

# 华中科技大学

## 大数据处理实验

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

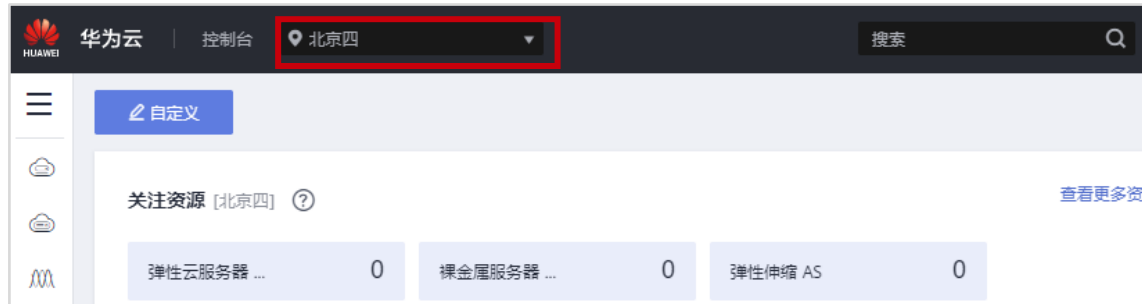
打开华为云官网首页 (<https://www.huaweicloud.com/>)，点击“登录”按钮后输入账号信息进行登录。



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。

| 组件名  | 版本    | 描述                                       |
|--|-------|--|
| <input type="checkbox"/> Presto            | 0.216 | 一种开源、分布式SQL查询引擎。                         |
| <input checked="" type="checkbox"/> Hadoop | 2.8.3 | 针对大数据集的分布式数据存储和处理框架，包含HDFS、YARN、MapRe... |

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。

① 软件配置

② 硬件配置

③ 高级配置

计费模式

包年/包月

按需计费

可用区

可用区1

可用区2

可用区3

可用区7

?

虚拟私有云

vpc-default

C

?

查看虚拟私有云

子网

subnet-default(192.168.0.0...

C

安全组

自动创建

C

?

管理安全组

弹性公网IP

暂不绑定

C

?

管理弹性公网IP

CPU 架构选择“鲲鹏计算”，集群节点默认。

CPU架构

x86计算

鲲鹏计算

集群节点

| 节点类型       | 计费模式 | 实例规格   | 实例数量   |
|------------|------|--|--|
| Master节点 ? | 按需计费 | 鲲鹏通用计算增强型<br>4 vCPUs   16 GB  <br>kc1.xlarge.4<br>系统盘<br>高IO 100 GB x 1<br>数据盘<br>高IO 200 GB x 1 | 2<br>集群高可用 <input checked="" type="checkbox"/> |
| 分析Core节点 ? | 按需计费 | 鲲鹏通用计算增强型<br>4 vCPUs   16 GB  <br>kc1.xlarge.4<br>系统盘<br>高IO 100 GB x 1<br>数据盘<br>高IO 100 GB x 1 | <div>- 3 +</div>                               |
| 分析Task节点 ? | 按需计费 |  |  |

点击“下一步”进入高级配置。

按需节点费用 ¥5.418/小时  
参考价格，具体扣费请以账单为准。 [了解计费详情](#)

上一步

下一步

标签、弹性伸缩、引导操作默认。

1 软件配置
2 硬件配置
3 高级配置

标签

标签键

标签值

你还可以添加10个标签。

弹性伸缩

请先返回上一步设置Task节点规格后再设置弹性伸缩策略。

引导操作

| 名称              | 执行节点 | 执行时机 | 操作 |
|-----------------|------|------|----|
| <div>添加 ?</div> |      |      |    |

引导操作添加的脚本个数不能超过18个。

委托、数据盘加密默认，告警“关闭”。

委托

暂不绑定

MRS\_ECS\_DEFAULT\_AGENCY

现有委托 ?

数据盘加密

关闭

开启 ?

告警

关闭

开启

集群运行异常或系统故障时，维护人员可根据告警信息定位问题原因，建议开启。

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）。

Kerberos认证

?

