实　　验　　报　　告

学　　号	1802003 2020	姓　　名	王子轩	专业班级	计算机科学与技术 2 班
课程名称	大数据导论			学期	**2021 年秋季学期**
任课教师	刘艳艳, 刘洁	完成日期	2021/10/2	上机课时间	周五 78 节
实 验 名 称	实验一 Hadoop 实验环境配置				

一、实验要求（10%）

1. VMware 软件的安装
2. 安装 Linux 系统
3. 配置模板虚拟机
4. 克隆虚拟机模板
5. 在 DataNode1 上安装 JDK
6. 在 DataNode1 安装 Hadoop
7. 完全分布式运行模式，SSH 无密登录配置，集群配置

二、实验内容及步骤（80%）

1. VMware 软件的安装

因本人电脑已经安装过 VMware 软件，故此步骤跳过。

2. 安装 Linux 系统

说明文档中有安装虚拟机的博客链接，里边内容很具体，因此在安装此步骤没有太大问题，就是在安装图形界面时遇到了问题，具体见下文中的心得总结。界面如下：

3. 配置模板虚拟机

修改 NameNode 的 ip 和主机名以及 Windows 中虚拟网卡中的 ip，教程给的步骤十分具体，没有遇到问题，按照教程即可完成。结果如下图：

```
[hadoop@NameNode ~]$ hostname
NameNode
[hadoop@NameNode ~]$ ifconfig
ens33: flags=4163<UP,BROADCAST,RUNNI
        inet 192.168.10.100  netmask
```

修改 ifcfg-ens33 文件，效果如下图：

```
hadoop@NameN
文件(F)  编辑(E)  查看(V)  搜索(S)  终端(T)
TYPE="Ethernet"
PROXY_METHOD="none"
BROWSER_ONLY="no"
BOOTPROTO="static"
DEFROUTE="yes"
IPV4_FAILURE_FATAL="no"
IPV6INIT="yes"
IPV6_AUTOCONF="yes"
IPV6_DEFROUTE="yes"
IPV6_FAILURE_FATAL="no"
IPV6_ADDR_GEN_MODE="stable-privacy"
NAME="ens33"
UUID="d4025ca1-fdb6-4d0f-aeb3-1b43ffce
DEVICE="ens33"
ONBOOT="yes"

IPADDR=192.168.10.100
GATEWAY=192.168.10.2
DNS1=192.168.10.2
```

安装相关工具包，如 epel-release、net-tool 等，在安装图形界面前就安装了，不在放效果图了。

修改/etc/sudoers 文件，效果图如下：

```
##
## Allow root to run any commands anywhere
root    ALL=(ALL)       ALL

## Allows members of the 'sys' group to run networking, software,
## service management apps and more.
# %sys ALL = NETWORKING, SOFTWARE, SERVICES, STORAGE, DELEGATING, PROCESSES,
LOCATE, DRIVERS

## Allows people in group wheel to run all commands
%wheel  ALL=(ALL)       ALL
hadoop  ALL=(ALL)       NOPASSWD: ALL
```

在/opt 目录下创建文件夹，并修改所属主和所属组，效果图如下：

```
[root@DataNode2 ~]# systemctl status firewalld
● firewalld.service - firewalld - dynamic firewall daemon
   Loaded: loaded (/usr/lib/systemd/system/firewalld.service; disabled; vendor p
reset: enabled)
   Active: inactive (dead)
     Docs: man:firewalld(1)
[root@DataNode2 ~]# cd /opt/
[root@DataNode2 opt]# ll
总用量 0
drwxr-xr-x. 5 hadoop hadoop  59 10月  1 00:04 module
drwxr-xr-x. 7 root   root   109 4月  20 04:09 openoffice4
drwxr-xr-x. 2 root   root     6 10月 31 2018 rh
drwxr-xr-x. 2 hadoop hadoop 137 9月  30 23:57 software
[root@DataNode2 opt]#
```

