華中科技大學

大数据处理实验报告

实验一：HDFS 的基本操作

姓　　名:

学　　号:

院　　系:

专　　业:

年　　级:

指导教师:

2021 年　月　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”。

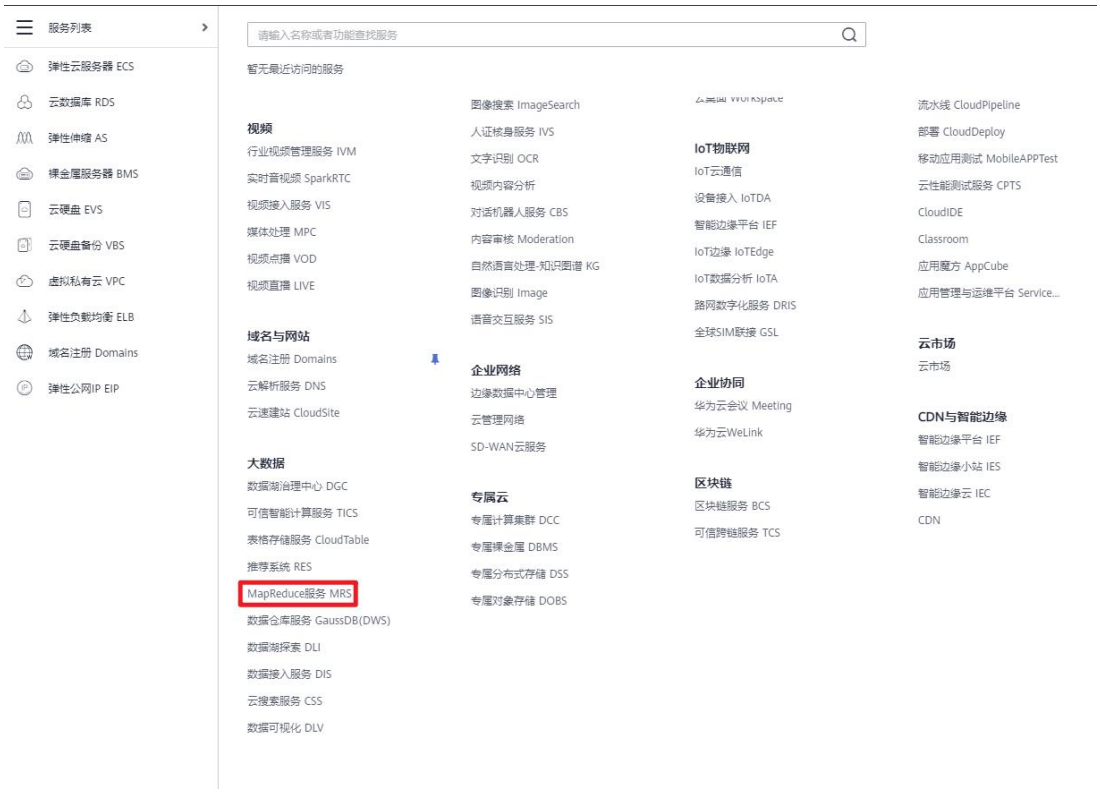
购买数量“2”，其他默认。

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”。

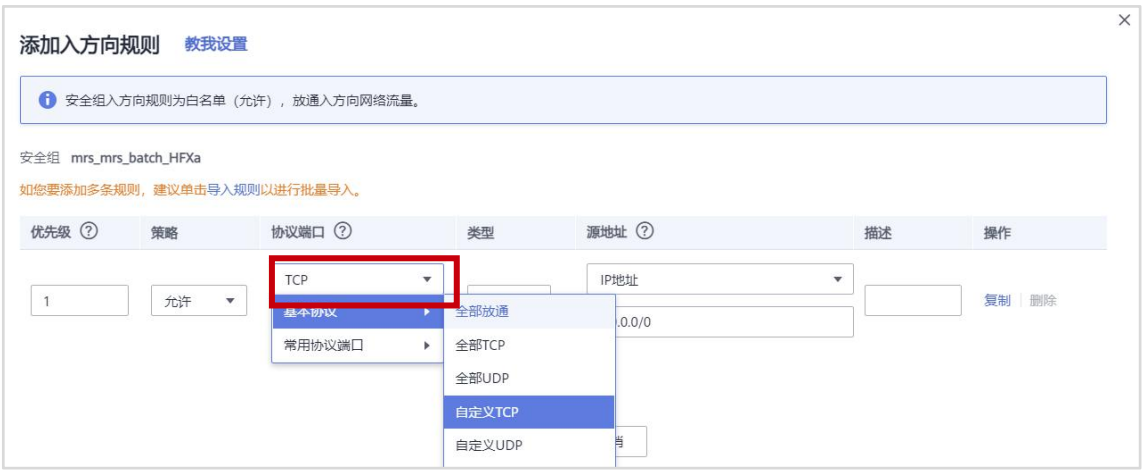
刷新后可以看到服务器已成功绑定一个弹性公网 IP。

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

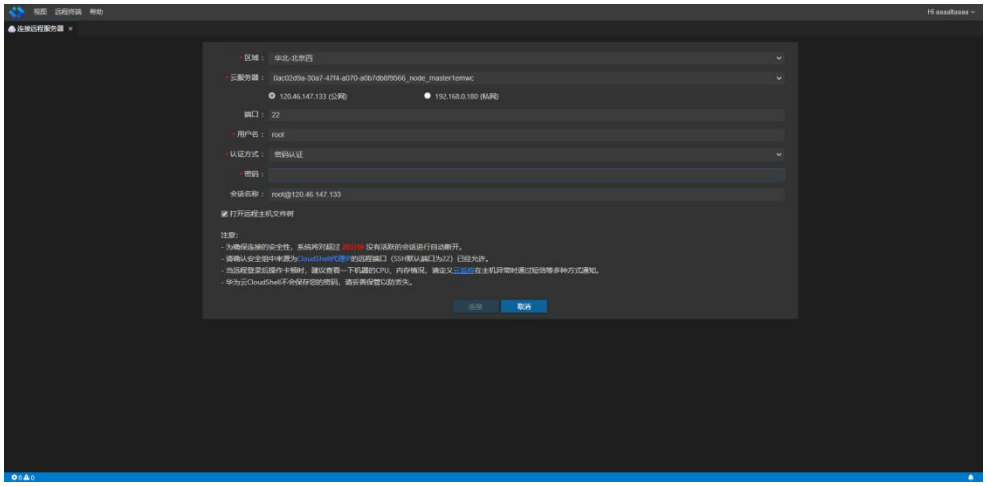
以下三种方法三选一即可：

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

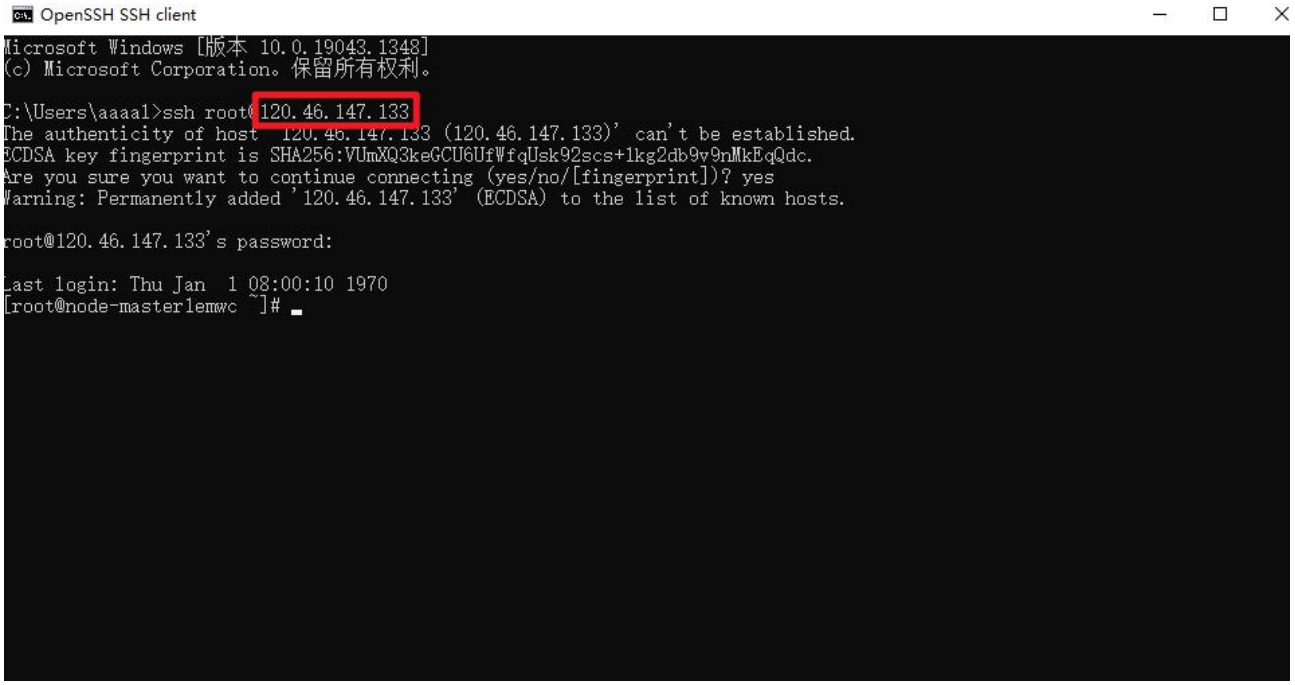
在弹出的界面中输入密码登录



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

Alternative binary files