用户名

admin

密码

.....

该密码用于登录集群管理页面。

确认密码

.....

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）。

登录方式

密码

密钥对

用户名

root

密码

.....

该密码用于远程登录ECS机器。

确认密码

.....

勾选“确认授权”。

通信安全授权 ☒ 确认授权

授权MRS集群开通相应的安全组规则，从而使得用户可以通过MRS管理控制台进行大数据组件部署和后续集群的使用、运维和管理等操作，此时不授权将无法创建集群，以下为需要开通的安全组规则。 [了解更多](#)

| 协议端口       | 类型   | 源地址               | 描述          |
|------------|------|-------------------|-------------|
| TCP : 9022 | IPv4 | 100.125.1.47/32   | MRS 默认安全组规则 |
| TCP : 9022 | IPv4 | 100.125.132.83/32 | MRS 默认安全组规则 |

点击“立即购买”。

按需节点费用 **¥5.418/小时**  
参考价格，具体扣费请以账单为准。 [了解计费详情](#)

上一步 **立即购买**

点击“返回集群列表”。

< 购买集群



您的mrs\_csbdb已开始创建。

**进入集群** 返回集群列表

等待集群创建。

现有集群   使用指南 **购买集群**

全部 请输入集群名称或ID 标签搜索 清除

| 名称/ID                                    | 集群版本      | 集群...  | 节点数 | 状态            | 计费类型        | 可用区  | 操作        |
|--|-----------|--------|-----|---------------|-------------|------|-----------|
| mrs_csbdb<br>e6b9b073-7d9d-494a-bb2a-... | MRS 1.9.2 | 分析集... | 5   | 启动中<br>创建虚拟机中 | <b>按需计费</b> | 可用区1 | 转包周期   删除 |

集群状态变为“运行中”，创建成功。

| 名称/ID                                    | 集群版本      | 集群类型 | 节点数 | 状态  | 计费类型        | 可用区  | 操作        |
|--|-----------|------|-----|-----|-------------|------|-----------|
| mrs_csbdb<br>e6b9b073-7d9d-494a-bb2a-... | MRS 1.9.2 | 分析集群 | 5   | 运行中 | <b>按需计费</b> | 可用区1 | 转包周期   删除 |

### 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。



弹性公网IP名称

批量购买弹性公网IP时，无法指定名称。

高级配置

带宽名称 | 标签

监控

☒ 默认开启基础监控

免费

☒ 免费提供分钟级粒度的流量监控

☒ 监控带宽流量波动、出入网带宽速率等指标详情

购买量

-

2

+

一次最多可以购买20个弹性公网IP。您还可以购买2个弹性公网IP，如需申请更多配额请点击[申请扩大配额](#)。

点击“立即购买”，进入确认页面点击“提交”。

弹性公网IP费用 **¥0.02/小时** + 公网流量费用 **¥0.80/GB**

立即购买

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。

弹性公网IP 

使用指南

购买弹性公网IP

诚邀您参加弹性公网IP使用体验调研，您宝贵的意见和建议是我们持续提升产品体验的动力，感谢您的参与！  
反馈即有机会获得2000码豆，码豆奖励规则：1. 码豆奖励将于问卷提交后的7个工作日内充值到账；2. 码豆到账情况、使用规则及兑奖详情请到会员中心查看。

