**大数据数据分析初步**

**一、数据获取：**

（一）、获取代码：

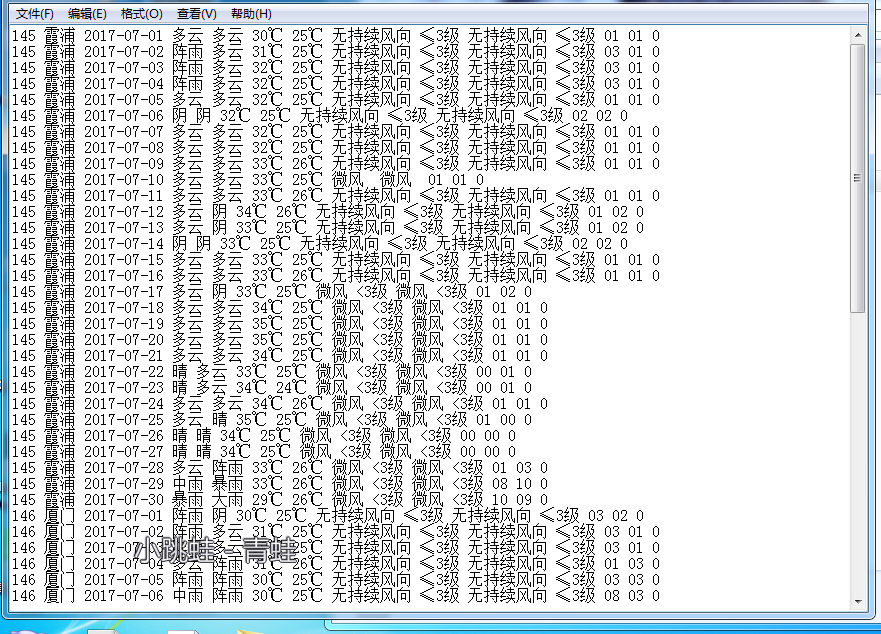
通过聚合网获取数据，根据给定的key值，地区和时间获取相应的天气：

部分代码如下：



以上代码获取的是地区编号为145,146 时间为2017-7 到2017-8月的历史天气；

