搭建Hadoop伪分布式大数据环境

[一、 实验目的 1](#_Toc31713305)

[二、 实验环境 1](#_Toc31713306)

[三、 实验要求 1](#_Toc31713307)

[四、 实验过程 1](#_Toc31713308)

[(一) 安装VMware虚拟机 1](#_Toc31713309)

[(二) 在虚拟机环境中安装Ubuntu操作系统 5](#_Toc31713310)

[(三) 在Ubuntu系统中安装JDK并完成环境变量配置 26](#_Toc31713311)

[(四) 安装Hadoop，实现伪分布式配置 28](#_Toc31713312)

[(五) 执行wordcount程序，统计指定目录下的词频 33](#_Toc31713313)

# 实验目的

搭建Hadoop伪分布式大数据环境

# 实验环境

操作系统：Ubuntu Server 16.04.6 LTS

实验工具：Hadoop 2.10.0

# 实验要求

1. 安装VMWare或VirtualBox虚拟机

2. 在虚拟机环境中安装Linux操作系统

3. 在Linux系统中安装JDK并完成环境变量配置

4. 安装Hadoop，实现伪分布式配置

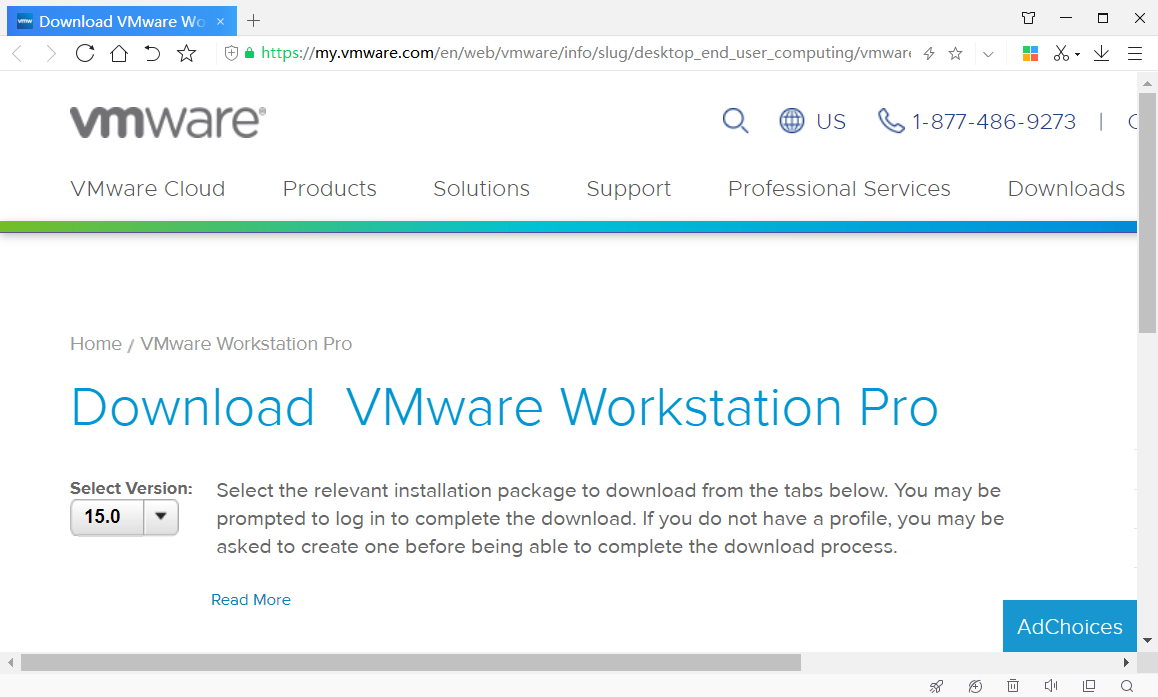
5. 执行wordcount程序，统计指定目录下的词频

6. 将1～5步实现步骤撰写为文档（Word或Jupyter），上传文档至GitHub

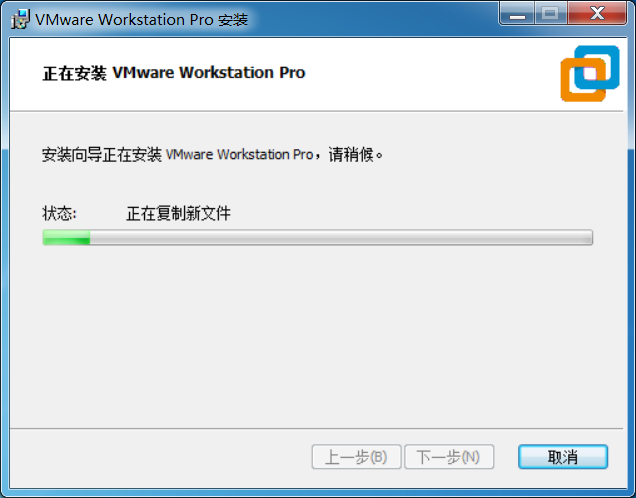
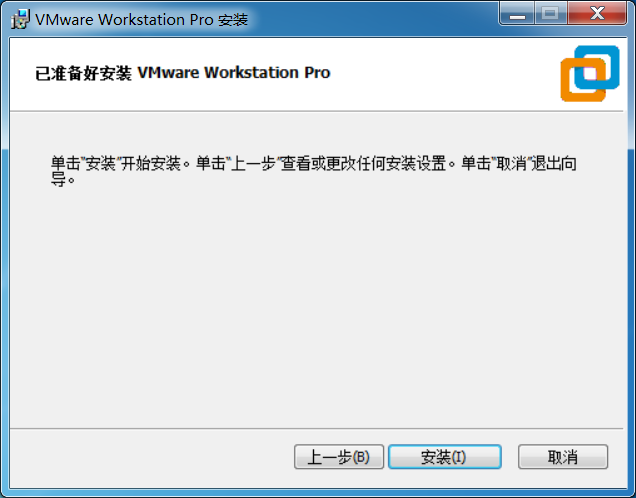
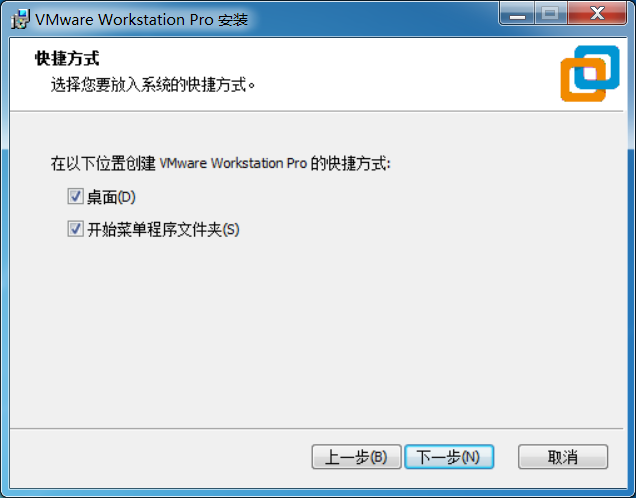
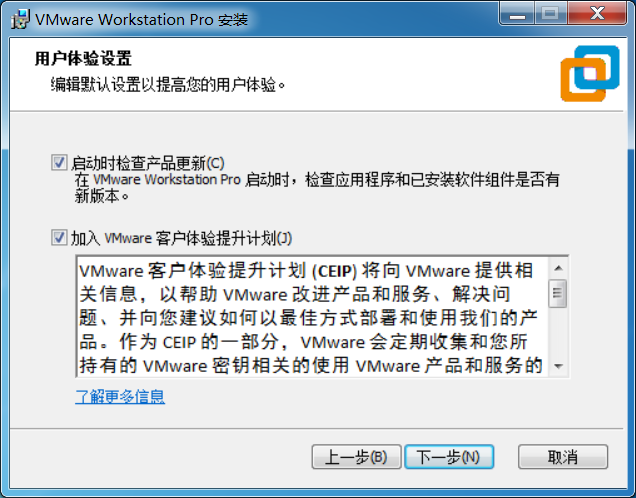
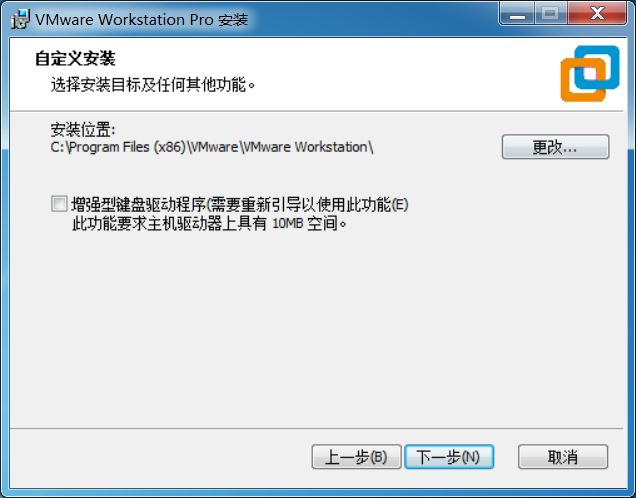
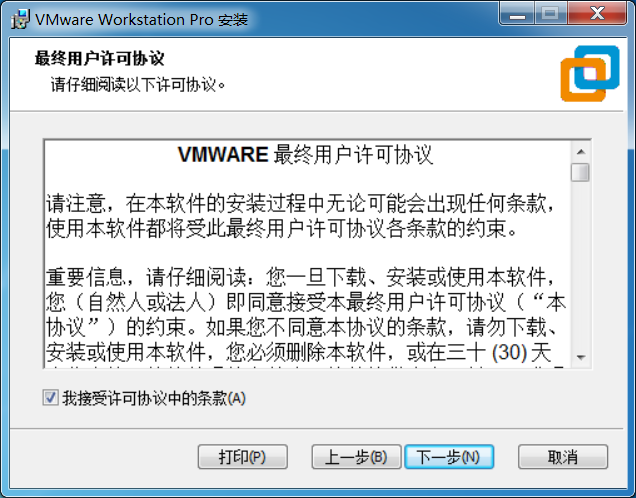
# 实验过程