刷新

停止带宽

设置

更多

所有状态

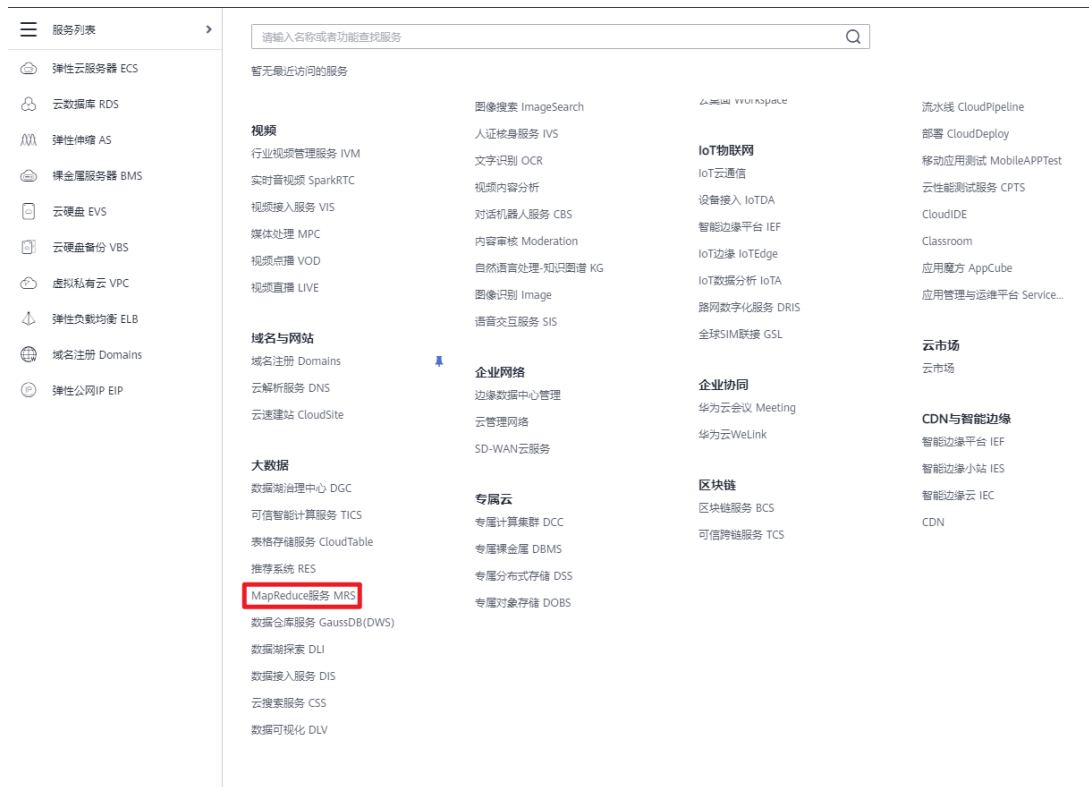
弹性公网IP

标签搜索

| <input type="checkbox"/> 弹性公网IP         | 监控 | 状态  | 类型     | 带宽                        | 带宽详情               | 已绑定实例 | 计费模式                         | 操作   |
|---|----|-----|--------|---------------------------|--------------------|-------|------------------------------|--|
| <input type="checkbox"/> 120.46.147.133 |    | 未绑定 | 金动芯BGP | Bandwidth_2021-11-28-17.. | 按流量计费<br>50 Mbit/s | ---   | 按周<br>2021/11/28 17:54:04 创建 | <a href="#">绑定</a> <a href="#">解绑</a> <a href="#">更多</a> |
| <input type="checkbox"/> 121.36.107.77  |    | 未绑定 | 金动芯BGP | Bandwidth_2021-11-28-17.. | 按流量计费<br>50 Mbit/s | ---   | 按周<br>2021/11/28 17:54:02 创建 | <a href="#">绑定</a> <a href="#">解绑</a> <a href="#">更多</a> |

