云南大学软件学院

实 验 报 告

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程： | 大数据分析及应用 | | | 任课教师： | | | | 何婧 | 实验指导教师（签名）： | | | | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 学号： | 2017112007720171120176 | 姓名： | | 王庆  黄峻嘉 | 专业： | | | 软件工程 |  | 日期： | 2019.10.7 | 成绩： | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

# 实验二 Hadoop平台搭建

## 一、 实验目的

1．掌握Ubuntu系统的安装与Linux的基本操作；

2．搭建Hadoop伪分布式平台；

3．了解MapReduce的原理并运行示例。

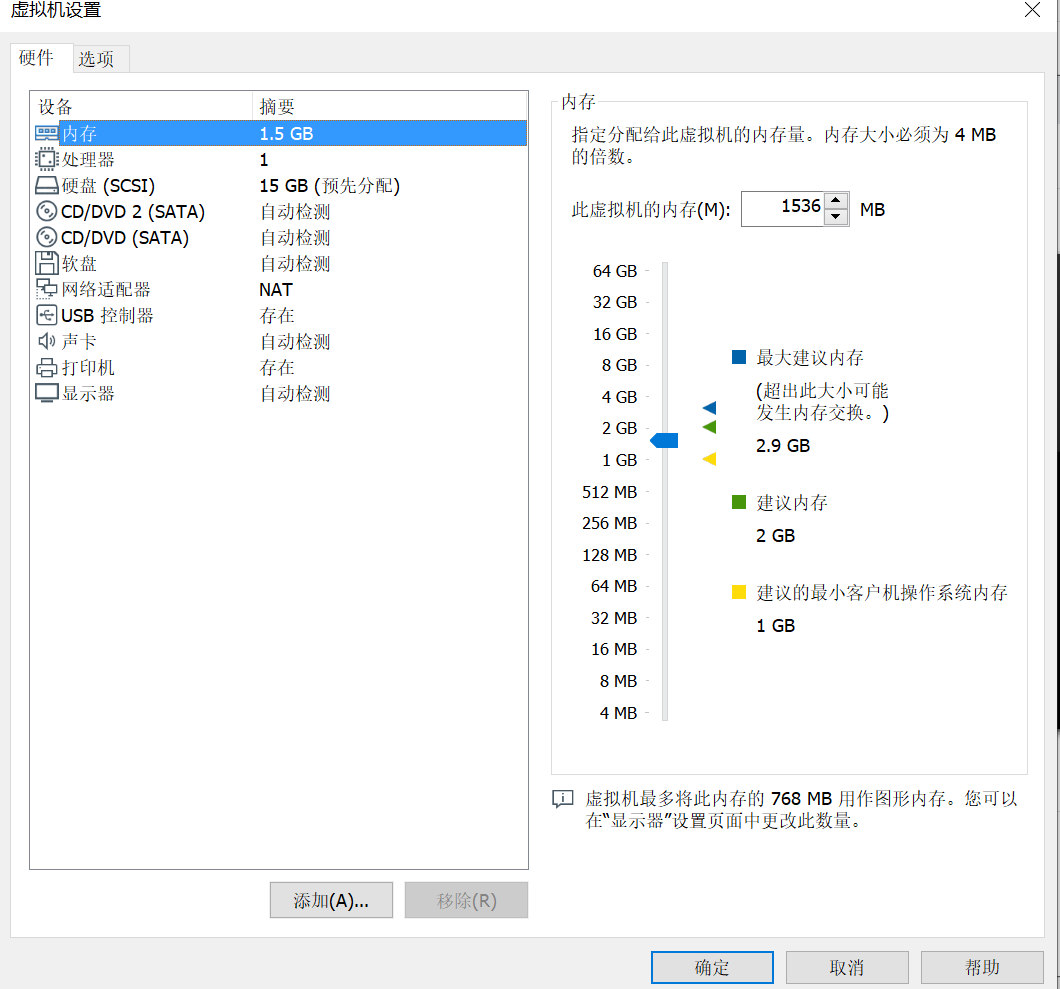
## 二、 实验内容及步骤

### 2.1、Ubuntu系统的安装

1、下载安装VMware



2、下载镜像文件并在VMware中新建虚拟机



3、打开虚拟机配置Ubuntu系统



### 2.2、Hadoop的安装及伪分布式配置

1、创建hadoop用户

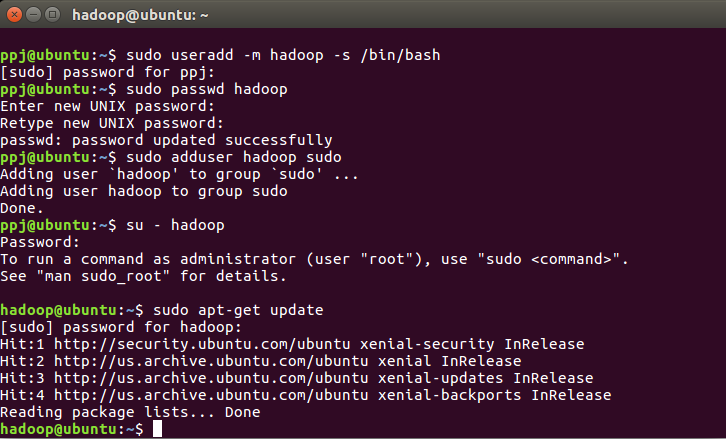
$ sudo useradd -m hadoop -s /bin/bash #创建hadoop用户，并使用/bin/bash作为shell

