

实验报告

报告标题：Hadoop 环境搭建

学号：21190630

姓名：黄艺杰

日期：2022 年 11 月 18 日

一、实验环境

1. 操作系统：Windows 10、Linux
2. 相关软件（含版本号）：VMWare15pro、FinalShell3.9、IntelliJ IDEA2019Professional
3. 其它工具：JDK1.8

二、实验内容及其完成情况

（针对上述实验内容逐一详述实验过程）

1. Linux（CentOS 7 发行版本）安装及网络配置：

按要求安装虚拟机，安装时确认网络适配器的工作模式为 NAT

使用 ipconfig 命令查看本地 IP 地址

```
以太网适配器 VMware Network Adapter VMnet8:

    连接特定的 DNS 后缀 . . . . . : 
    本地链接 IPv6 地址. . . . . : fe80::3d83:39b8:823c:da67%14
    IPv4 地址 . . . . . : 192.168.110.5
    子网掩码 . . . . . : 255.255.255.0
    默认网关. . . . . : 192.168.110.2
```

图 1.1 本地 IP 地址

在虚拟编辑器中查看 VMnet8 的 IP 地址、子网掩码和默认网关，在 NAT 设置中将网关 IP 设置为一个同一子网下不同于主机 IP 的值



图 1.2 虚拟编辑器 VMnet8 的 IP 地址、子网掩码

在网络适配器中更改 VMnet8 的 IP 地址、子网掩码和默认网关。



图 1.3 虚拟编辑器 VMnet8 的 IP 地址、子网掩码和默认网关

在虚拟机中更改 IP 地址、子网掩码和默认网关，DNS。

取消(C) 有线 应用(A)

详细信息 身份 IPv4 IPv6 安全

IPv4 Method

☐ 自动 (DHCP) ☐ 仅本地链路

☒ 手动 ☐ Disable

Addresses

地址	子网掩码	网关
192.168.110.101	255.255.255.0	192.168.110.2

DNS 自动 打开

114.114.114.114

Separate IP addresses with commas

路由 自动 打开

地址	子网掩码	网关	Metric

图 1.4 虚拟机中 IP 地址、子网掩码和默认网关，DNS

使用 FinalShell 选用 SSH 连接，连接 CentOS 7，虚拟机的 IP 地址为 192.168.110.101，连接成功，完成网络配置

```
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)
[hyj@master ~]$
连接断开
连接主机...
连接主机成功
Last login: Sun Nov 20 12:26:10 2022
[hyj@master ~]$
```

图 1.5 使用 FinalShell 连接虚拟机

2. JDK1.8 的安装配置

将 JDK 安装包上传至虚拟机中，并使用 tar 命令进行解压缩，将解压后的文件夹放到 /usr/java/jdk1.8 中

```
[hyj@master java]$ ll
总用量 0
drwxr-xr-x. 7 10 143 245 7月 7 2018 jdk1.8
```

图 2.1 将 JDK 安装包解压

使用 vim 编辑 .bash_profile，在 PATH 中加入 JDK 的路径

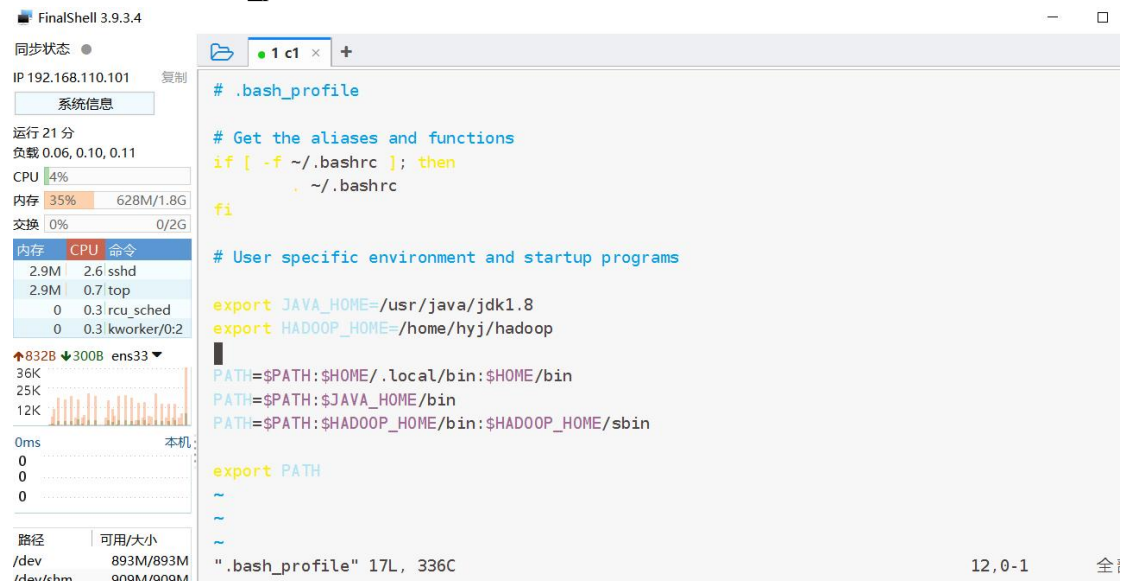


图 2.2 在.bash_profile 加入 JDK 路径

使用命令查看 java 版本

```
[hyj@master ~]$ java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) 64-Bit Server VM (build 25.181-b13, mixed mode)
```