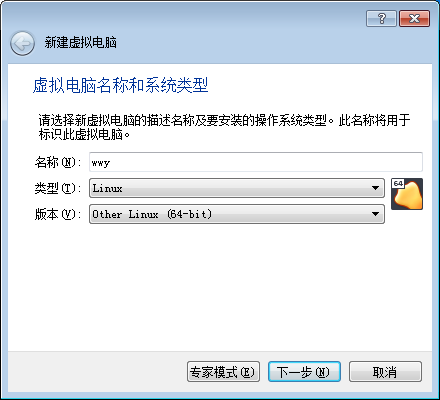
1. 、获取到的数据部分展示：



**二、搭建环境**

（一）、单机环境：

1．安装好centos后新虚拟机：

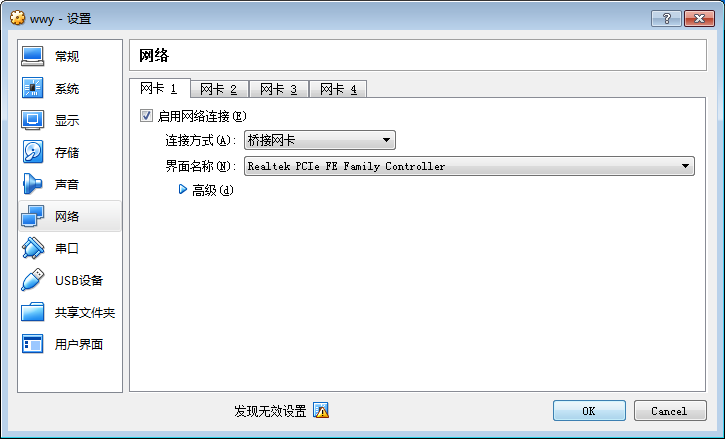


2.内存选择2g



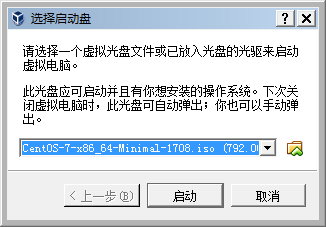
