项目二 Hive环境搭建



《Hive数据仓库应用教程》

目 录 / CONTENTS

1 任务2-1 Linux环境搭建

2 任务2-2 Hadoop环境搭建

3 任务2-3 Hive环境搭建

4 任务2-4 Hive体验

任务 2-1 Linux环境搭建

2.1.1 任务描述

本任务将学会如何使用虚拟机工具软件VMware Workstation来创建、克隆和启动Linux系统，同时在虚拟机中对Linux系统的网络、主机名等进行设置，为后续Hadoop和Hive的部署打下基础。







2.1.2 任务准备知识

1. Linux

Linux全称GNU/Linux，是一套免费使用和自由传播的类Unix操作系统，是多用户、多任务、支持多线程和多CPU的操作系统。伴随着互联网的发展，Linux得到了来自全世界软件爱好者、组织、公司的支持。它除了在服务器方面保持着强劲的发展势头以外，在个人电脑、嵌入式系统上都有着长足的进步。

2. VMware Workstation

VMware Workstation是一款功能强大的桌面虚拟计算机软件，提供用户可在单一的桌面上同时运行不同的操作系统，和进行开发、测试 、部署新的应用程序的最佳解决方案。

1．虚拟机的安装与创建

步骤1：单击图2-2中的【创建新的虚拟机】选项，打开【新建虚拟机向导】对话框，选择【自定义（高级）】选项，然后单击【下一步】按钮，如图2-3所示。

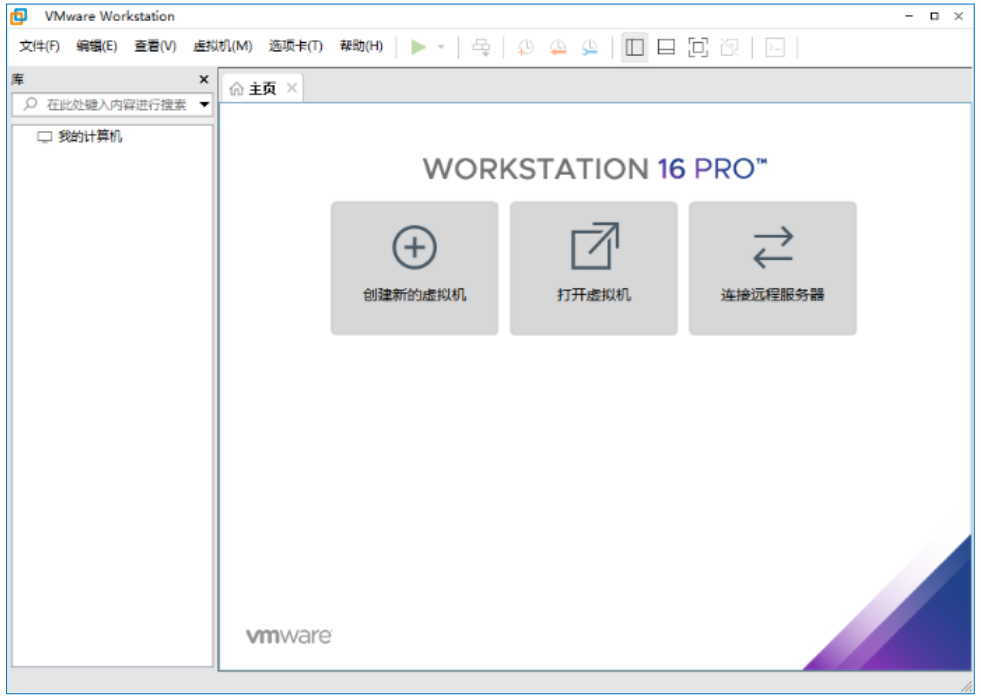


图2-2



图2-3

1．虚拟机的安装与创建

步骤2：选择【稍后安装操作系统】，然后单击【下一步】按钮，如图2-4所示。

步骤3：选择Linux，版本下拉选择【CentOS 7 64位】，然后单击【下一步】按钮，如图2-5所示

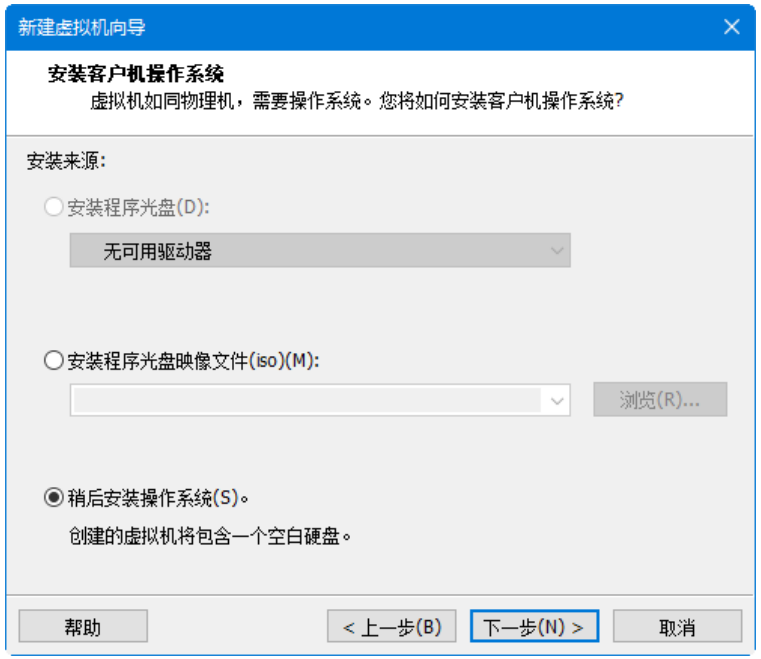


图2-4

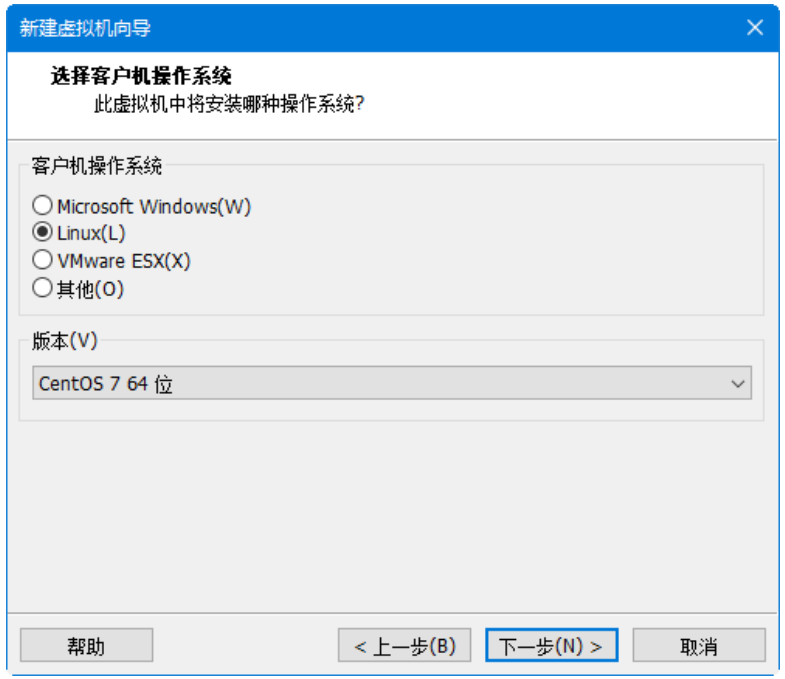


图2-5

2.1.3 任务实施

1．虚拟机的安装与创建

步骤4：填写虚拟机名node1，选择虚拟机文件的位置，然后单击【下一步】按钮，如图2-6所示。

步骤5：选择处理器数量为1，每个处理的内核数量为2，然后单击【下一步】按钮，如图2-7所示。

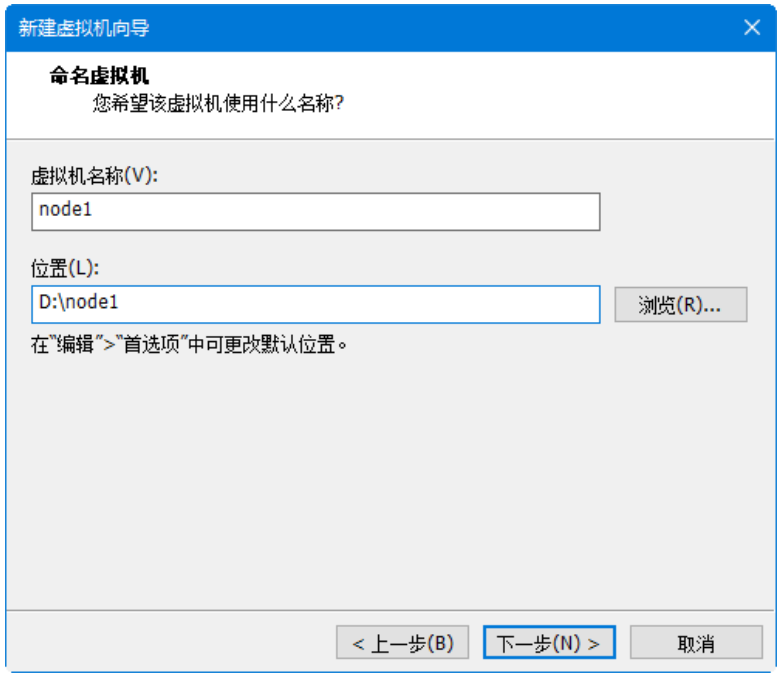


图2-6

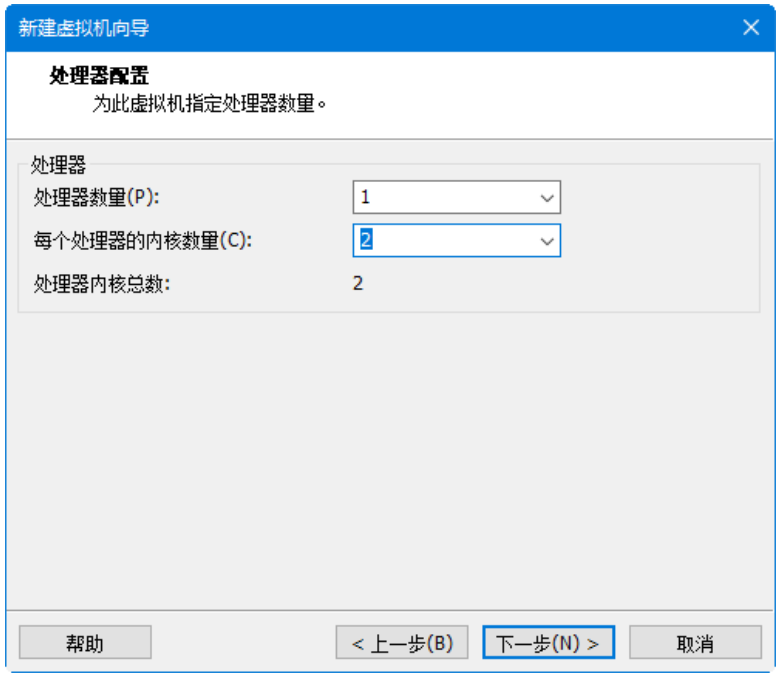


图2-7

1．虚拟机的安装与创建

步骤6：根据自身电脑硬件配置情况，设置内存为2048MB，也即是2GB，然后单击【下一步】按钮，如图2-8所示

步骤7：选择【使用网络地址转换（NAT）】选项，然后单击【下一步】按钮，如图2-9所示。

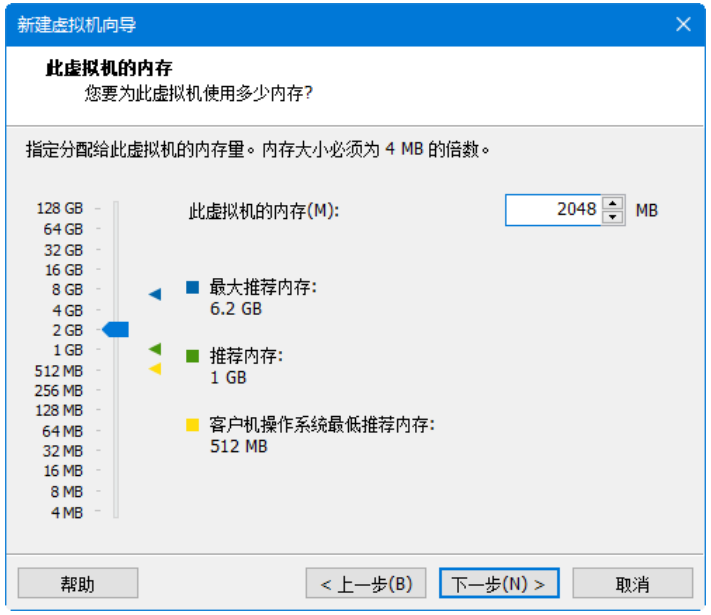


图2-8

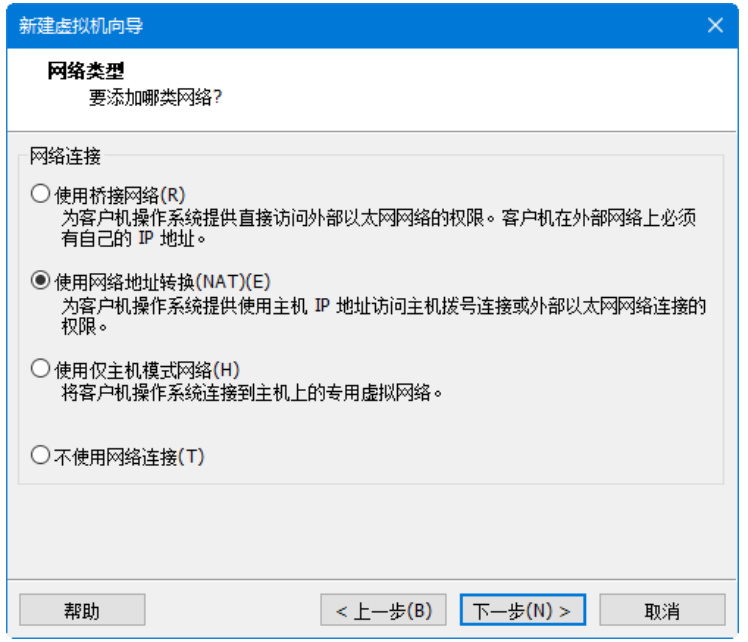


图2-9

1．虚拟机的安装与创建

步骤8：选择默认的【LSI Logic(L)】选项，然后单击【下一步】按钮，如图2-10所示。

步骤9：此处选择默认的【SCSI(S)】选项，然后单击【下一步】按钮，如图2-11示。

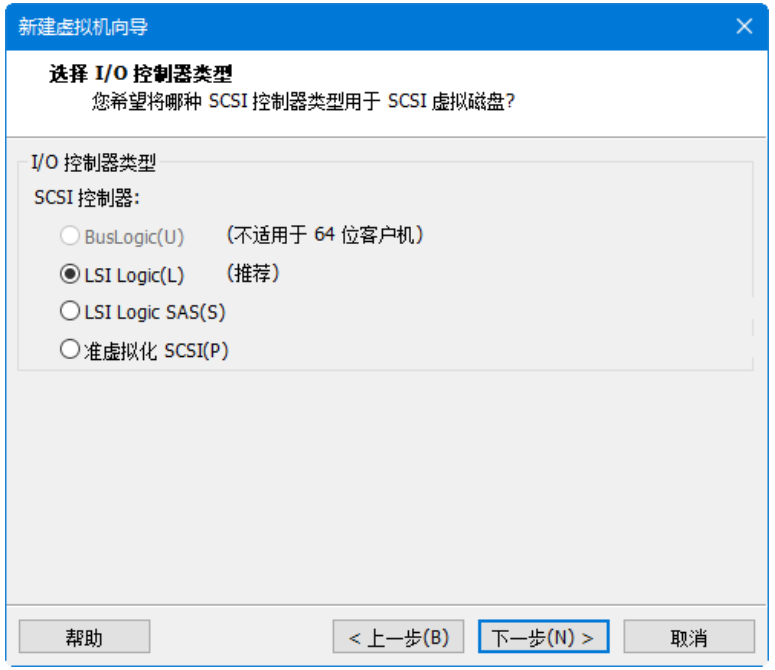


图2-10



图2-11

1．虚拟机的安装与创建