图 2.3 查看 Java 版本

3. 单机版 Hadoop 的安裝配置

首先将 hadoop 压缩包上传到虚拟机，并解压 hadoop-2.6.0-cdh5.7.0.tar

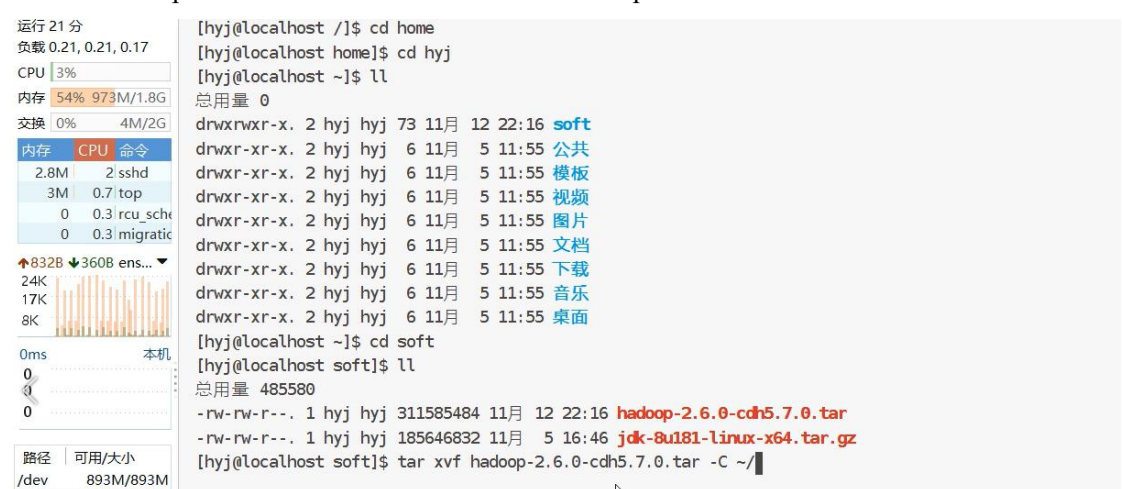


图 3.1 解压 hadoop 压缩包

配置 Hadoop 环境变量，使用 vim 编辑 .bash_profile 后执行命令重载

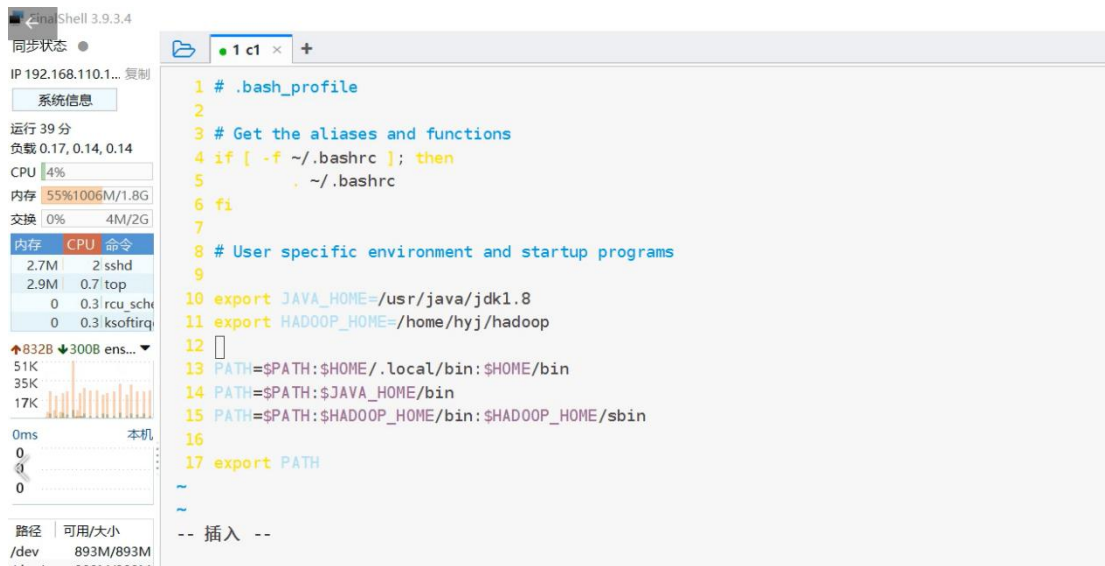


图 3.2 编辑.bash_profile

编辑 Hadoop 环境配置文件 hadoop-env.sh



图 3.3 编辑 Hadoop 环境配置文件

配置 Hadoop 核心文件



配置 HDFS 文件

图 3.4 配置 Hadoop 核心文件


```
[hyj@localhost hadoop]$ vim hdfs-site.xml

<configuration>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:///home/hyj/hadoop/tmp/hdfs/name</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
```

图 3.5 配置 HDFS 文件

执行初始化 NameNode 命令

```
[hyj@localhost hadoop]$ vim hdfs-site.xml
[hyj@localhost hadoop]$ hdfs namenode -format
```