### 3.1.4 绑定公网 ip

从服务列表切换到 MRS。



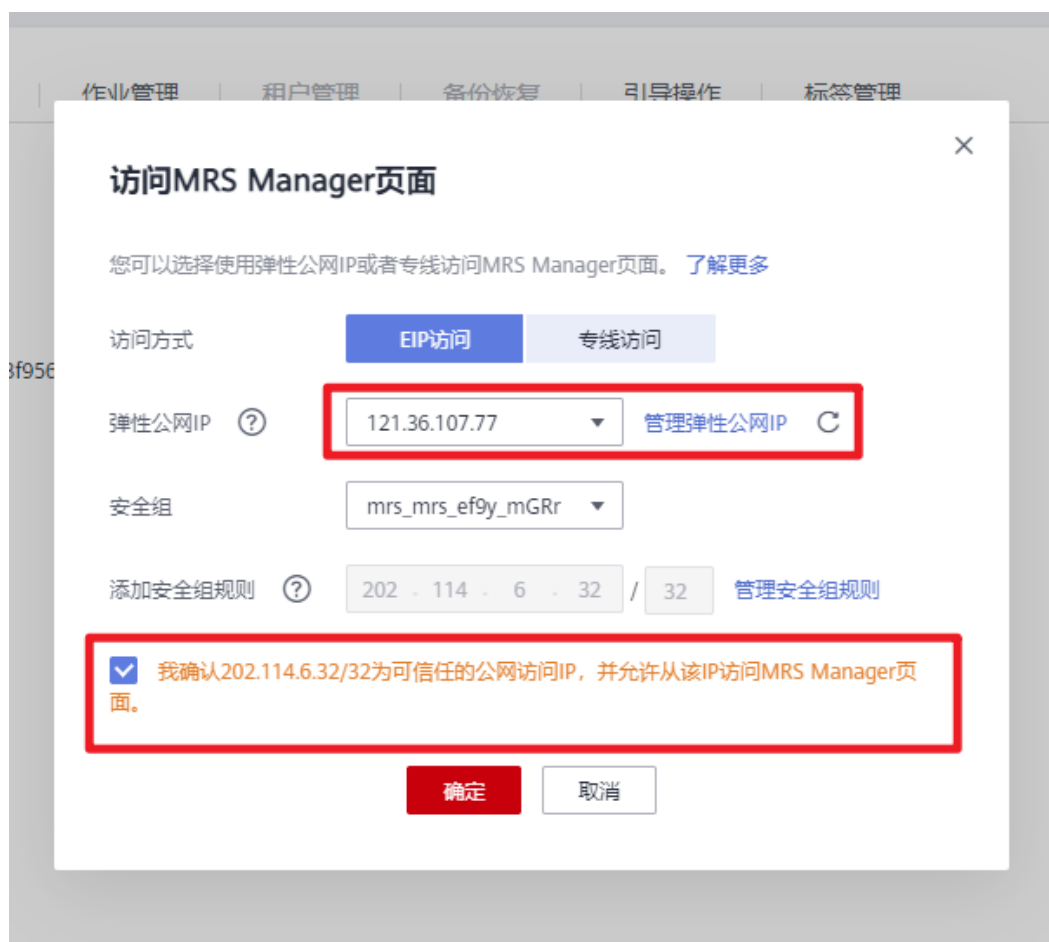
点击新建的集群。



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。



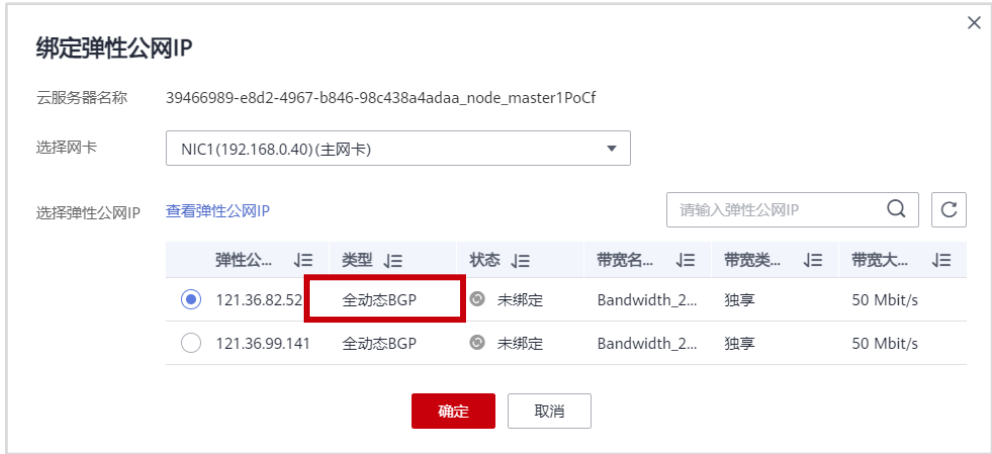
在现有集群列表页面中点击 MRS 集群名称 mrs\_batch 进入集群页,切换到“节点管理”,  
展开 master 节点组并点击其中的节点名称,进入 master 节点的弹性云服务器控制台。