The installer packages above will provide versions of all of these (except PuTTYtel),

(Not sure whether you want the 32-bit or the 64-bit version? Read the FAQ entry.)

putty.exe (the SSH and Telnet client itself)

32-bit:	putty.exe	(or by FTP)	(signature)
64-bit:	putty.exe	(or by FTP)	(signature)

pscp.exe (an SCP client, i.e. command-line secure file copy)

32-bit:	pscp.exe	(or by FTP)	(signature)
64-bit:	pscp.exe	(or by FTP)	(signature)

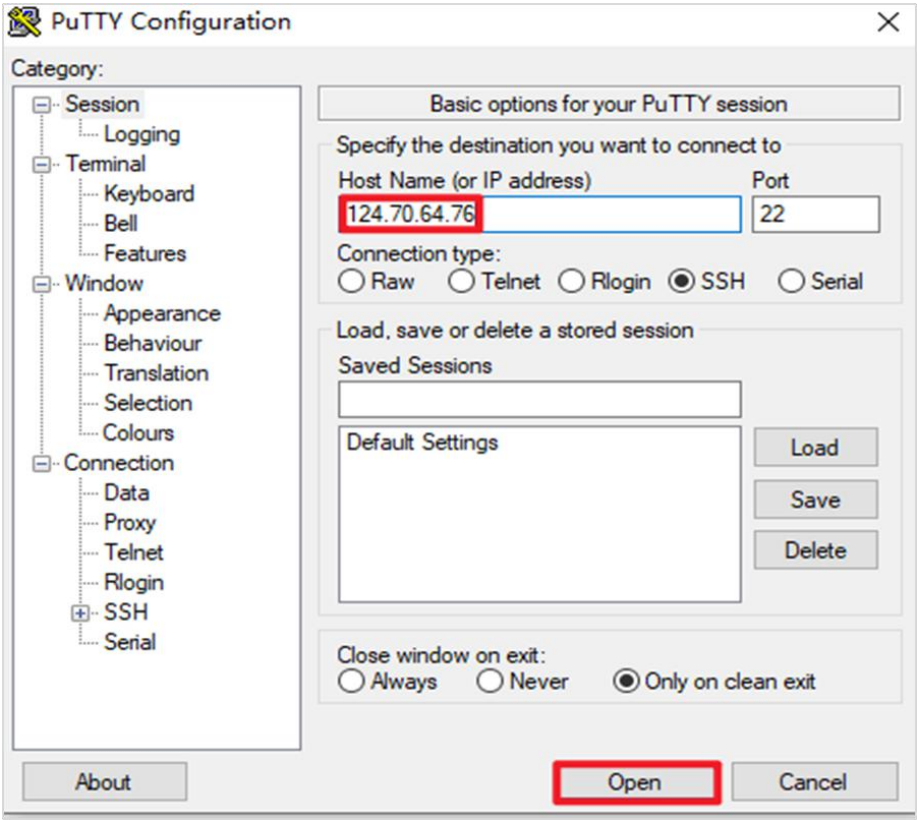
psftp.exe (an SFTP client, i.e. general file transfer sessions much like FTP)

32-bit:	psftp.exe	(or by FTP)	(signature)
64-bit:	psftp.exe	(or by FTP)	(signature)

puttytel.exe (a Telnet-only client)

32-bit:	puttytel.exe	(or by FTP)	(signature)
64-bit:	puttytel.exe	(or by FTP)	(signature)

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 filename 是个人学号。(10')

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

```
[root@node-master1emwc ~]# dd if=/dev/zero of=filename bs=200M count=1
1+0 records in
1+0 records out
209715200 bytes (210 MB, 200 MiB) copied, 0.45018 s, 466 MB/s
[root@node-master1emwc ~]# ll
total 204808
-rw-------. 1 root root        79 Mar  8  2020 env_file
-rw-r-----. 1 root root 209715200 Nov 29 23:03 filename
```