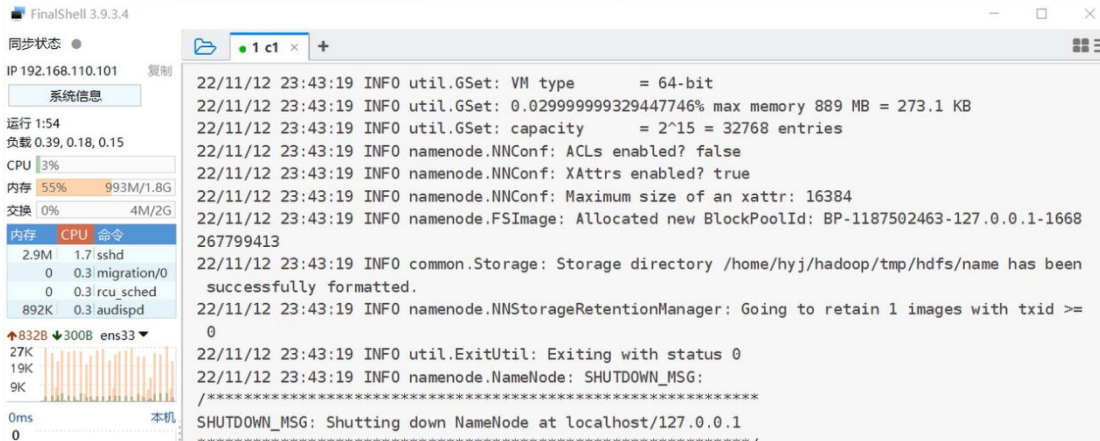


图 3.6 执行初始化 NameNode 命令，出现 successfully formatted

添加 hostname 为 master

```
[hyj@localhost hadoop]$ sudo hostnamectl set-hostname master
[hyj@localhost hadoop]$ ls
bin          cloudera    examples    include     libexec     NOTICE.txt /sbin      src
bin-mapreduce1  etc        examples-mapreduce1  lib         LICENSE.txt  README.txt  share     tmp
[hyj@localhost hadoop]$
连接断开
连接主机...
连接主机成功
Last login: Sat Nov 12 22:55:15 2022 from 192.168.110.5
[hyj@master ~]$
```

图 3.7 成功添加 hostname

启动 Hadoop，执行 start-dfs.sh 命令，之后执行 jps，可以看到 NameNode，则运行成功。

```
Last login: Sat Nov 12 22:55:15 2022 from 192.168.110.5
[hyj@master ~]$ start-dfs.sh
```