步骤10：选择默认的【创建新虚拟磁盘】选项，然后单击【下一步】按钮，如图2-12所示。。

步骤11：设置最大磁盘大小为20GB，然后单击【下一步】按钮，如图2-13所示。

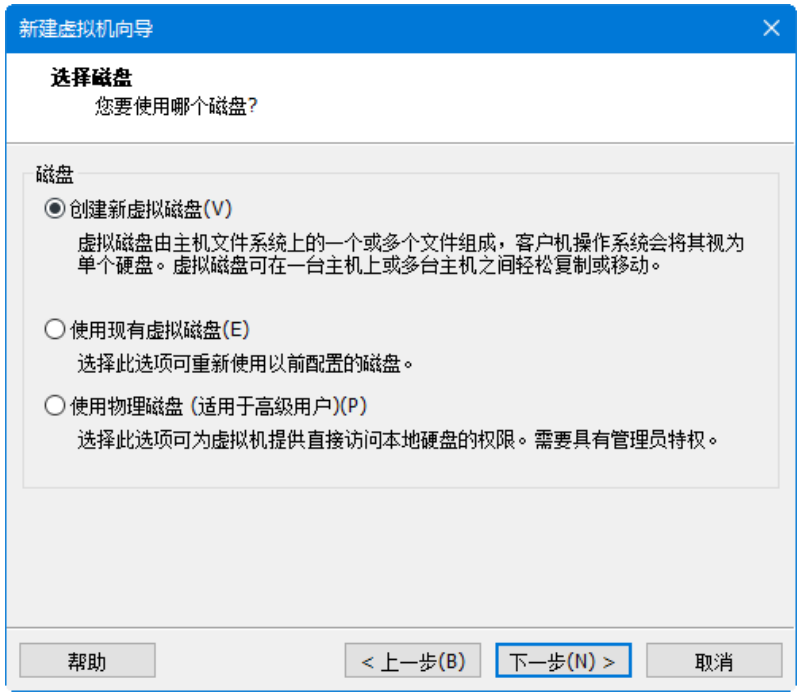


图2-12

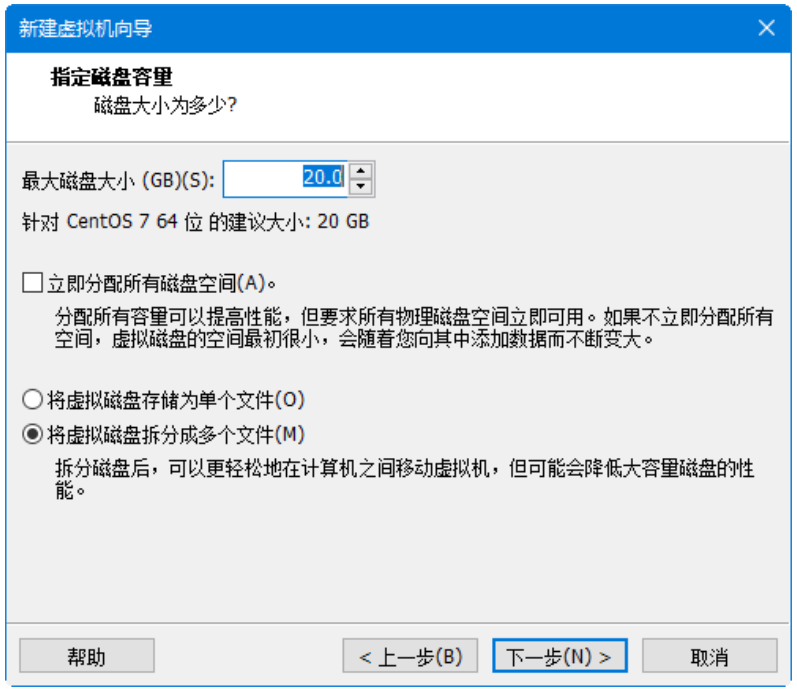


图2-13

1．虚拟机的安装与创建

步骤12：设置磁盘文件名为node1.vmdk，然后单击【下一步】按钮，如图2-14所示。

步骤13：打开【已准备好创建虚拟机】对话框，确认无误后单击【完成】按钮，如图2-15所示。

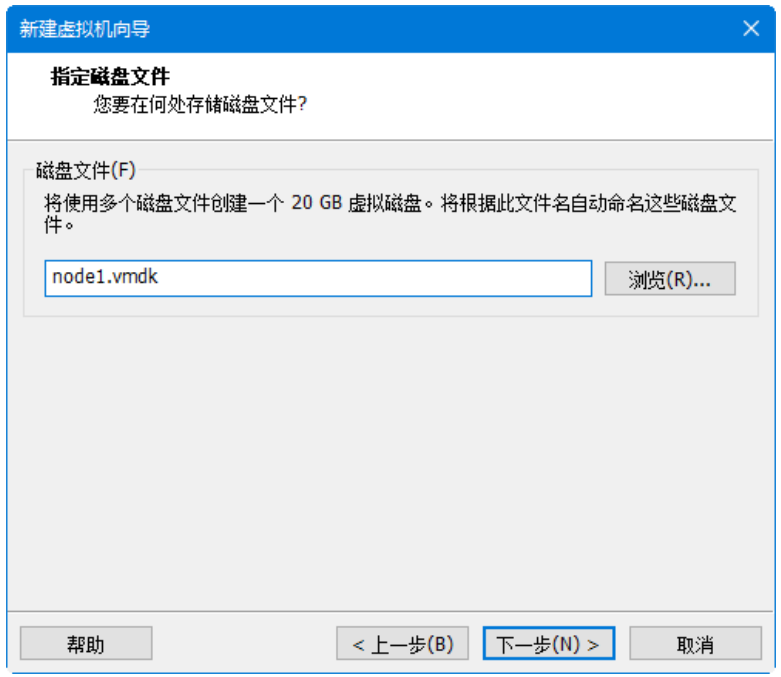


图2-14

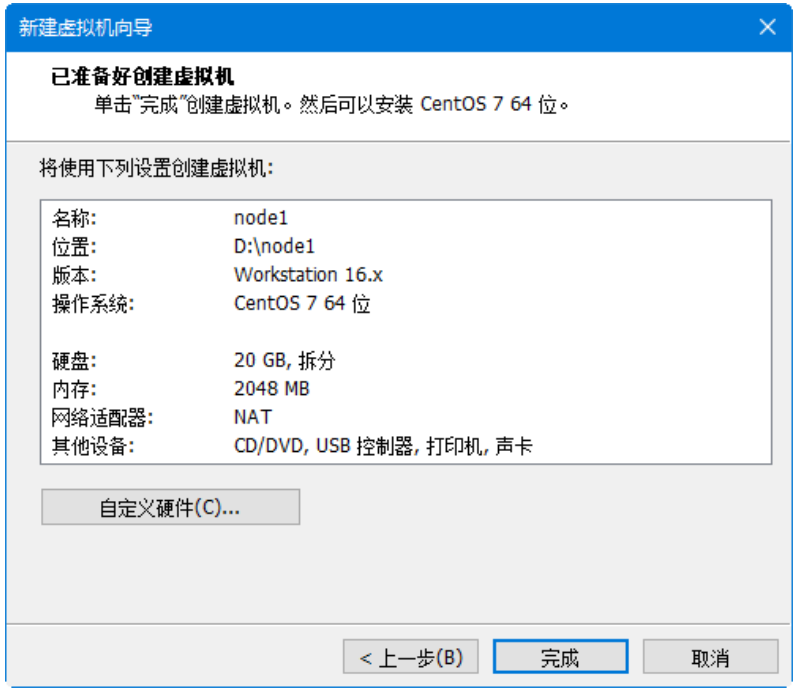


图2-15

1．虚拟机的安装与创建

步骤14：最终完成新建虚拟机的创建，并在VMWare主页右侧打开node1选项卡，其中显示node1虚拟机的配置情况，同时在左侧【我的计算机】下出现node1菜单选项，如图2-16所示。



图2-16

2．Linux安装

步骤1：在上图2-16中点击【编辑选项卡设置】，选择本机Linux的镜像文件，单击【确定】按钮，如图2-17所示。

步骤2：VMWare重新回到了如图2-16的界面，此时选择【开启此虚拟机】，进入安装界面，选择【Install CentOS 7】，按键盘回车键，开始在虚拟机中安装Linux，如图2-18所示。

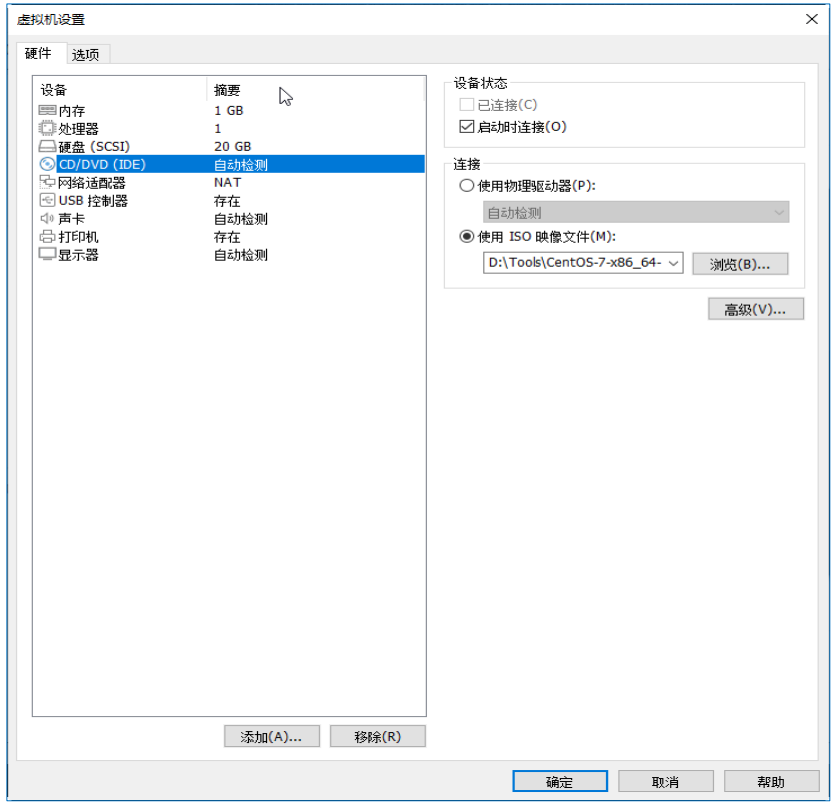


图2-17

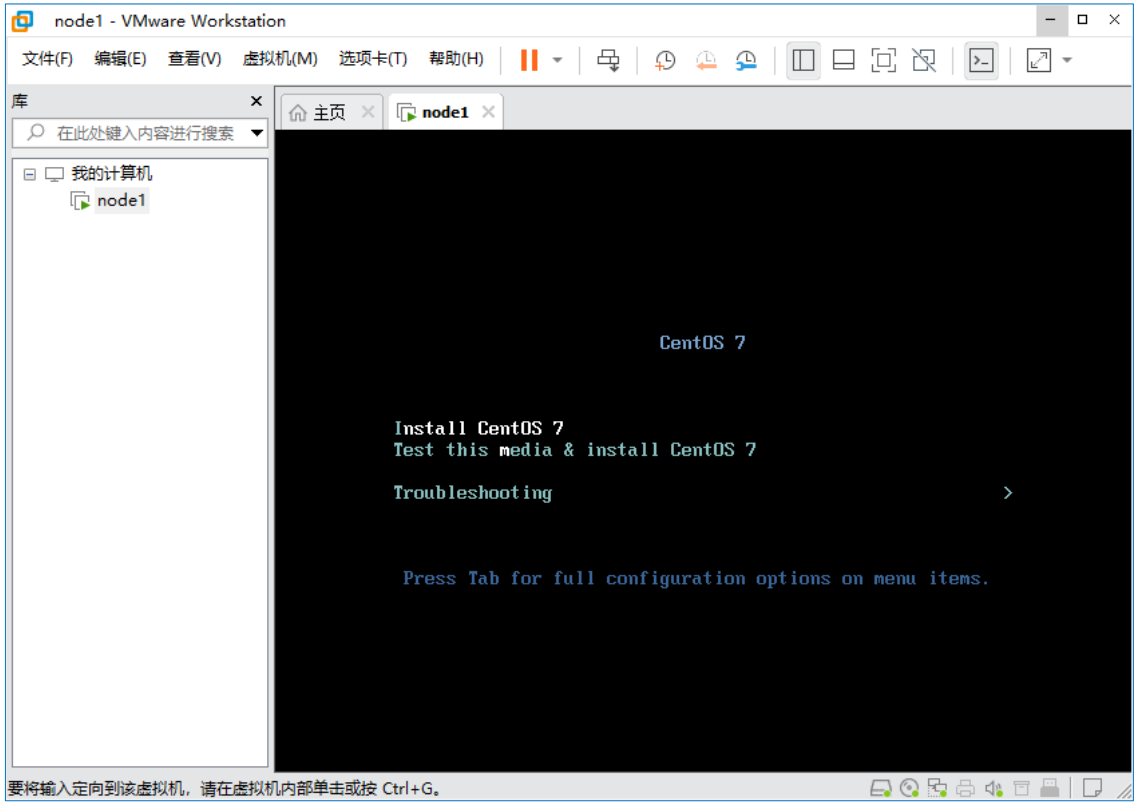


图2-18

2．Linux安装

步骤3：选择语言为默认的English(UnitedStates),然后点击继续【Continue】按钮，如图2-19所示

步骤4：设置时间与时区，点击【DATE&TIME】，如图2-20所示。。

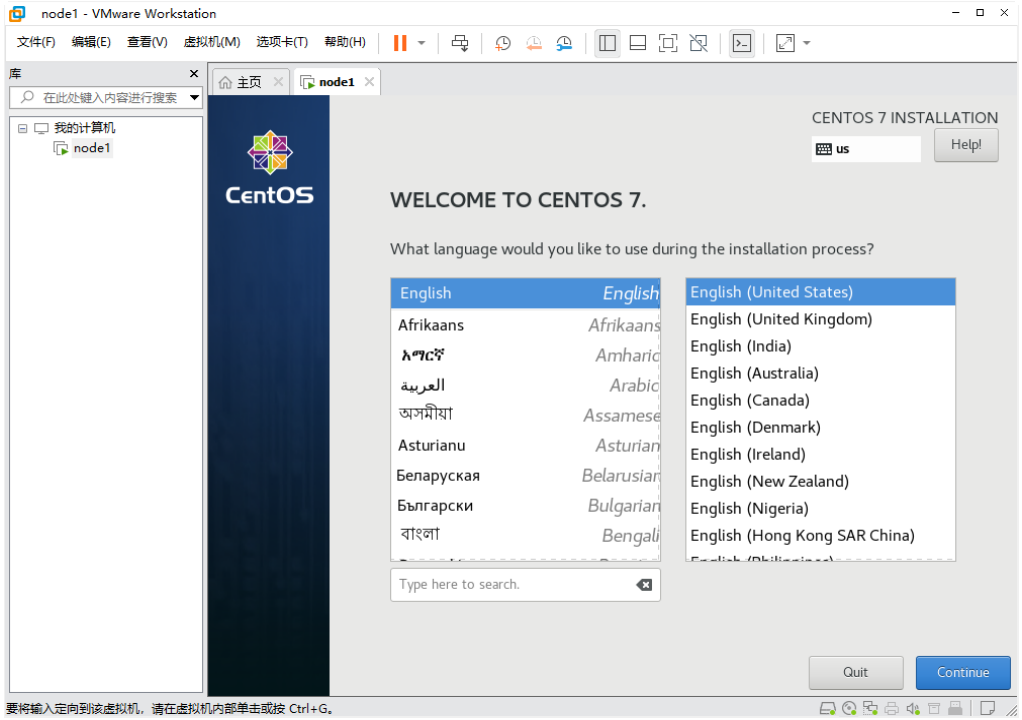


图2-19

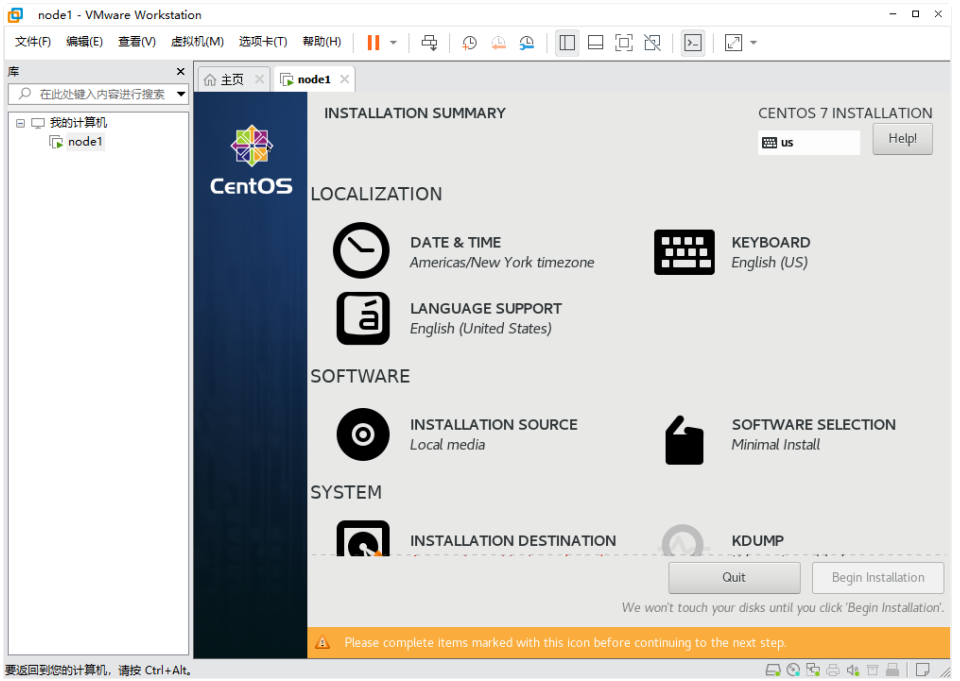


图2-20

2．Linux安装

接步骤4，打开时间与时区界面【DATE&TIME】，选择区域为亚洲【Asia】，城市为上海【Shanghai】，同时可以设置时间与日期，完成设置后点击完成按钮【Done】，如图2-21。

步骤5：回到如图2-20的安装概览界面，选择并打开【SOFTWARE SELECTION】界面，选择默认的最小化安装【Minimal Install】，完成设置后点击完成按钮【Done】，如图2-22所示。

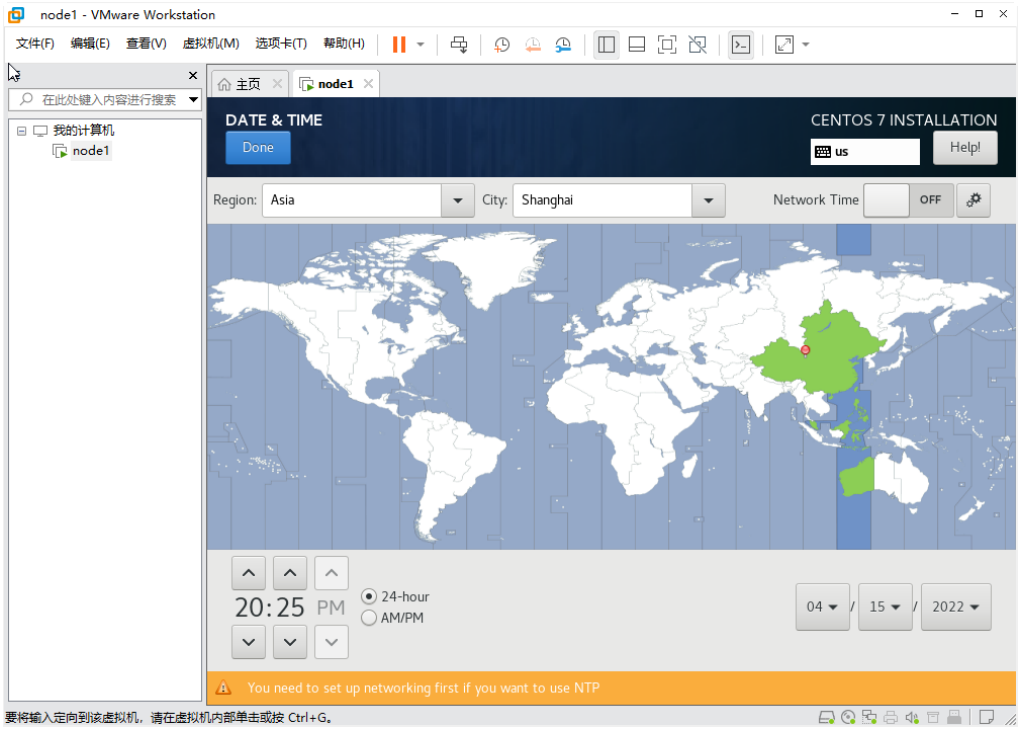


图2-21

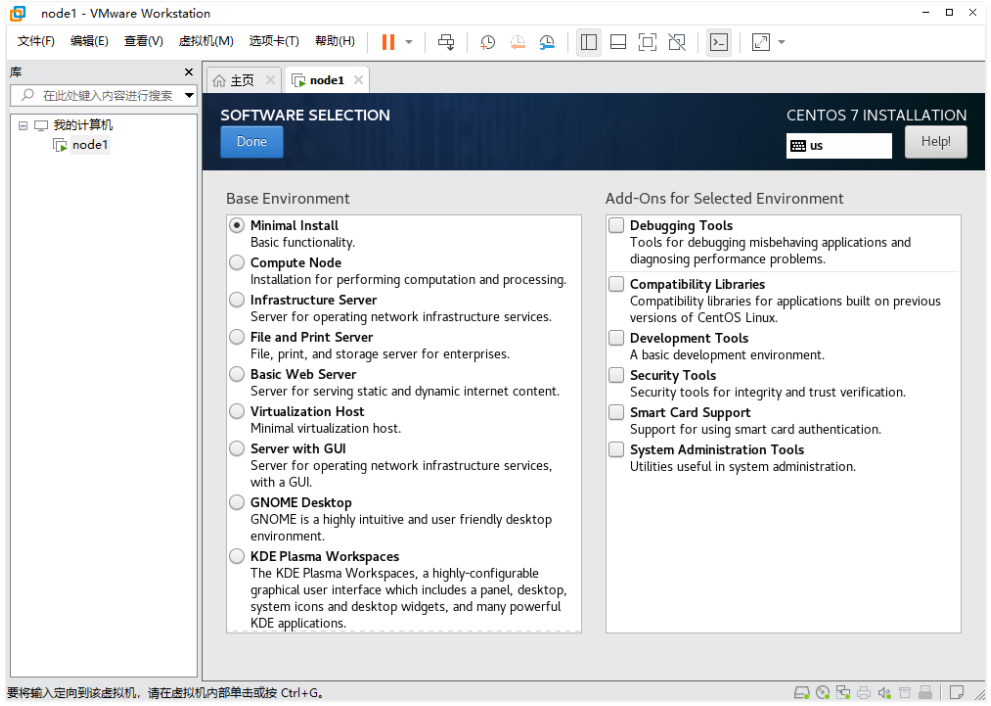


图2-22

2．Linux安装

步骤6：回到如图2-20的安装概览界面，选择并打开【INSTALL DESTINATION】界面，选择默认的自动配置分区【Automatically configure partitioning】，然后点击【Done】，如图2-23所示。

步骤7：回到如图2-20的安装概览界面，选择并打开【NETWORK & HOST NAME】界面，开启网络为【ON】状态，设置主机名为【node1】，点击【Apply】按钮，完成设置后点击完成按钮【Done】，如图2-24所示。

