# Hadoop大数据技术应用

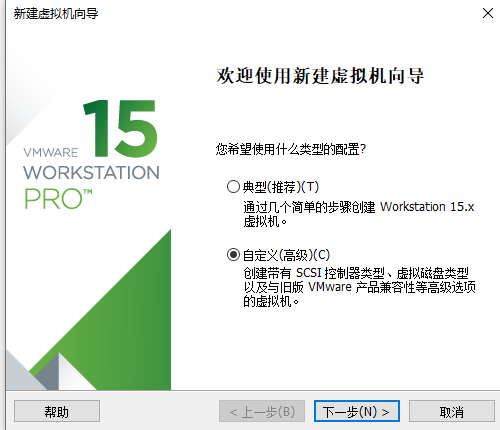
## 安装VMware虚拟机

下载VMware pro 15安装包，安装VMware pro 15。

进入主页选择创建新的虚拟机。



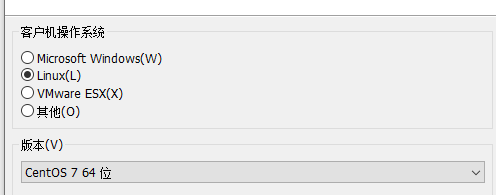
选择自定义（高级），点击下一步。



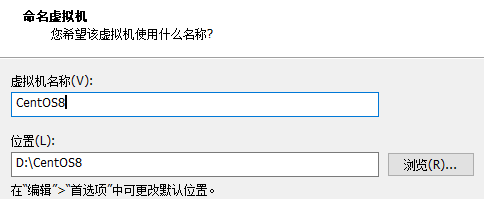
选择稍后安装操作系统，下一步即可。



不知道为啥版本没有CentOS 8.