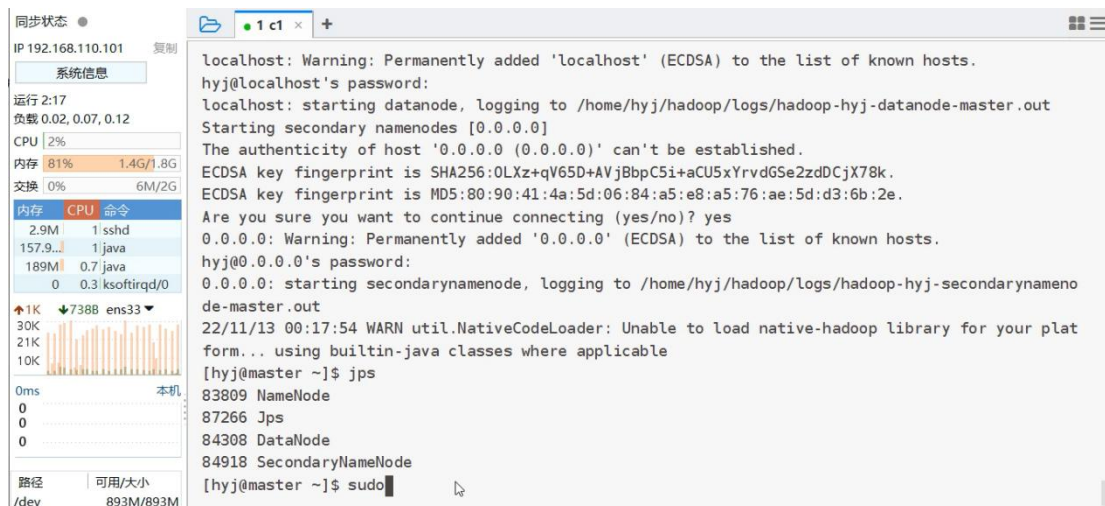


图 3.8 成功启动 Hadoop

设置防火墙允许访问 50070 端口，通过浏览器查看 CentOS 7 的 50070 端口

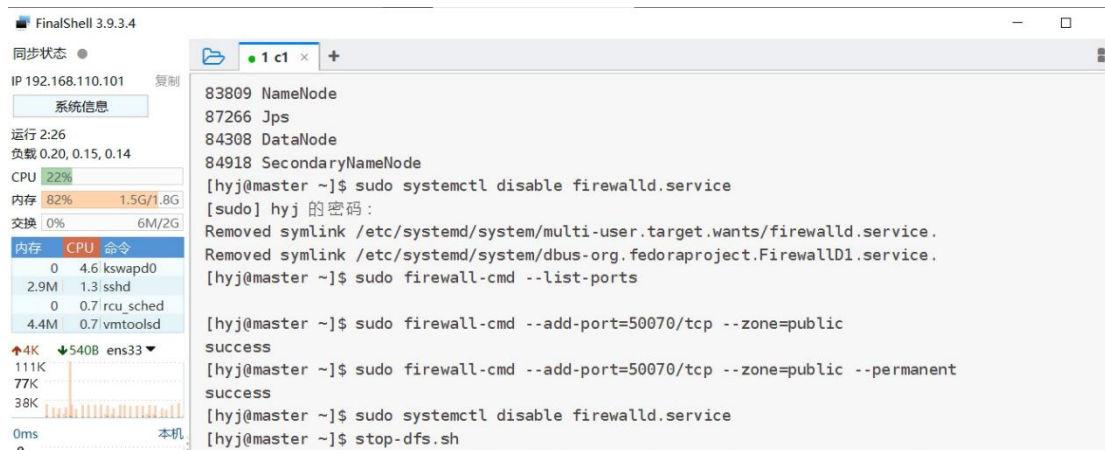


图 3.9 设置防火墙允许访问 50070 端口

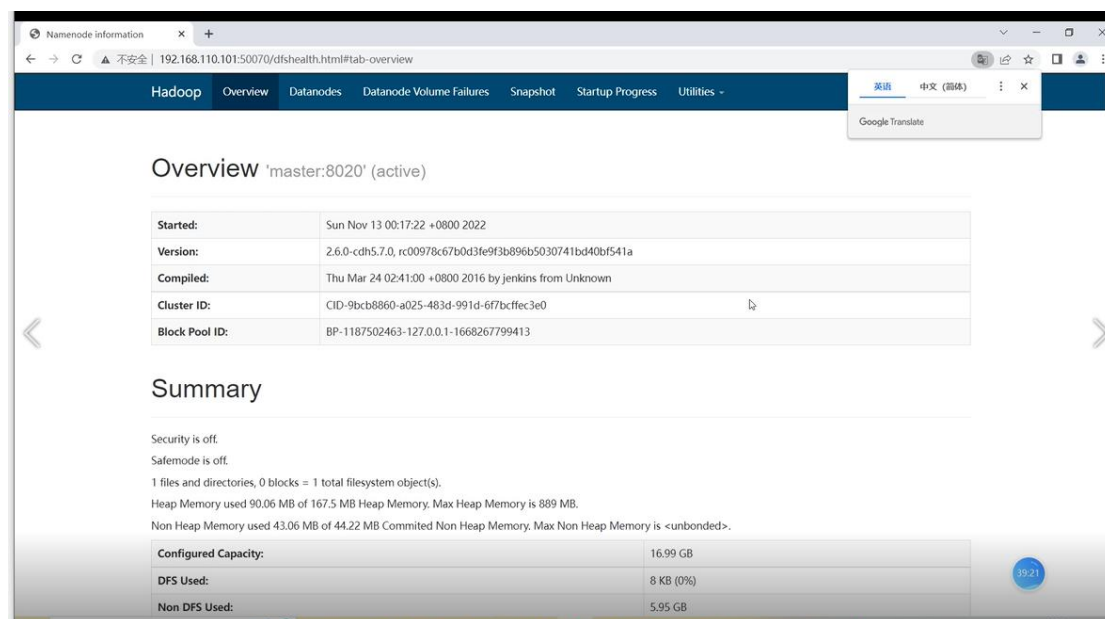