## 安装VMware虚拟机

### 下载VMware Workstation Pro



### 安装VMware Workstation Pro

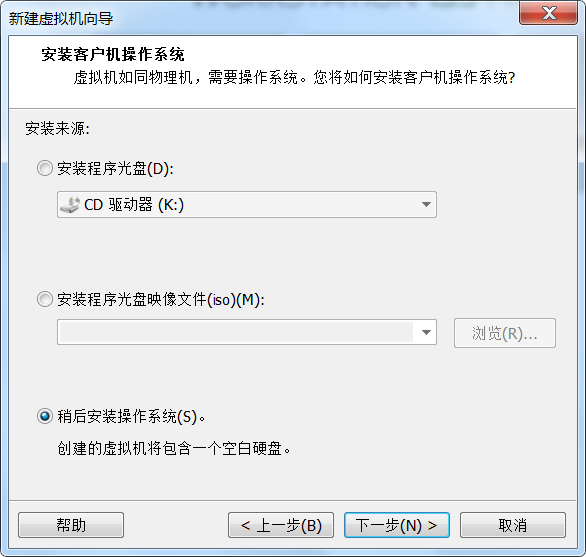


## 在虚拟机环境中安装Ubuntu操作系统

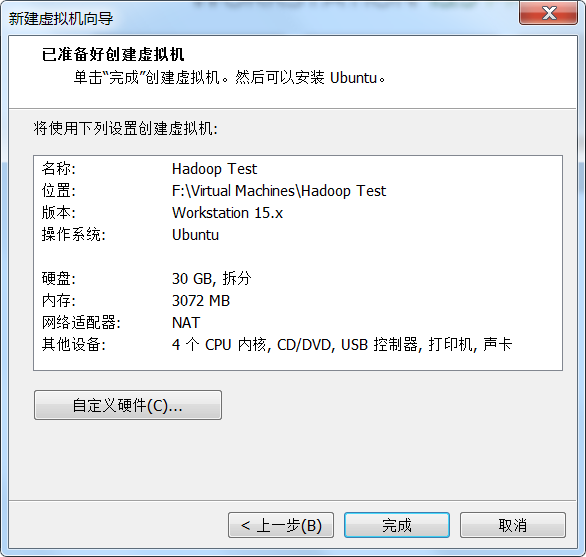
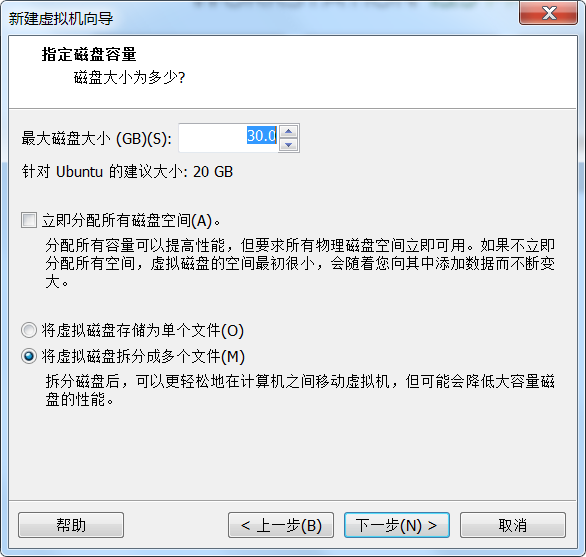
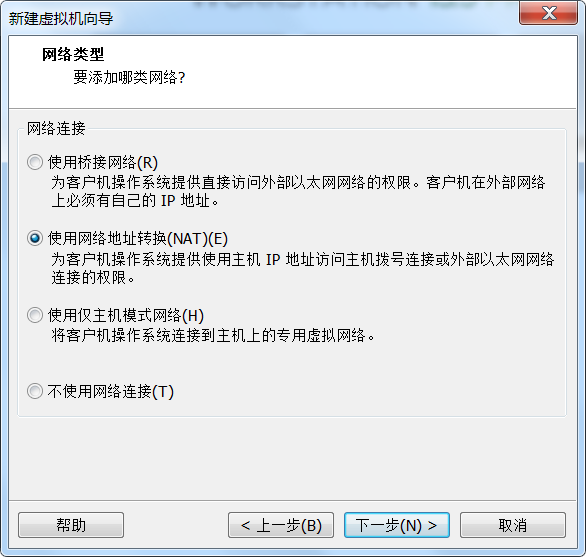
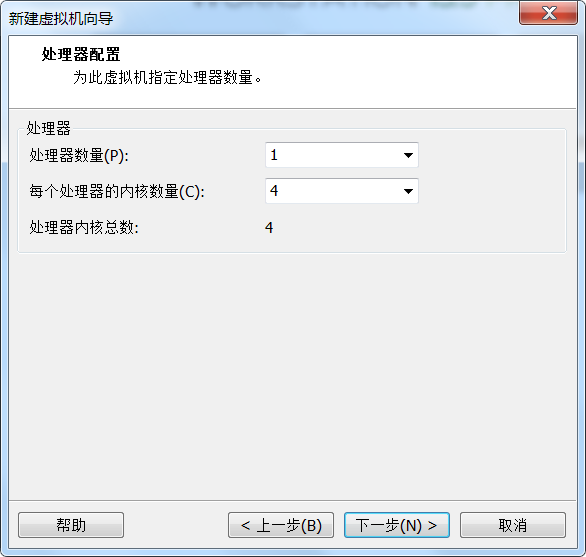
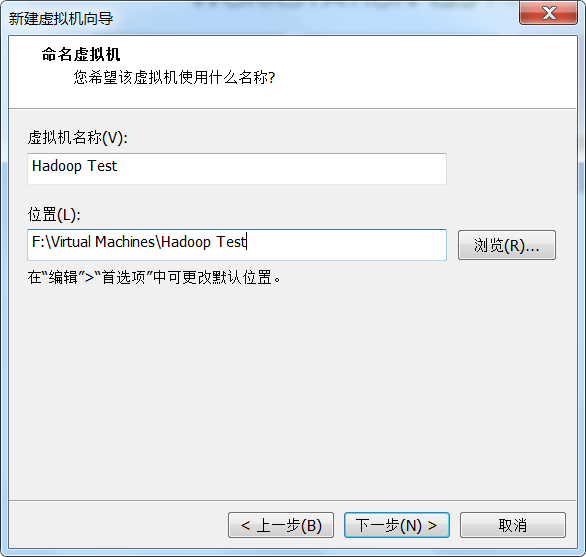
### 打开VMware Workstation Pro→文件→新建虚拟机



