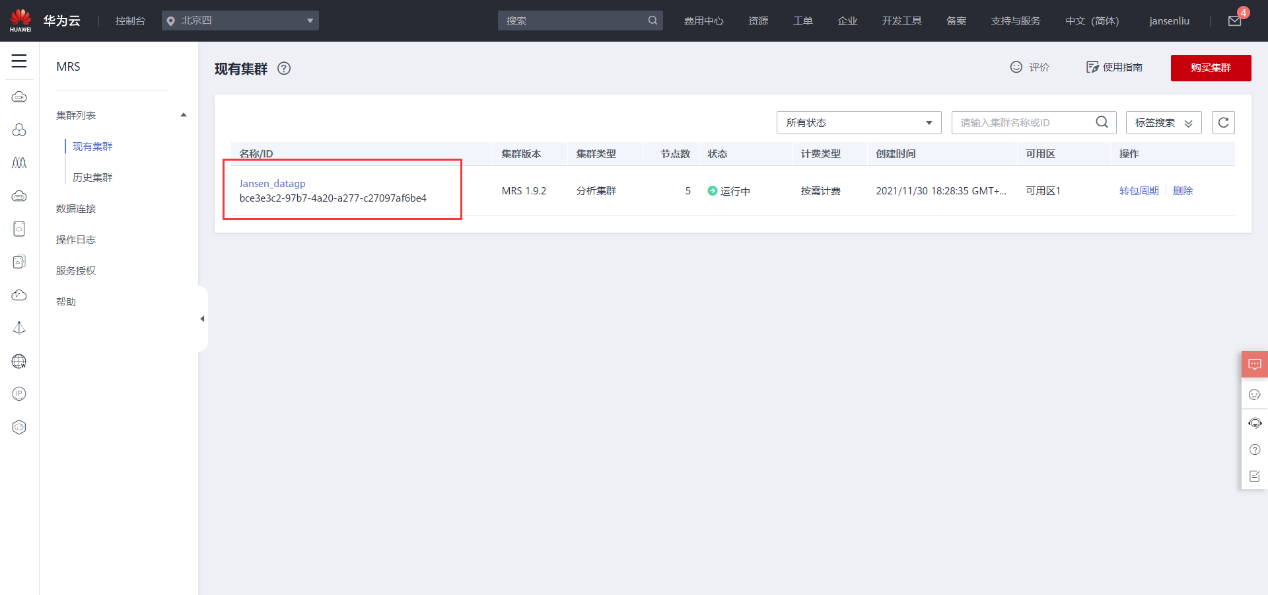


### 4.3 DataNode故障模拟(40’)（附上实验过程截图以及必要的文字分析）