图 3.10 通过浏览器查看 CentOS 7 的 50070 端口

接下来因在启动过程中需要多次输入用户密码，可以配置 SSH 免输入密码
先产生密钥，在复制一份公钥，将密钥下载并设置使用

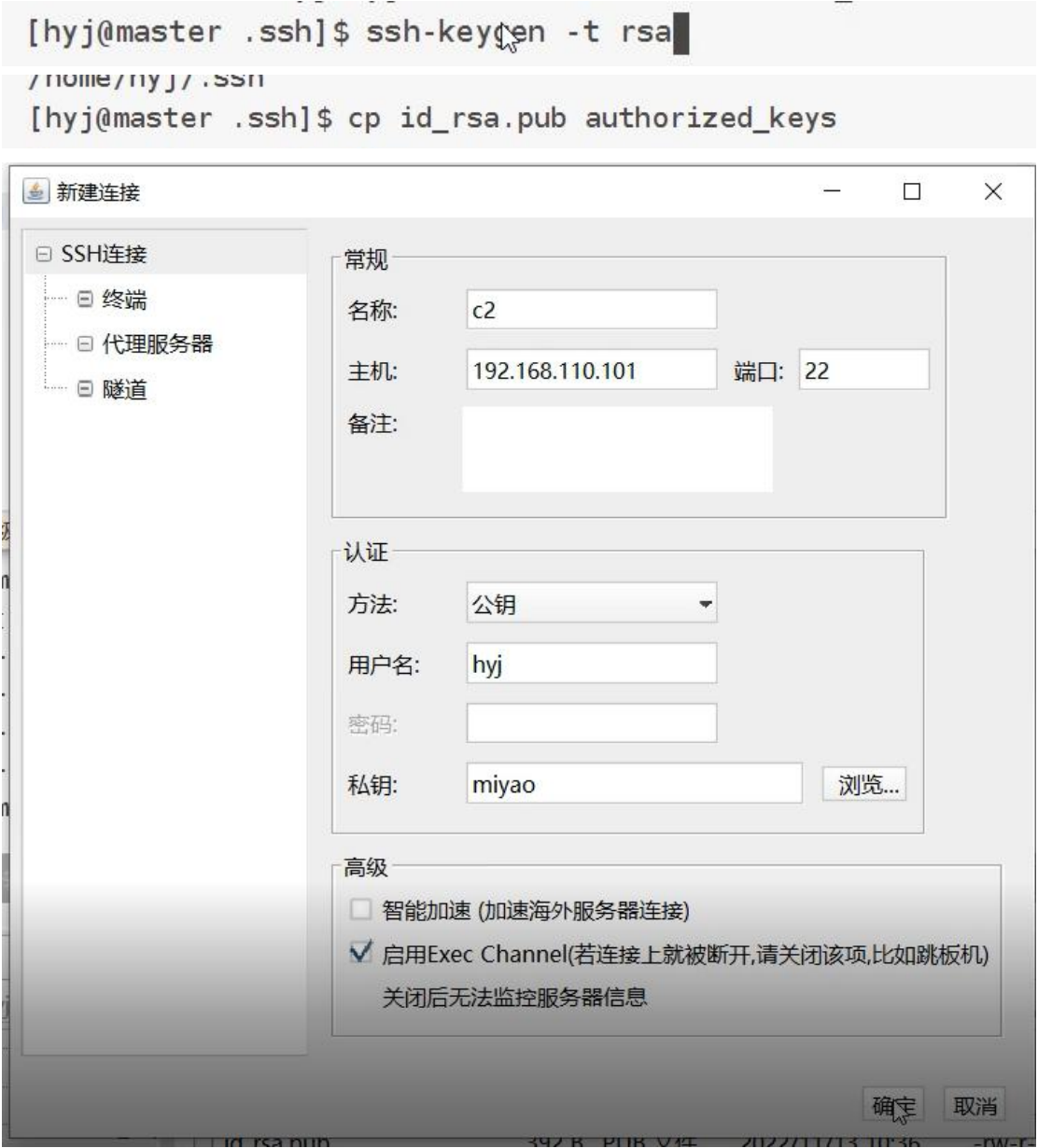
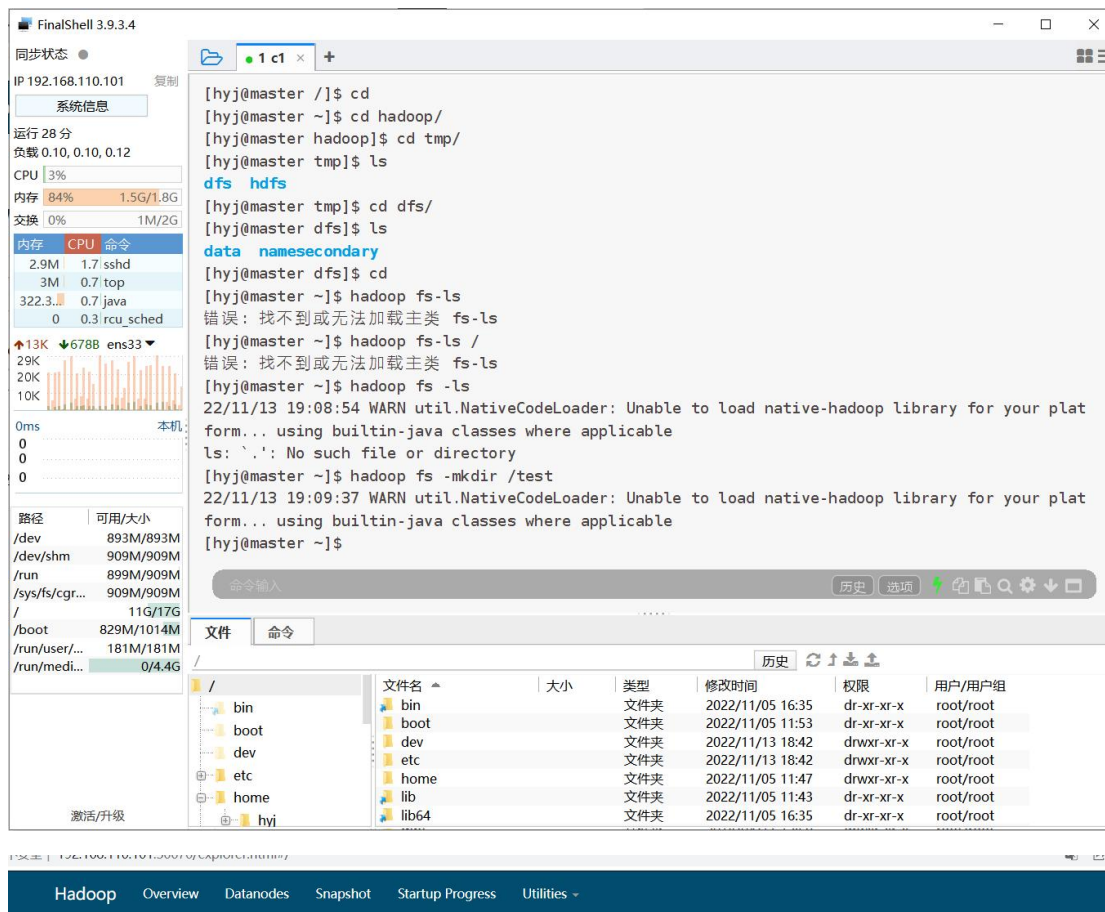