卸载虚拟机自带的 JDK，如下图：

```
[root@NameNode ~]# rpm -qa | grep java
java-1.7.0-openjdk-headless-1.7.0.261-2.6.22.2.el7_8.x86_64
java-1.7.0-openjdk-1.7.0.261-2.6.22.2.el7_8.x86_64
tzdata-java-2021a-1.el7.noarch
python-javapackages-3.4.1-11.el7.noarch
java-1.8.0-openjdk-headless-1.8.0.302.b08-0.el7_9.x86_64
java-1.8.0-openjdk-1.8.0.302.b08-0.el7_9.x86_64
javapackages-tools-3.4.1-11.el7.noarch
```

```
[root@NameNode ~]# yum -y remove copy-jdk-configs-3.3-10.el7_5.noarch
已加载插件：fastestmirror, langpacks
正在解决依赖关系
--> 正在检查事务
```

```
依赖关系解决

================================================================================
 Package                  架构    版本                          源          大小
================================================================================
正在删除:
 copy-jdk-configs         noarch  3.3-10.el7_5                  @base       16 k
为依赖而移除:
 icedtea-web              x86_64  1.7.1-4.el7_9                 @updates   2.3 M
 java-1.7.0-openjdk       x86_64  1:1.7.0.261-2.6.22.2.el7_8    @base      679 k
 java-1.7.0-openjdk-headless
                          x86_64  1:1.7.0.261-2.6.22.2.el7_8    @base       95 M
 java-1.8.0-openjdk       x86_64  1:1.8.0.302.b08-0.el7_9       @updates   646 k
 java-1.8.0-openjdk-headless
                          x86_64  1:1.8.0.302.b08-0.el7_9       @updates   110 M
 jline                    noarch  1.0-8.el7                     @base       72 k
 rhino                    noarch  1.7R5-1.el7                   @base      1.1 M

事务概要
================================================================================
移除  1 软件包 (+7 依赖软件包)
```

```
删除:
  copy-jdk-configs.noarch 0:3.3-10.el7_5

作为依赖被删除:
  icedtea-web.x86_64 0:1.7.1-4.el7_9
  java-1.7.0-openjdk.x86_64 1:1.7.0.261-2.6.22.2.el7_8
  java-1.7.0-openjdk-headless.x86_64 1:1.7.0.261-2.6.22.2.el7_8
  java-1.8.0-openjdk.x86_64 1:1.8.0.302.b08-0.el7_9
  java-1.8.0-openjdk-headless.x86_64 1:1.8.0.302.b08-0.el7_9
  jline.noarch 0:1.0-8.el7
  rhino.noarch 0:1.7R5-1.el7

完毕！
```

```
[root@NameNode ~]# rpm -qa|grep java
tzdata-java-2021a-1.el7.noarch
python-javapackages-3.4.1-11.el7.noarch
javapackages-tools-3.4.1-11.el7.noarch
```

关闭 NameNode 的防火墙，并且关闭开机自启，效果图如下：

```
[root@DataNode2 ~]# systemctl status firewalld
● firewalld.service - firewalld - dynamic firewall daemon
   Loaded: loaded (/usr/lib/systemd/system/firewalld.service; disabled; vendor p
reset: enabled)
   Active: inactive (dead)
     Docs: man:firewalld(1)
```

4. 在 NameNode 上安装 JDK

在 NameNode 上安装 JDK 之前，先下载 winscp 来将压缩包传到虚拟机中，但是在使用 winscp 连接 NameNode 时，一直显示连接超时，一直连接不上，具体解决过程见下面心得总结，在这里放几张图片：