3.其他点击下一步；

进入设置：将网卡调整成桥接

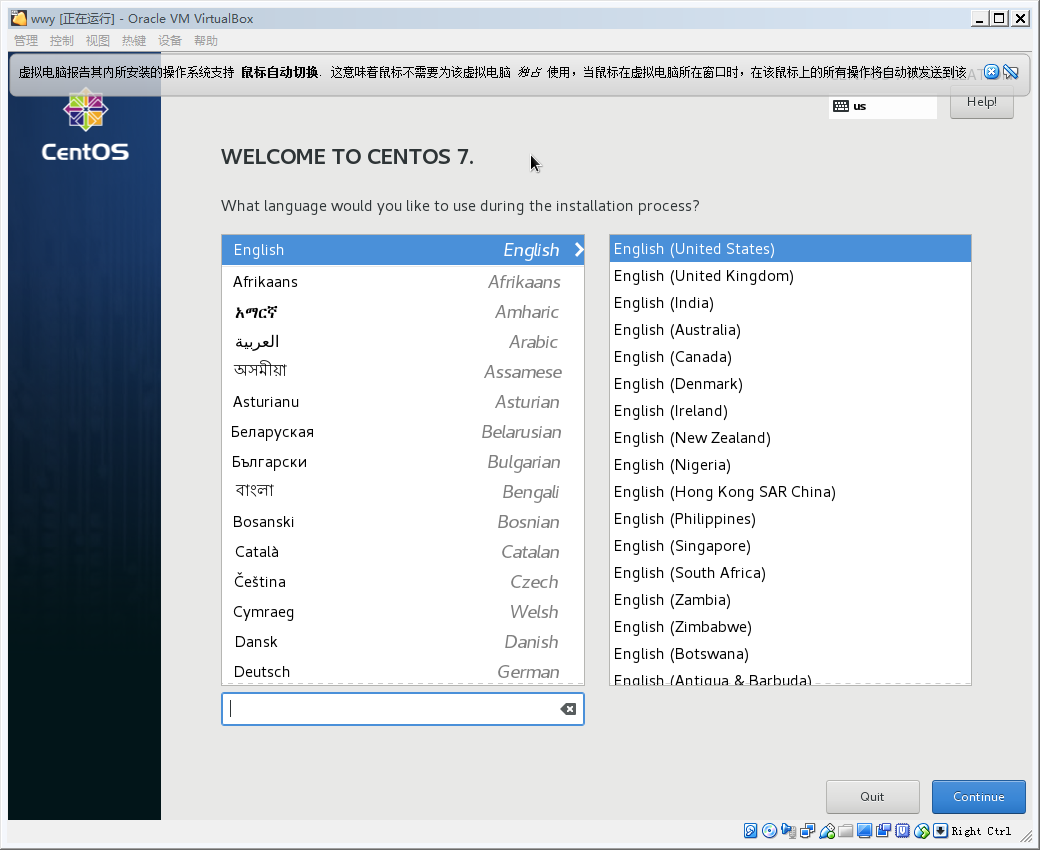


4.启动虚拟机：

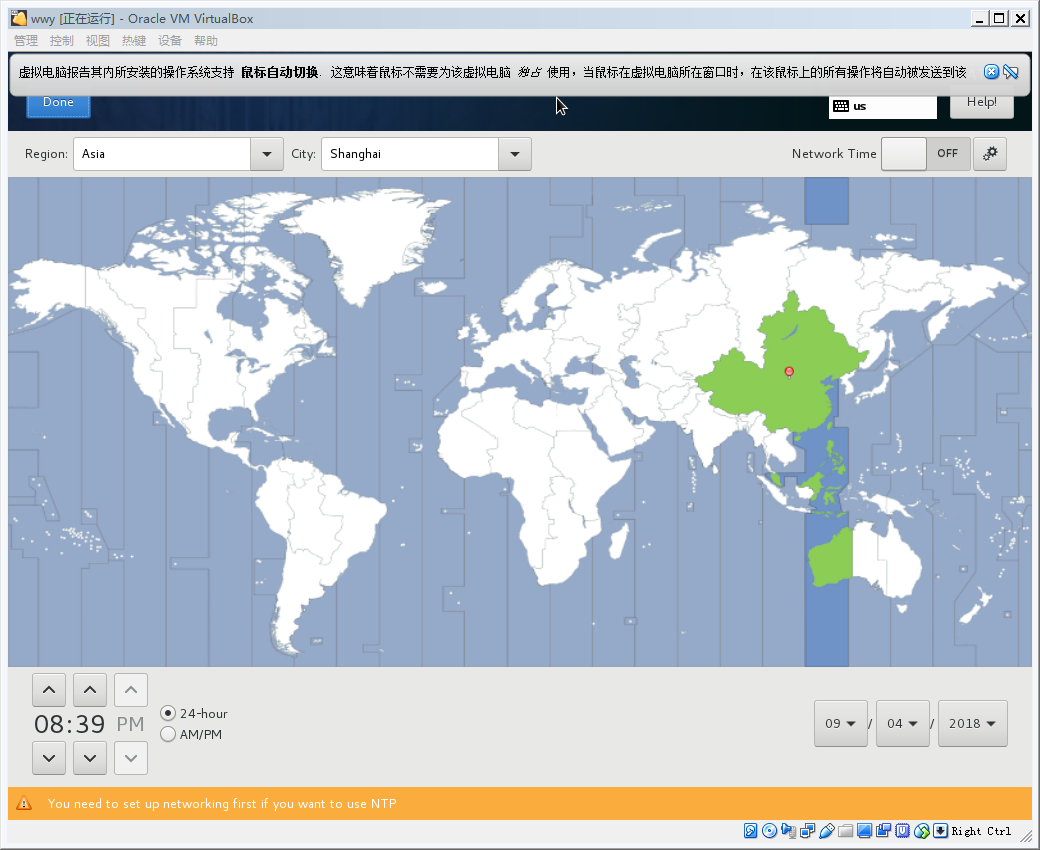
选择启动盘：



5.选择英语



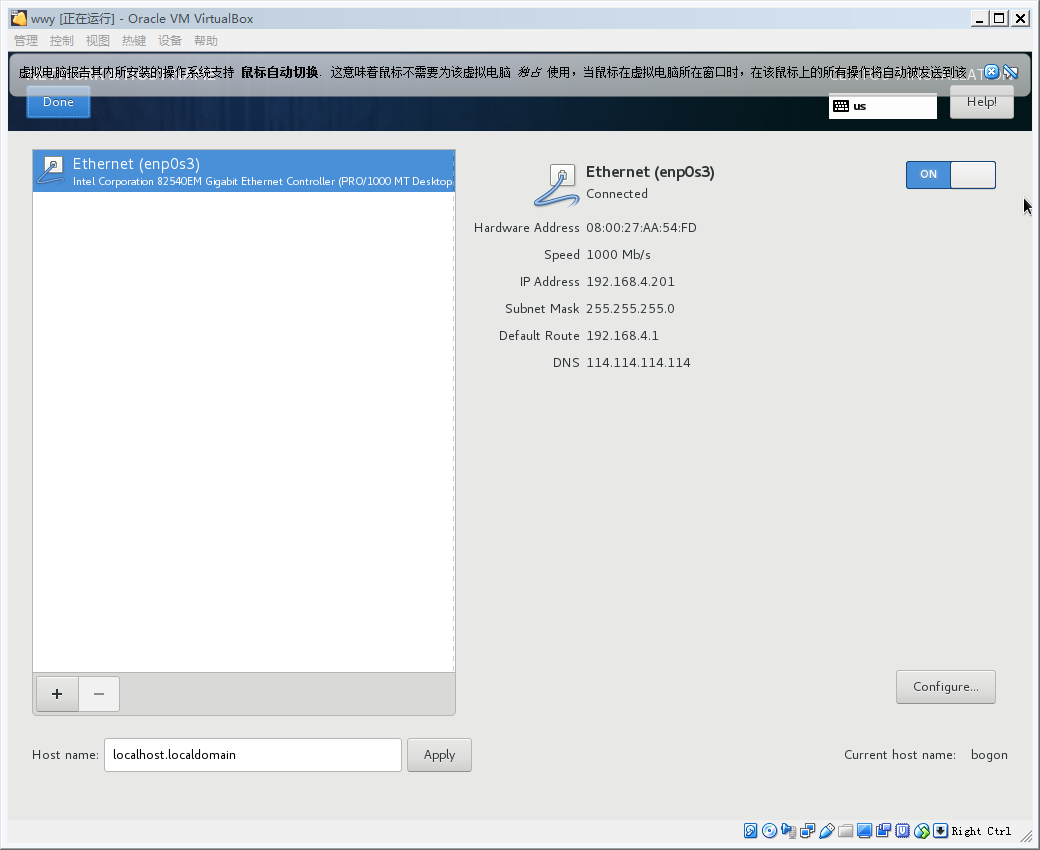
6.时区选择上海：