图 3.11 配置 SSH

最后是利用 hadoop 创建文件夹，在 HDFS 中可查看



Browse Directory

/								Go!
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 19:09:37 +0800 2022	0	0 B	test	

Hadoop, 2014.

图 3.12 利用 hadoop 创建文件夹，在 HDFS 中查看

4. Hadoop 的 Java 编程

在下载下来的 setting.xml 文件中修改镜像为阿里云镜像

```
<mirror>
  <id>alimaven</id>
  <name>aliyun maven</name>
  <url>http://maven.aliyun.com/nexus/content/repositories/central/</url>
  <mirrorOf>central</mirrorOf>
</mirror>
```

图 4.1 修改 setting.xml 为阿里云镜像

在初次打开 IDEA 时,配置环境,选择通过 maven 环境管理创建,选择 Project SDK 为 JDK1.8,勾选 Create from archetype 框并选择 maven-archetype-quickstart
在 maven settings 中修改的配置

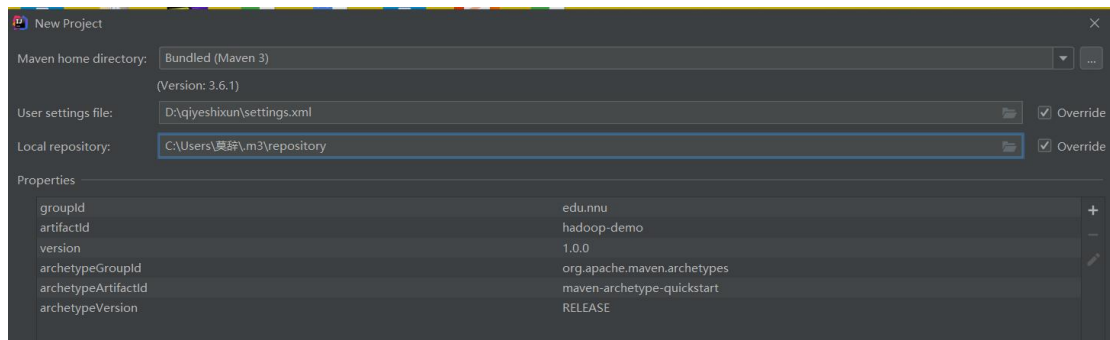


图 4.2 配置 IDEA 环境

创建项目后,在 pom.xml 添加 hadoop 版本、hadoop-common 依赖、hadoop-hdfs 依赖

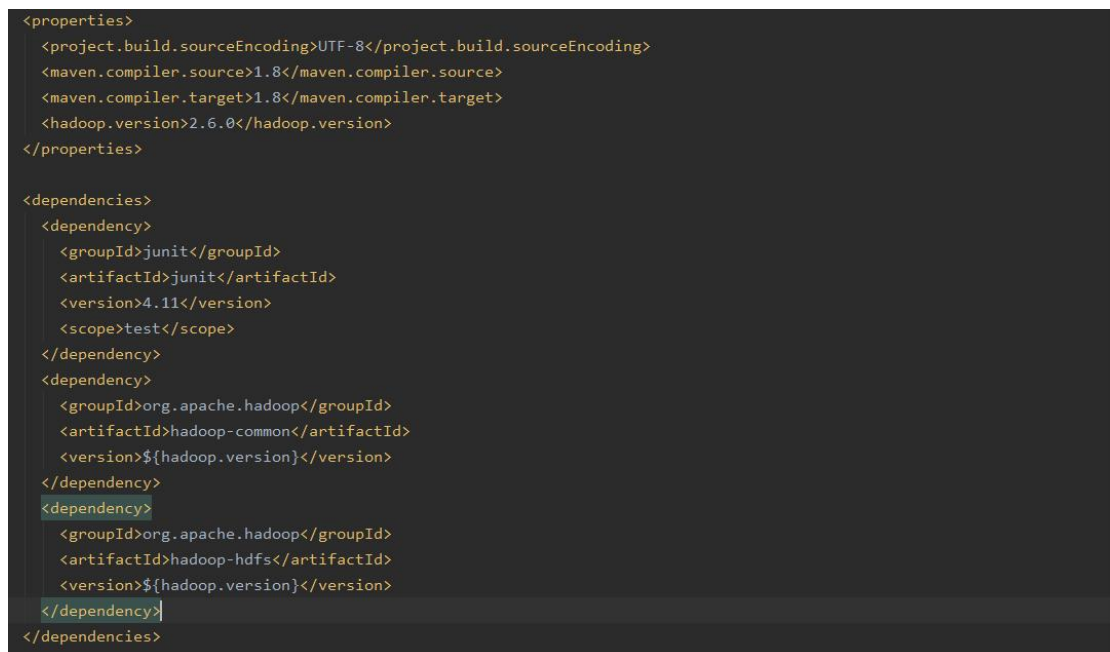


图 4.3 修改 pom.xml 中的配置

删除 test 文件夹中的文件,在 App.java 中完成一些基本链接和 config 配置

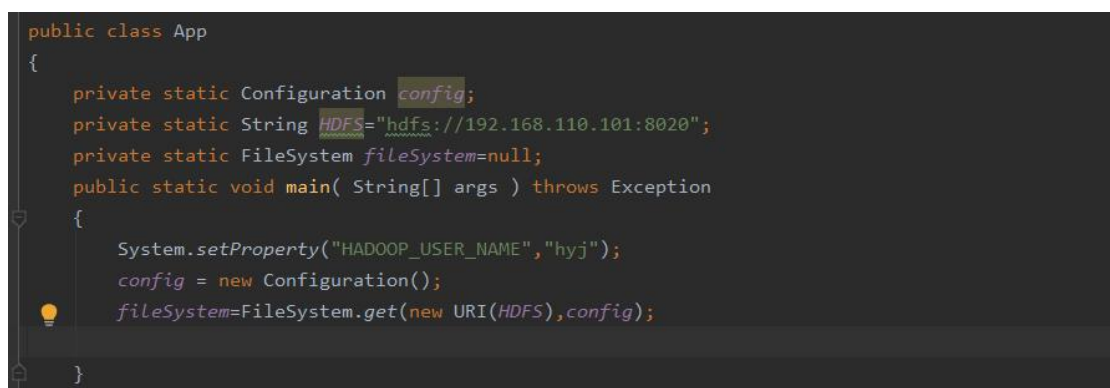


图 4.4 完成 HDFS 和 config 的配置

测试新建文件夹，分别是 demo 和多级文件夹 abc/efg/hjk

```
/*
新建文件夹
*/
private static void mkdir(String path) throws IOException {
    boolean ret=fileSystem.mkdirs(new Path(path));
    System.out.println(ret);
}
```

Browse Directory

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 22:09:56 +0800 2022	0	0 B	abc
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 22:05:58 +0800 2022	0	0 B	demo
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 19:09:37 +0800 2022	0	0 B	test