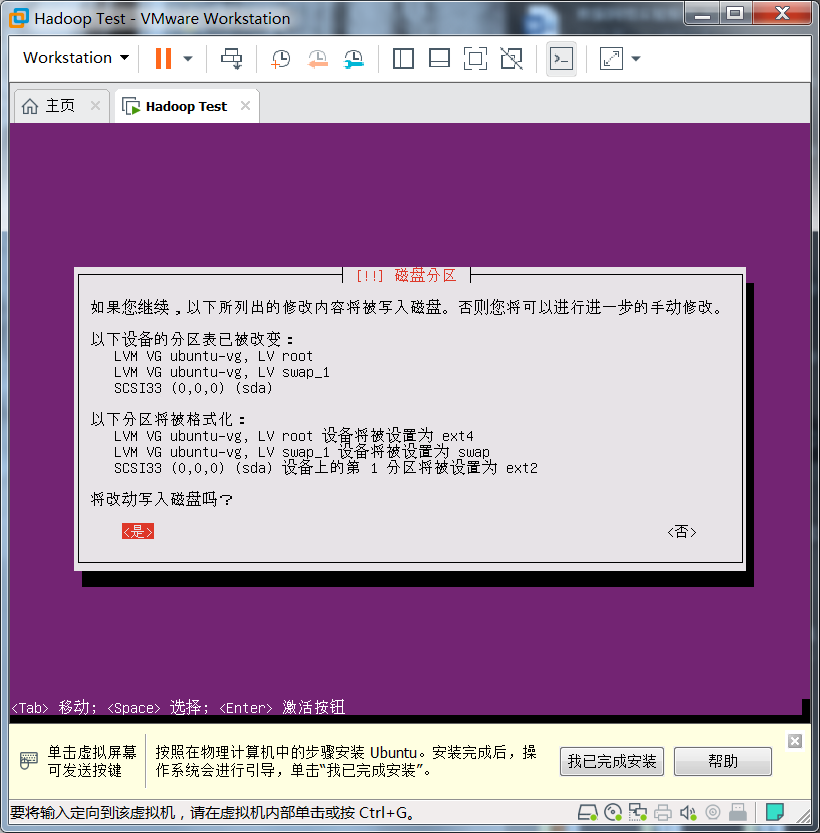
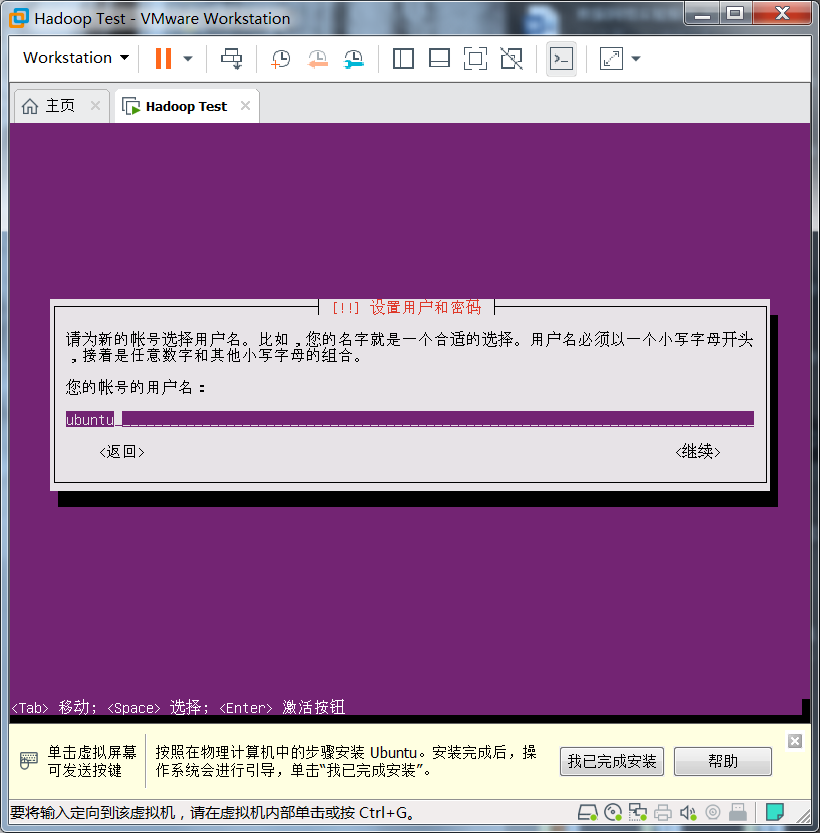
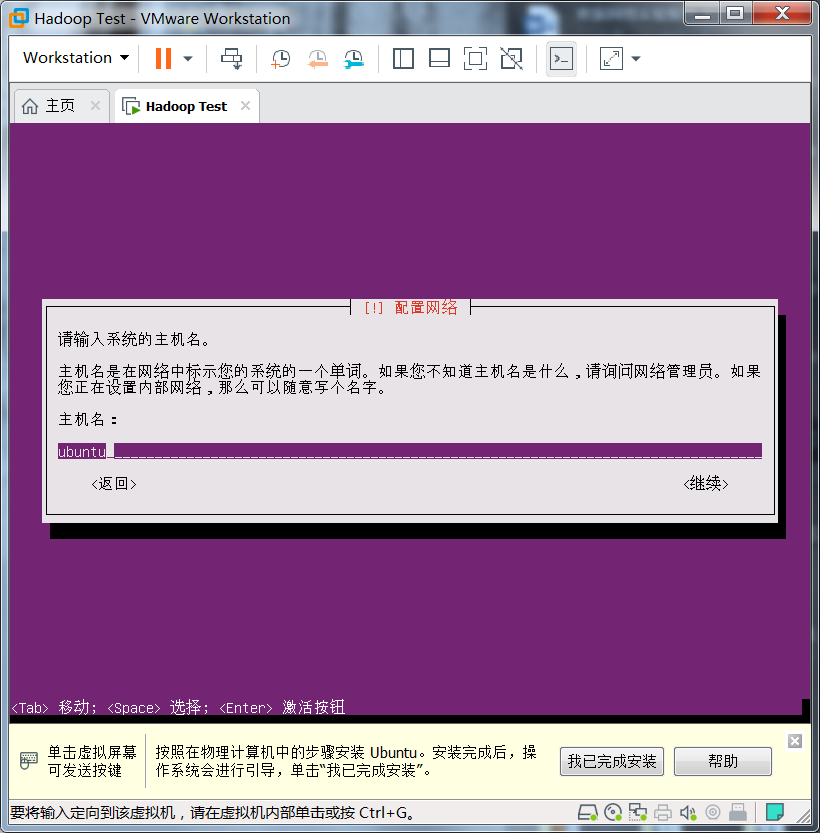
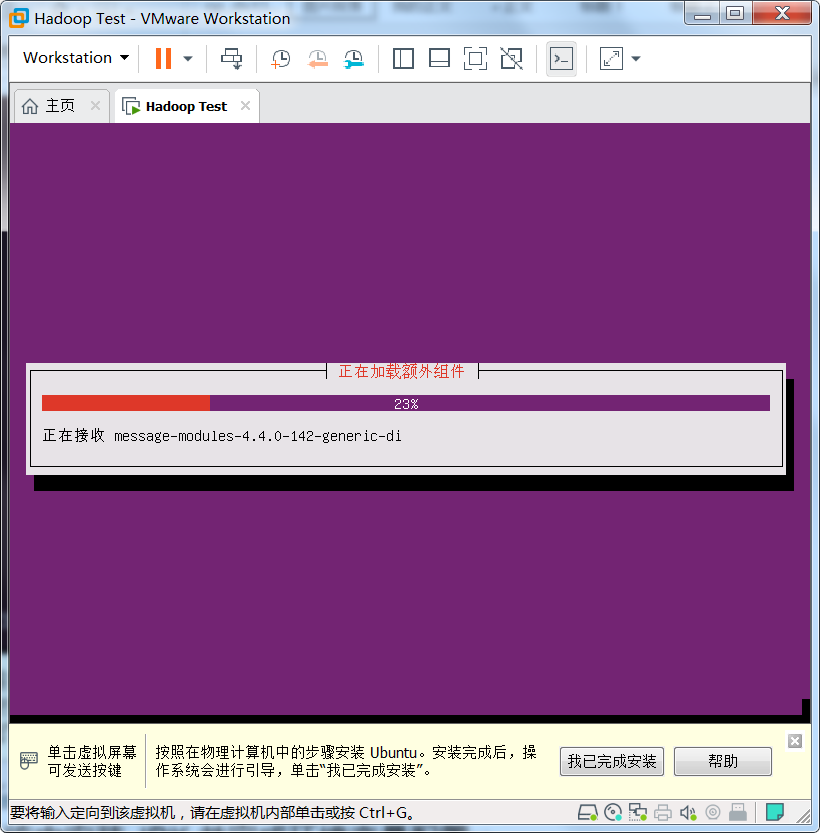
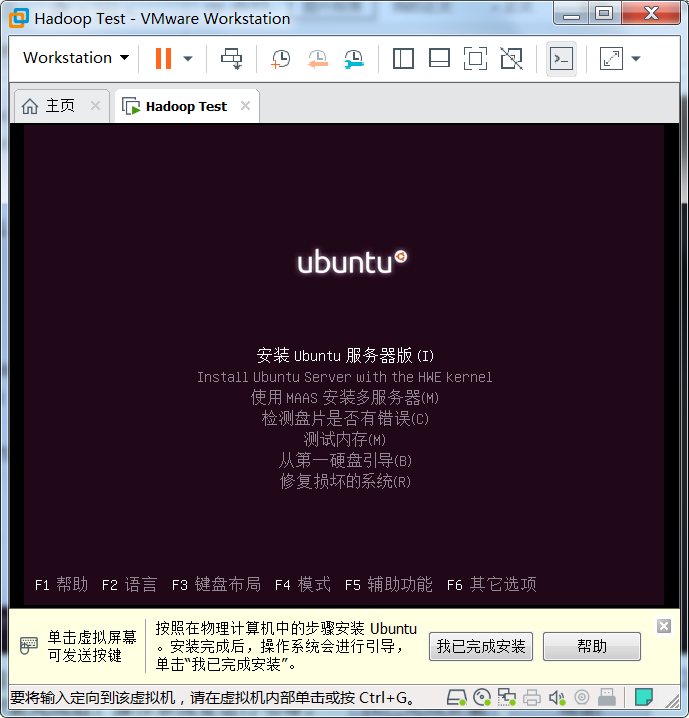
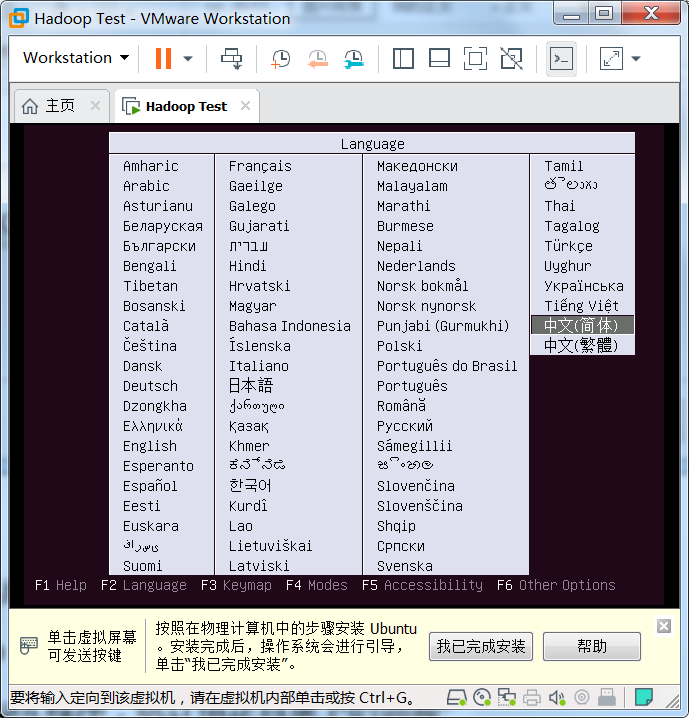
### 选择“稍后安装操作系统”