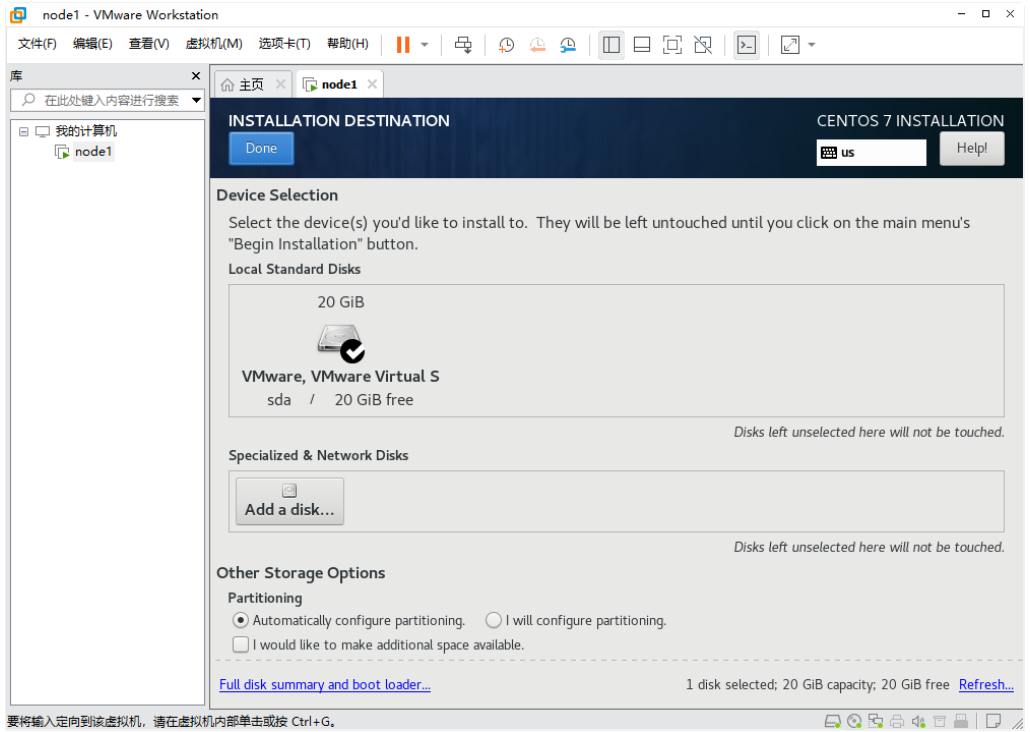


图2-23

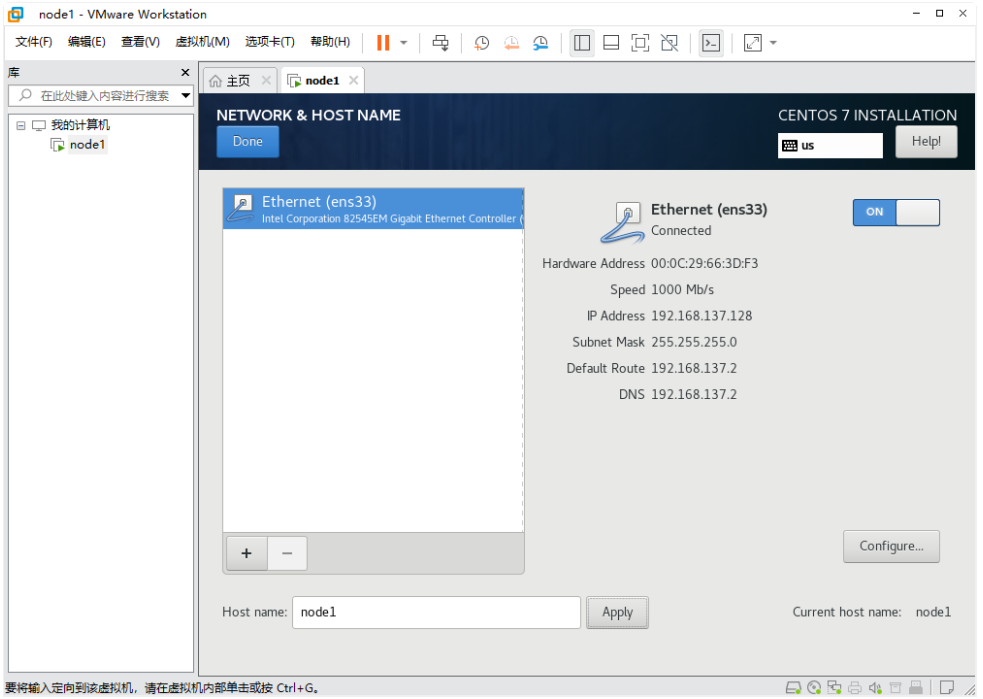


图2-24

2．Linux安装

步骤8：完成配置后点击开始安装按钮【Begin Installation】，如图2-25所示。

步骤9：安装过程中设置root密码，点击【ROOTPASSWORD】，如图2-26所示。

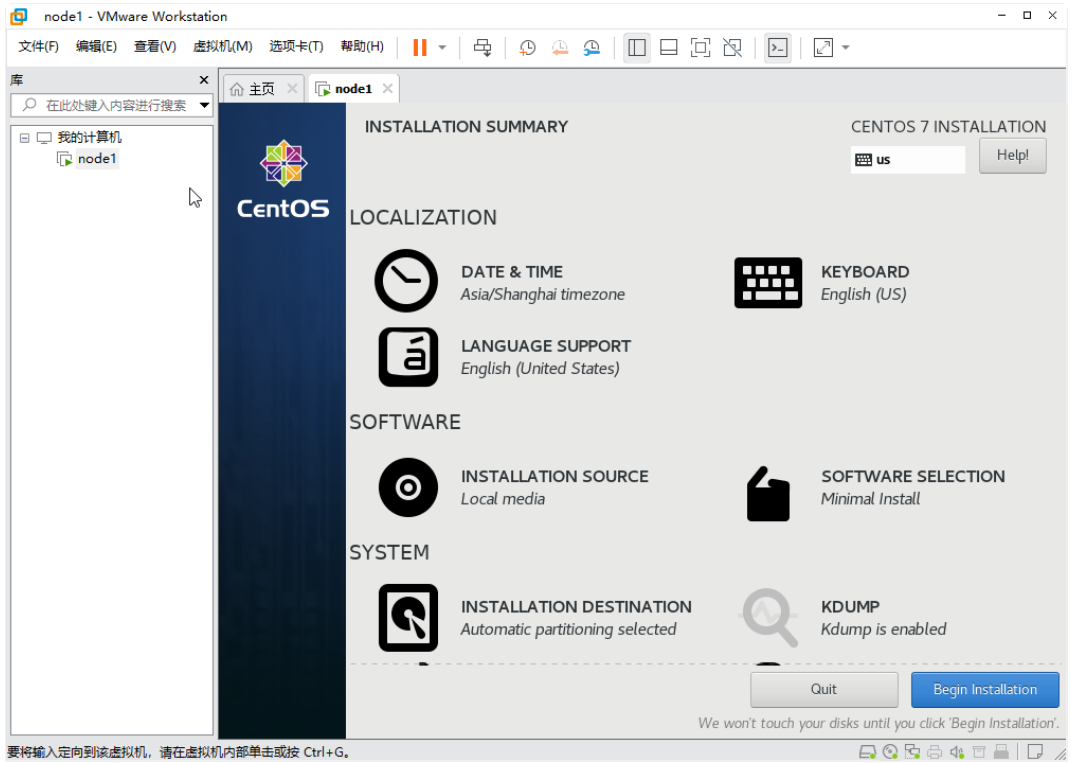


图2-25

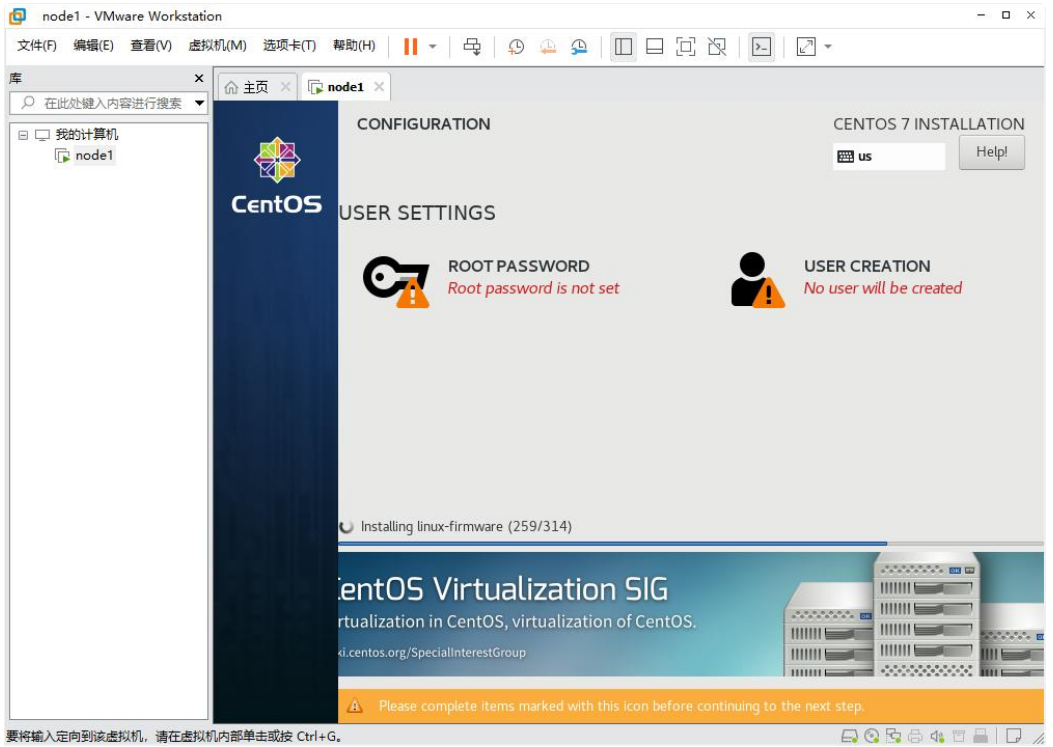


图2-26

2．Linux安装

接步骤9，设置密码123456，以及相同的确认密码，设置后点击【Done】，如图2-27所示。

步骤10：安装完成后在【CONFIGURATION】界面中点击【Reboot】按钮，将重启Linux系统，如图2-28所示

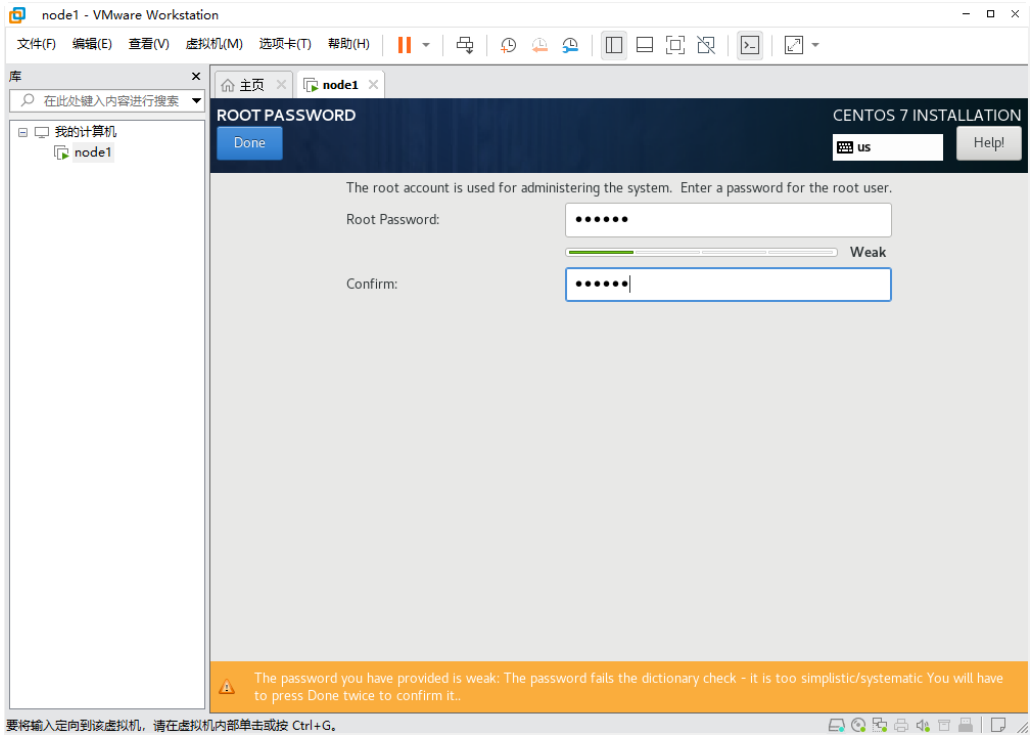


图2-27

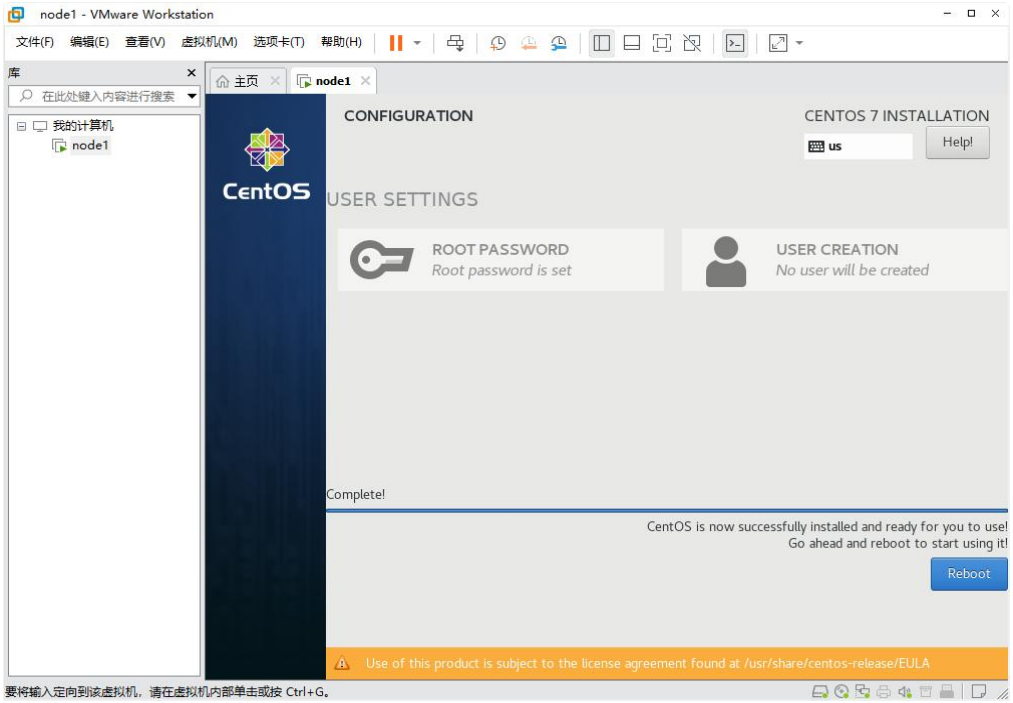


图2-28

2．Linux安装

步骤11：重启Linux系统后输入用户名root和密码123456，出现Linux的提示符[root@node1 ~]#，如图2-29所示，安装Linux成功。

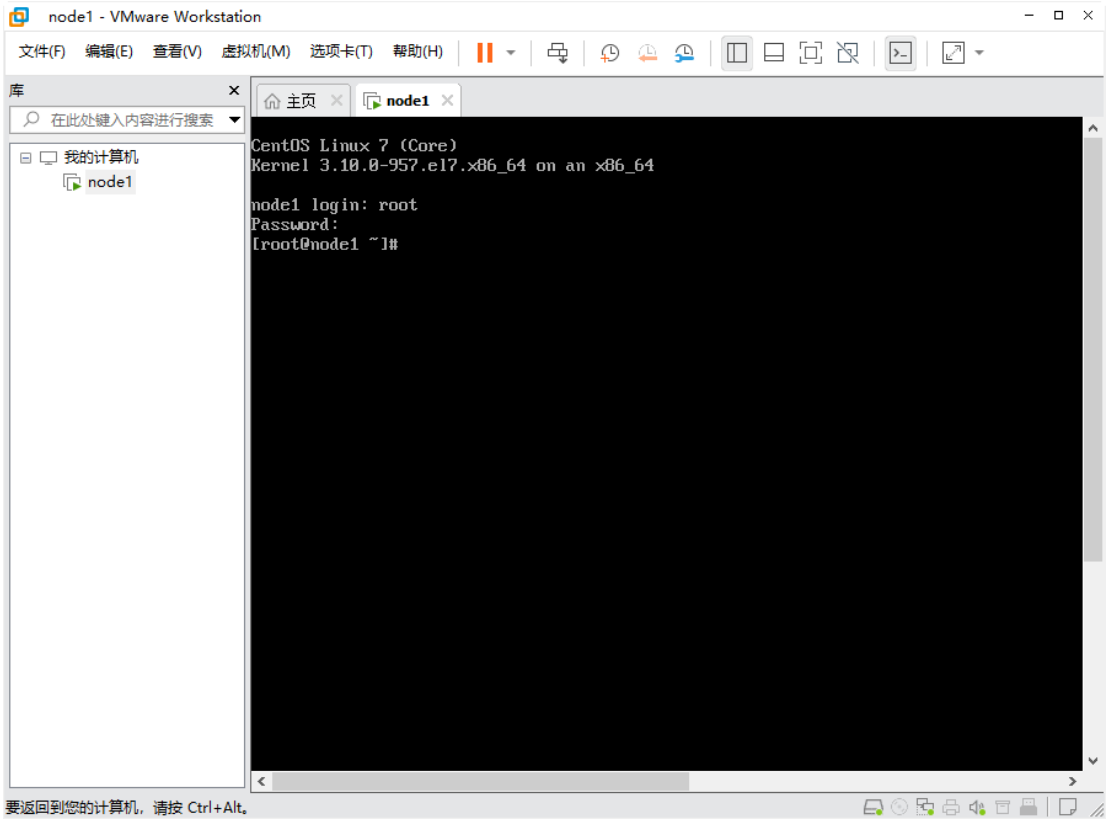


图2-29

3．Linux克隆

步骤1：在如图选中的node1，快捷键【Ctrl+E】关闭虚拟机，点击【关机】在如图2-30所示，

步骤2：右键node1虚拟机，弹出快捷菜单，选择【管理】→【克隆】菜单，在如图2-31所示，

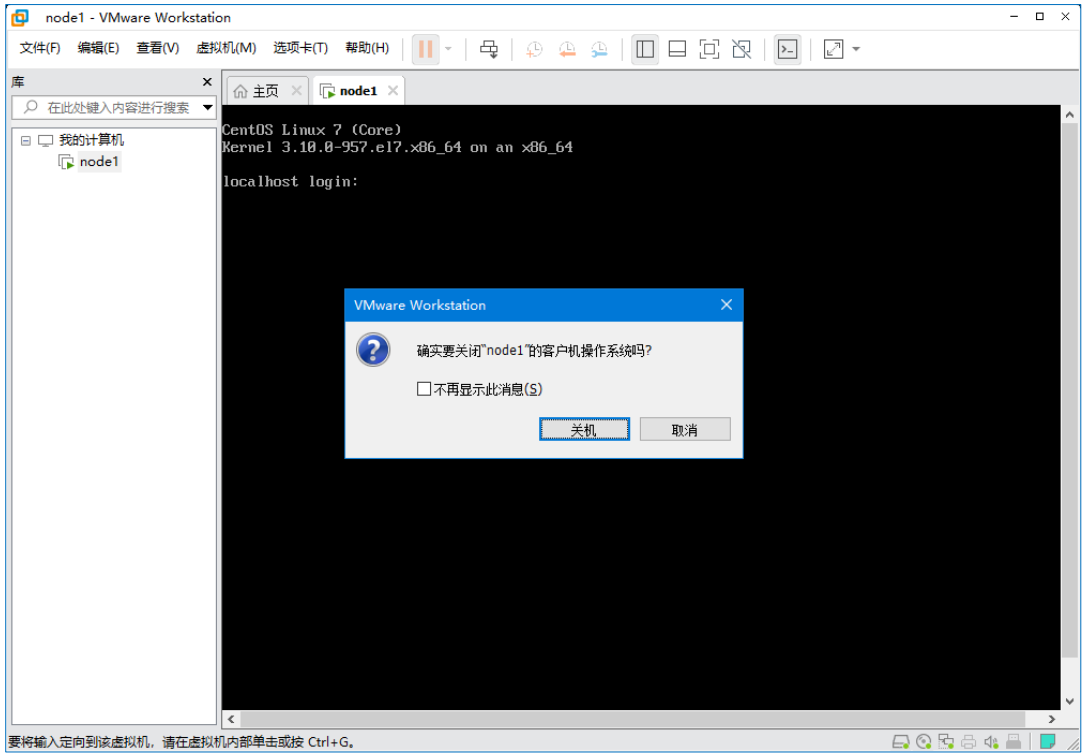