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

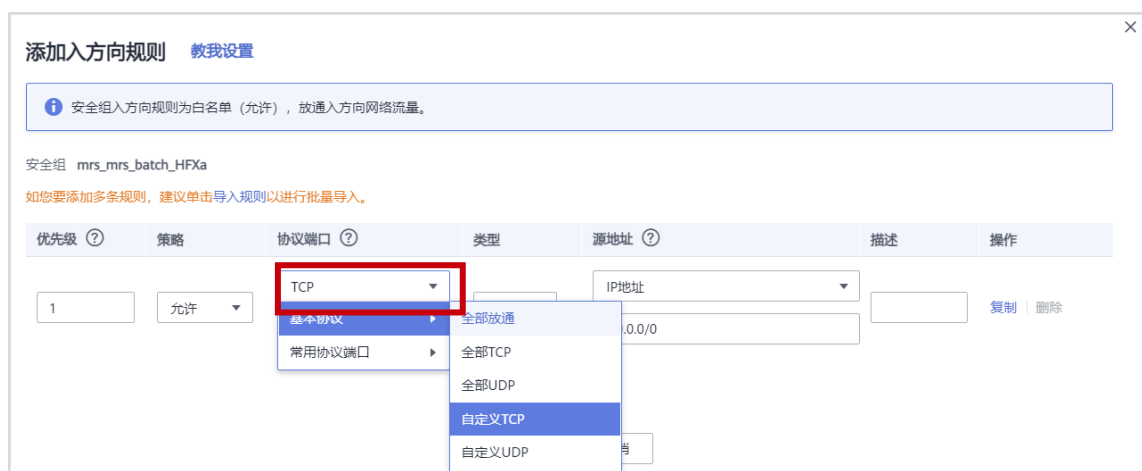
配置好 IP 后需要修改网络安全组，否则无法登录到 master 服务器。点击“安全组”，选择“更改安全组规则”。