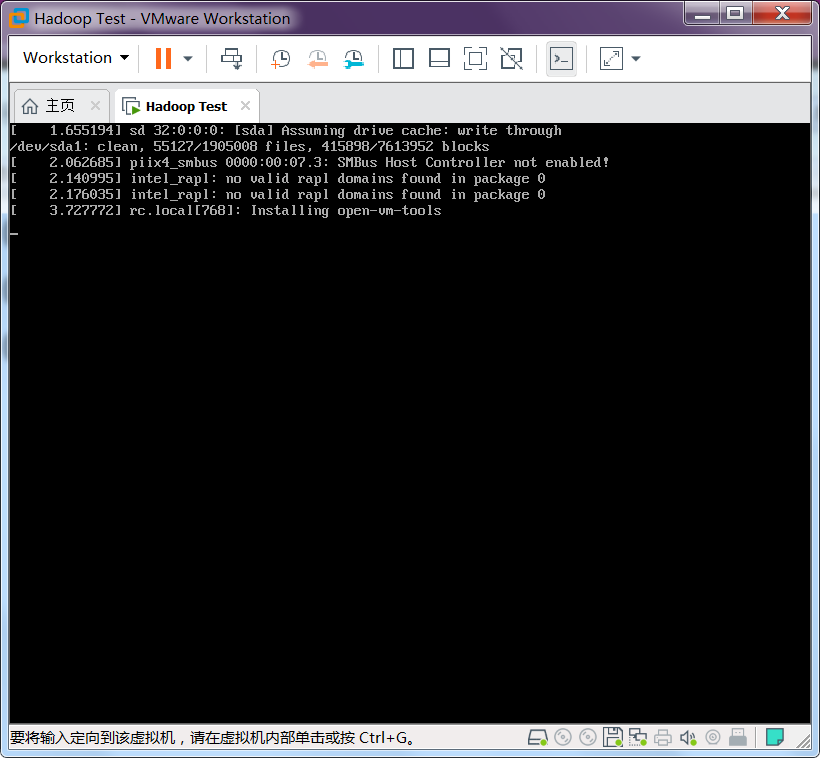
### 设置虚拟机位置和名称



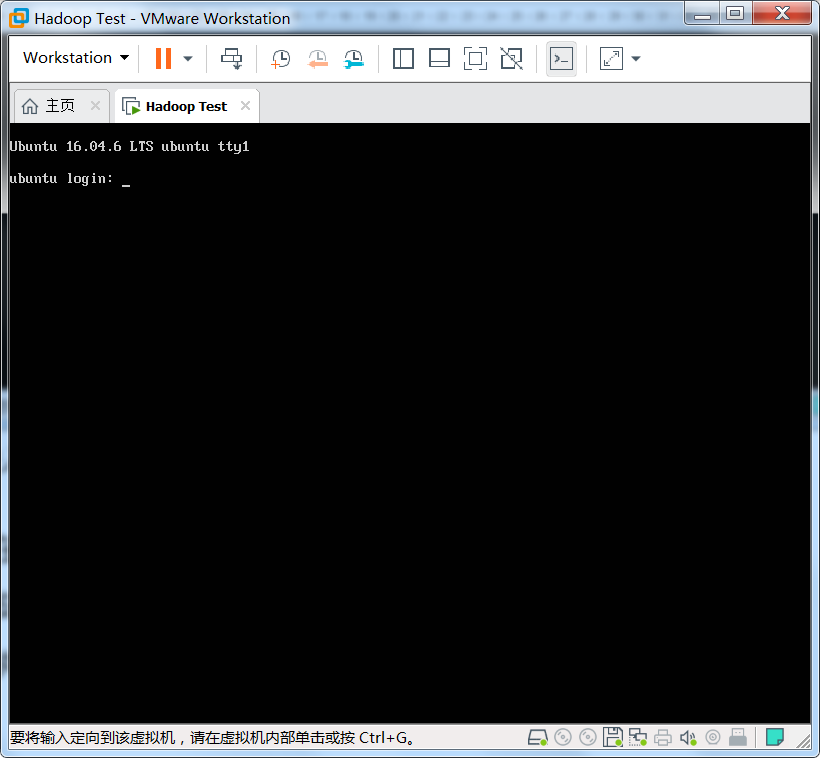
### 开启虚拟机



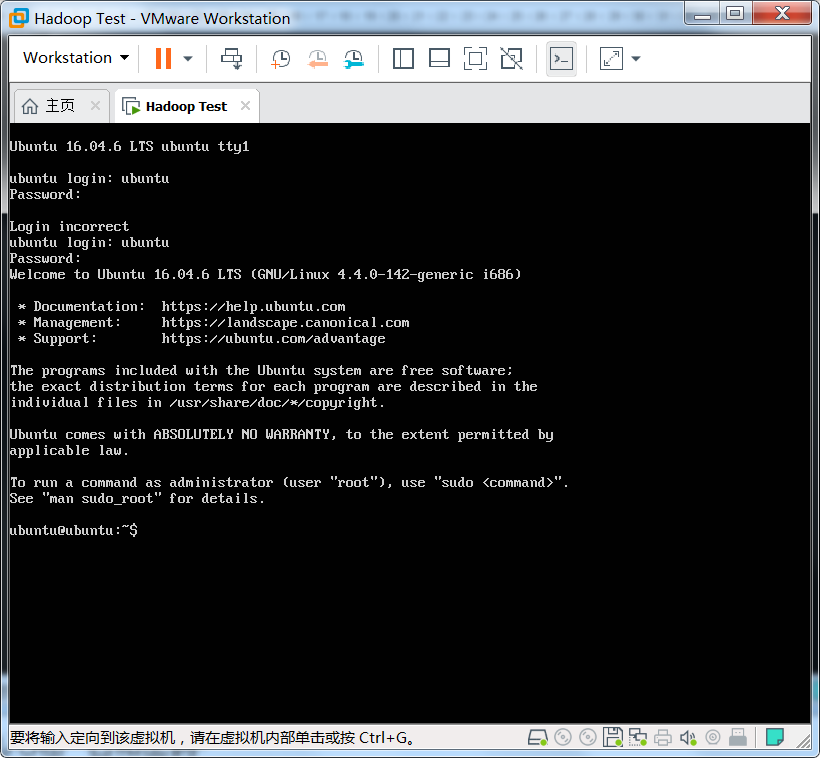
### 安装完成，重新启动



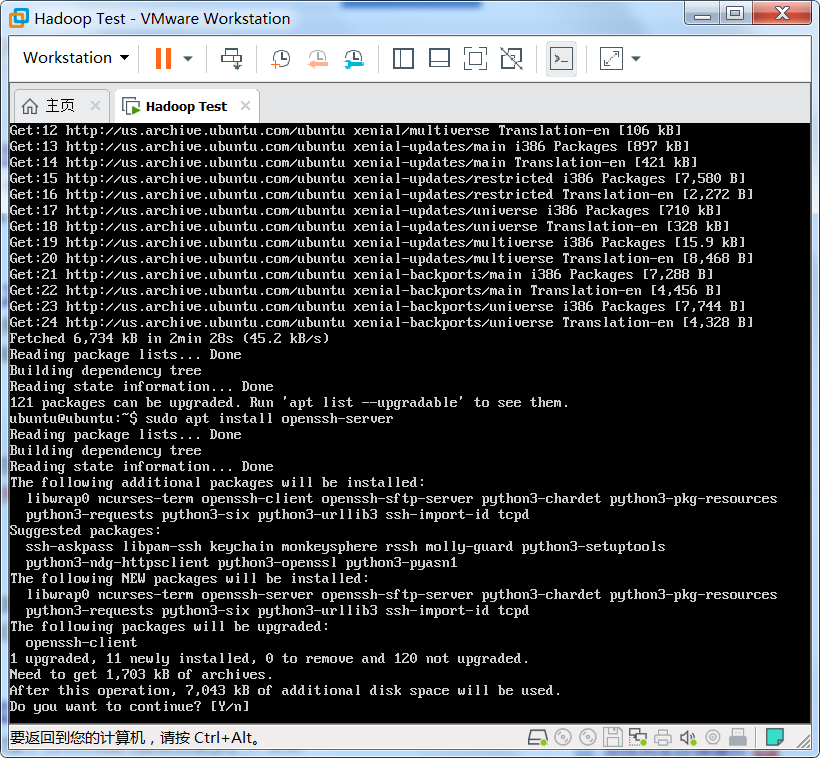
### 进入Ubuntu系统



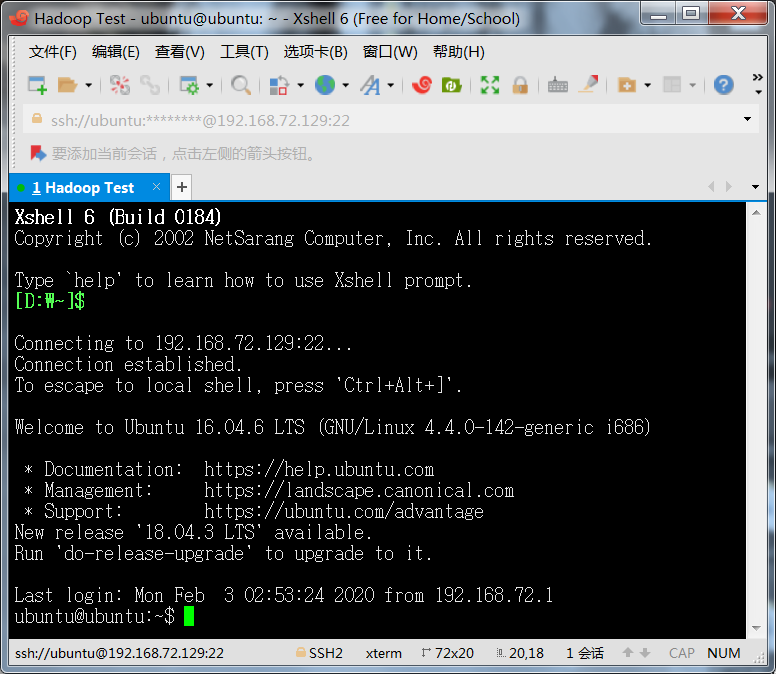