图2-30

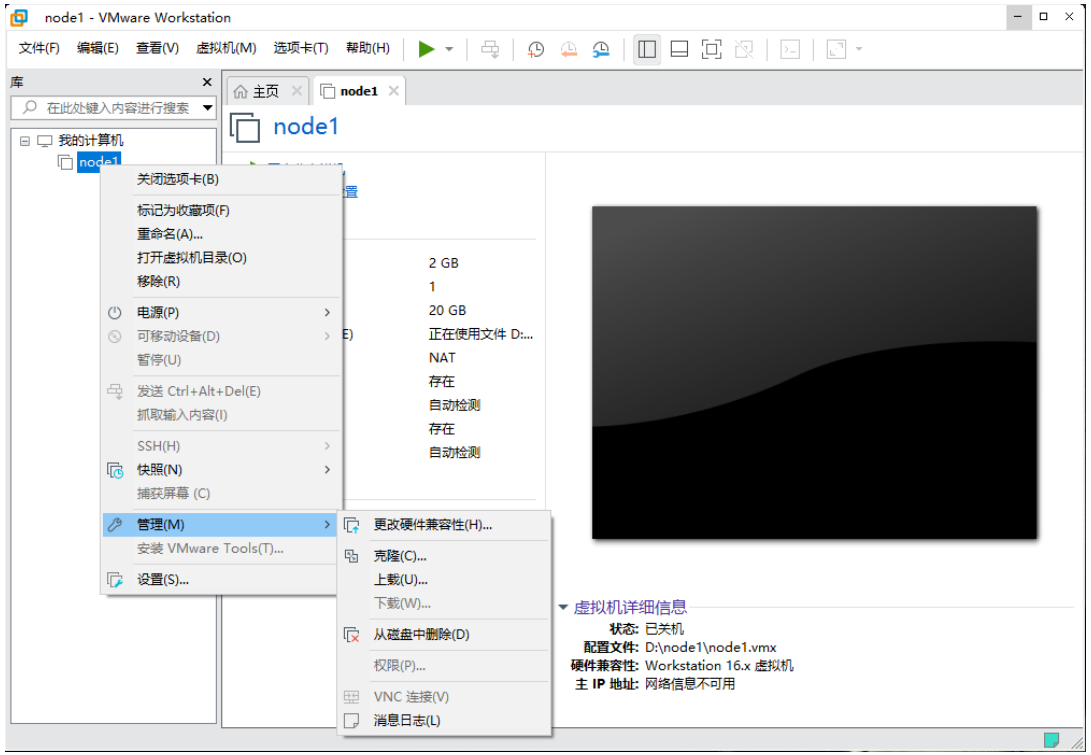


图2-31

3．Linux克隆

接步骤2， 弹出克隆虚拟机向导界面，如图2-32所示，点击【下一页】按钮。

步骤3：步骤3：在克隆源页面选择【虚拟机中的当前状态(C)】，如图2-33所示



图2-32

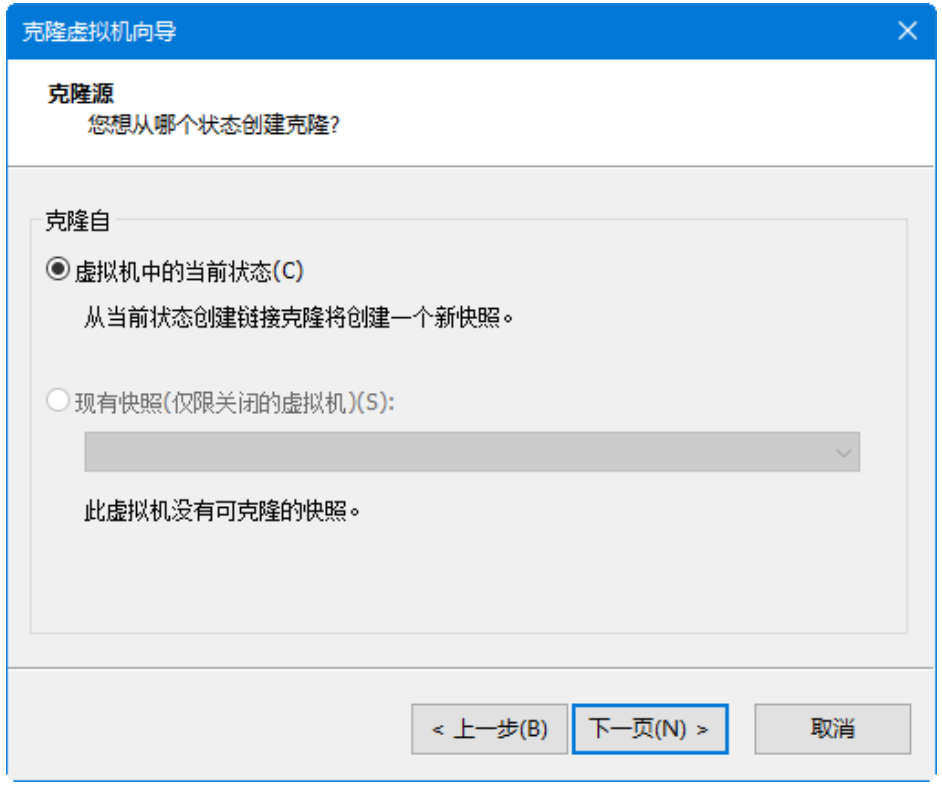


图2-33

3．Linux克隆

步骤4：在克隆类型页面选择【创建完整克隆(F)】，然后点击【下一页】按钮，如图2-34所示。

步骤5：填写虚拟机名称为【node2】，位置填写【D:\node2】，点击【完成】按钮，如图3-35。

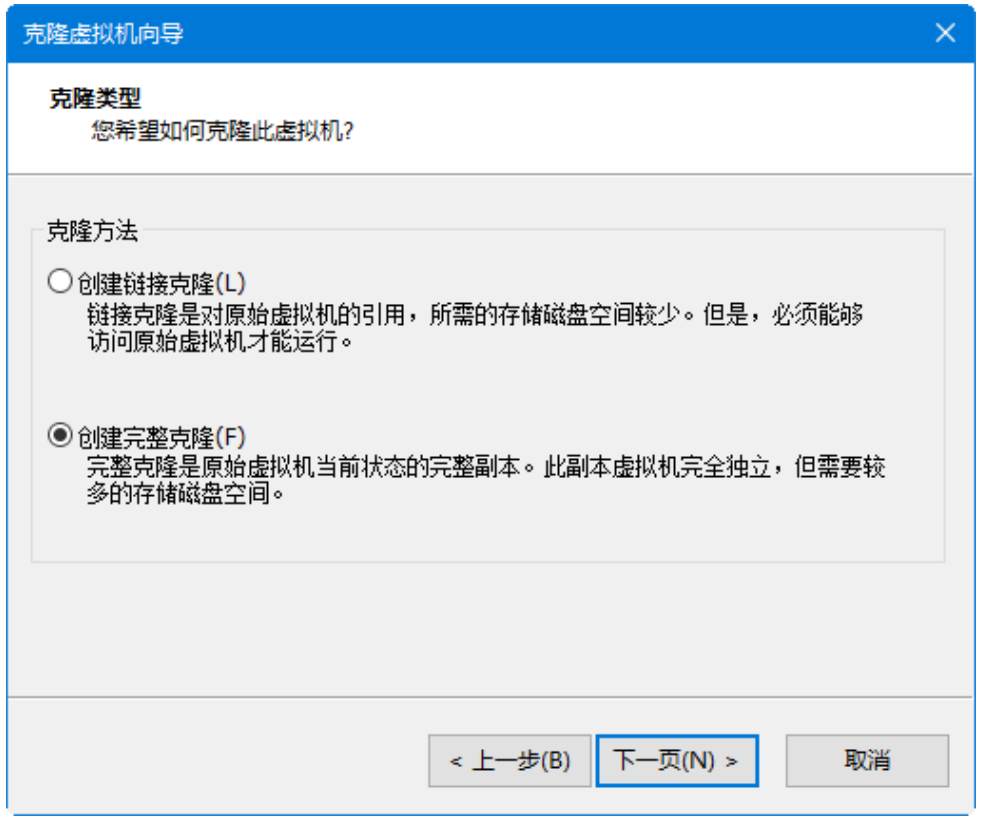


图2-34

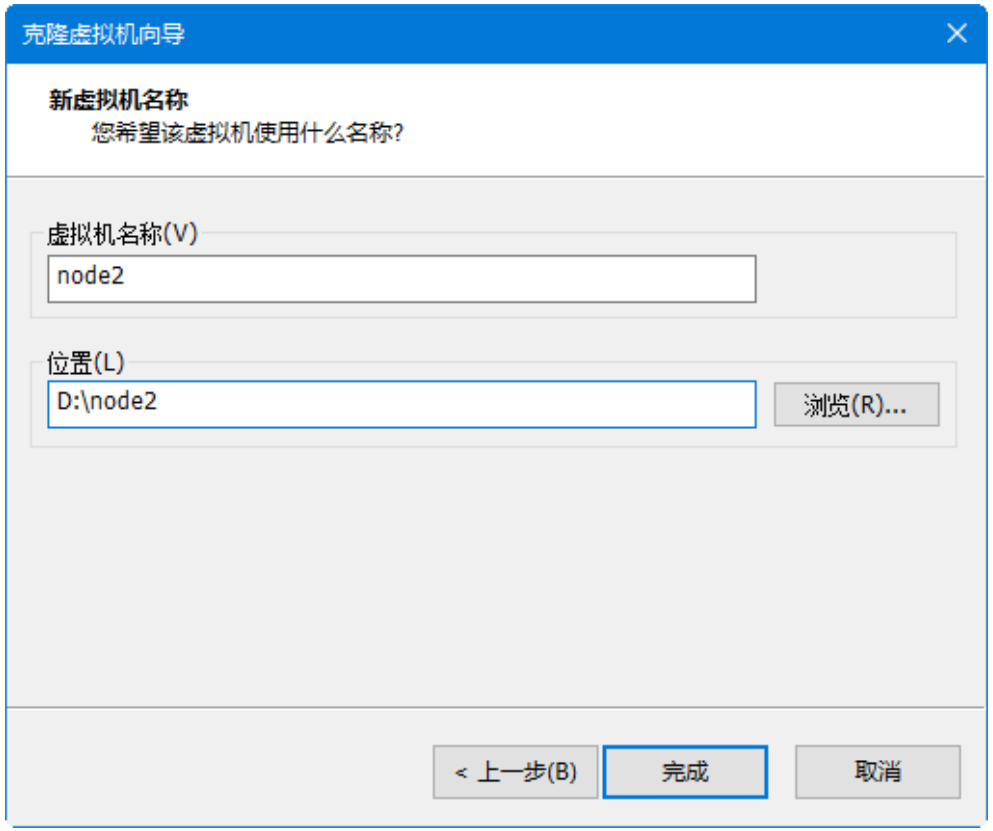


图2-35

3．Linux克隆

步骤6：用同样的方法，重复上述步骤，即可完成node3虚拟机的创建，最终三台虚拟机创建完成的情况如图3-36所示。



图2-36

4．Linux配置

步骤1：修改主机名，启动名为node2的虚拟机，输入用户名root和密码123456登录系统，在Linux提示符输入修改主机名的命令，然后在重启虚拟机，以便设置的新主机名node2生效，具体命令如图2-37所示。

[root@node1 ~]#hostnamectl set-hostname node2

[root@node1 ~]#reboot

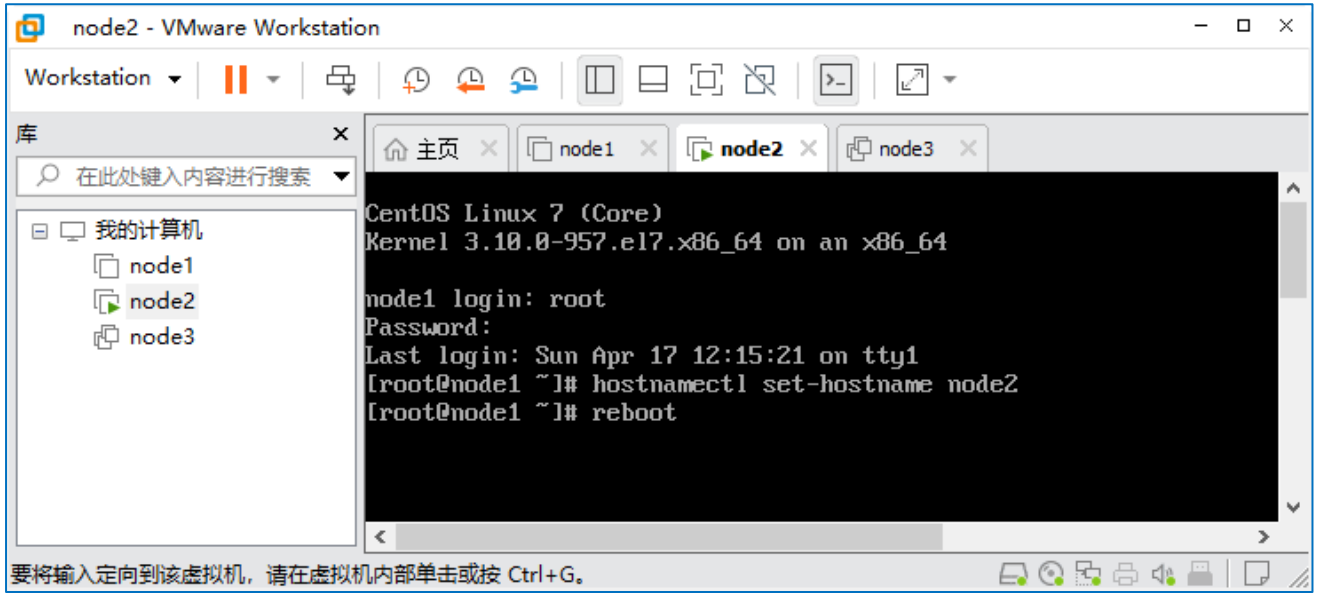


图2-37

4．Linux配置

接步骤1，重启node2虚拟机的主机名被修改成了node2，如图2-38所示，用同样的方法也把node3虚拟机的主机名改成node3.