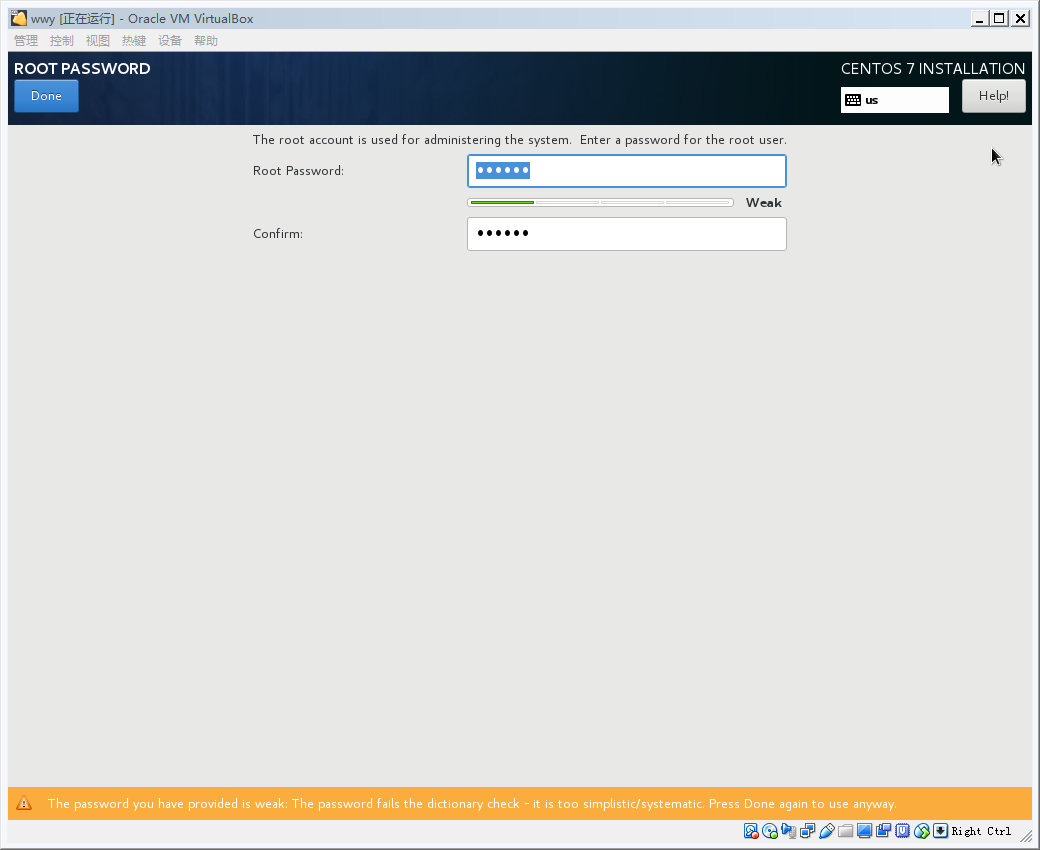


7.该选项点击进去后选择done

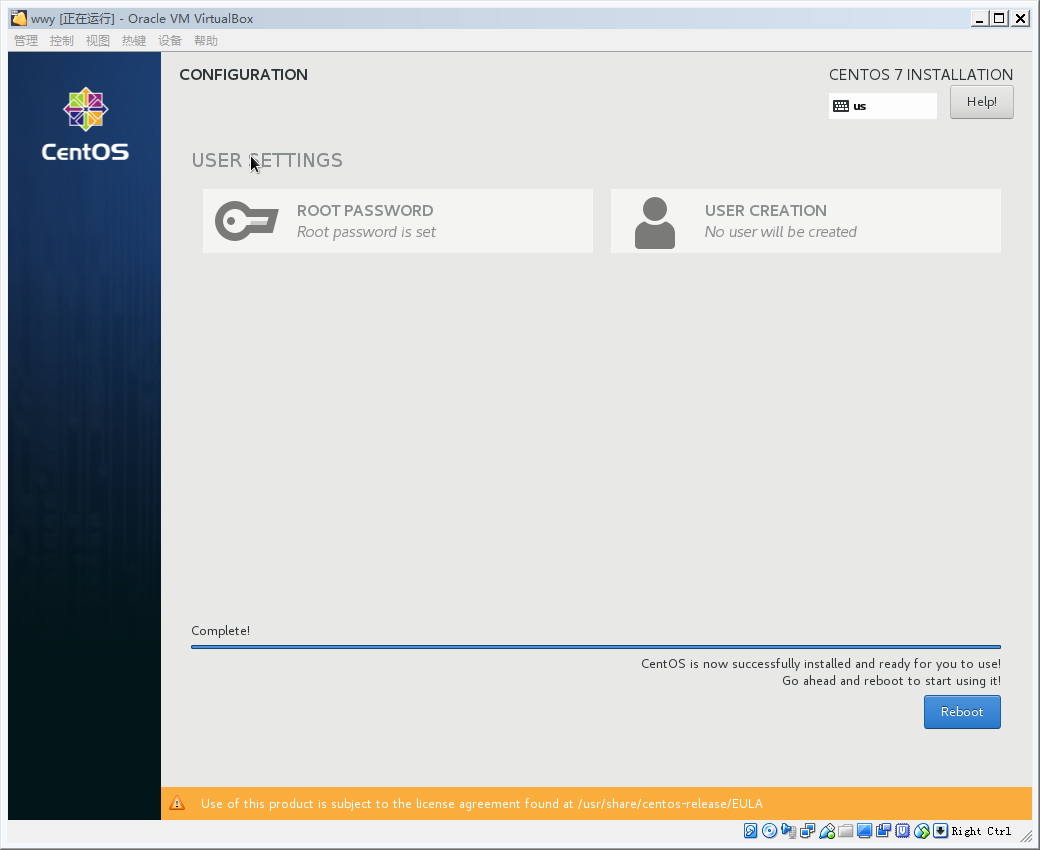


Network打开：

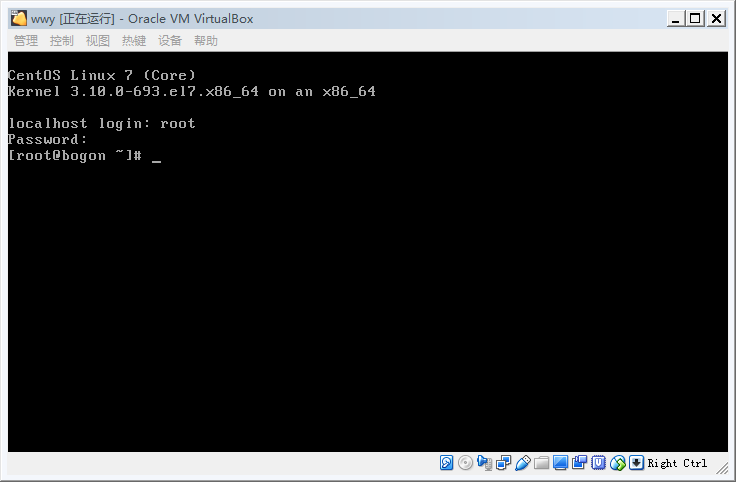


8.设置密码：

9.安装成功后重启：