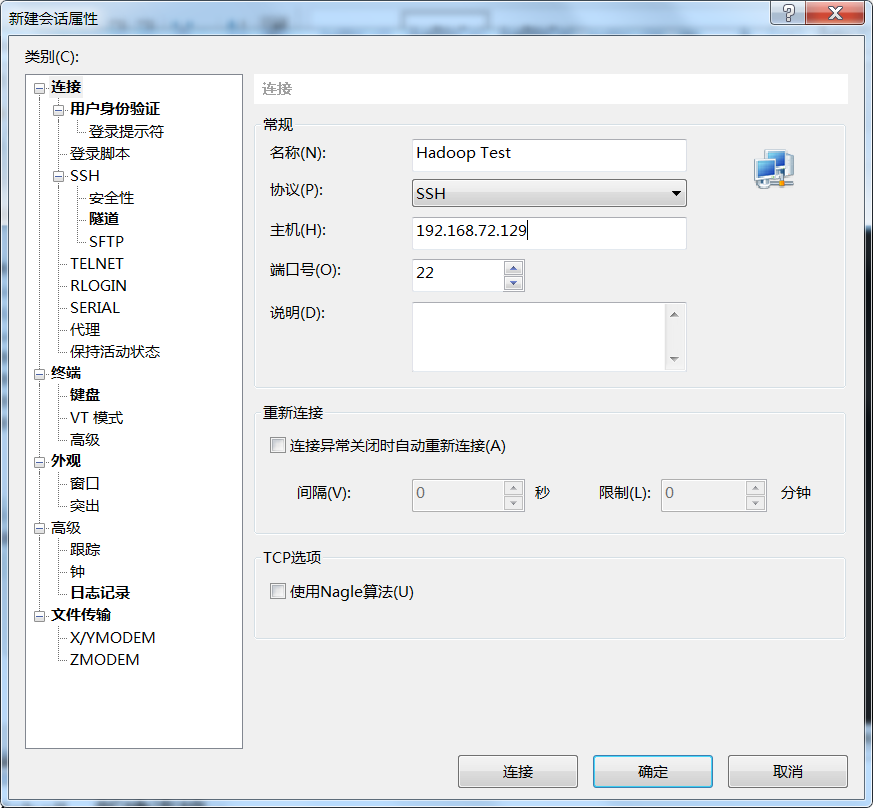
### 输入用户名和密码



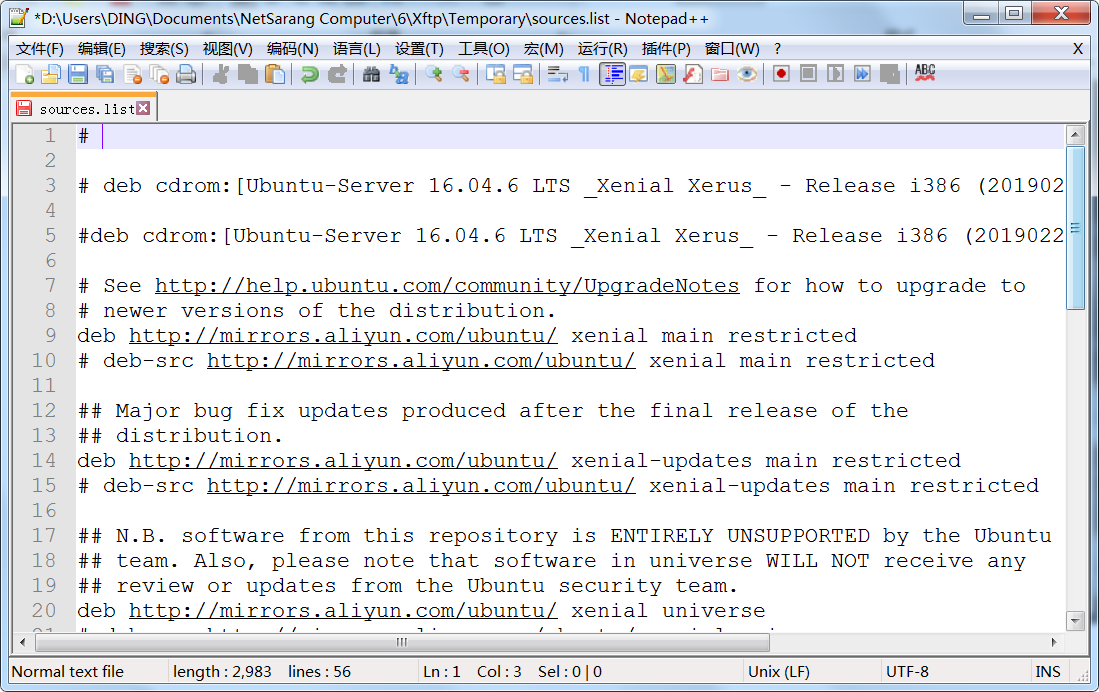
### 安装openssh-server



### 打开Xshell，新建连接（ssh连接：Socket error Event: 32 Error: 10053. - 足各火丁 - 博客园 https://www.cnblogs.com/hjj801006/p/10898355.html）