连接超时：

虚拟机可以 ping 通 windows：

```
[root@NameNode ~]# ping www.baidu.com
ping: www.baidu.com: 未知的名称或服务
[root@NameNode ~]# ping 192.168.43.85
PING 192.168.43.85 (192.168.43.85) 56(84) bytes of data.
64 bytes from 192.168.43.85: icmp_seq=1 ttl=128 time=0.756 ms
64 bytes from 192.168.43.85: icmp_seq=2 ttl=128 time=1.82 ms
64 bytes from 192.168.43.85: icmp_seq=3 ttl=128 time=0.836 ms
64 bytes from 192.168.43.85: icmp_seq=4 ttl=128 time=2.03 ms
64 bytes from 192.168.43.85: icmp_seq=5 ttl=128 time=1.95 ms
^Z
[3]+  已停止               ping 192.168.43.85
```

NameNode 的 ssh 服务已经打开：

```
[root@NameNode ~]# systemctl status sshd.service
● sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor pres
et: enabled)
   Active: active (running) since 四 2021-09-30 09:57:30 CST; 50min ago
     Docs: man:sshd(8)
           man:sshd_config(5)
 Main PID: 1083 (sshd)
    Tasks: 1
   CGroup: /system.slice/sshd.service
           └─1083 /usr/sbin/sshd -D
```

在 NameNode 上添加 20 端口：

```
[root@NameNode ~]# iptables -nL
Chain INPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     udp  --  0.0.0.0/0            0.0.0.0/0           udp dpt:53
ACCEPT     tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:53
ACCEPT     udp  --  0.0.0.0/0            0.0.0.0/0           udp dpt:67
ACCEPT     tcp  --  0.0.0.0/0            0.0.0.0/0           tcp dpt:67

Chain FORWARD (policy ACCEPT)
target     prot opt source               destination
ACCEPT     all  --  0.0.0.0/0            192.168.122.0/24    ctstate RELATED
,ESTABLISHED
ACCEPT     all  --  192.168.122.0/24     0.0.0.0/0
ACCEPT     all  --  0.0.0.0/0            0.0.0.0/0
REJECT     all  --  0.0.0.0/0            0.0.0.0/0           reject-with icm
p-port-unreachable
REJECT     all  --  0.0.0.0/0            0.0.0.0/0           reject-with icm
p-port-unreachable

Chain OUTPUT (policy ACCEPT)
target     prot opt source               destination
ACCEPT     udp  --  0.0.0.0/0            0.0.0.0/0           udp dpt:68
```

```
[root@NameNode ~]# firewall-cmd --zone=public --add-port=20/tcp --permanent
success
[root@NameNode ~]# firewall-cmd --add-port=20/tcp
success
[root@NameNode ~]# firewall-cmd --reload
success
[root@NameNode ~]# firewall-cmd --zone=public --query-port=20/tcp
yes
[root@NameNode ~]# firewall-cmd --list-ports
20/tcp
```

```
[root@NameNode hadoop]# netstat -ntpl | grep 22
tcp        0      0 192.168.122.1:53        0.0.0.0:*               LISTEN
1396/dnsmasq
tcp        0      0 0.0.0.0:22              0.0.0.0:*               LISTEN
1073/sshd
tcp6       0      0 :::22                   :::*                    LISTEN
1073/sshd
```

最后发现是需要重启一波 windows 里的对应虚拟网卡，可以成功连接。然后使用 winscp 将 jdk 和 hadoop 的压缩包传输到/pot/software 中去。

将 jdk 压缩包解压到 module 文件夹中，效果图片如下：

```
[root@NameNode ~]# tar -xvf /opt/software/exe/jdk-8u212-linux-x64.tar.gz -C /opt
/software/anzhuang
```

```
jdk1.8.0_212/bin/javah
jdk1.8.0_212/bin/javac
jdk1.8.0_212/bin/jvisualvm
jdk1.8.0_212/bin/jcontrol
jdk1.8.0_212/release
[root@NameNode ~]#
```