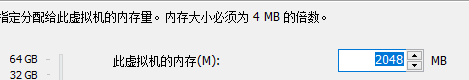
对虚拟机的名称和安装位置进行设置。



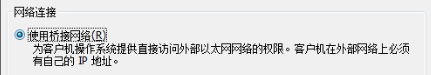
对虚拟机的处理器进行配置。



对虚拟机的内存进行配置。



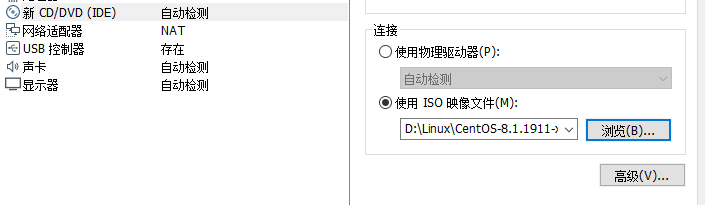
对虚拟机的网络类型进行配置。



对虚拟机的硬件进行配置。



移除打印机硬件，同时选择光盘驱动的ISO映像文件。

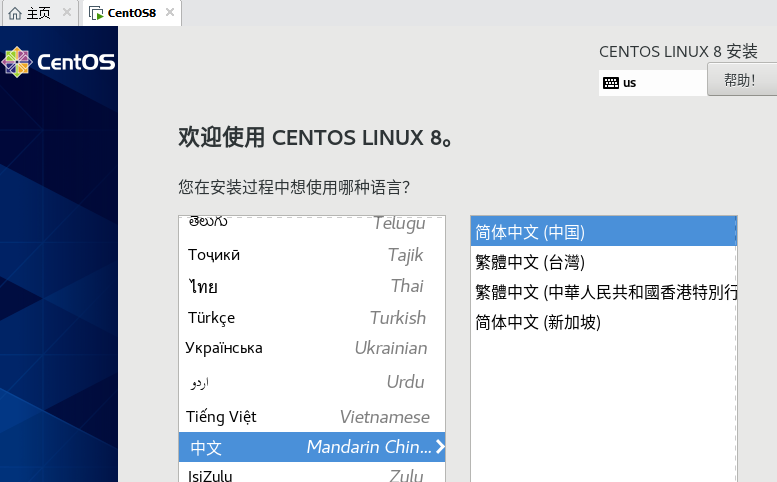


点击关闭，再点击完成，至此虚拟机创建完毕。

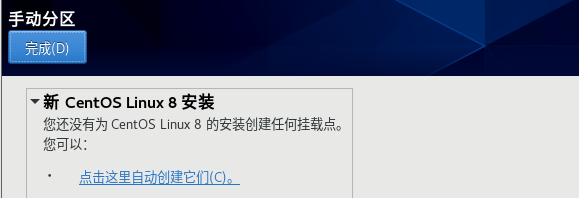
## 安装Linux操作系统

下载好要安装的Linux操作系统的镜像文件，本次安装的为CentOS8版本。

点击开启此虚拟机，选择语言。



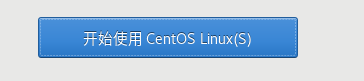
设置安装目的地。



关闭KDUMP，打开网络连接，点击完成后开始安装。。。。。。安装完毕。

设置好用户名和密码，开始进入系统界面。



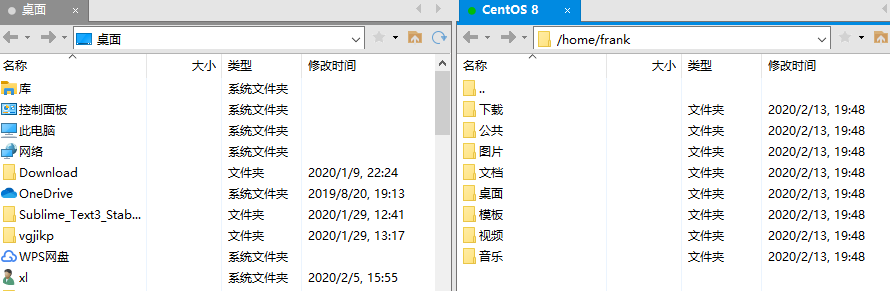
## 搭建Java开发环境

## 采用xftp6工具，在本地下载好然后上载到centos8指定的文件夹.

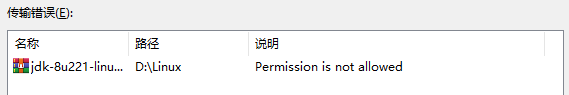
## 使用xfpt6连接虚拟机，设置名称，主机IP和用户名密码。



连接成功界面如下：



传文件时遇到这个问题。



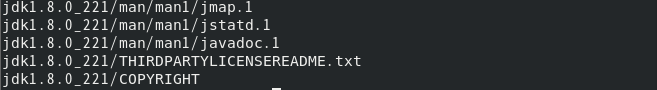
解决方法是在linux系统终端以root的权限输入以下命令行，r表是读 (Read) 、w表示写 (Write) 、x表示执行 (eXecute)  
读、写、运行三项权限可以用数字表示，就是r=4,w=2,x=1，777就是rwxrwxrwx



Centos查看下文件是否存在

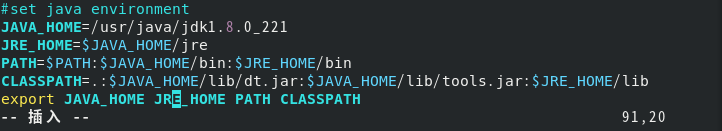


解压文件夹

解压完成

查看java文件夹，发现多了一个文件jdk1.8.0\_221

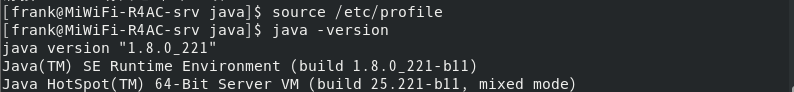
输入 vim /etc/profile进行JDK的配置，点击i进入输入模式，在最后添加如下内容：