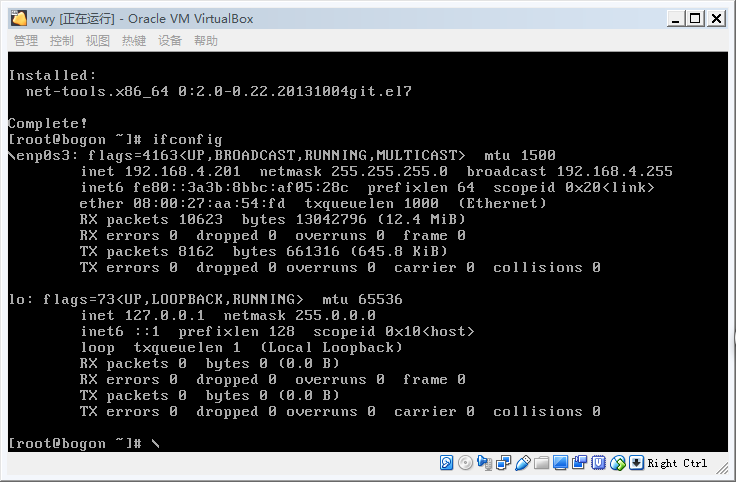


10.登录完成：



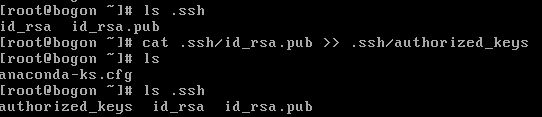
11.下载net-tools

Yum install net-tools下载完成后即可运行ifconfig

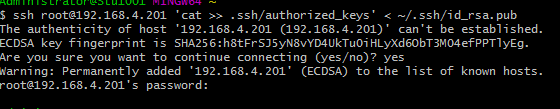


12.输入ssh keygen创建公钥

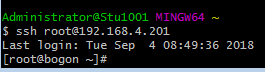
将公钥移至authorized\_keys中



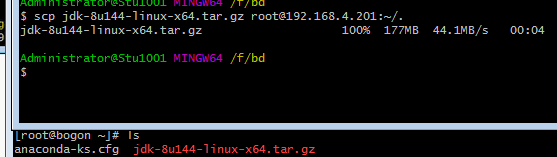
13.在gitbuh中输入下图中命令，将公钥放入authorized\_keys中，可实现免密码登录