配置 jdk 环境变量并查看是否安装成功，效果图如下：

```
[root@NameNode hadoop]# vim /etc/profile.d/my_env.sh
[root@NameNode hadoop]# sourse /etc/profile
bash: sourse: 未找到命令...
[root@NameNode hadoop]# source /etc/profile
[root@NameNode hadoop]# java -version
java version "1.8.0_212"
Java(TM) SE Runtime Environment (build 1.8.0_212-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.212-b10, mixed mode)
[root@NameNode hadoop]#
```

5. 在 NameNode 安装 Hadoop

将压缩包解压到 module 文件夹中：

```
[root@NameNode hadoop]# tar -zxvf /opt/software/hadoop-3.1.3.tar.gz -C /opt/
module/
hadoop-3.1.3/share/doc/hadoop/hadoop-yarn/hadoop-yarn-server/hadoop-yarn-
server-resourcemanager/apidocs/org/apache/hadoop/yarn/server/resourcemanager/
```

查看是否解压成功：

```
hadoop-3.1.3/include/StringUtils.hh
hadoop-3.1.3/include/TemplateFactory.hh
[root@NameNode ~]# ls /opt/module/hadoop-3.1.3
bin  include  libexec      NOTICE.txt  sbin
etc  lib      LICENSE.txt  README.txt  share
```

将 hadoop 添加到环境变量并查看是否安装成功：

```
[root@NameNode ~]# vi /etc/profile.d/my_env.sh
[root@NameNode ~]# vi /etc/profile.d/my_env.sh
[root@NameNode ~]# vi /etc/profile.d/my_env.sh
[root@NameNode ~]# source /etc/profile
[root@NameNode ~]# hadoop version
Hadoop 3.1.3
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r ba63
1c436b806728f8ec2f54ab1e289526c90579
Compiled by ztang on 2019-09-12T02:47Z
Compiled with protoc 2.5.0
From source with checksum ec785077c385118ac91aadde5ec9799
This command was run using /opt/module/hadoop-3.1.3/share/hadoop/common/hadoo
p-common-3.1.3.jar
[root@NameNode ~]#
```

6. 克隆虚拟机模板

克隆虚拟机比较简单，直接点击克隆，选择完整克隆，即可克隆出两台虚拟机，以防万一，在克隆之前保存个快照。

因为在克隆之前就在模板机上安装了新的 JDK 和 Hadoop，因此只需改变克隆出的两台机子的主机名和 ip，DataNode2 如下图：

```
[root@DataNode2 ~]# hostname
DataNode2
[root@DataNode2 ~]# ifconfig
ens33: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.10.102  netmask 255.255.255.0  broadcast 192.168.10.255
```

修改 windows 的主机映射文件（host 文件），比较简单，不在放图。

7. 完全分布式运行模式，SSH 无密登录配置，集群配置

scp 安全拷贝，案例实操，效果图如下：

```
[root@DataNode2 scp-test]# cd
[root@DataNode2 ~]# scp -r /home/hadoop/桌面/scp-test hadoop@DataNode1:/home/
hadoop/桌面
ssh: connect to host datanode1 port 22: No route to host
lost connection
[root@DataNode2 ~]# systemctl status sshd
● sshd.service - OpenSSH server daemon
   Loaded: loaded (/usr/lib/systemd/system/sshd.service; enabled; vendor pres
et: enabled)
   Active: active (running) since 六 2021-10-02 00:22:50 CST; 16min ago
```

```
[root@DataNode2 ~]# scp -r /home/hadoop/桌面/scp-test hadoop@DataNode1:/home/
hadoop/桌面
The authenticity of host 'datanode1 (192.168.10.101)' can't be established.
ECDSA key fingerprint is [illegible].
ECDSA key fingerprint is [illegible].
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'datanode1,192.168.10.101' (ECDSA) to the list of
known hosts.
hadoop@datanode1's password:
scp-test.txt                                  100%   31    12.0KB/s   00:00
```

配置 SSH：

```
[root@DataNode2 ~]# ssh DataNode1
root@datanode1's password:
Last login: Sat Oct  2 00:43:23 2021
[root@DataNode1 ~]# exit
登出
Connection to datanode1 closed.
```