# 期间本来应该按exit 然后输入 :wq保存，但是遇到 E45: 已设定选项 'readonly' (请加 ! 强制执行)

输入:wq!，又出现E212：无法打开并写入文件，解决方法是使用w ! sudo tee %,输入密码进行文件的保存。

编辑完成，查看Java的版本，如下图，至此完成Java的开发环境的配置。



安装Eclipse

下载好Eclipse压缩包，上传到虚拟机的CentOS 8系统。下得着实有点头秃。。速度太慢了

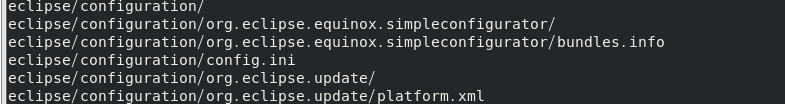


安装目录： /usr/local/src

解压文件



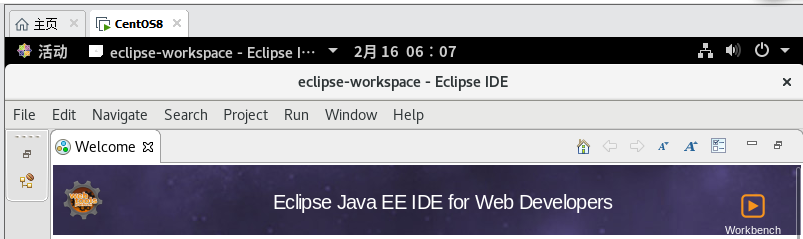
解压完成



进入eclipse目录下，输入命令：cd eclipse

执行eclipse可执行程序，输入命令： ./eclipse

启动完成界面：



## 搭建Hadoop开发环境

安装目录: /usr/local/hadoop

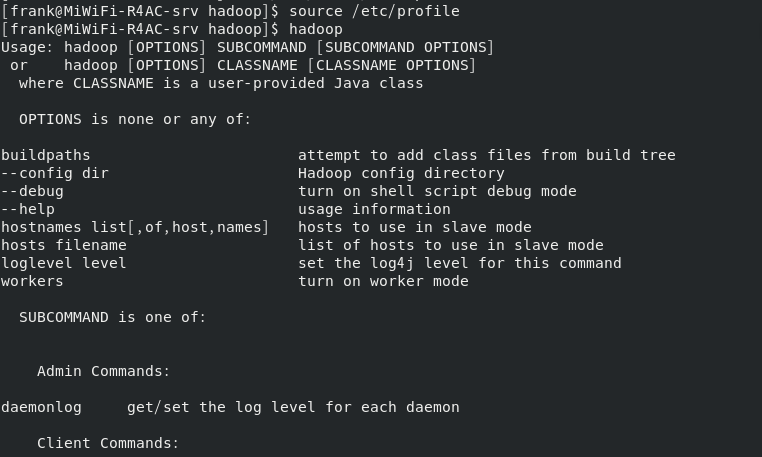
进入hadoop所在目录，进行解压。



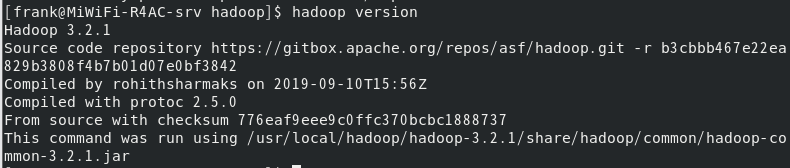
## 配置hadoop环境变量，输入命令vi /etc/profile，并且输入i进入编辑模式。