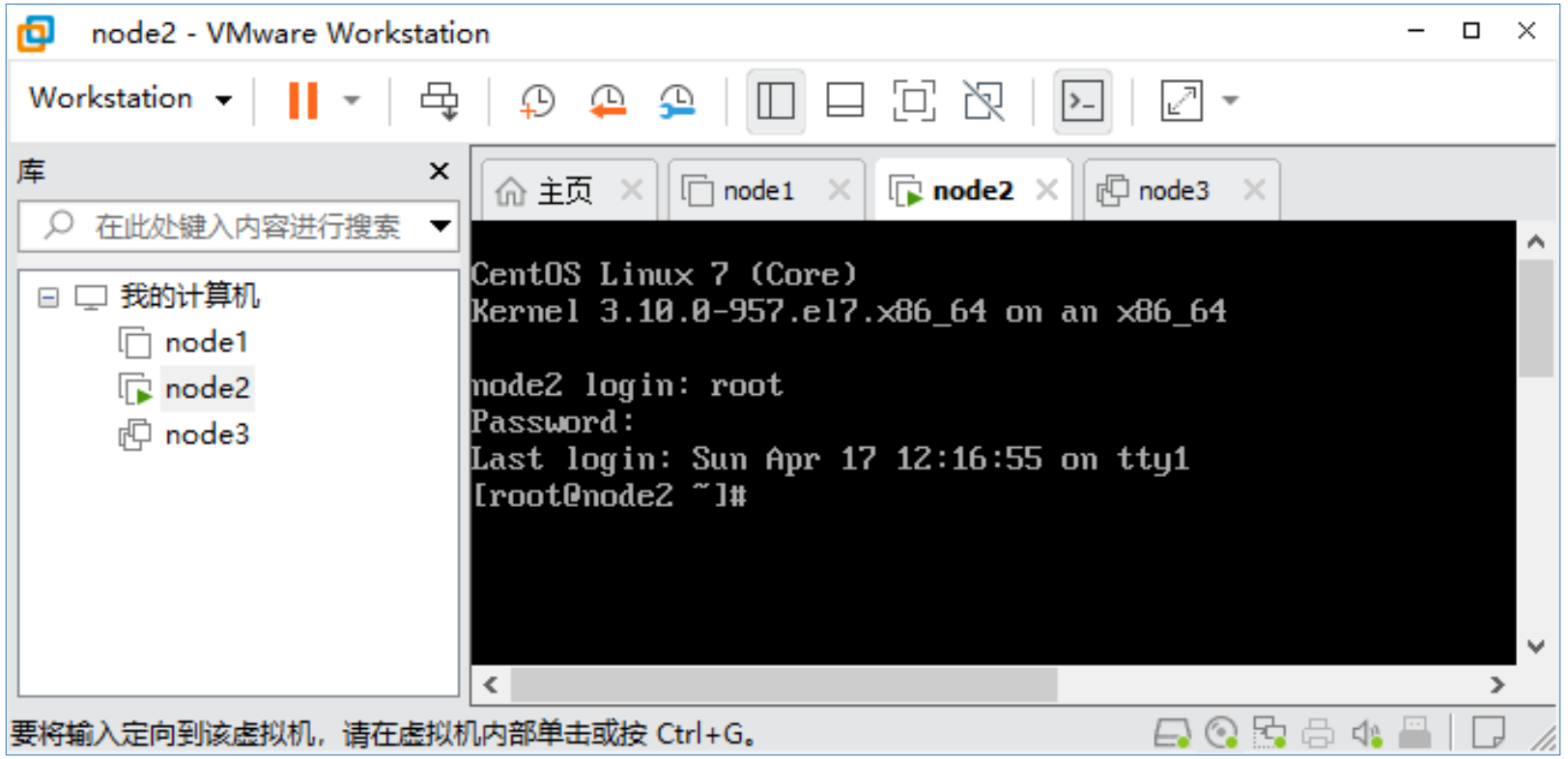


图2-38

4．Linux配置

步骤2：修改网络配置，启动node2虚拟机，输入查看网络配置文件目录的命令，具体如下

[root@node2 ~]#ls /etc/sysconfig/network-scripts/

在此目录下可以看到网络配置文件ifcfg-ens33，如图2-39所示

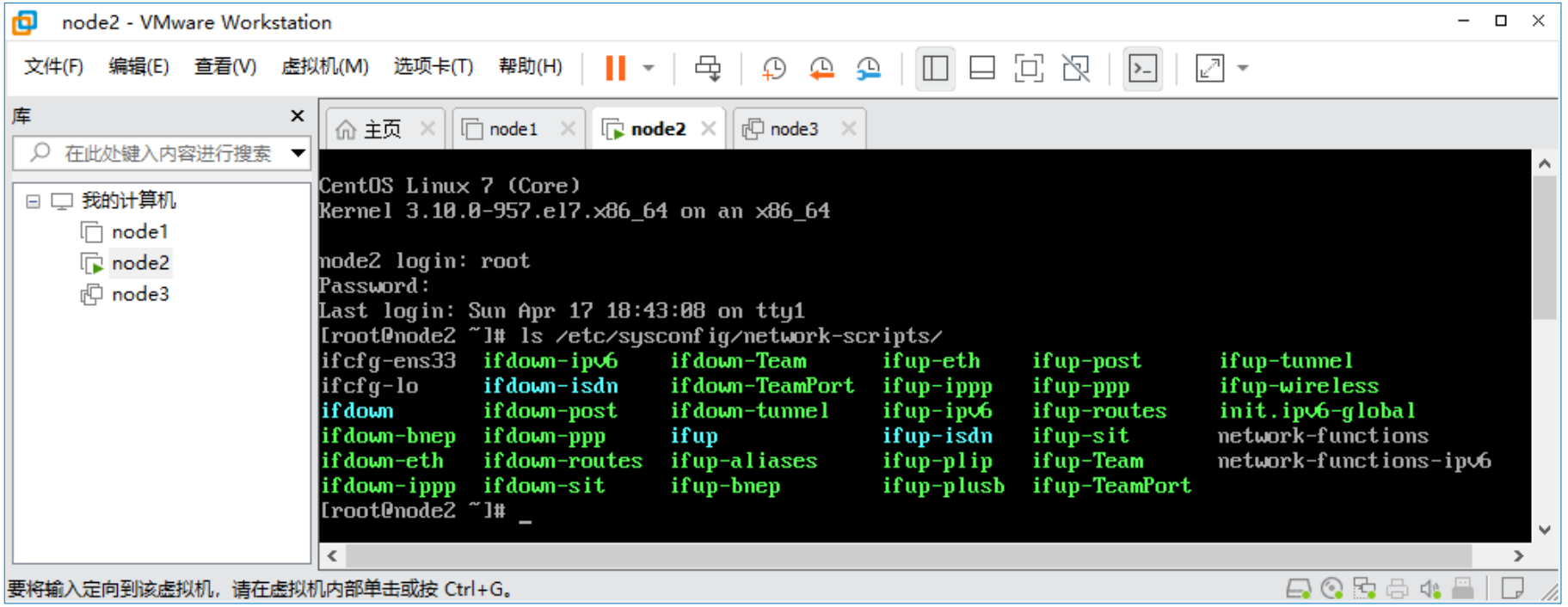


图2-39

4．Linux配置

使用vi编辑命令修改网络配置文件，具体命令如下，如图2-40所示。

[root@node2 ~]#vi /etc/sysconfig/network-scripts/ifcfg-ens33

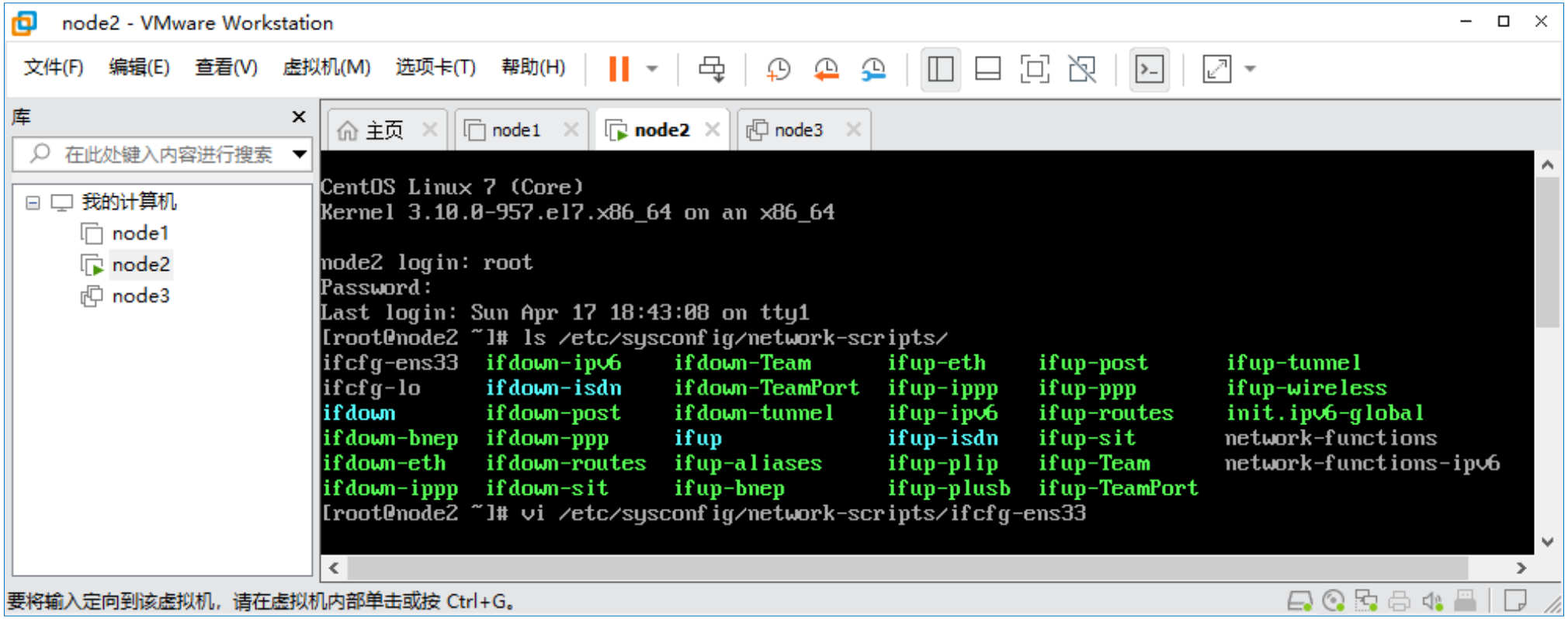


图2-40

4．Linux配置

使用vi编辑命令修改网络配置文件，修改完保存配置并退出vi编辑器，网络配置如图2-41所示。

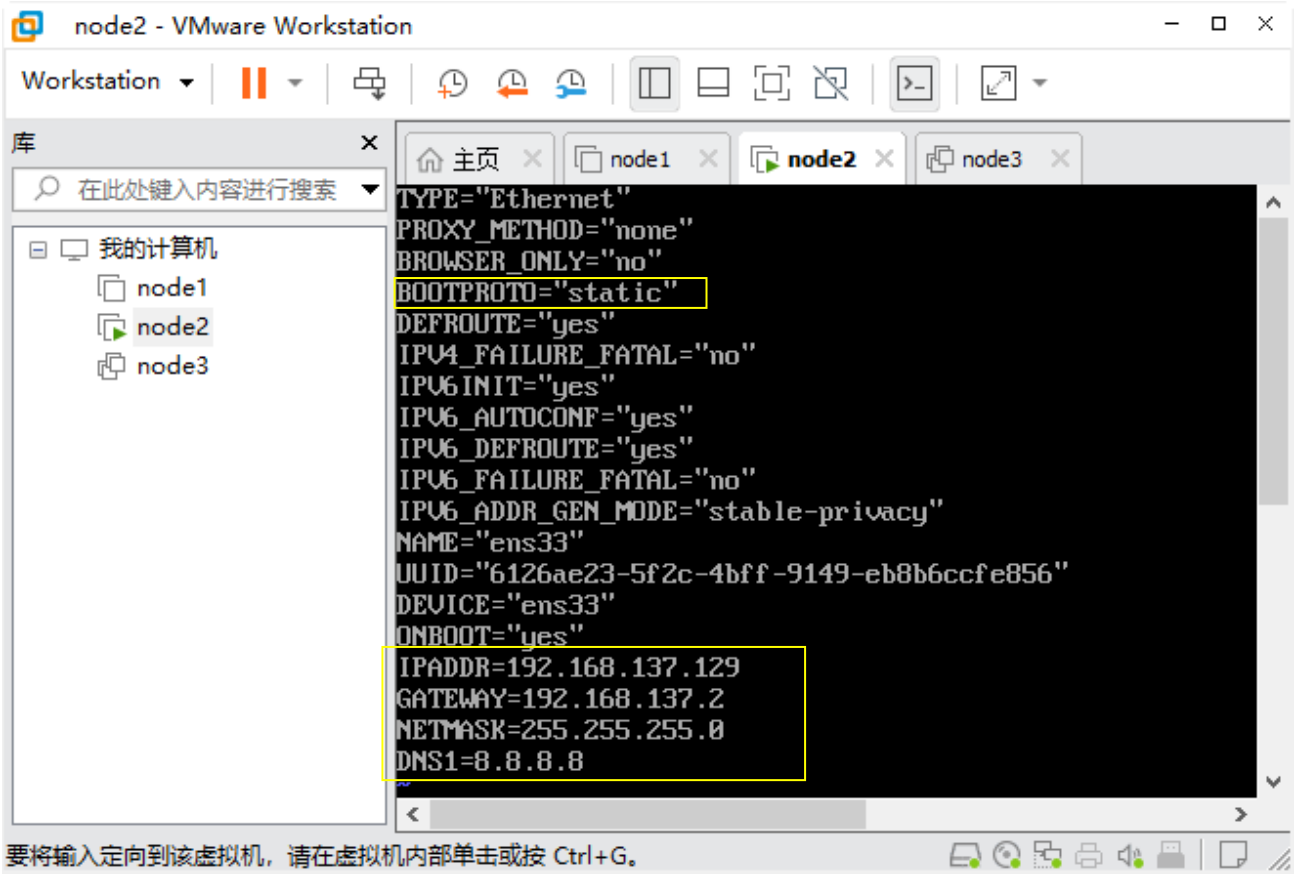


图2-41

4．Linux配置

- 由于node2是完全克隆node1、网络配置中的UUID也完全相同，UUID是虚拟机的唯一标识码，通过命令来重新生成UUID，具体命令如下：

```
[root@node2 ~]#sed -i '/UUID=/c\UUID='`uuidgen`' /etc/sysconfig/network-scripts/ifcfg-ens33
```

注意区分上述命令中的单引号“'”与反引号“`”。

- 完成上述node2的网络设置与UUID的重生成后一定要执行service network restart命令，使其生效。

- 使用同样的方法设置node1的ip为192.138.137.128，node3的ip为192.168.137.130，其他网络配置相同，同时也将node3的UUID重生成。

5．Xshell连接Linux虚拟机

步骤1：打开Xshell工作窗口，点击【文件】→【新建会话属性】菜单，打开如图【新建会话属性】窗口，在名称处输入【node1】，主机输入【192.168.137.128】，然后点击【确定】按钮，如图所示。

步骤2：新建连接后，打开【会话】窗口，选中node1，点击【连接】按钮，如图43所示。

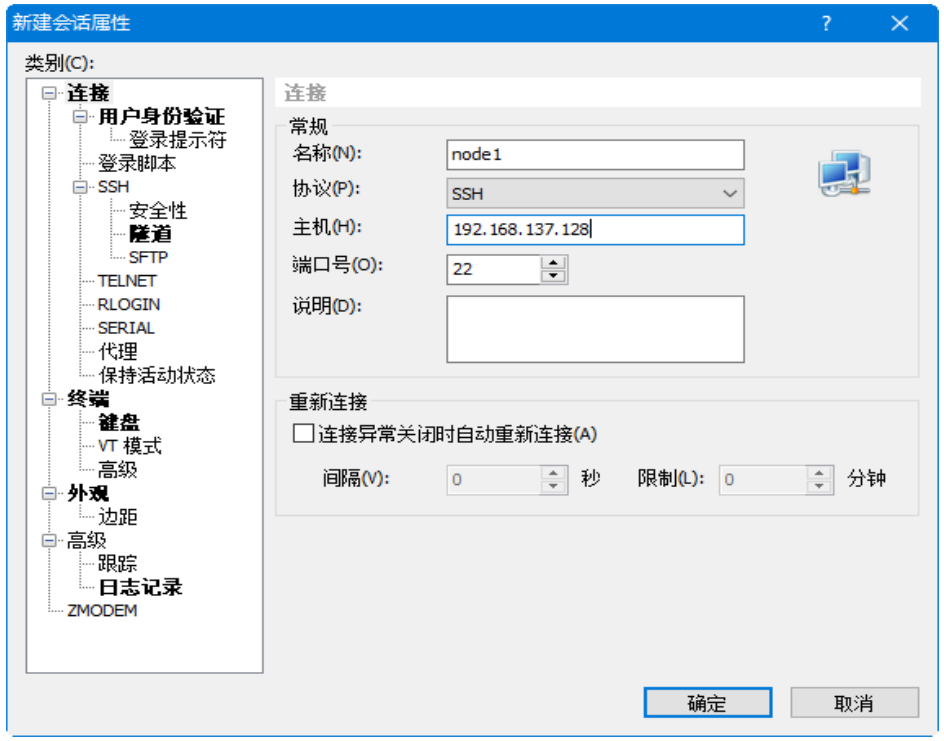


图2-42

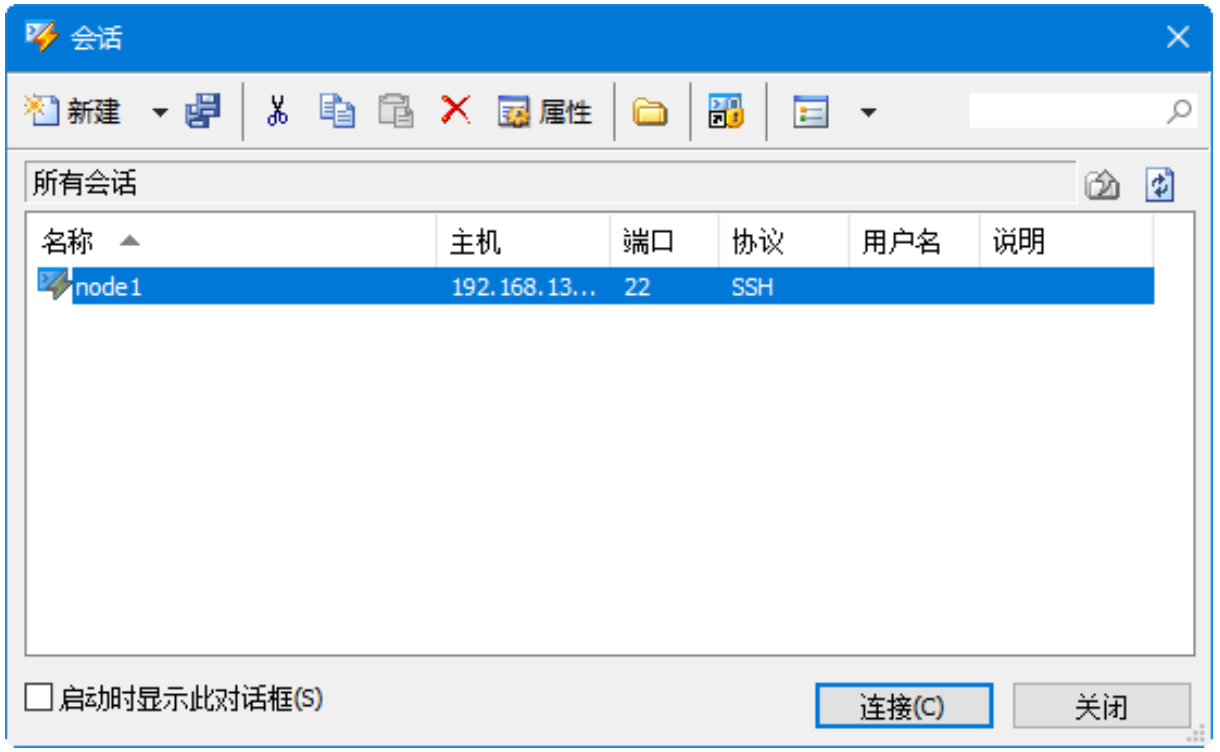


图2-43

5．Xshell连接Linux虚拟机

步骤3：在打开【SSH用户名】窗口中输入用户名root，勾选【记住用户名】然后点击【确定】按钮。如图2-44所示

步骤4：在打开【SSH用户身份验证】窗口中输入密码123456，勾选【记住密码】，如图45所示

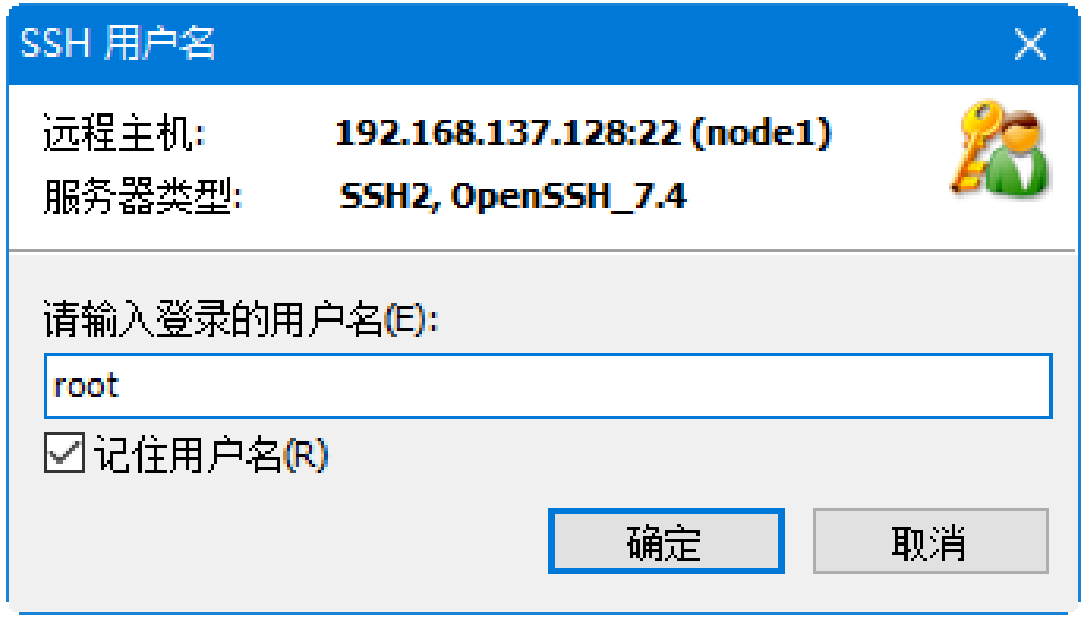


图2-44



图2-45

5．Xshell连接Linux虚拟机

用同样的方法新建node2、node3连接，最终Xshell连接三台虚拟机的效果如图2-46、图2-47、图2-48所示，后面就可以通过Xshell来连接虚拟机完成Hadoop、Hive等相关操作。

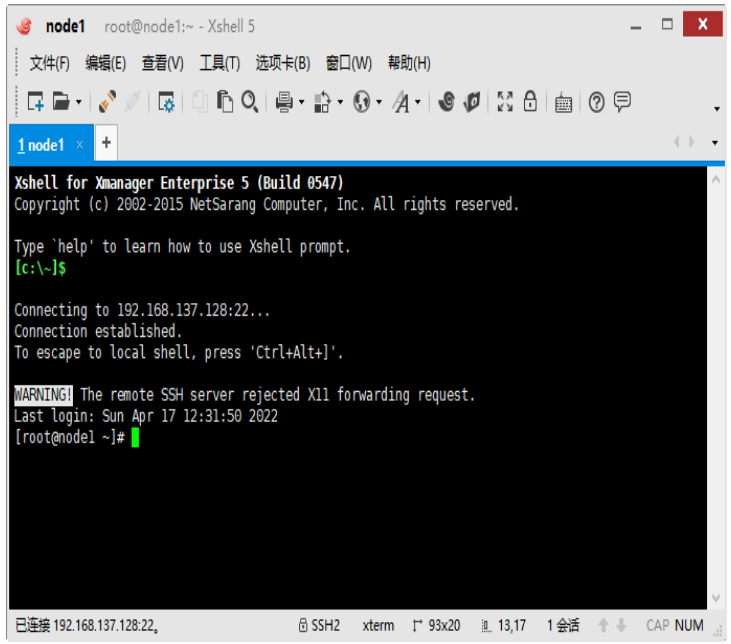


图2-46

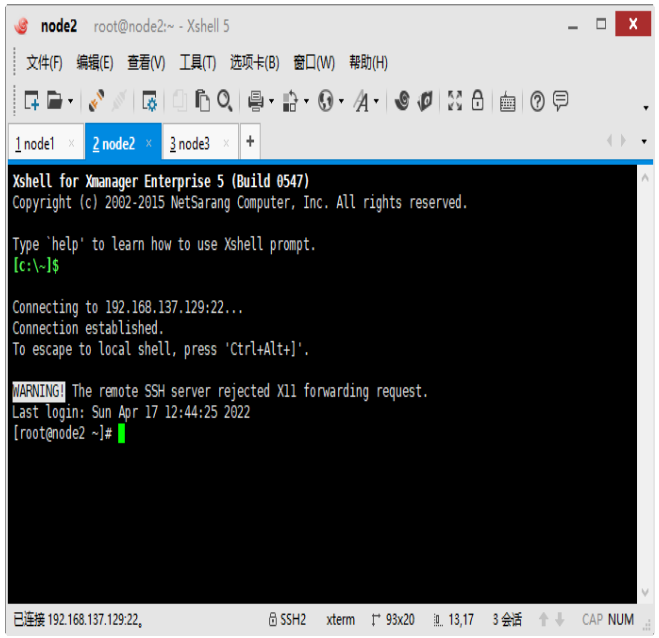


图2-47

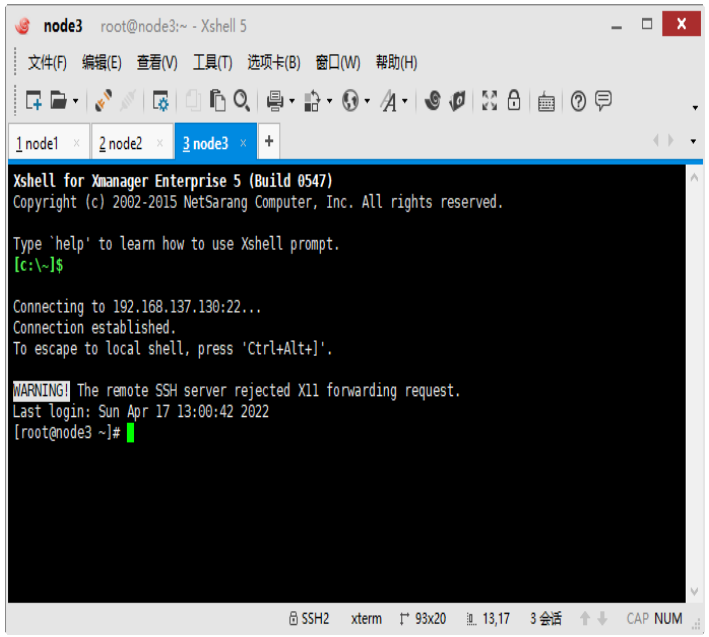


图2-48

任务 2-2 Hadoop环境搭建

2.2.1 任务描述

Hive数据仓库底层的数据存储依赖的是Hadoop平台的分布式文件系统（HDFS）。因此在搭建Hive之前需要先部署Hadoop环境，本任务将学会如何搭建全分布式Hadoop环境，为后续搭建Hive数据仓库打好基础。







2.2.2 任务准备知识

Hadoop搭建有三种模式，分别为单机模式、伪分布模式、全分布模式。

- 单机（本地）模式：这种模式在一台单机上运行，没有分布式文件系统，而是直接读写本地操作系统的文件系统。
- 伪分布式模式：也是在一台单机上运行，但用不同的Java进程模仿分布式运行中的各类结点
- 全分布式模式：全分布模式通常被用于生产环境，使用n台主机组成一个Hadoop集群，Hadoop守护进程运行在每台主机之上