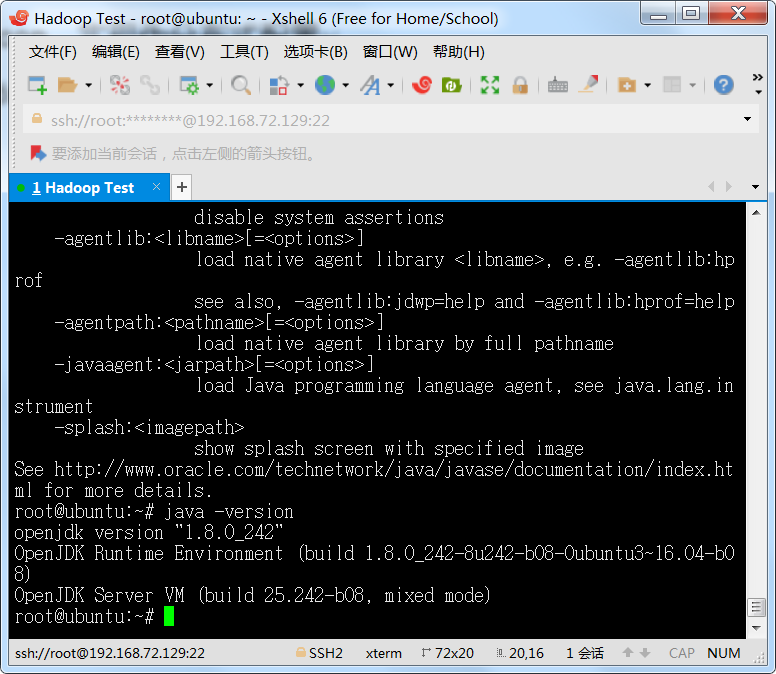


### 将apt源改为国内镜像源



## 在Ubuntu系统中安装JDK并完成环境变量配置

### 安装openjdk（apt install openjdk-8-jdk）



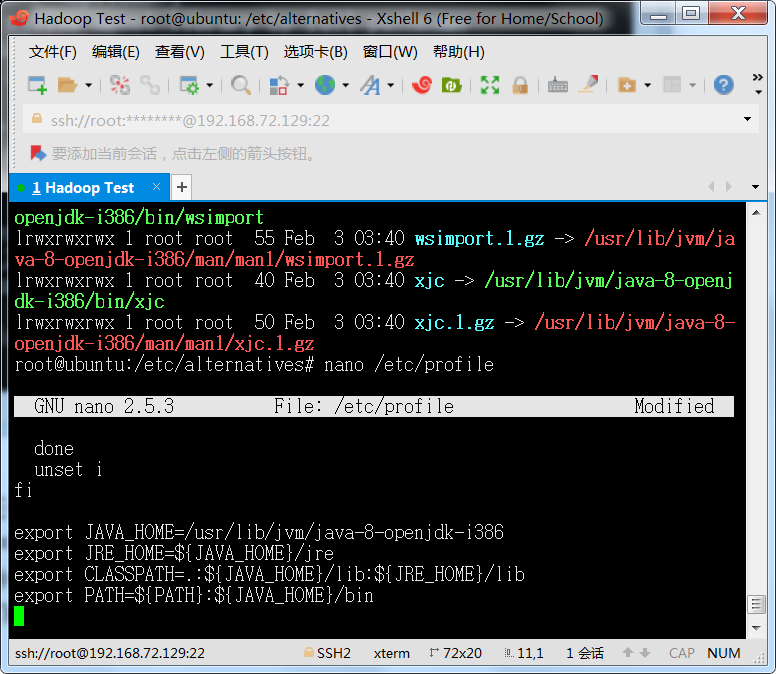
### 配置环境变量

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-i386

export JRE\_HOME=${JAVA\_HOME}/jre

export CLASSPATH=.:${JAVA\_HOME}/lib:${JRE\_HOME}/lib

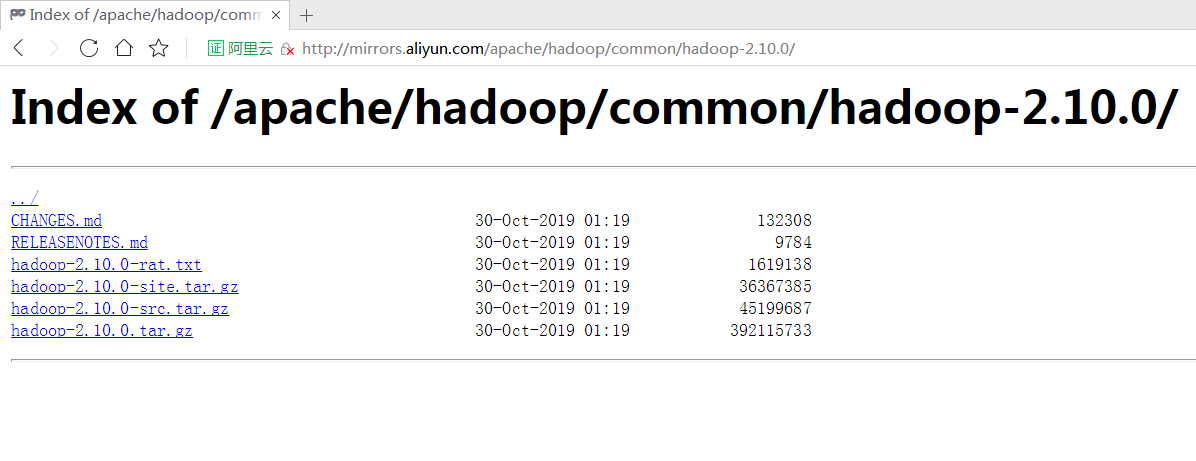
export PATH=${PATH}:${JAVA\_HOME}/bin



## 安装Hadoop，实现伪分布式配置

参考：Hadoop的伪分布式的安装及部署\_xujingran的博客-CSDN博客 https://blog.csdn.net/xujingran/article/details/83898140