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 filename /test

hdfs dfs -ls /test

```
[root@node-master1emwc ~]# hdfs dfs -mkdir /test
[root@node-master1emwc ~]# hdfs dfs -put filename /test
[root@node-master1emwc ~]# hdfs dfs -ls /test
Found 1 items
-rw-r--r--   2 root ficommon  209715200 2021-11-29 23:08 /test/filename
```

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/filename -files -blocks -replicaDetails

```
[root@node-master1emwc ~]# hdfs fsck /test/filename -files -blocks -replicaDetails
Connecting to namenode via http://node-master1emwc.mrs-9cyd.com:9870/fsck?ugi=root&files=1&blocks=1&replicadetails=1&path=%2Ftest%2Ffilename
FSCK started by root (auth:SIMPLE) from /192.168.0.180 for path /test/filename at Mon Nov 29 23:12:05 CST 2021
/test/filename 209715200 bytes, 2 block(s):  OK
0. BP-217980482-192.168.0.180-1638180801861:blk_1073742471_1647 len=134217728 Live_repl=2 [DatanodeInfoWithStorage[192.168.0.140:9866,DS-2667d3f1-2c87-4fa0-baad-1a08e5d9a82a,DISK](LIVE), DatanodeInfoWithStorage[192.168.0.242:9866,DS-5e6f569d-598c-4df7-add1-764ac7d46603,DISK](LIVE)]
1. BP-217980482-192.168.0.180-1638180801861:blk_1073742472_1648 len=75497472 Live_repl=2 [DatanodeInfoWithStorage[192.168.0.140:9866,DS-2667d3f1-2c87-4fa0-baad-1a08e5d9a82a,DISK](LIVE), DatanodeInfoWithStorage[192.168.0.242:9866,DS-5e6f569d-598c-4df7-add1-764ac7d46603,DISK](LIVE)]

Status: HEALTHY
 Total size:    209715200 B
 Total dirs:    0
 Total files:   1
 Total symlinks:                0
 Total blocks (validated):      2 (avg. block size 104857600 B)
 Minimally replicated blocks:   2 (100.0 %)
 Over-replicated blocks:        0 (0.0 %)
 Under-replicated blocks:       0 (0.0 %)
 Mis-replicated blocks:         0 (0.0 %)
 Default replication factor:    2
 Average block replication:     2.0
 Corrupt blocks:                0
 Missing replicas:              0 (0.0 %)
 Number of data-nodes:          3
 Number of racks:               1
FSCK ended at Mon Nov 29 23:12:05 CST 2021 in 1 milliseconds
```

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@xx.xxx.xx.xx

```
[root@node-master1emwc ~]# ssh root@192.168.0.140
Warning: Permanently added '192.168.0.140' (ECDSA) to the list of known hosts

root@192.168.0.140's password:

Last login: Mon Nov 29 23:19:07 2021 from 192.168.0.180
```

4.2.3 查找该块，文件名为块 ID，后缀为.meta。(5’)