1．准备工作

步骤1：配置主机名与IP的映射，修改node1的/etc/hosts文件，命令如下：

[root@node1 ~]# vi /etc/hosts

在文件末尾添加如图2-49内容，然后保存退出。node2、node3都做同样的修改。

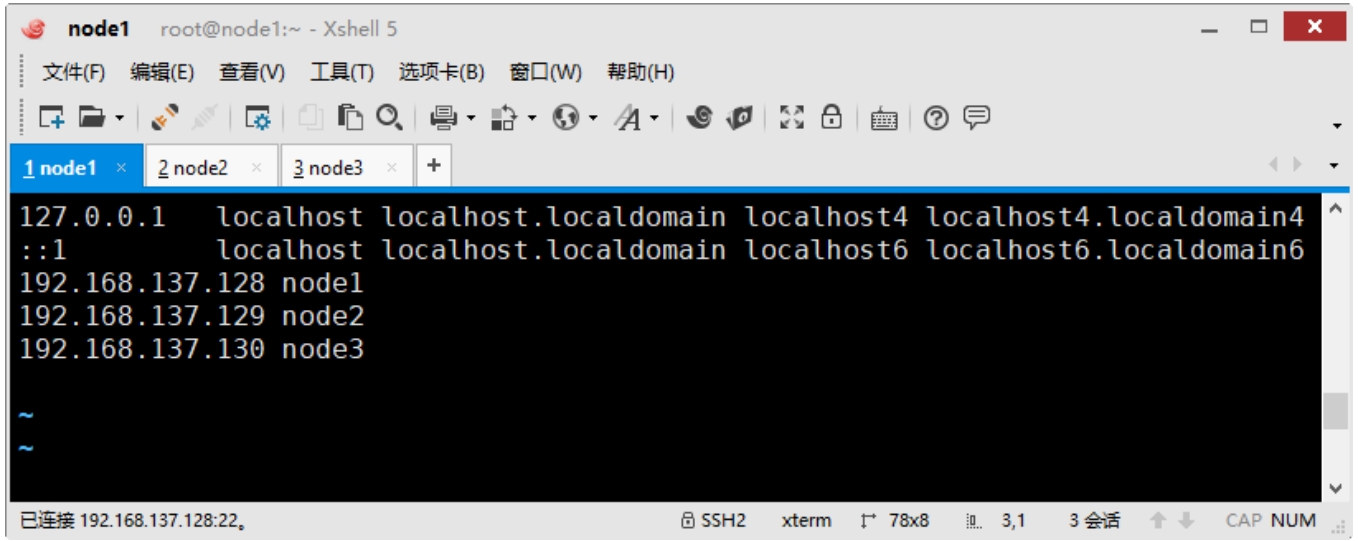


图2-49

1．准备工作

步骤2：SSH免密登录，在完全分布式模式下,集群内任意一台主机可免密登录集群内所有主机,即实现了两两免密登录。以node1免密登录集群内其他所有主机为例进行操作。

1.在node1生成密钥对，使用“ssh-keygen -t rsa”命令生成密钥对，其中rsa表示加密算法,键入上面一条命令后连续敲击四次回车键，如图2-50所示：在/root/.ssh目录下生成公钥(id_rsa.pub)和私钥(id_rsa)，可通过命令“ls /root/.ssh”查看，如图2-51所示。

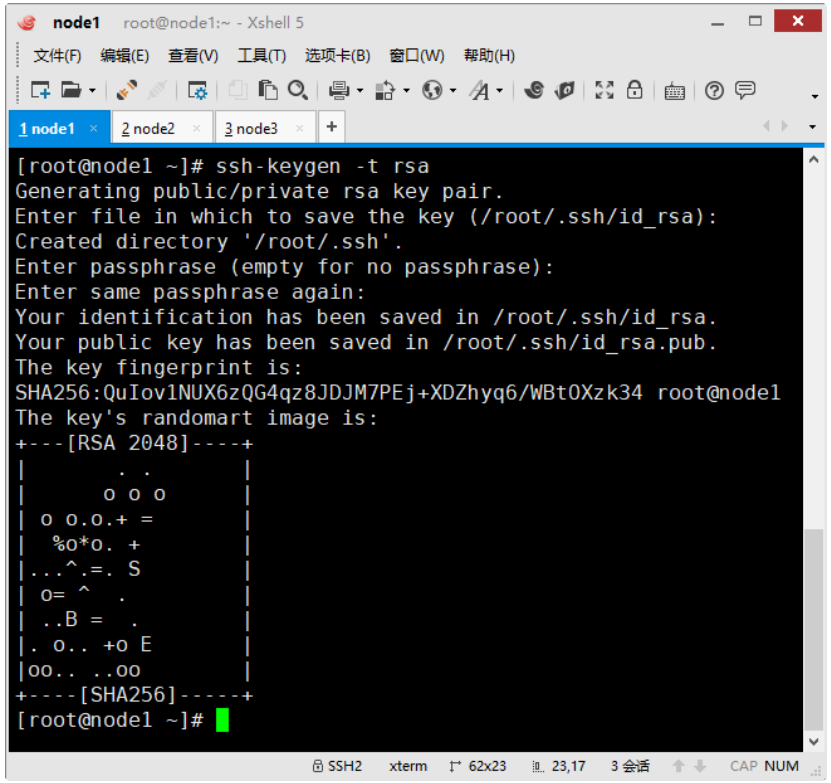


图2-50

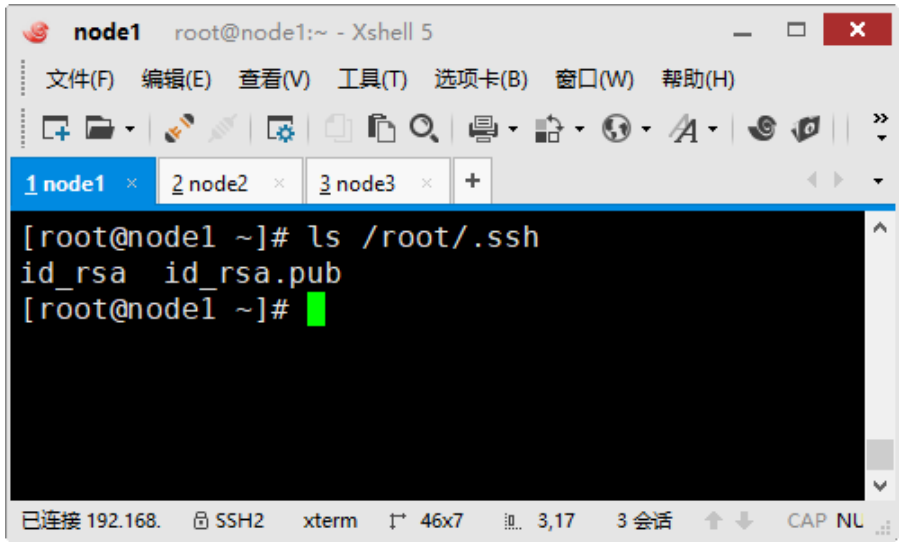


图2-51

1．准备工作

接步骤2，2. 将node1公钥id_rsa.pub复制到node1、node2和node3主机上，使用“ssh-copy-id”命令来复制公钥，命令如下，

```
[root@node1 ~]# ssh-copy-id node1
[root@node1 ~]# ssh-copy-id node2
[root@node1 ~]# ssh-copy-id node3
```

执行命令过程中会提示输入密码，此处密码都为123456。

3. 验证免密登录，在node1主机输入“ssh node2”、“ssh node2”命令验证，验证结果如图2-52所示。

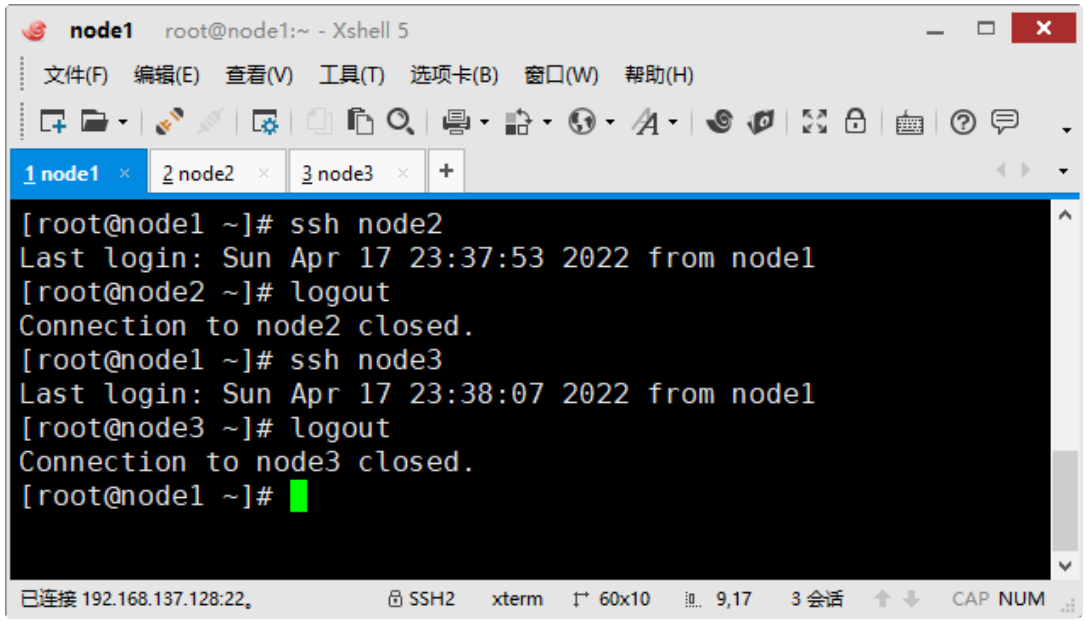


图2-52

2．安装JDK

步骤1：解压JDK，在node1使用mkdir命令创建/tools目录，使用Xftp新建连接node1 的会话，如图2-53所示。

将本机事先从Oracal官网下载的jdk-8u221-linux-x64.tar上传到node1的tools目录下，如图2-54所示：



图2-53

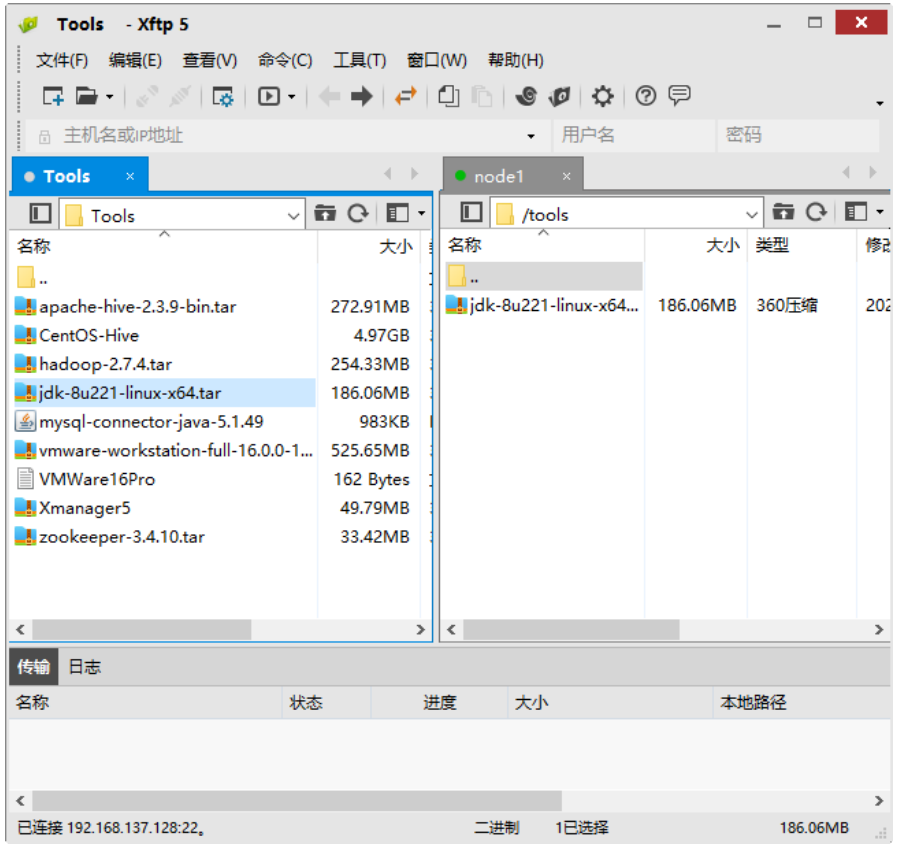


图2-54

2．安装JDK

将当前目录切换到/tools目录，可以通过ls命令查看通过Xftp上传的到/tool目录下的jdk安装包，然后使用tar命令将jdk安装包解压到tools文件内，如图2-55所示：使用mv命令将jdk1.8.0_221安装目录重命名，以便后面进行配置的时候更加简洁，重命名如图2-56所示

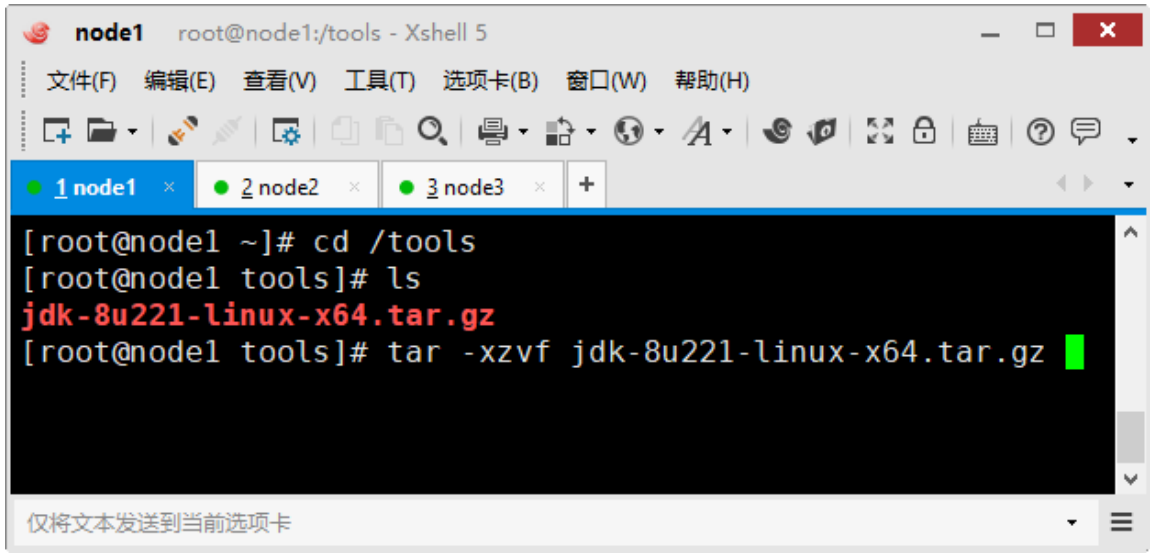


图2-55

```
[root@node1 tools]# mv jdk1.8.0_221 jdk
[root@node1 tools]# ls
jdk  jdk-8u221-linux-x64.tar.gz
```

图2-56

2．安装JDK

步骤2：配置环境变量，使用vi /etc/profile命令编辑系统的环境变量文件，在文件尾部添加如下内容，如图2-57所示。

export JAVA_HOME=/tools/jdk

export PATH=$PATH:$JAVA_HOME/bin

export CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar

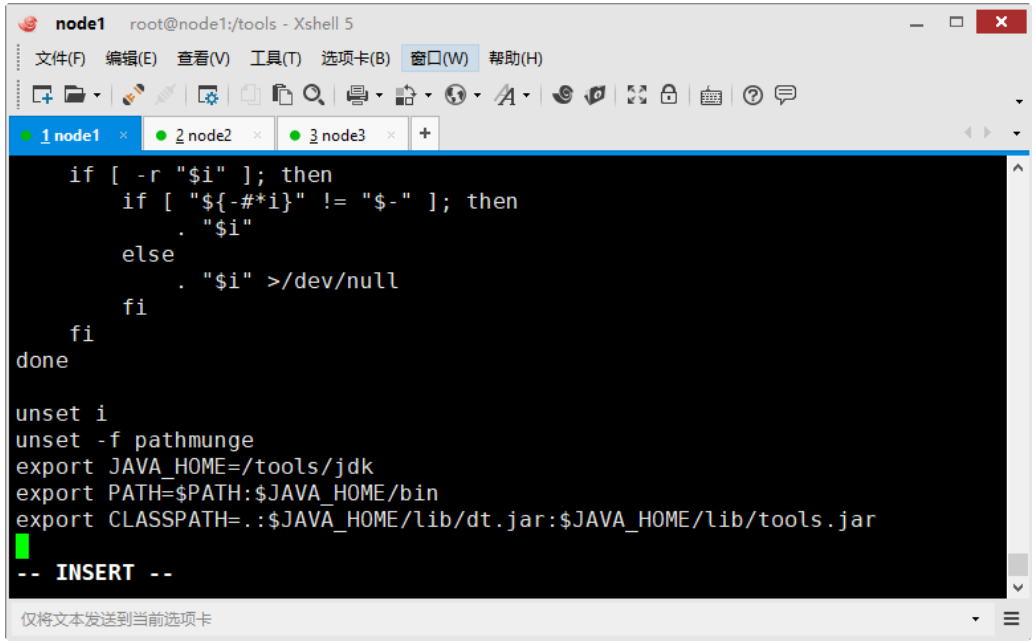


图2-57

配置环境变量后记住一定要使用**source /etc/profile**命令使得配置生效，然后使用java -version命令验证JDK环境是否配置成功，如图2-58所示

```
[root@node1 tools]# vi /etc/profile
[root@node1 tools]# source /etc/profile
[root@node1 tools]# java -version
java version "1.8.0_221"
Java(TM) SE Runtime Environment (build 1.8.0_221-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.221-b11, mixed mode)
```

图2-58

2．安装JDK

步骤3：远程拷贝JDK与配置文件，使用scp命令将node1的JDK安装包与配置文件远程拷贝到node2与node3虚拟机里，拷贝前在node2、node3虚拟机的根目录下新建tools目录，命令如下:

[root@node1 /]#scp -r /tools/jdk/ root@node2:/tools

[root@node1 /]#scp -r /tools/jdk/ root@node3:/tools

远程拷贝配置文件的命令如下:

[root@node1 /]#scp /etc/profile root@node2:/etc/profile

[root@node1 /]#scp /etc/profile root@node3:/etc/profile

配置文件拷贝后一定要使用**source /etc/profile**命令使配置在node2、node3虚拟机生效，同时再使用java -version命令验证一下JDK是否安装成功。

3．部署Hadoop(完全分布模式)

步骤1：将事先下载的Hadoop安装包hadoop-2.7.4.tar.gz使用Xftp工具上传到node1的tools目录下，如图2-59所示.