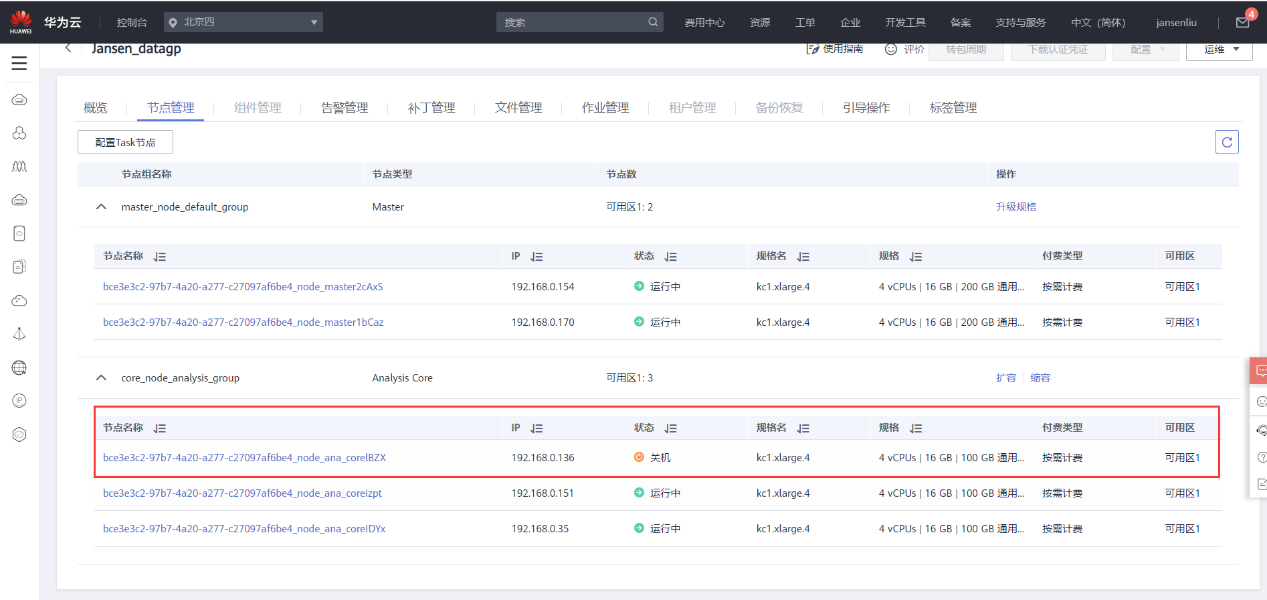
4.3.1 进入华为云MRS。(5’)



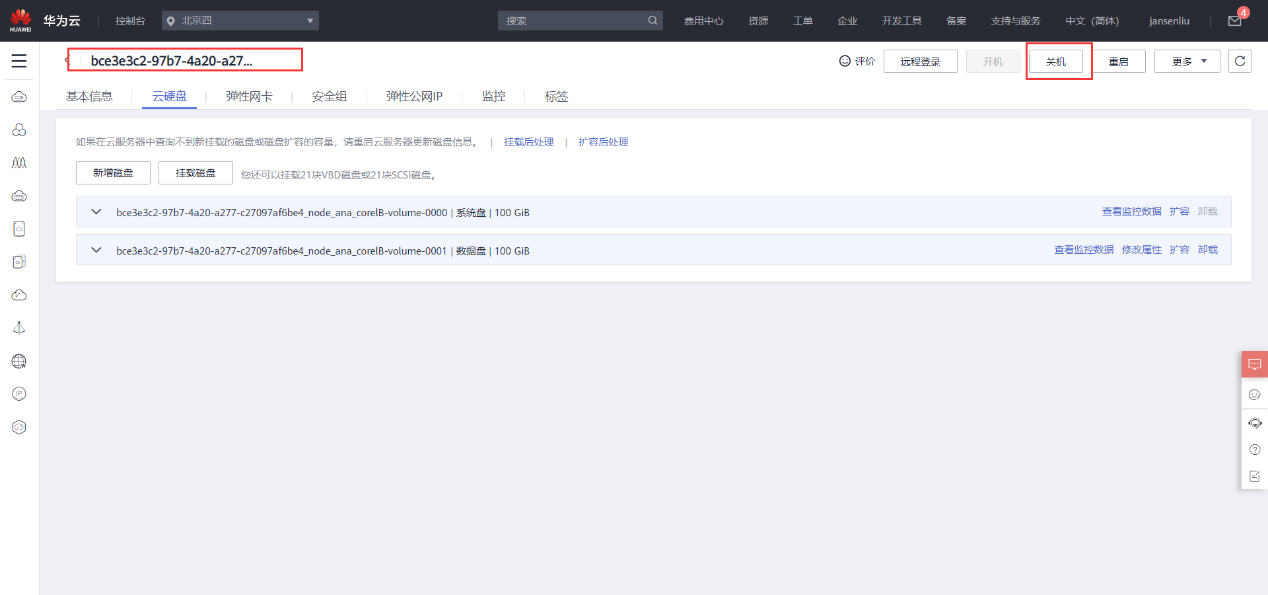
4.3.2 进入集群。(5’)