保存强制退出后，输入命令：source /etc/profile 使文件生效

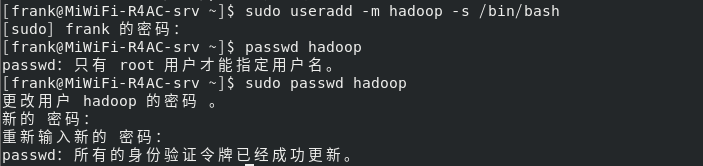
输入命令:hadoop 查看环境变量是否配置完成。出现如下信息即可：



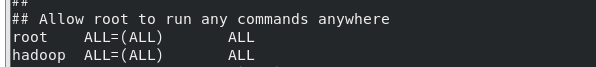
检查hadoop版本，输入：hadoop version 出现如下信息。



创建hadoop用户



给hadoop用户增加管理员权限



修改hadoop的配置文件。

进入hadoop目录：



1、配置core-site.xml 输入命令： vi core-site.xml 配置如下：

<configuration>

　　<!--指定HADOOP所使用的文件系统schema（URI），HDFS的老大（NameNode）的地址-->

　　<property>

　　　　<name>fs.defaultFS</name>

　　　　<value>hdfs://localhost:9000</value>

　　</property>

　　<!--指定HADOOP运行时产生文件的存储目录-->  
　　<property>

　　　　<name>hadoop.tmp.dir</name>

　　　　<value>/usr/local/hadoop/hadoop-3.2.1/tmp</value>

　　</property>

</configuration>

2、配置hdfs-site.xml 输入命令： vi hdfs-site.xml

<configuration>

　　<!--指定HDFS副本的数量-->

　　<property>

　　　　<name>dfs.replication</name>

　　　　<value>1</value>

　　</property>

</configuration>

3、配置mapred-site.xml输入命令： vi mapred-site.xml

<configuration>

<property>

　　 <name>mapreduce.framework.name</name>

　　 <value>yarn</value>

</property>

</configuration>

4、配置yarn-site.xml输入命令： vi yarn-site.xml

<configuration>

<!-- 指定YARN的老大（ResourceManager）的地址-->

<property>

<name>yarn.resourcemanager.hostname</name>

<value>master</value>

</property>

<!-- 指定reducer获取数据的方式-->

<property>

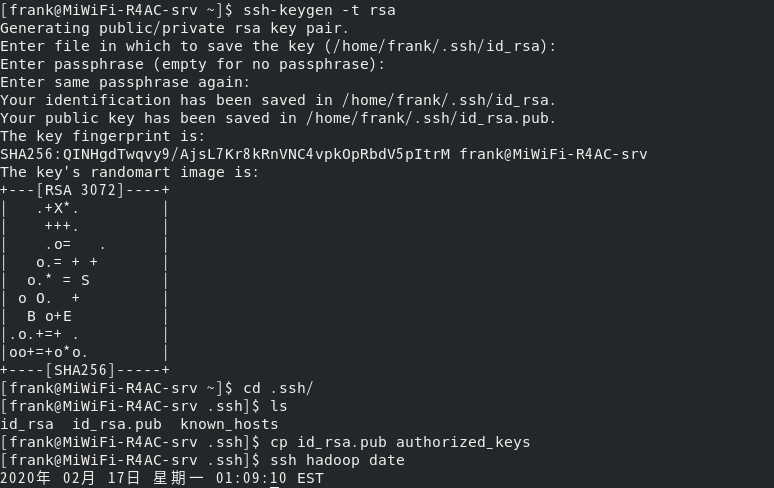
<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

</configuration>

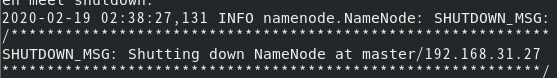
解决互信问题



设置本地主机名为master

格式化NameNode

在 bin 目录下运行 hdfs namenode -format



使用root配置的hadoop并启动会出现报错：

Starting namenodes on [localhost]

ERROR: Attempting to operate on hdfs namenode as root  
ERROR: but there is no HDFS\_NAMENODE\_USER defined. Aborting operation.  
Starting datanodes

ERROR: Attempting to operate on hdfs datanode as root  
ERROR: but there is no HDFS\_DATANODE\_USER defined. Aborting operation.

Starting secondary namenodes [master]

ERROR: Attempting to operate on hdfs secondarynamenode as root

ERROR: but there is no HDFS\_SECONDARYNAMENODE\_USER defined. Aborting operation.

解决方法：

在/hadoop/sbin路径下：   
将start-dfs.sh，stop-dfs.sh两个文件顶部添加以下参数  
HDFS\_DATANODE\_USER=root  
HADOOP\_SECURE\_DN\_USER=hdfs  
HDFS\_NAMENODE\_USER=root  
HDFS\_SECONDARYNAMENODE\_USER=root  
start-yarn.sh，stop-yarn.sh顶部也需添加以下  
YARN\_RESOURCEMANAGER\_USER=root  
HADOOP\_SECURE\_DN\_USER=yarn  
YARN\_NODEMANAGER\_USER=root

添加上述配置后出现以下错误

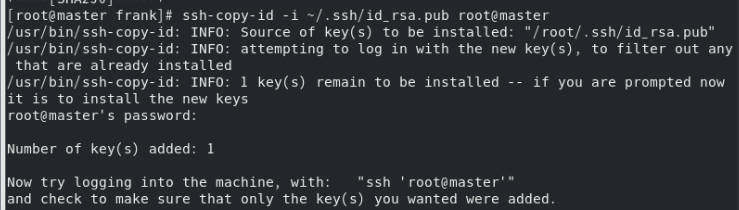
WARNING: HADOOP\_SECURE\_DN\_USER has been replaced by HDFS\_DATANODE\_SECURE\_USER. Using value of HADOOP\_SECURE\_DN\_USER.  
Starting namenodes on [localhost]  
mylinux\_1: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).  
Starting datanodes  
localhost: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).  
Starting secondary namenodes [master]  
mylinux\_1: Permission denied (publickey,gssapi-keyex,gssapi-with-mic,password).