在node1中将当前目录切换到/tools目录，可以通过ls命令查看通过Xftp上传的到/tool目录下的Hadoop安装包，然后使用tar命令将Hadoop安装包解压到tools文件夹内，如图2-60所示：

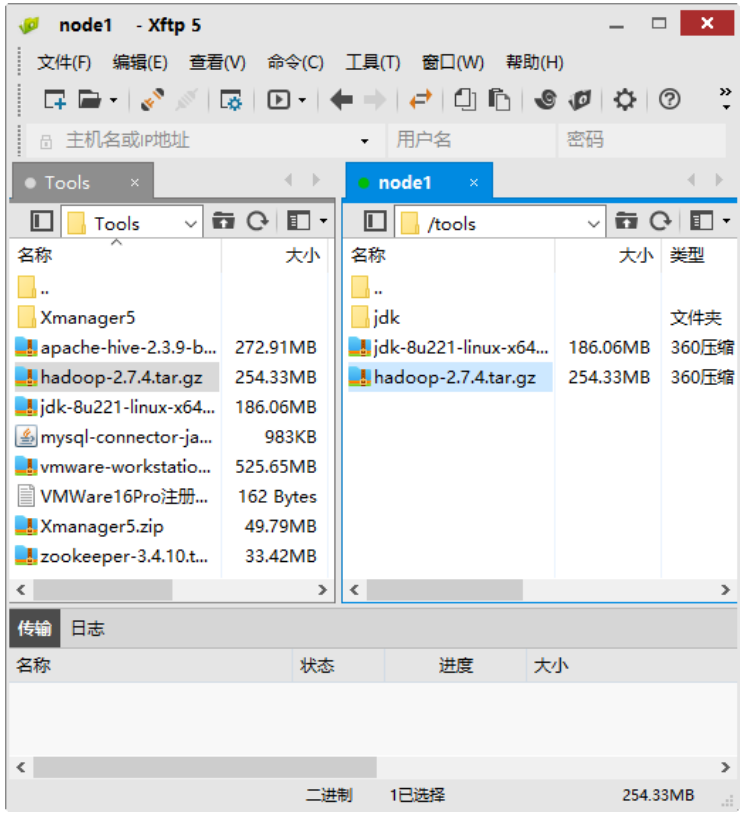


图2-59

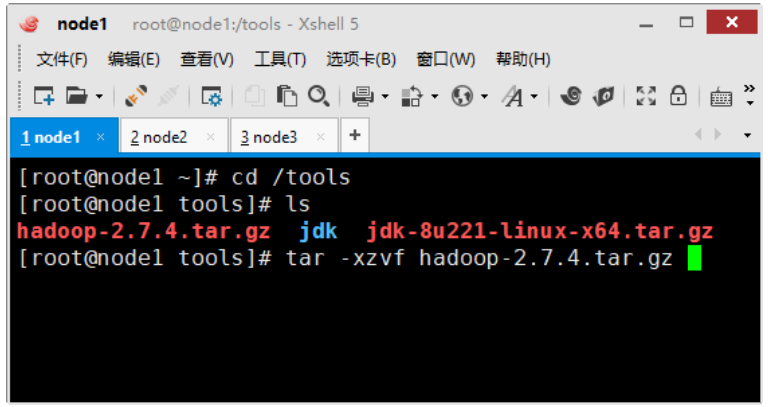


图2-60

3．部署Hadoop(完全分布模式)

接步骤1，使用mv命令将hadoop-2.7.4安装目录重命名为hadoop，以便后面进行配置的时候更加简洁，重命名如图2-61所示

步骤2：配置Hadoop

用vi /etc/profile命令编辑系统的环境变量文件，在文件尾部添加如下内容，如图2-62所示。

export HADOOP_HOME=/tools/hadoop

export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

```
[root@node1 tools]# mv hadoop-2.7.4 hadoop
[root@node1 tools]# ls
hadoop                  jdk
hadoop-2.7.4.tar.gz     jdk-8u221-linux-x64.tar.gz
```

图2-61

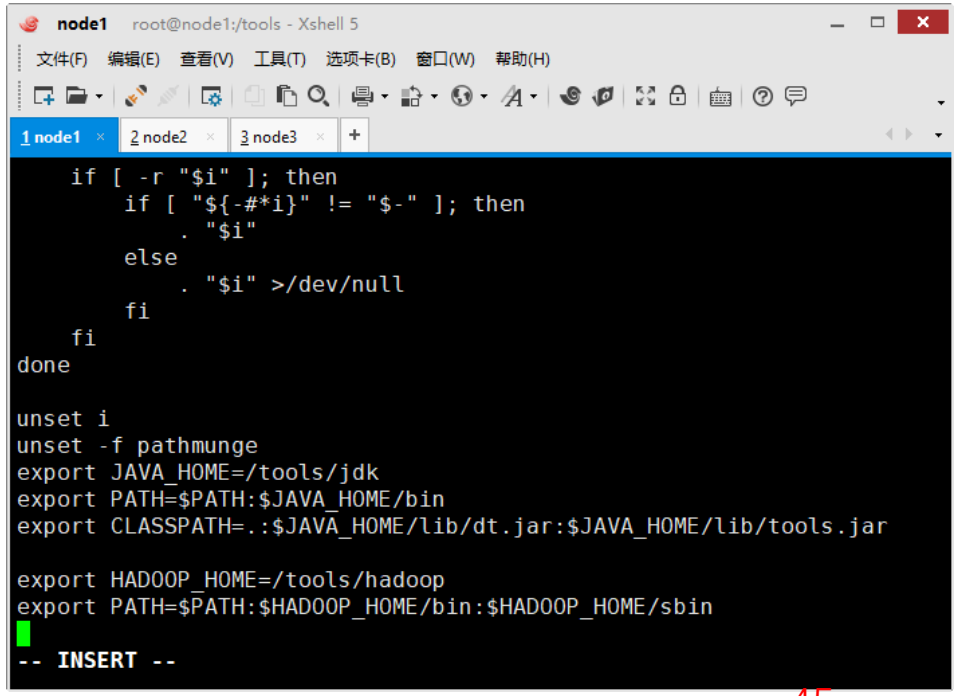


图2-62

3．部署Hadoop(完全分布模式)

接步骤2，配置环境变量后记住一定要使用source /etc/profile命令使得配置生效，同时使用hadoop version命名验证Hadoop环境是否配置成功，如图2-63所示

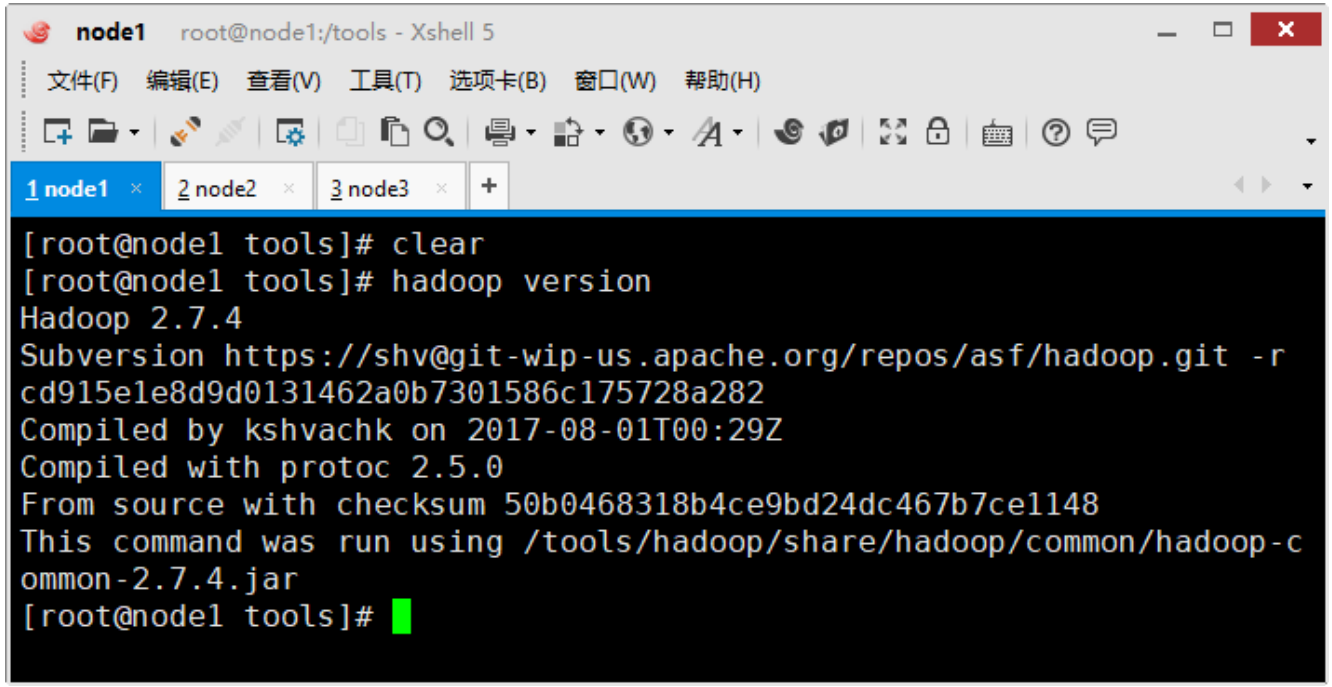


图2-63

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

1.编辑hadoop-env.sh文件

用Xshell连接node1，hadoop-env.sh在hadoop安装目录的etc/hadoop目录下，用vi命令编辑hadoop-env.sh，指定Hadoop运行的Java环境变量，将JAVA_HOME设置为node1中JDK的安装路径，如图2-64所示。

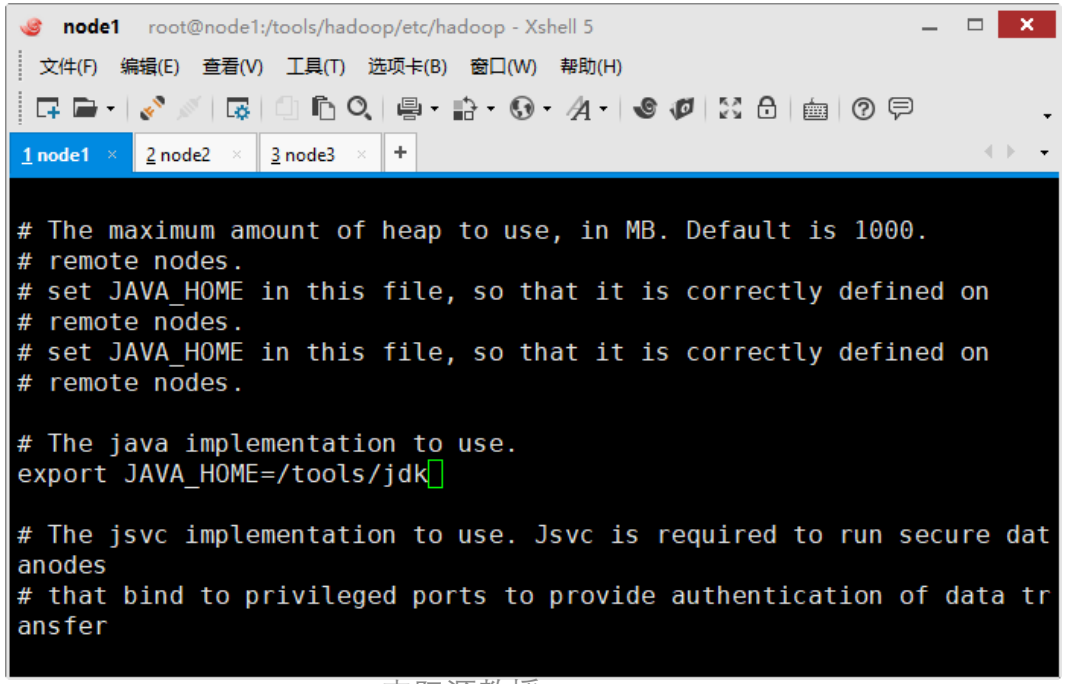


图2-64

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

2.编辑yarn-env.sh文件

yarn-env.sh在hadoop安装目录的etc/hadoop目录下，用vi命令编辑yarn-env.sh，指定Yarn运行的Java环境变量，将JAVA_HOME设置为node1中JDK的安装路径，如图2-65所示

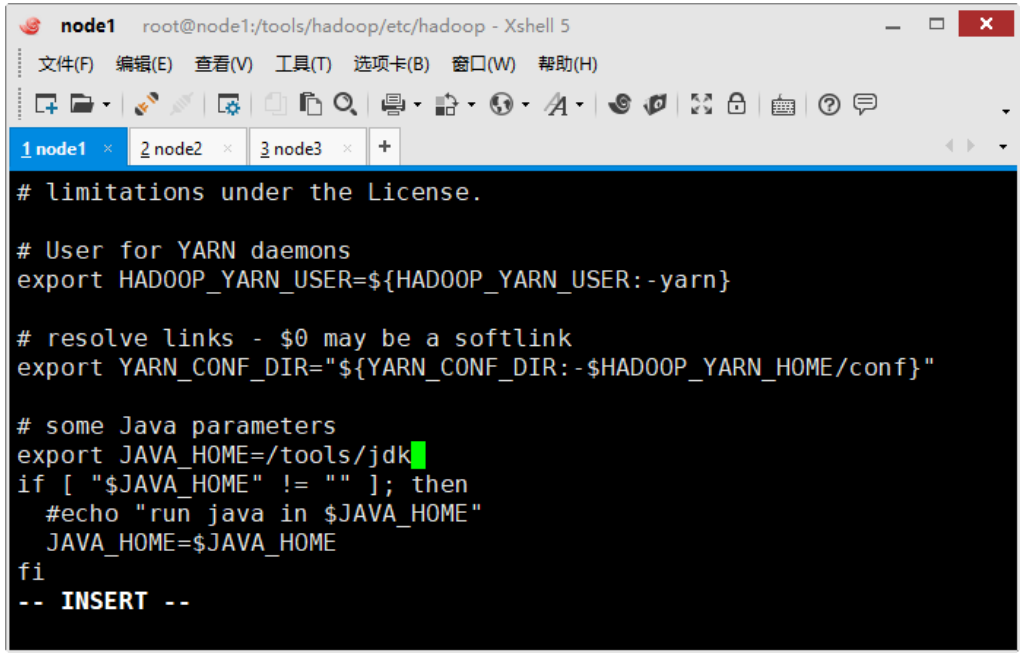


图2-65

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

3.编辑core-site.xml文件

core-site.xml在hadoop安装目录的etc/hadoop目录下，是Hadoop的核心配置文件，用vi命令编辑core-site.xml，指定namenode的地址在node1上，同时设置Hadoop的临时目录（事先使用mkdir命令在hadoop目录下新建tmp目录），在<configuration>标签内添加内容如下，完成配置后如图2-66所示。

```
<property>
    <name>fs.defaultFS</name>
    <value>hdfs://node1:9000</value>
</property>
<property>
    <name>hadoop.tmp.dir</name>
    <value>/tools/hadoop/tmp</value>
</property>
```

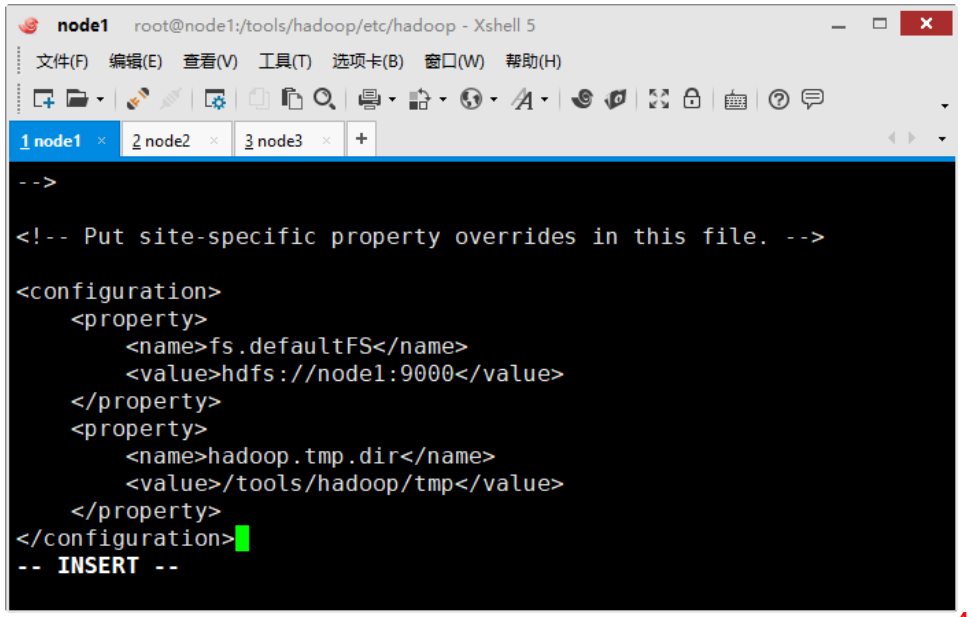


图2-66

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

4.编辑hdfs-site.xml文件

hdfs-site.xml在hadoop安装目录的etc/hadoop目录下，是HDFS的核心配置文件，用vi命令编辑hdfs -site.xml，指定HDFS的数量，同时设置secondary namenode 所在主机的和端口，在<configuration>标签内添加内容如下，完成配置后如图2-67所示。

```
 <property>
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.namenode.secondary.http-address</name>
    <value>node2:50090</value>
  </property>
```

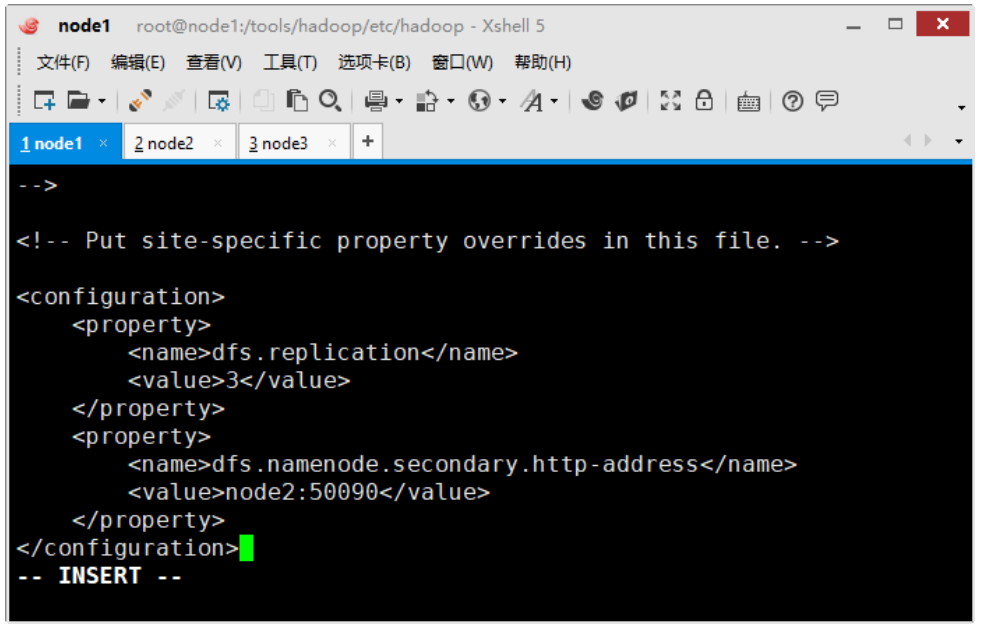


图2-67

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

5.编辑mapred-site.xml文件

在hadoop安装目录的etc/hadoop目录下只有mapred-site.xml.template文件，使用cp mapred-site.xml.template mapred-site.xml命令拷贝模板来新建mapred-site.xml文件，用vi命令编辑mapred-site.xml，指定MapReduce运行时的框架，这里指定在YARN上，默认在local，在<configuration>标签内添加内容如下，完成配置后如图6-68所示：

```
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
```

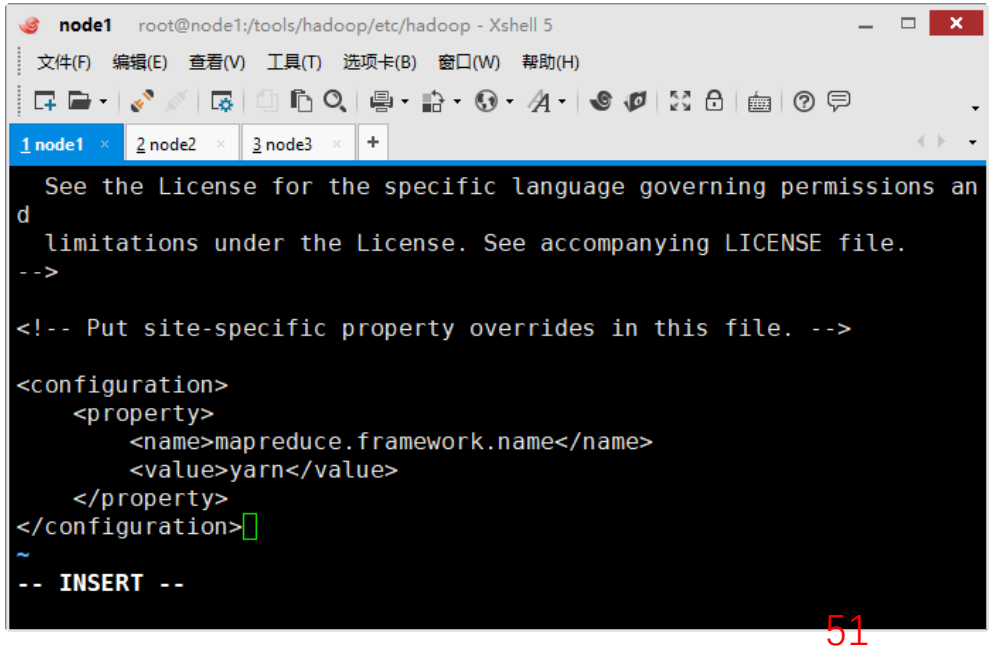


图2-68

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

6.编辑yarn-site.xml文件

yarn-site.xml在hadoop安装目录的etc/hadoop目录下，用vi命令编辑yarn -site.xml，指定YARN集群的管理者（ResourceManager）的地址，在<configuration>标签内添加内容如下，完成配置后如图2-69所示：

```
<property>
        <name>yarn.resourcemanager.hostname</name>
        <value>node1</value>
    </property>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
```