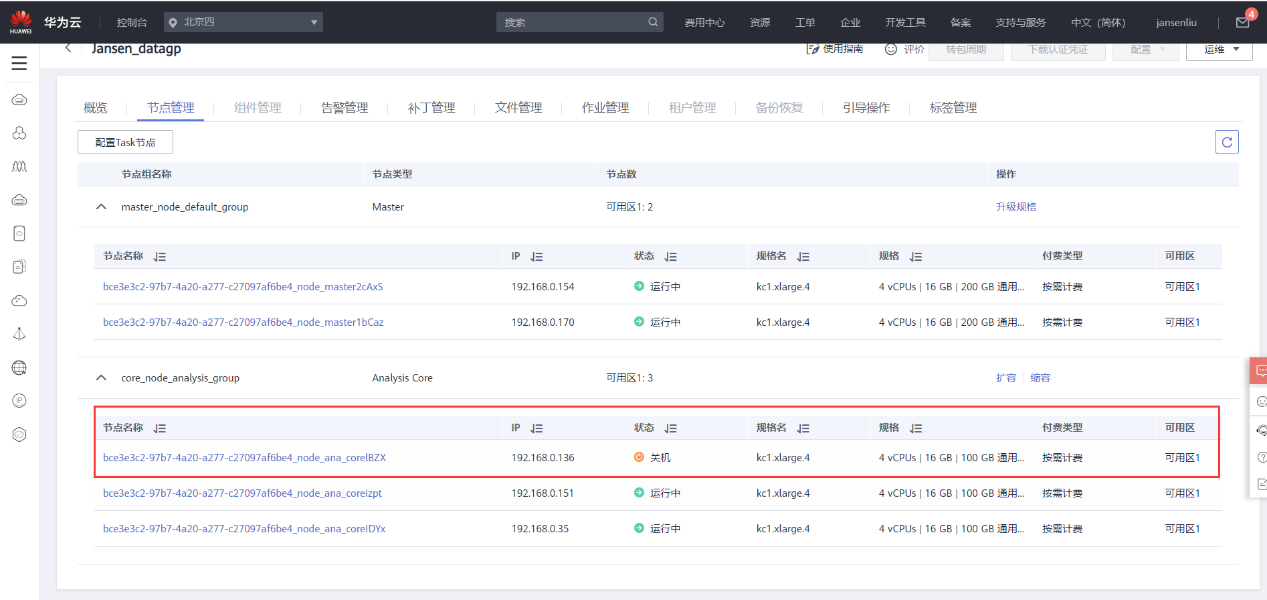
4.3.3在core节点中找到上述文件0号块第一个副本所在的数据节点。(5’)