### 下载Hadoop



### 执行tar -zxvf hadoop-2.10.0.tar.gz -C /usr/local（将hadoop-2.10.0.tar.gz解压到/usr/local）

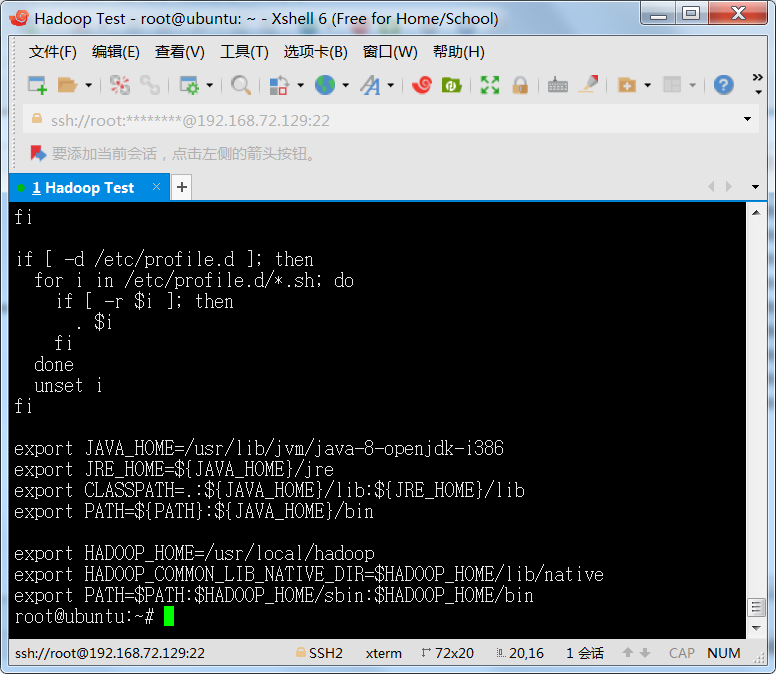
### 执行mv hadoop-2.10.0/ hadoop

### 配置hadoop的环境变量

export HADOOP\_HOME=/usr/local/hadoop

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin



### 修改core-site.xml文件

<configuration>

<property>

<name>hadoop.tmp.dir</name>

<value>file:/usr/local/hadoop/tmp</value>

<description>Abase for other temporary directories.</description>

</property>

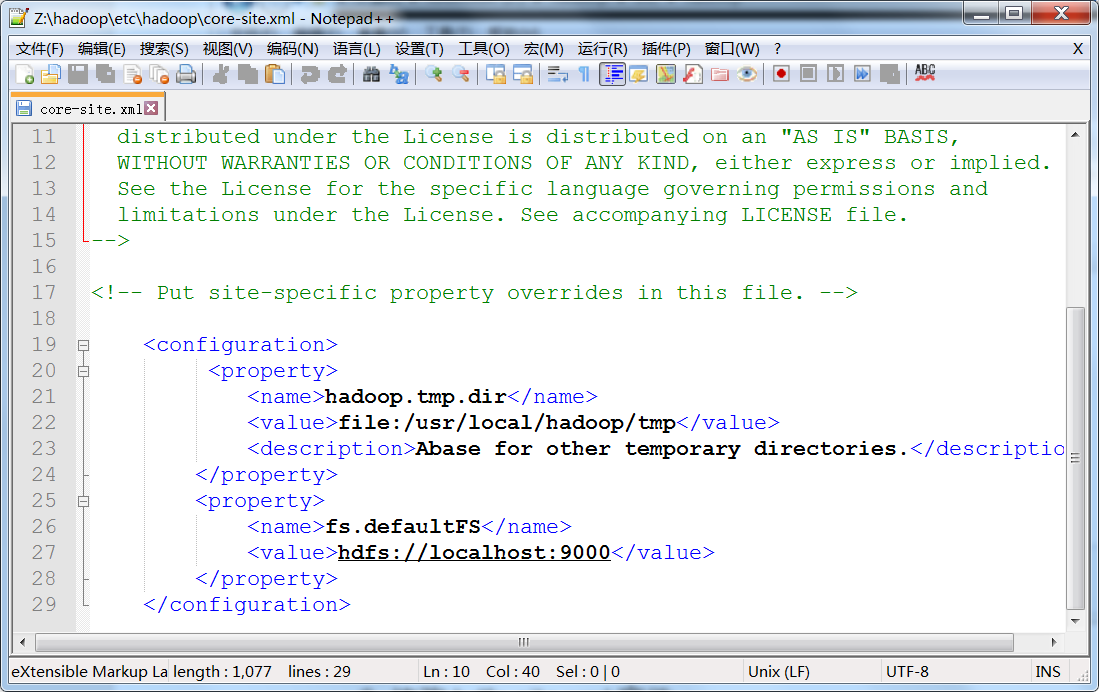
<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>



### 修改hdfs-site.xml文件

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>file:/usr/local/hadoop/tmp/dfs/name</value>

</property>

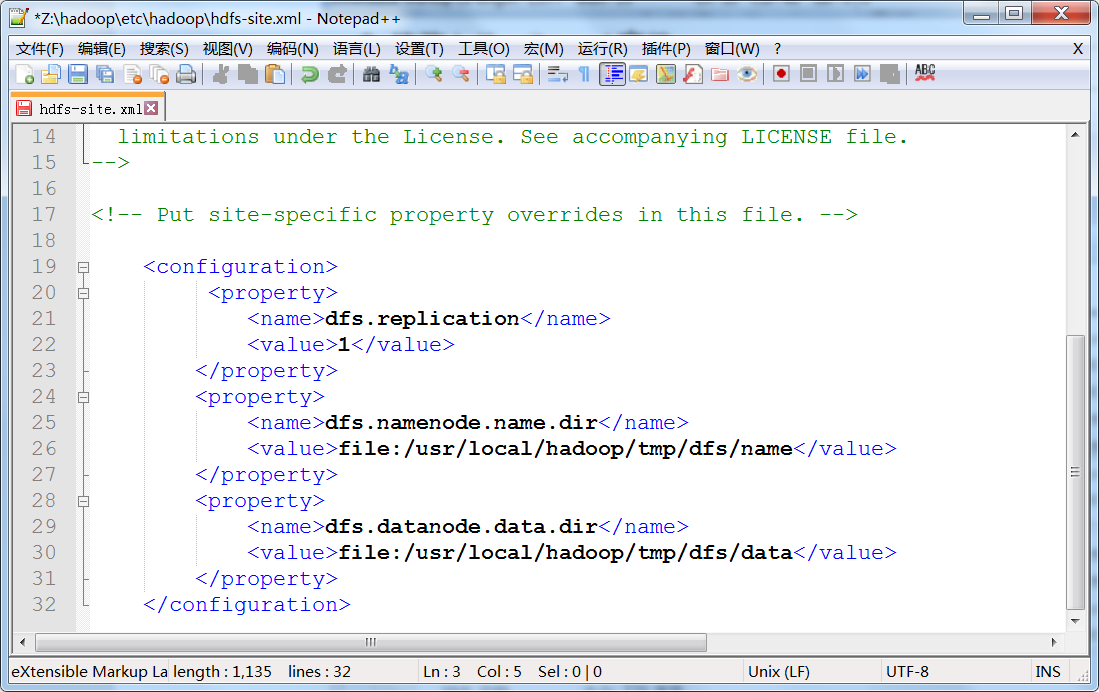
<property>

<name>dfs.datanode.data.dir</name>

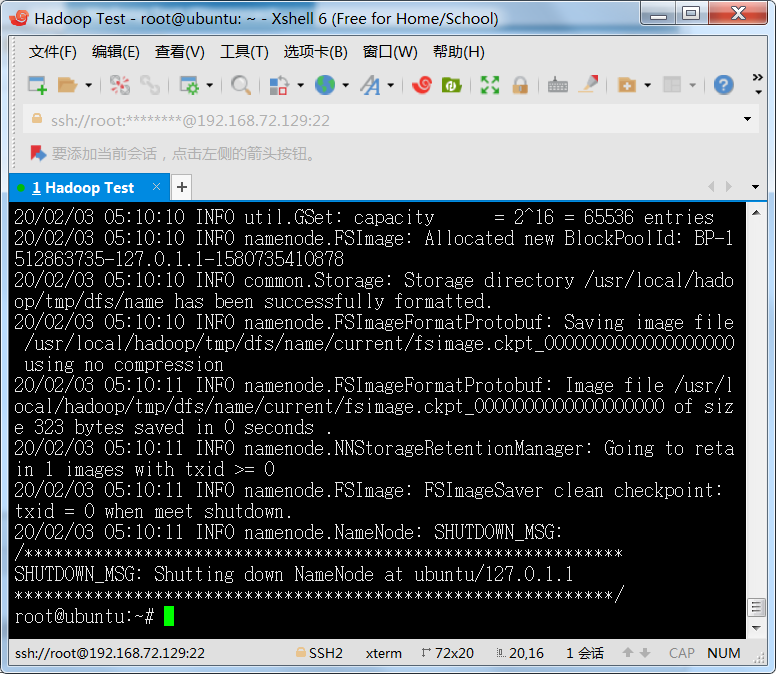
<value>file:/usr/local/hadoop/tmp/dfs/data</value>