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



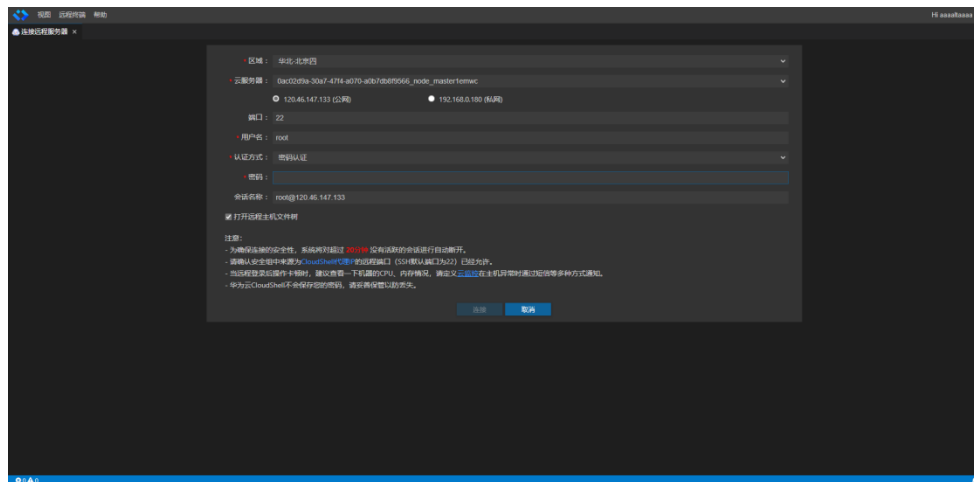
点击右上角远程登陆



## 点击 CloudShell 登录



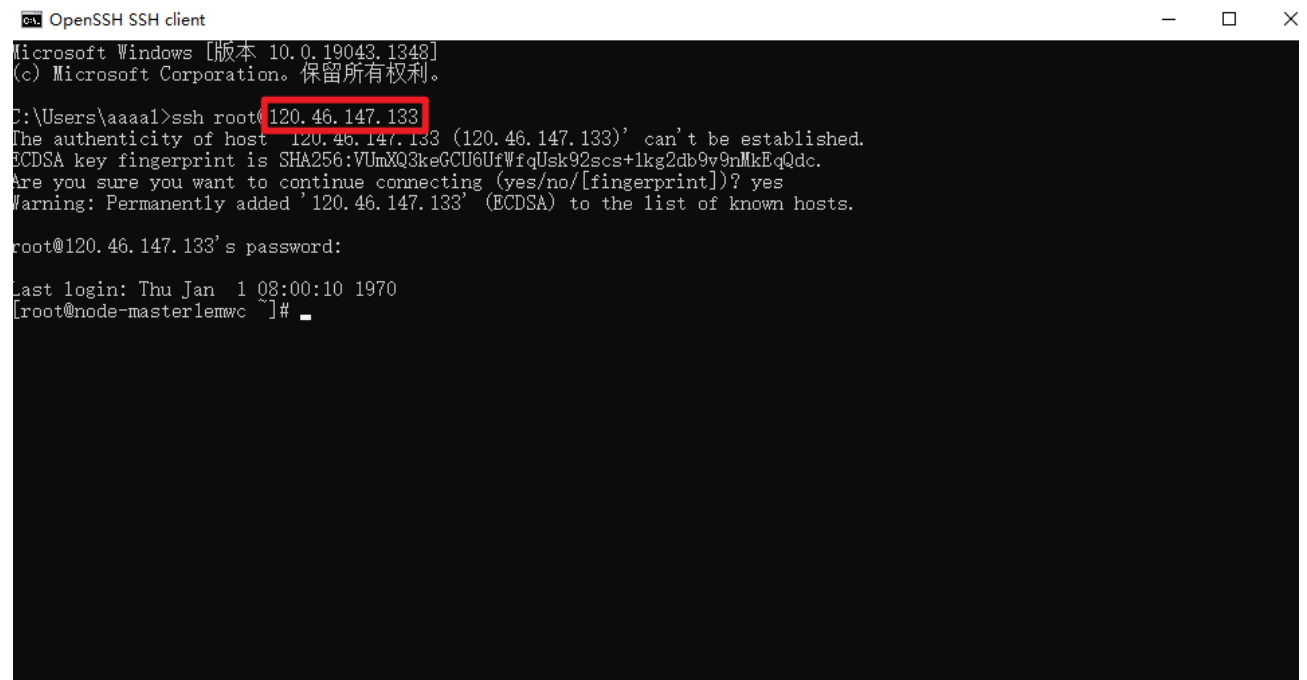
## 在弹出的界面中输入密码登录



## 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: | <a href="#">putty.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |
| 64-bit: | <a href="#">putty.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |

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

|         |                          |                             |                             |
|---------|--------------------------|-----------------------------|-----------------------------|
| 32-bit: | <a href="#">pscp.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |
| 64-bit: | <a href="#">pscp.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |

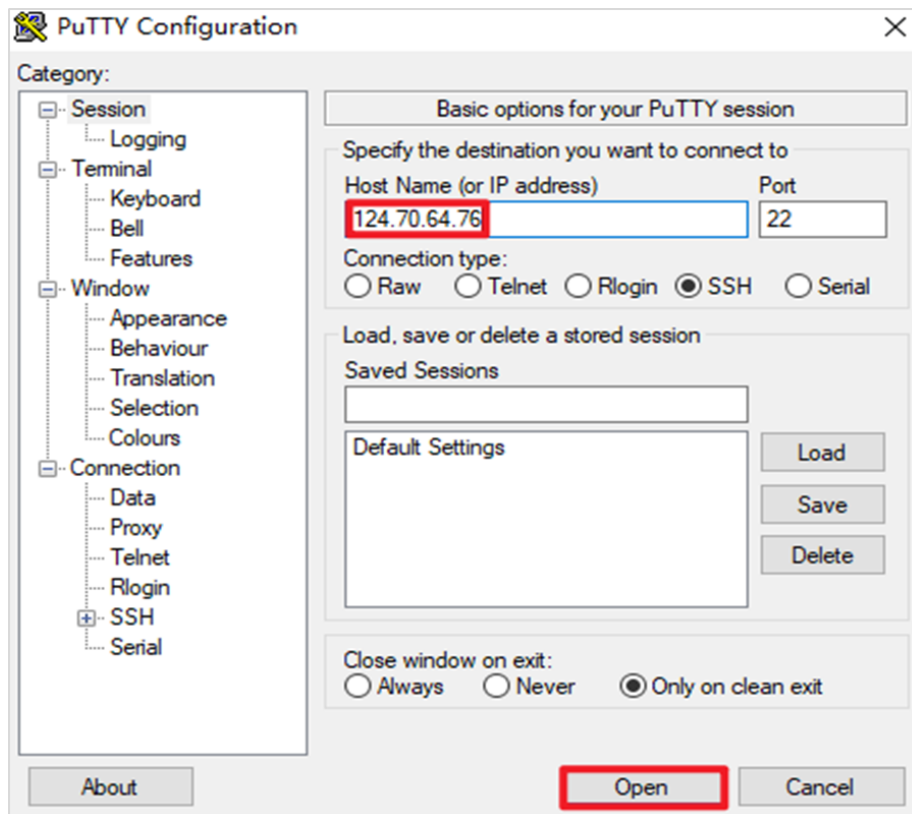
### psftp.exe (an SFTP client, i.e. general file transfer sessions much like FTP)

|         |                           |                             |                             |
|---------|---------------------------|-----------------------------|-----------------------------|
| 32-bit: | <a href="#">psftp.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |
| 64-bit: | <a href="#">psftp.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |

### puttytel.exe (a Telnet-only client)

|         |                              |                             |                             |
|---------|------------------------------|-----------------------------|-----------------------------|
| 32-bit: | <a href="#">puttytel.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |
| 64-bit: | <a href="#">puttytel.exe</a> | <a href="#">(or by FTP)</a> | <a href="#">(signature)</a> |

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=16blocks=16replicadetails=16path=/test%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

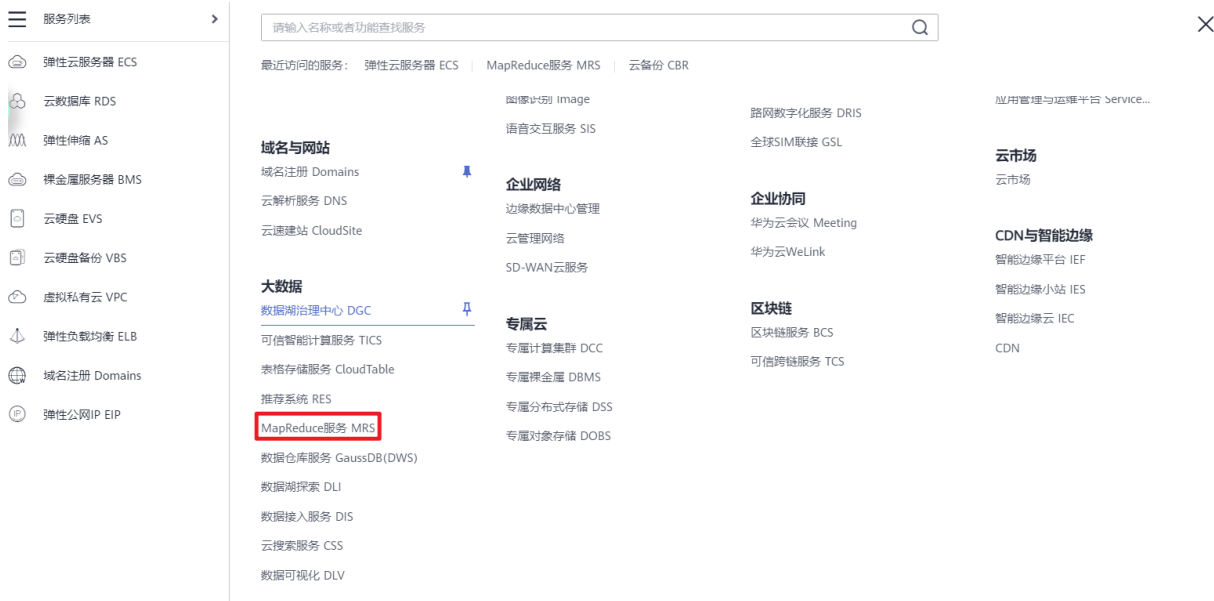
```
[root@node-ana-corekbkN ~]# find /srv -name blk_1073742471_1647.meta
/srv/BigData/hadoop/data1/dn/current/BP-217980482-192.168.0.180-1638180801861/current/finalized/subdir0/subdir2/blk_1073742471_1647.meta
```