find /srv -name blk_xxxxxxxxxx_xxxx.meta

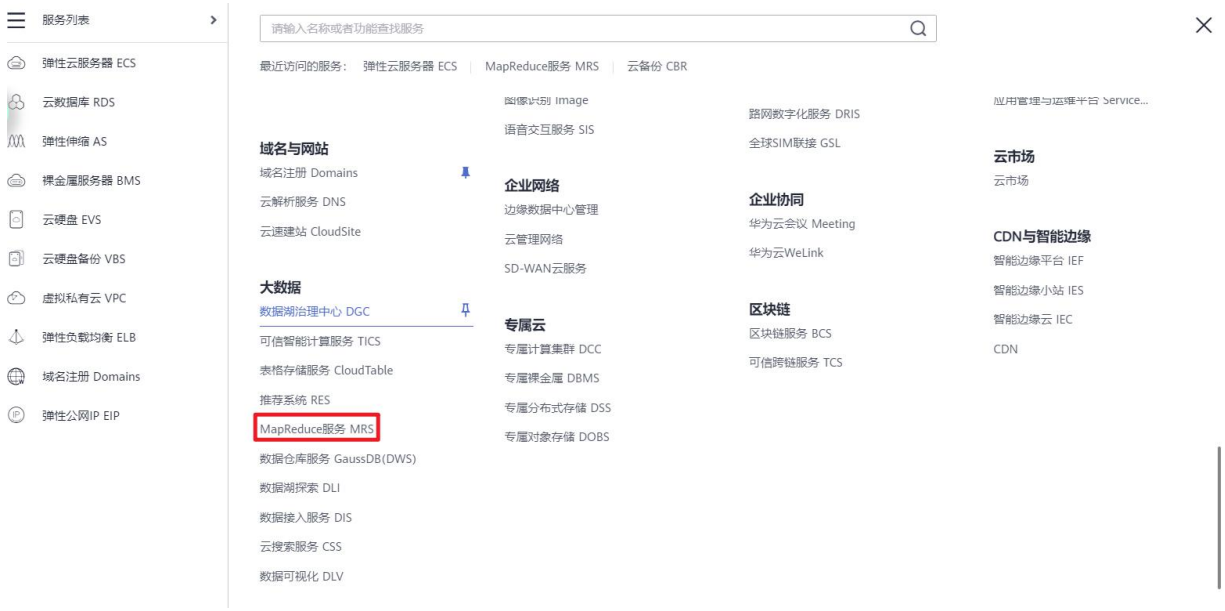
```
[root@node-ana-corekbKN ~]# find /srv -name blk_1073742471_1647.meta
/srv/BigData/hadoop/data1/dn/current/BP-217980482-192.168.0.180-1638180801861/current/finalized/subdir0/subdir2/blk_1073742471_1647.meta
```

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’)

```
[root@node-ana-corekbKN ~]# cd /srv/BigData/hadoop/data1/dn/current/BP-217980482-192.168.0.180-1638180801861/current/finalized/subdir0/subdir2/
[root@node-ana-corekbKN subdir2]# ll
total 338516
-rw--------. 1 omm wheel 134217728 Nov 29 22:10 blk_1073742401
-rw--------. 1 omm wheel   1048583 Nov 29 22:10 blk_1073742401_1577.meta
-rw--------. 1 omm wheel 134217728 Nov 29 23:08 blk_1073742471
-rw--------. 1 omm wheel   1048583 Nov 29 23:08 blk_1073742471_1647.meta
-rw--------. 1 omm wheel  75497472 Nov 29 23:08 blk_1073742472
-rw--------. 1 omm wheel    589831 Nov 29 23:08 blk_1073742472_1648.meta
-rw--------. 1 omm wheel        14 Nov 29 23:25 blk_1073742492
-rw--------. 1 omm wheel        11 Nov 29 23:25 blk_1073742492_1668.meta
```

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’)

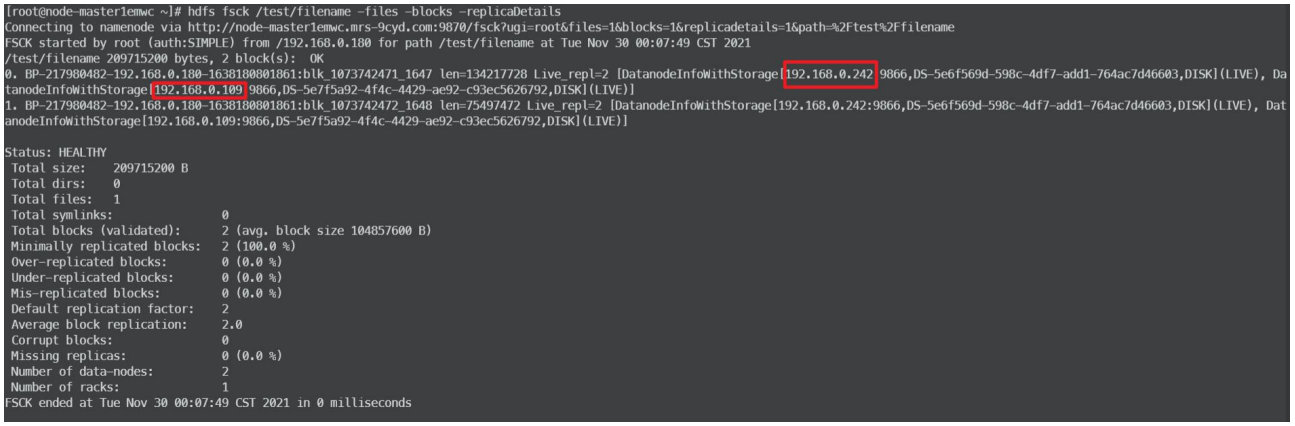


4.3.2 进入集群。(5’)

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’)

附录：

1 HDFS 基本命令

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

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