</property>

</configuration>

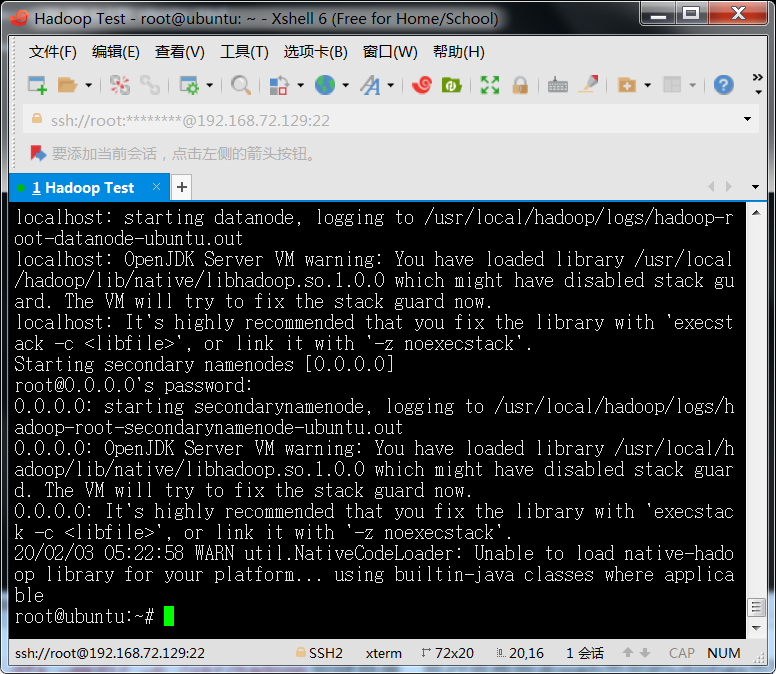


### 将NameNode格式化：执行hdfs namenode -format命令

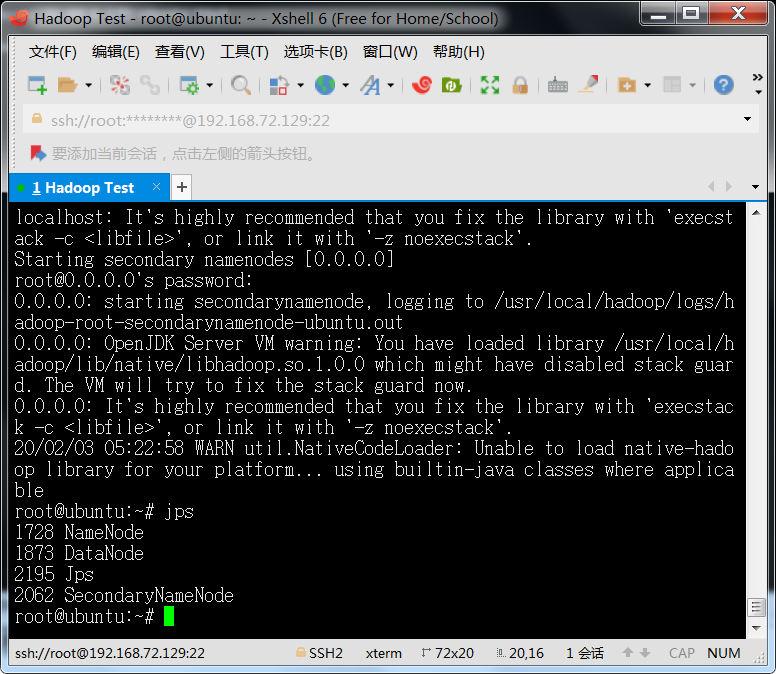


### 开启NameNode进程和DataNode进程，命令为：start-dfs.sh

Hadoop启动报Error: JAVA\_HOME is not set and could not be found解决办法 - 火星十一郎 - 博客园 https://www.cnblogs.com/hxsyl/p/6020879.html



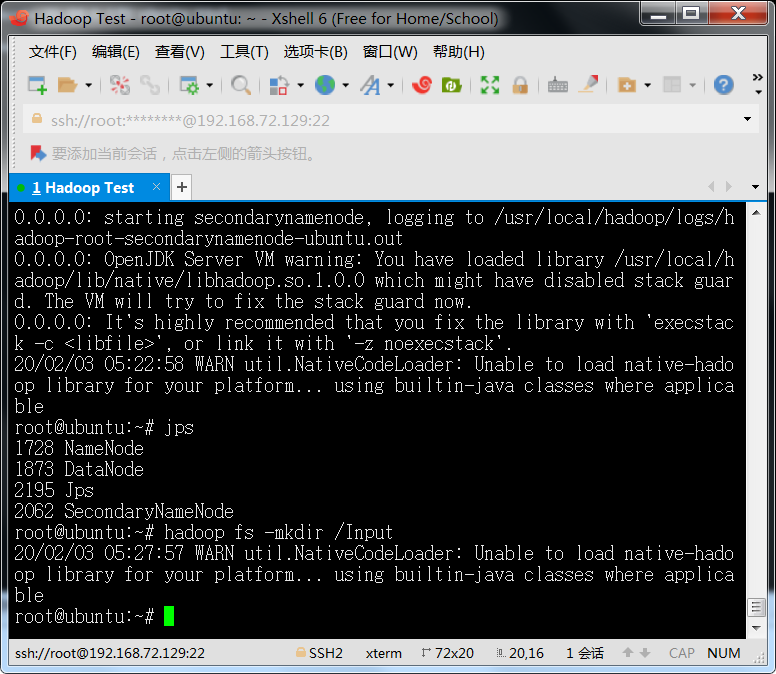
### 执行jps命令



有Jps、SecondaryNameNode、NameNode和DataNode，说明配置成功。

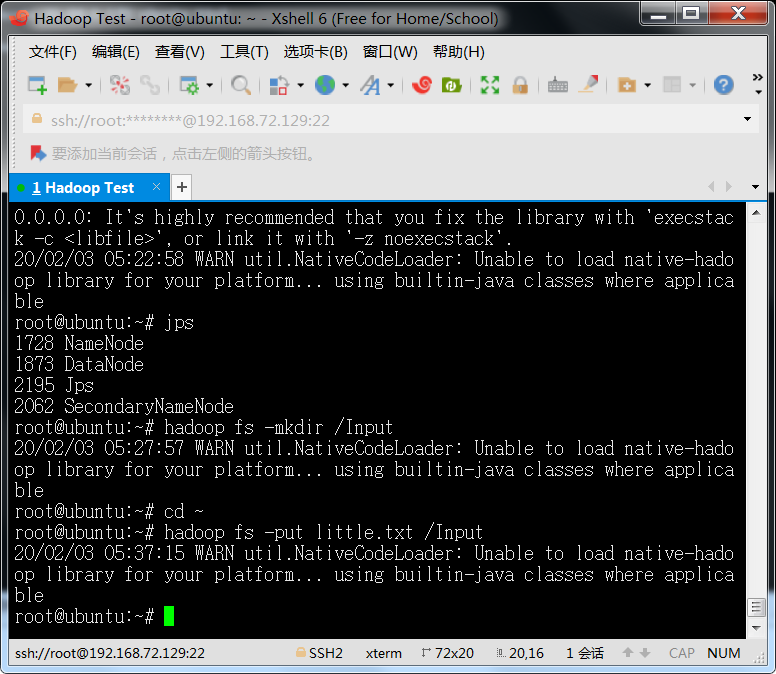
## 执行wordcount程序，统计指定目录下的词频

### 在Hadoop文件系统下创建文件夹Input



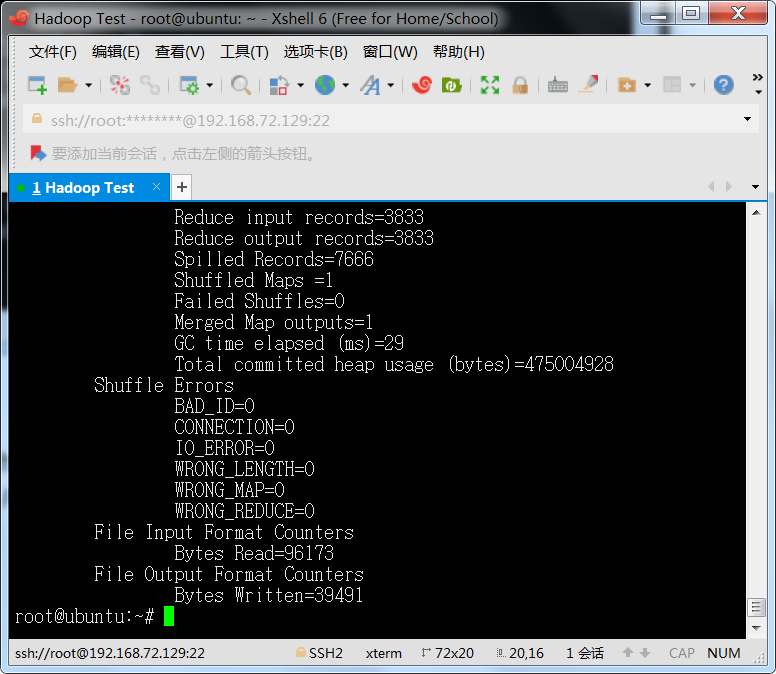
### 将测试文件移动到该文件夹下

hadoop fs -put little.txt /Input



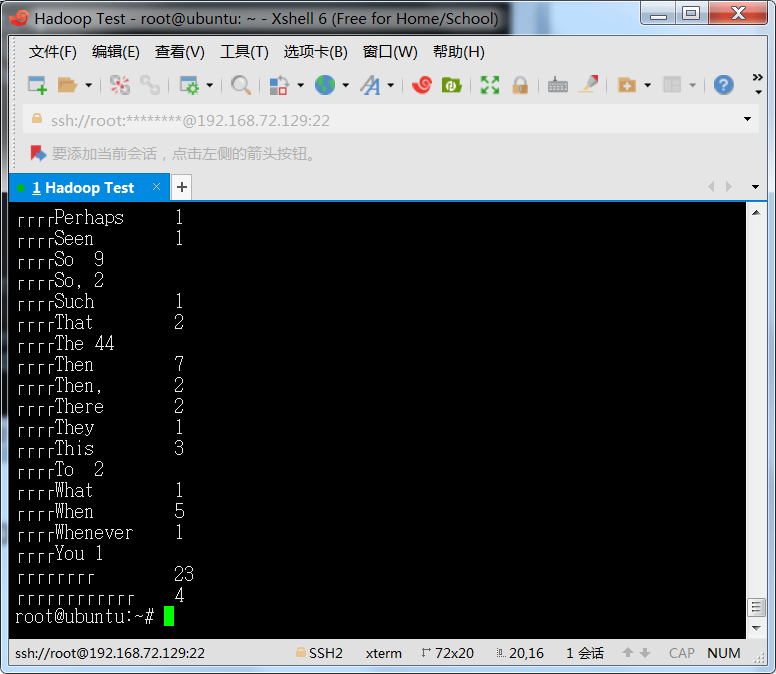
### 运行wordcount

hadoop jar /usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.10.0.jar wordcount /Input /Output



### 查看结果

hadoop fs -cat /Output/part-r-00000



结果表明，测试文件中单词的数量得到了统计，实验成功。