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 22:09:56 +0800 2022	0	0 B	hjk

图 4.5 新建文件夹并查看运行结果

分别删除 abc/efg/hjk，删除 abc，在 HDFS 的 Browse 上分别看到 abc 下的 efg 中为空，根目录下没有了 abc。

```
/*
*删除文件或者文件夹方法
*/
private static void rm(String path) throws IOException {
    boolean ret=fileSystem.delete(new Path(path), b: true);
    System.out.println(ret);
}
```

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name

Hadoop, 2014.

Browse Directory

Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 22:05:58 +0800 2022	0	0 B	demo
drwxr-xr-x	hyj	supergroup	0 B	Sun Nov 13 19:09:37 +0800 2022	0	0 B	test

Hadoop, 2014.

图 4.6 删除文件夹并查看运行结果

用 put 上传同一个本地文件 1.txt，一个不改名，一个改为 a.txt，可以看到它们除了姓名信息不同，其他信息是一样的

```
/*
*上传文件
*/
private static void put(String src, String dest) throws IOException {
    fileSystem.copyFromLocalFile(new Path(src),new Path(dest));
    System.out.println("put ok!");
}
```

Browse Directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hyj	supergroup	16 B	Sun Nov 13 22:31:03 +0800 2022	3	128 MB	1.txt
-rw-r--r--	hyj	supergroup	16 B	Sun Nov 13 22:33:01 +0800 2022	3	128 MB	a.txt

Hadoop, 2014.

图 4.7 上传文件夹并查看运行结果

从 demo 中使用 get()下载 1.txt 到本地指定路径。

```
/*
*下载文件
*/
private static void get(String src, String dest) throws IOException {
    fileSystem.copyToLocalFile( delSrc: false,new Path(src),new Path(dest), useRawLocalFileSystem: true);
    System.out.println("get ok!");
}
```