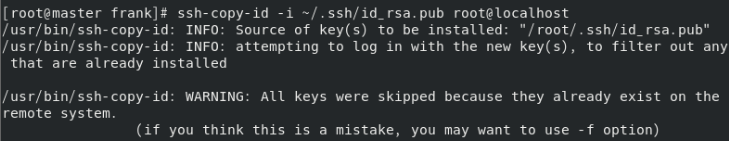
解决办法：配置免密登录（注意：对本机也需要配置）

ssh-keygen -t rsa

ssh-copy-id -i ~/.ssh/id\_rsa.pub root@localhost

ssh-copy-id -i ~/.ssh/id\_rsa.pub root@master

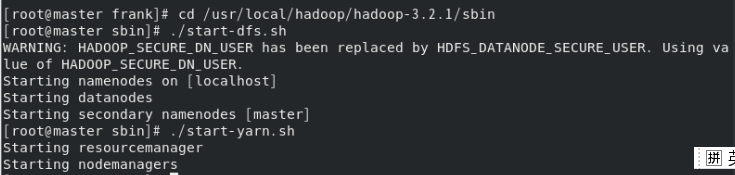




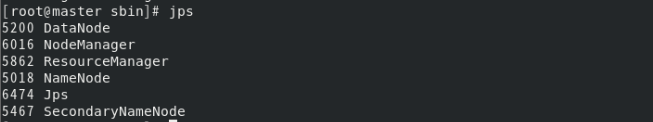
在hadoop目录下的sbin下执行以下命令：

./start-dfs.sh 和 ./start-yarn.sh

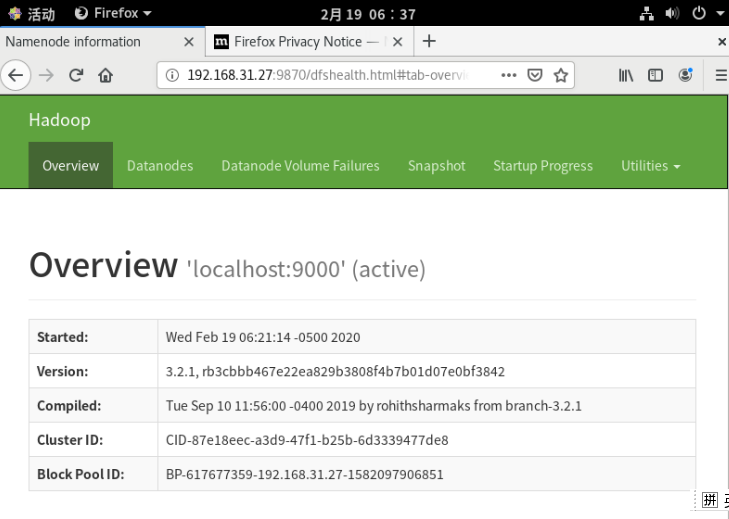
出现下图结果则为启动成功。



输入jps查看启动的进程：



登录系统浏览器输入：http://192.168.31.21:9870可出现以下界面

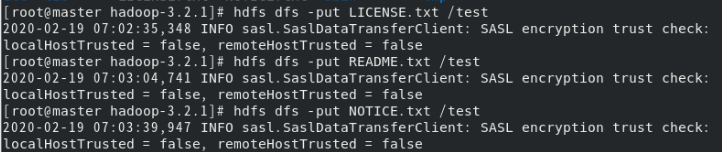


## 执行程序，统计词频

选用系统自带的MapReduce案例，进行wordcount测试。

新建test文件

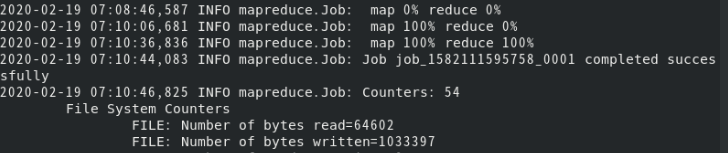
把自带的LICENSE.txt,README.txt,NOTICE.txt都放入hdfs的test文件中去，



输入以下命令：



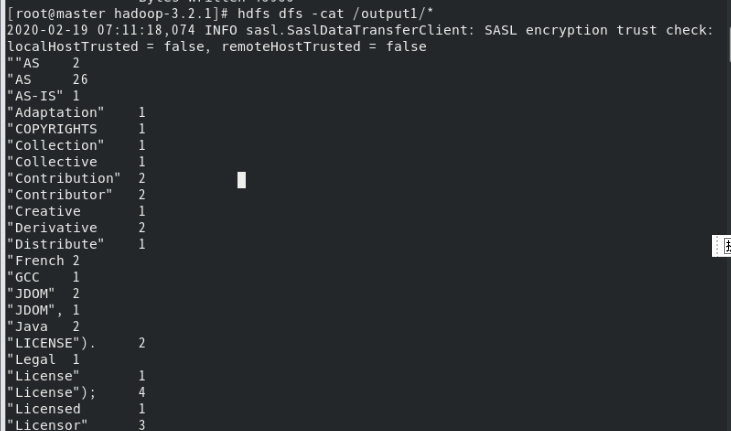
等待测试结束：



查看测试结果，输入以下命令：



结果如下（只截取前面部分）：



至此结束程序。