无秘钥配置：

```
[hadoop@DataNode2 root]$ cd /home/hadoop/.ssh
bash: cd: /home/hadoop/.ssh: 没有那个文件或目录
[hadoop@DataNode2 root]$ cd
[hadoop@DataNode2 ~]$ cd /home/hadoop/.ssh
bash: cd: /home/hadoop/.ssh: 没有那个文件或目录
[hadoop@DataNode2 ~]$ .ssh-keygen -t rsa
bash: .ssh-keygen: 未找到命令...
相似命令是： 'ssh-keygen'
[hadoop@DataNode2 ~]$ ssh-keygen -t rsa

Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa): Created direc
tory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa.
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub.
The key fingerprint is:
```

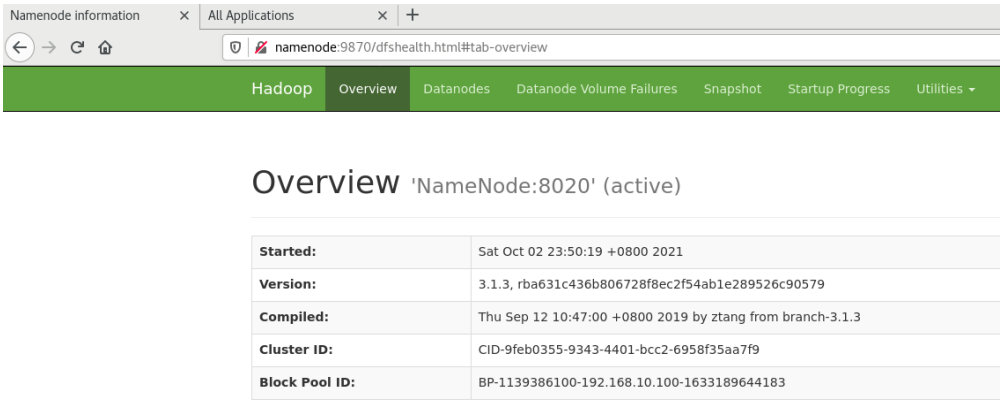
拷贝公钥不在放图了，在下面的集群配置中会给出免密登录的效果图。

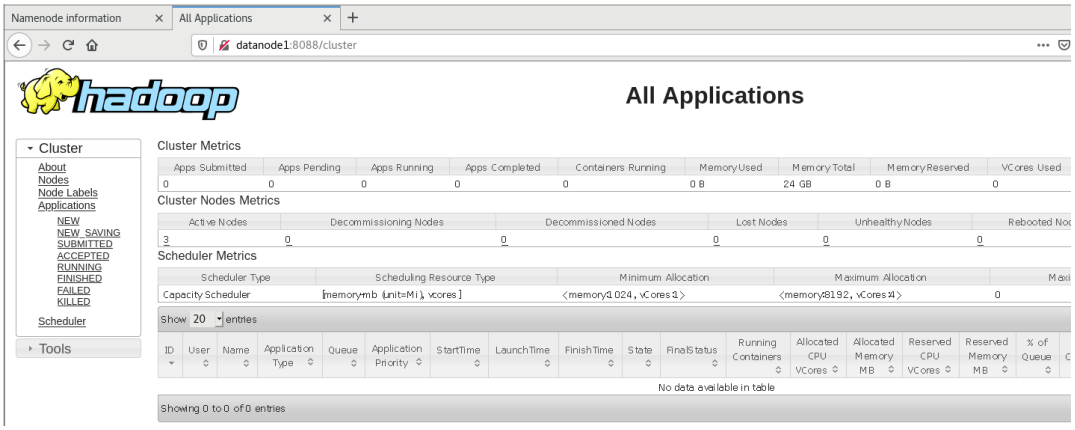
配置四个.xml 文件也比较简单，将教程导入三台虚拟机中，使用 vim 命令打开编辑配置文件，复制教程中的内容到相应的配置文件当中即可。