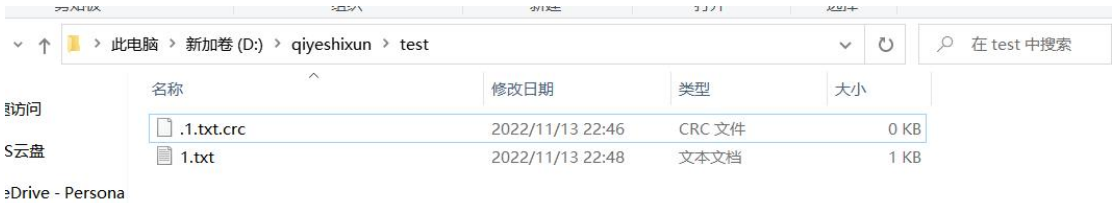


图 4.8 下载文件并查看运行结果

用 list()方法查看根目录和 demo 文件夹中详细信息。

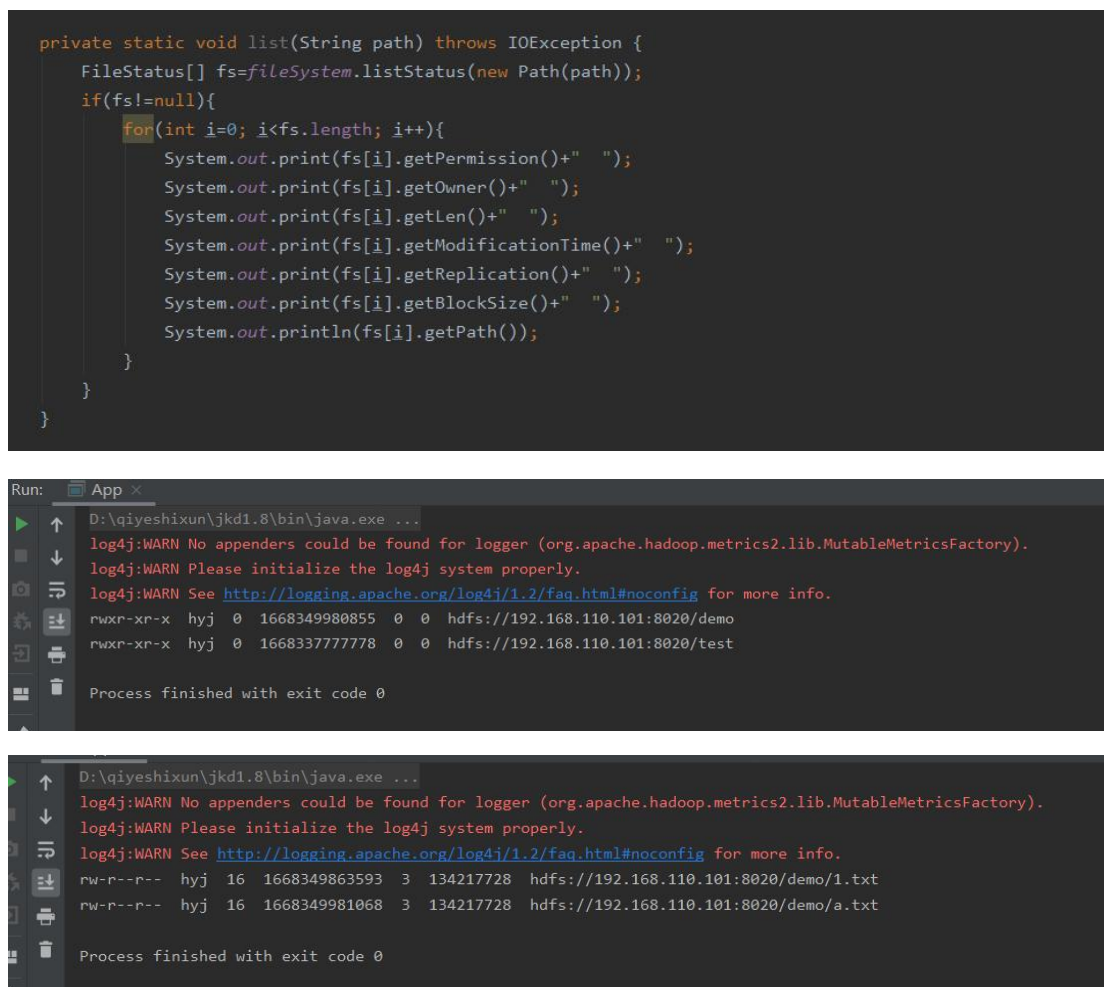


图 4.9 查看根目录和 demo 文件夹中详细信息

最后用 `put()` 函数上传四大名著文件到 HDFS 的 `/book` 目录中。

Browse Directory

/book							Go!
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hyj	supergroup	2.35 MB	Sun Nov 20 17:00:11 +0800 2022	3	128 MB	hongloumeng-utf8.txt
-rw-r--r--	hyj	supergroup	1.71 MB	Sun Nov 20 17:00:11 +0800 2022	3	128 MB	sanguoyanyi-utf8.txt
-rw-r--r--	hyj	supergroup	2.42 MB	Sun Nov 20 17:00:12 +0800 2022	3	128 MB	shuihu-utf8.txt
-rw-r--r--	hyj	supergroup	2.05 MB	Sun Nov 20 17:00:12 +0800 2022	3	128 MB	xiyouji-utf8.txt

Hadoop, 2014.

图 4.10 上传四大名著到 HDFS

三、实验总结

(可以总结实验中出现问题以及解决的思路，也可以列出没有解决的问题)

1. 在用浏览器打开 HDFS 页面前，用以下指令在虚拟机中关闭防火墙。

关闭防火墙：`sudo systemctl stop firewalld`

打开 hdfs 服务：`start-dfs.sh`

2. 对于某些文件，有些是 root 下的文件，并不属于所有者，这时候对文件进行操作时记得要用 `sudo` 提高权限。
3. 在 IDEA 上设定不同路径时，Windows 宿主机文件目录路径中的 ‘/’ 要使用 ‘\’ 来进行引导。