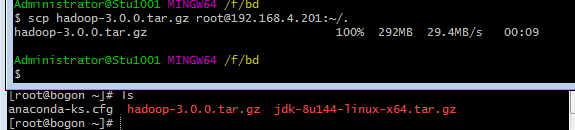
登录结果如下：



14.将jdk传到虚拟机中



15.将hadoop传到虚拟机中：





16.在/etc/profile文件中添加





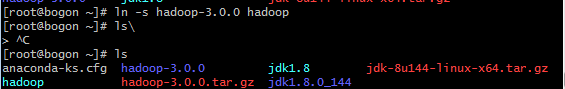
17.在/etc/hosts文件中添加

192.168.4.201 master

然后解压hadoop







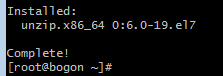
18.将hadoopfile移至虚拟机中



19.下载unzip



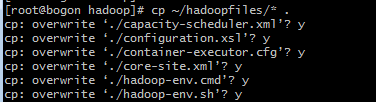
完成



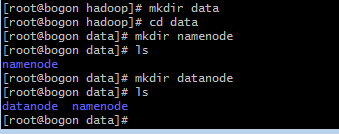
20.解压hadoopfile



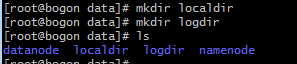
21.将文件复制到hadoopfile中



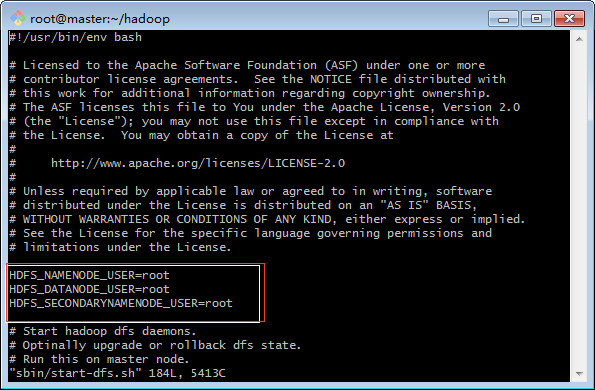
22.在hadoop文件夹中创建data文件夹，并在其中创建namenode和datanode两个文件夹



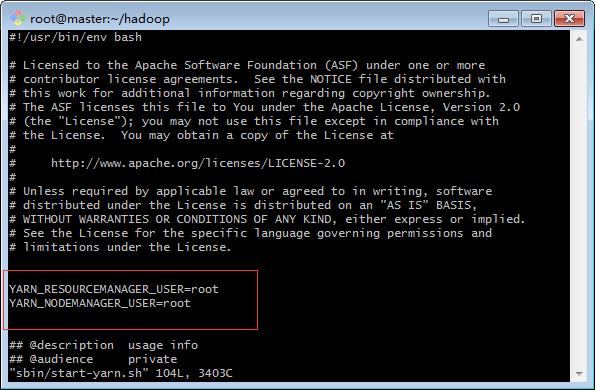
23.再创建localdir和logdir两个文件夹



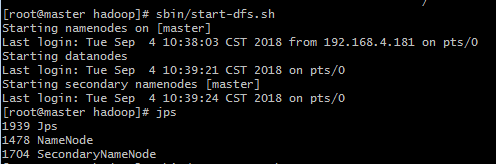
在sbin/start-dfs-sh和中sbin/stop-dfs-sh添加



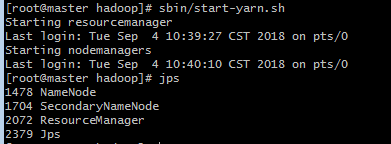
在sbin/start-yarn-sh和中sbin/stop-yarn-sh添加



24.启动start-dfs

= 

启动start-yarn



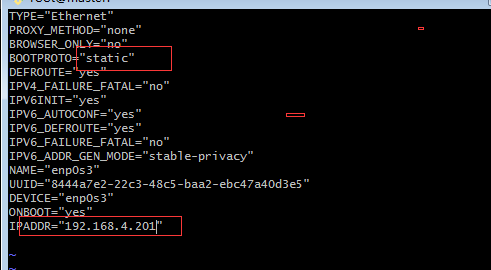
单机配置完成

（二）、联机配置如下(以master为例，slave1和slave2相同)