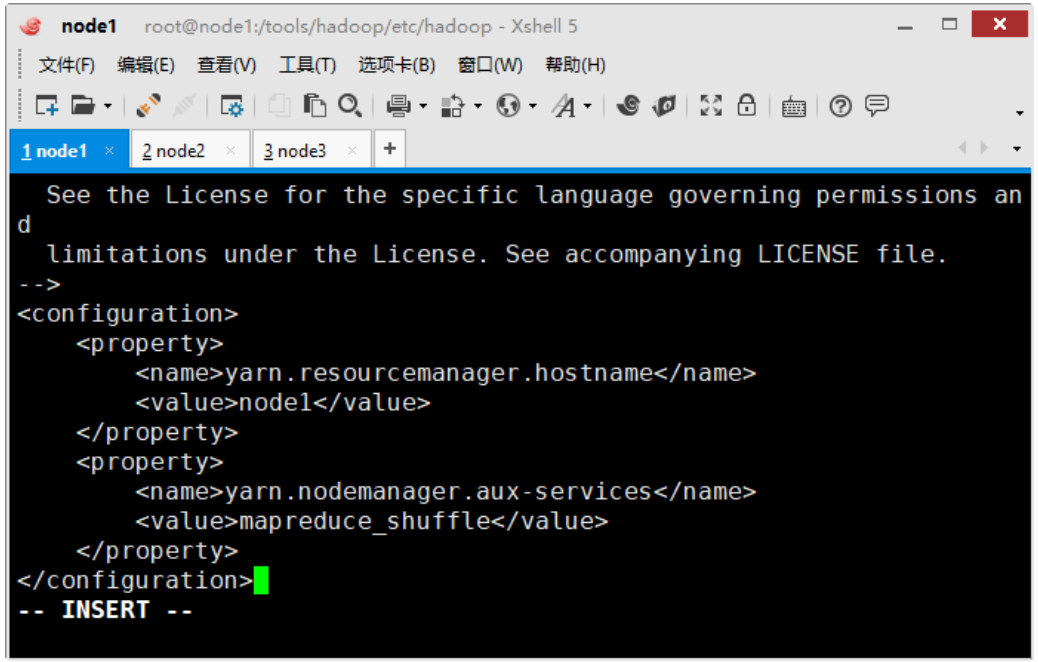


图2-69

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

7.编辑slaves文件

slaves在hadoop安装目录的etc/hadoop目录下，用vi命令编辑slaves，将文件中的localhost删除，添加主节点和子节点的主机名称，此处主节点为node1，子节点为node2和node3，完成配置后如图2-70所示

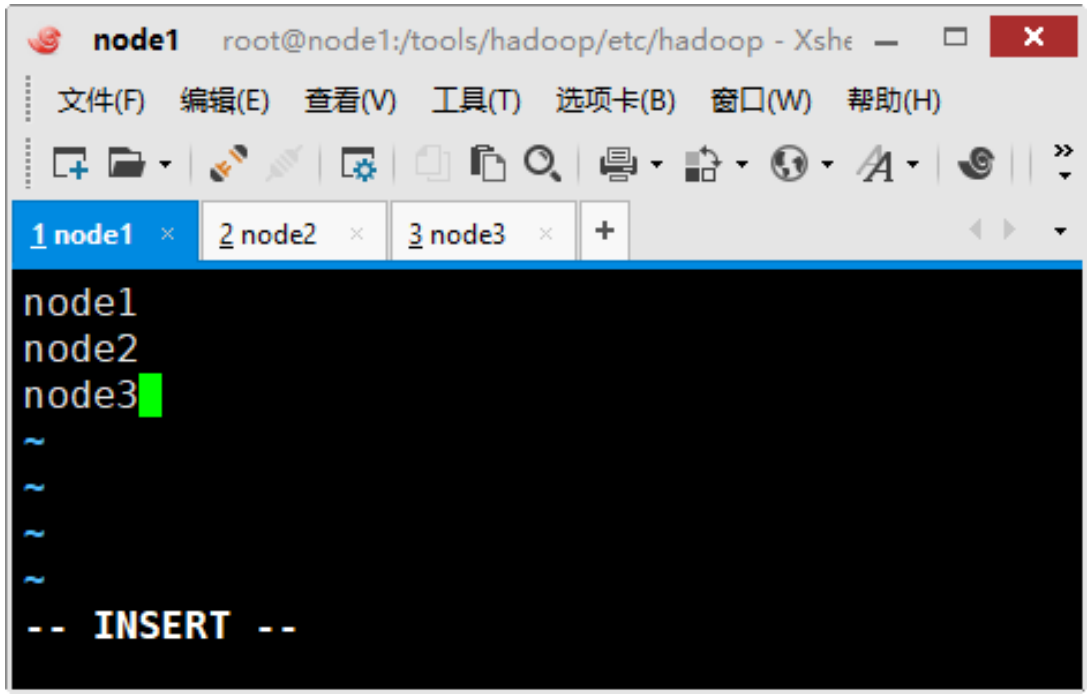


图2-70

3．部署Hadoop(完全分布模式)

步骤3：配置Hadoop集群

8. 远程拷贝Hadoop安装包与配置文件

使用scp命令将node1的Hadoop安装包与配置文件远程拷贝到node2与node3虚拟机里，远程拷贝Hadoop安装包的命令如下。

[root@node1 /]#scp -r /tools/hadoop/ root@node2:/tools

[root@node1 /]#scp -r /tools/hadoop/ root@node3:/tools

远程拷贝配置文件的命令如下：

[root@node1 /]#scp /etc/profile root@node2:/etc/profile

[root@node1 /]#scp /etc/profile root@node3:/etc/profile

配置文件拷贝后一定要使用**source /etc/profile**命令使配置在node2、node3虚拟机生效。

3．部署Hadoop(完全分布模式)

步骤4：主节点文件系统格式化，在node1上执行hdfs namenode –format 命令，对主节点上的文件系统进行格式化，此命令只在初次启动Hadoop的时候执行，出现successfully formatted信息则表明格式化成功，如图2-71所示。

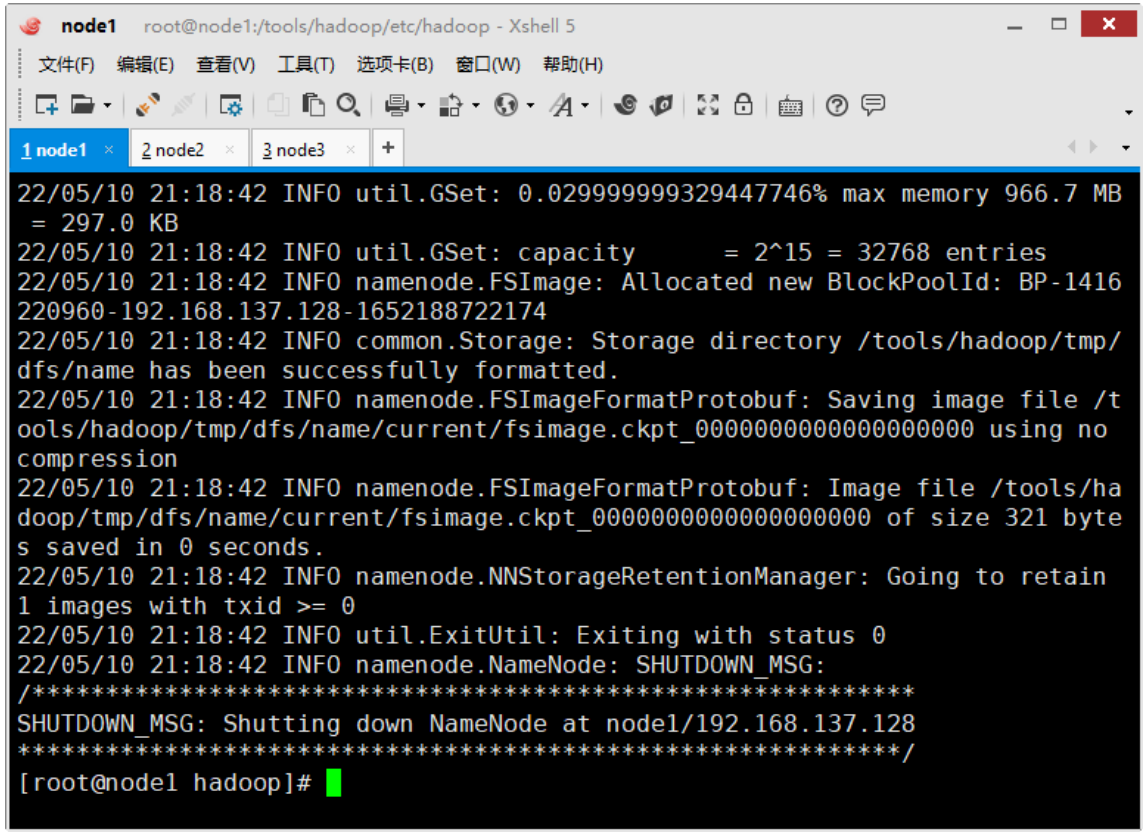


图2-71

3．部署Hadoop(完全分布模式)

步骤5：配置Hadoop集群

使用start-dfs.sh 命令启动Hadoop集群的HDFS，node1和node2的NameNode,以及node1、node2和node3的DataNode都会被启动。

使用start-yarn.sh命令启动Hadoop集群的Yarn，node1的ResourceManager和node1、node2和node3的NodeManager都会被启动。

两个命令及其执行结果如图2-72所示

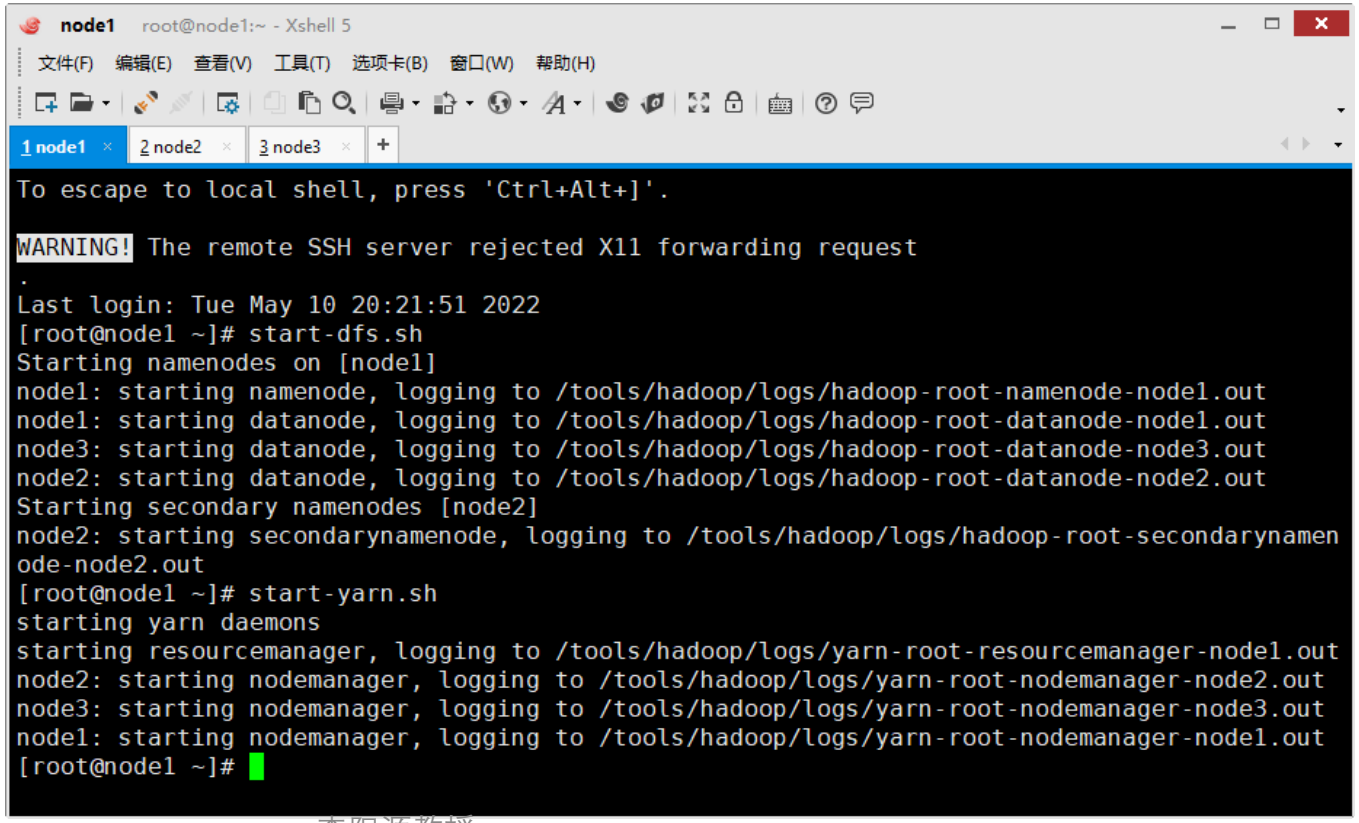


图2-72

3．部署Hadoop(完全分布模式)

步骤6：使用jps命令查看进程

在node1、node2和node3上分别执行jps命令，查看完全分布模式Hadoop集群是否启动成功

效果如图2-73、图2-74、图2-75所示。

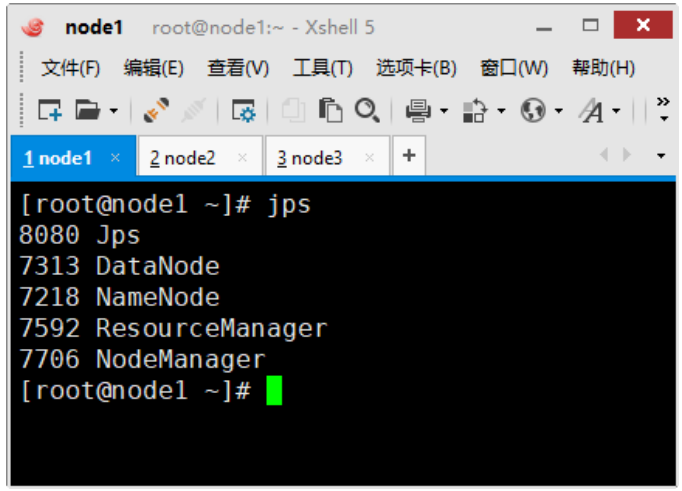


图2-73

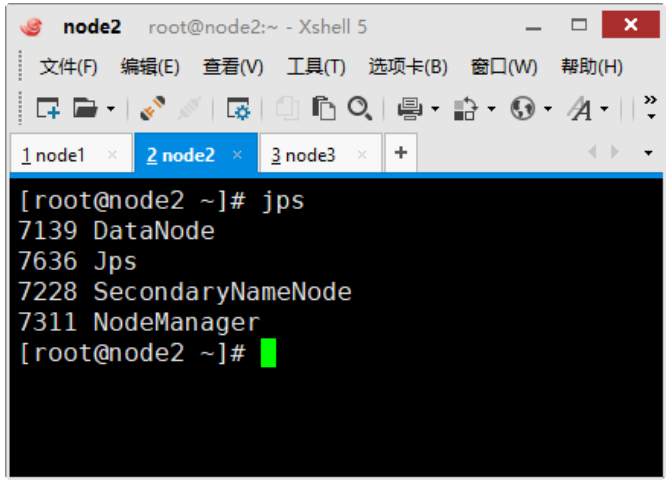


图2-74

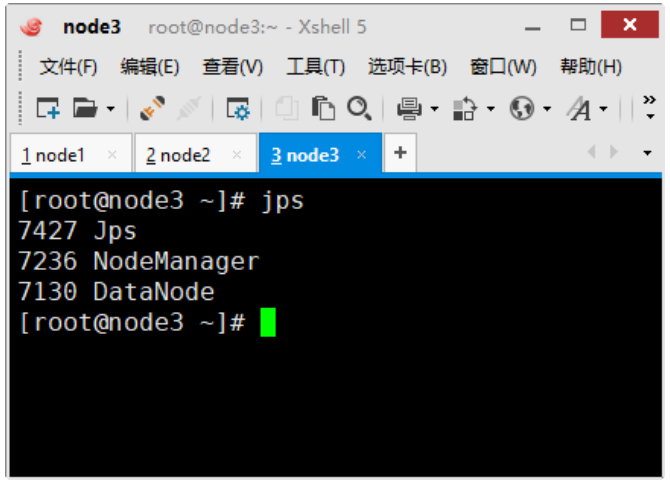


图2-75

3．部署Hadoop(完全分布模式)

步骤7：关闭防火墙

默认情况下CentOS 7 是开启防火墙的，在node1、node2和node3上分别执行关闭防火墙服务和禁止防火墙开机启动命令，如下所示

[root@node1 /]#systemctl stop firewalld

[root@node1 /]#systemctl disable firewalld

步骤8：在Windows系统下，访问http://192.168.137.128:50070，查看HDFS集群状态，如图2-76所示

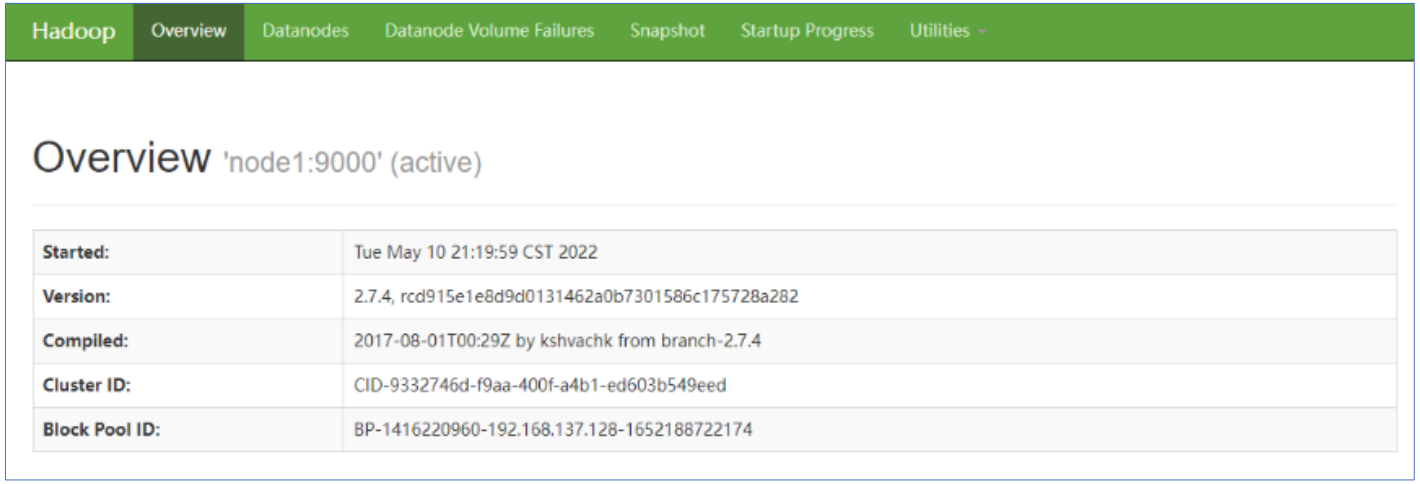


图2-76

3．部署Hadoop(完全分布模式)

步骤9：在Windows系统下，访问http://192.168.137.128:8088，查看Yarn集群状态，如图2-77所示

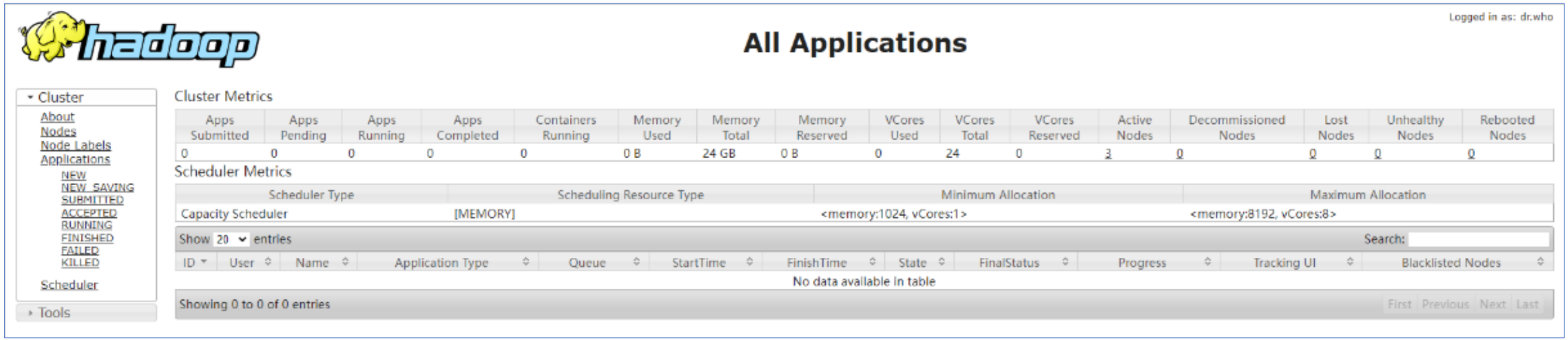


图2-77

任务 2-3 搭建Hive环境

2.3.1 任务描述

1. 기본 시작점 : C:\Users\ebi\Documents\Virtual Machines\Spark16Ram100GB20251212.ova

2. 설치 指南 :
   - Hive 설치 : C:\Users\ebi\Documents\중국강의\강좌명_Hive\Hive설치및설정파일\Hive_Installation_Guide.docx
   - Apache Web Server 설치: C:\Users\ebi\Documents\중국강의\강좌명_Hive\Hive설치및설정파일\ Apache_phpMyAdmin_Installation_Guide.docx