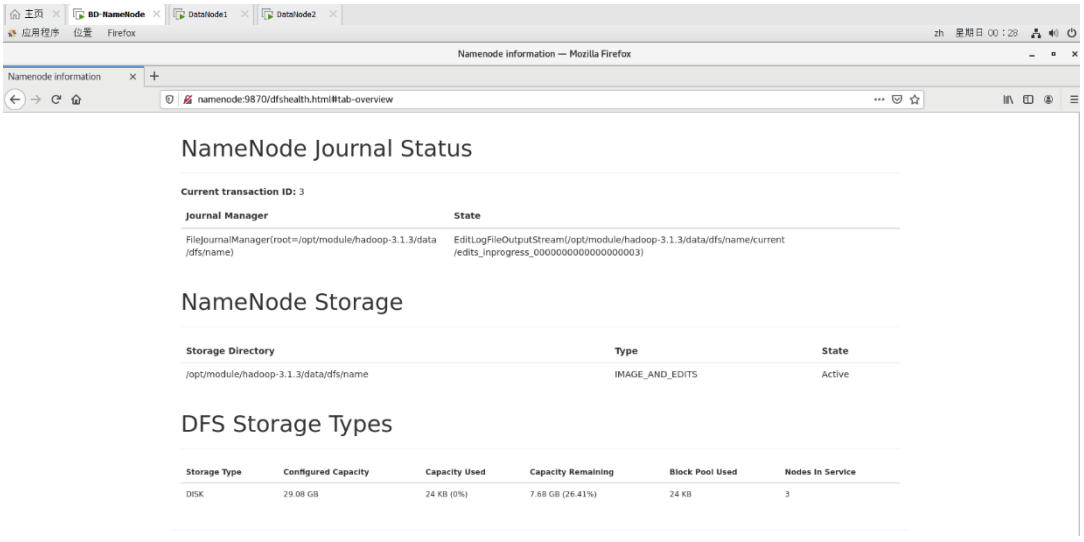
在分发配置文件的过程中出现问题，分发文件的位置好像出现问题，具体情况在下面的心得总结中有体现。

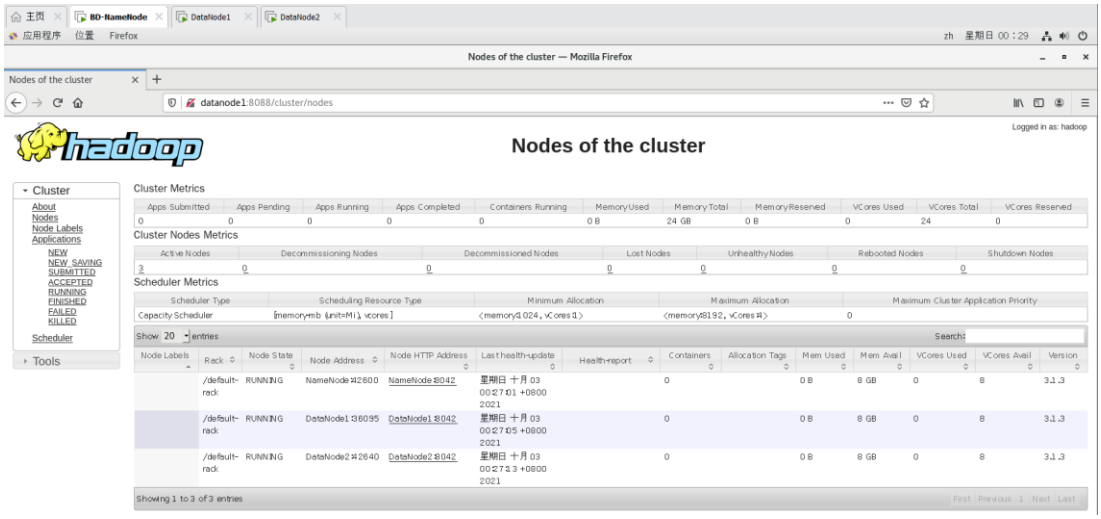
最后附上启动集群后的几张效果图：

```
[hadoop@NameNode ~]$ cd /opt/module/hadoop-3.1.3
[hadoop@NameNode hadoop-3.1.3]$ hdfs namenode -format
WARNING: /opt/module/hadoop-3.1.3/logs does not exist. Creating.
2021-10-02 23:47:16,473 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
```