1.网络设置成桥接网络

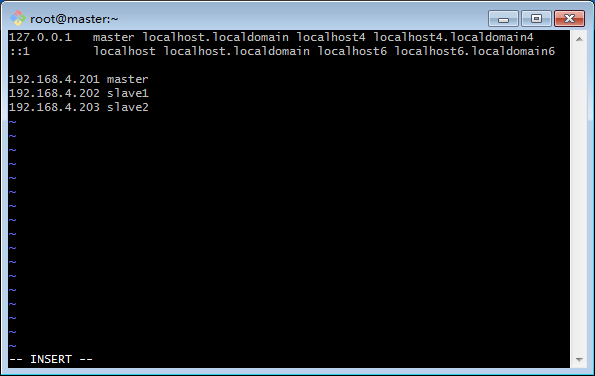
2.修改静态ip





3.将三台机子的ip分配





4.master将公钥分给slave1和slave2





5．Slave1和slave2将公钥发给master





6．配置完成，三台机子可以无密码访问

Master和slave1：





Master和slave2：

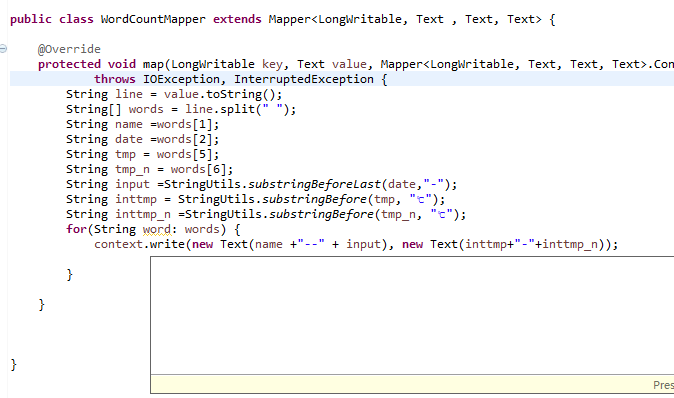




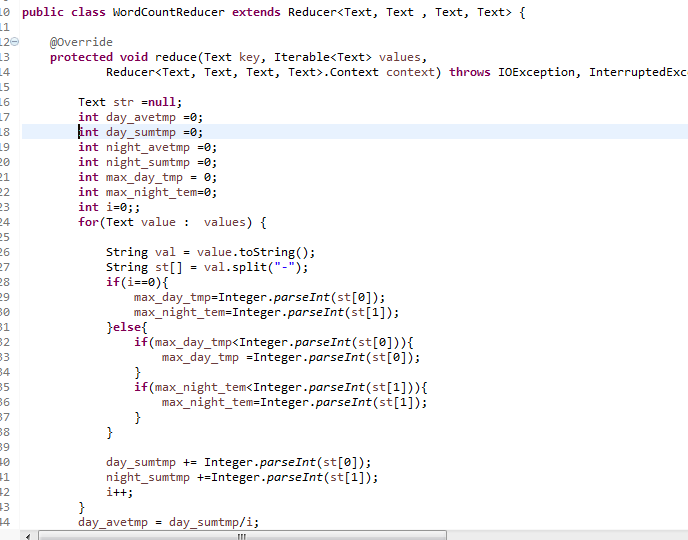
**三、数据分析:**

（一）、代码准备：

WordMapper中



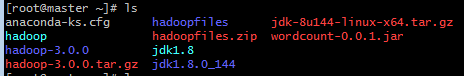
WordCountReduce中：



（二）、数据准备

1.将jar包导入到~目录下





2.将数据文件导入到hadoop文件下





3.创建名为weather1的文件夹，并加入将data1.txt放入



4.运行



（三）分析结果展示：

展示结果：