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-r--r-- 1 omm wheel 134217728 Nov 29 22:10 blk_1073742401
-rw-r--r-- 1 omm wheel 1048583 Nov 29 22:10 blk_1073742401_1577.meta
-rw-r--r-- 1 omm wheel 134217728 Nov 29 23:08 blk_1073742471
-rw-r--r-- 1 omm wheel 1048583 Nov 29 23:08 blk_1073742471_1647.meta
-rw-r--r-- 1 omm wheel 75497472 Nov 29 23:08 blk_1073742472
-rw-r--r-- 1 omm wheel 589831 Nov 29 23:08 blk_1073742472_1648.meta
-rw-r--r-- 1 omm wheel 14 Nov 29 23:25 blk_1073742492
-rw-r--r-- 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')



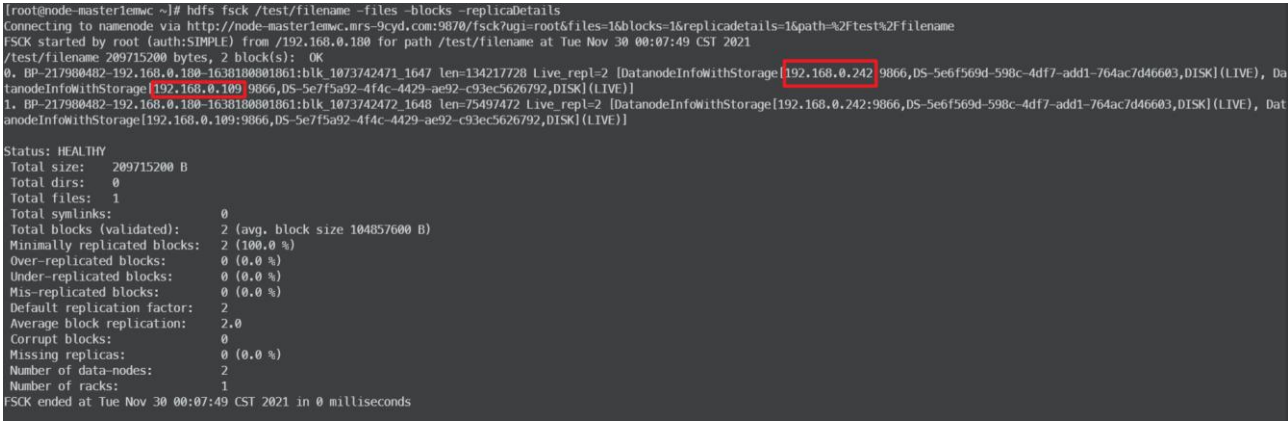
### 4.3.3 在 core 节点中找到上述文件 0 号块第一个副本所在的数据节点。(5')



### 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')



## 附录：

### 1 HDFS 基本命令

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

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