这张图片可以显示出 ssh 免密登录的效果：

```
[hadoop@NameNode ~]$ ssh DataNode2
Last login: Sun Oct  3 00:37:27 2021 from namenode
[hadoop@DataNode2 ~]$ jps
5698 NodeManager
5460 SecondaryNameNode
5276 DataNode
6382 Jps
[hadoop@DataNode2 ~]$ exit
登出
Connection to datanode2 closed.
[hadoop@NameNode ~]$ jps
6305 NodeManager
5588 NameNode
5742 DataNode
7599 Jps
[hadoop@NameNode ~]$ ssh DataNode1
Last login: Sat Oct  2 22:28:36 2021 from namenode
[hadoop@DataNode1 ~]$ jps
5939 NodeManager
5349 DataNode
7480 Jps
5774 ResourceManager
```

三、心得总结（写出自己在完成实验过程中遇到的问题、解决方法，以及体会、收获等）（10%）

在安装好第一台虚拟机后，尝试安装图形界面，但是第一次在下载是，中途显示没有去其他可尝试的镜像，安装失败；然后按照网上教程更新了一波内核，删除缓存，再重新下载安装图形界面，最终安装成功，后面克隆的两台虚拟机也因此都有图形界面。

在下载使用 winscp 和 Xshell 时，死活连接不上虚拟机中的 centos，一直显示连接超时。在网上找解决方法，首先是关闭 Linux 的防火墙，这个在安装 Linux 的时候就做了，为了以防万一还是查看了防火墙的状态，确定是关闭的，没有解决问题；然后有博客说是

在 Linux 上启动 ssh 服务，查看 sshd 状态后，发现是在 running，仍未解决问题；再然后是说看看 Linux 是否开启 20 端口，我发现我的 NameNode 并没有这个端口，于是添加这个端口，并且开启端口（期间需要打开 firewalld），最后发现还是显示超时；之后又查看了 VMnet8 中 ip 和默认网关与虚拟机中虚拟网络编辑器中的网关和子网 IP 是否对应（其实应该本来就是对应的，在第一步中就有设置这两个地方的），仍未解决问题；最后是将 VMnet8 这个网卡禁用再重新启用后解决问题。

解压 jdk 压缩包出现问题，发现是给的教程文档里边的命令有问题，应该加上 jdk 所在位置，只有文件名是找不到文件的，命令应为“tar -xvf /opt/software/exe/jdk-8u212-linux-x64.tar.gz -C /opt/module”，解压成功。后来发现是我没有先进入 software 文件夹，命令没有给错。

配置好某台机子的集群文件后要分发到剩下两台机子上，但是分发命令好像有问题，将配置好的 hadoop 文件夹给分发到目的机的 hadoop 文件下了，应该是取代这个文件夹的才对。

最后在 NameNode 启动 hdfs 和在 DataNode1 启动 yarn 时，电脑基本卡得动都动不了，经过五六分钟左右才稍微好点，可以再虚拟机的浏览器中查看 HDFS 上的数据信息和 YARN 运行中的 Job 信息，也不会知道往后的实验会不会更卡。

总之，实验一就遇到了不少问题，虽然花费大量时间在网上查找解决方法（网上内容太繁杂，筛选出正确解答方法很费时间），但最终还是都能解决，最后实验完成时的成就感还是很强的。后面的实验也一定会遇到各种各样的问题，但我还是会尽力去解决问题的。