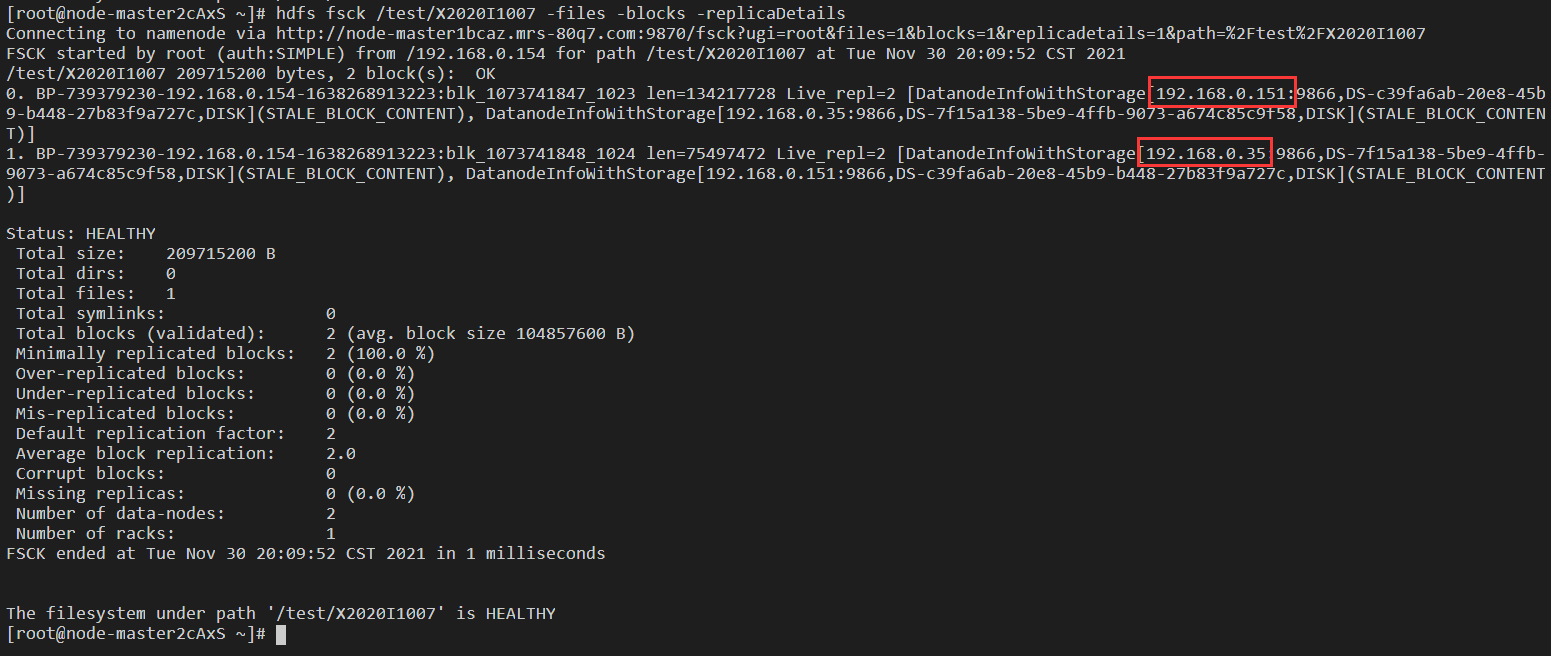
4.3.4 强制关闭该节点模拟数据离线。(10’)







4.3.5 等待一段时间后，连接master节点查看文件详细信息，观察到在另外的节点上生成了新的副本，以保证副本数量不变。(15’)



四：实验总结(10’)

本次实验遵循实验指导手册以及在助教耐心辅导下顺利完成，该实验目的是在华为云上创建MapReduce(MRS)服务集群也就是Hadoop集群，并且借用华为云上的HDFS分布式文件系统进行实现文件的分布式管理。这让我对HDFS分布式文件系统有了更深层次的认知和熟悉。在华为云上创建了集群以后，进入集群里面创建2个“主节点”（NameNode）名称节点和3个“从节点”（DataNode）数据节点。而分布式文件系统的物理结构是由多个主和从节点构成的。这对应的是HDFS的主从（Master/Slave）结构模型，也就是一个HDFS集群包含一个名称节点（NameNode）和若干个数据节点（DataNode）。在本次实验过程中，未免会出现一些小问题，其中有些是因为个人的急于求成的心态导致漏了一些步骤的截图。例如在关闭DataNode数据节点使节点模拟数据离线时忘记截图了，导致需要重新恢复节点运行并重复关闭节点操作。

通过本次实验，加深了我对HDFS基本命令的使用，让我对HDFS命令在linux系统上操作得更加得心应手。这次大数据处理实验把我带入了分布式文件系统操作的大门，我会在下一次实验中更加细心地检查每一步操作去完成每一部分的实验内容。

**附录：**

1 HDFS基本命令

<https://blog.csdn.net/WQY992/article/details/89002269>

<https://blog.csdn.net/aohuang8877/article/details/101116099>