$ sudo passwd hadoop #为hadoop用户设置密码

$ sudo adduser hadoop sudo #为hadoop用户增加管理员权限

$ su - hadoop #切换当前用户为用户hadoop

$ sudo apt-get update #更新hadoop用户的apt



2、安装SSH，设置SSH无密码登陆

$ sudo apt-get install openssh-server #安装SSH server

$ ssh localhost #登陆SSH

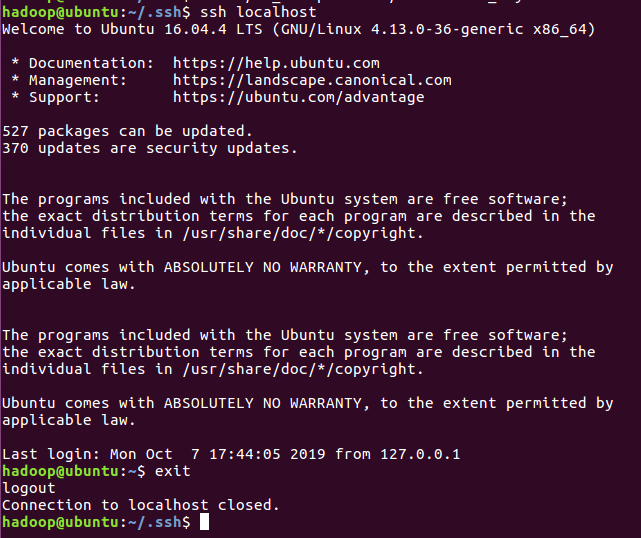
$ exit #退出登录的ssh localhost

$ cd ~/.ssh/

$ ssh-keygen -t rsa

$ cat ./id\_rsa.pub >> ./authorized\_keys #加入授权

$ ssh localhost



3、安装java jdk

（1）下载jdk

（2）解压并添加环境变量

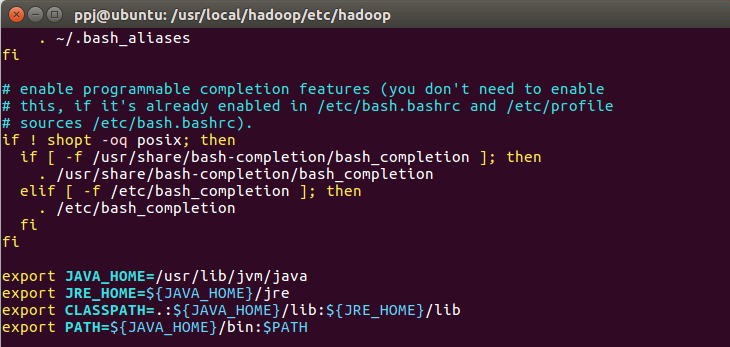
在.bashrc文件添加如下指令：

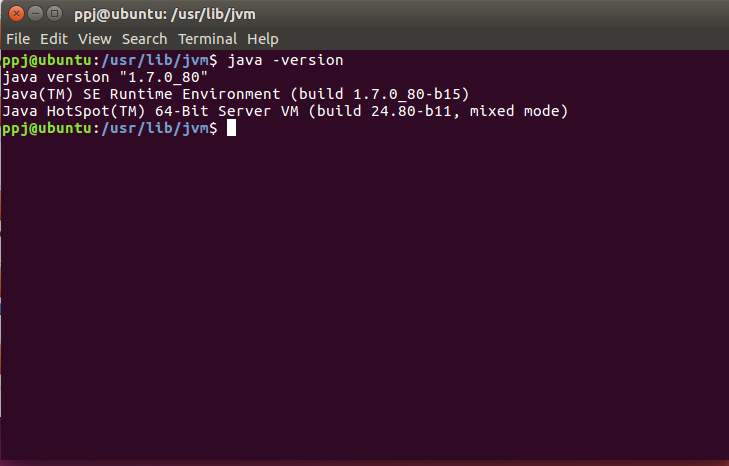
export JAVA\_HOME=/usr/lib/jvm/java

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

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

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





4、安装Hadoop

（1）下载Hadoop

（2）解压并添加环境变量

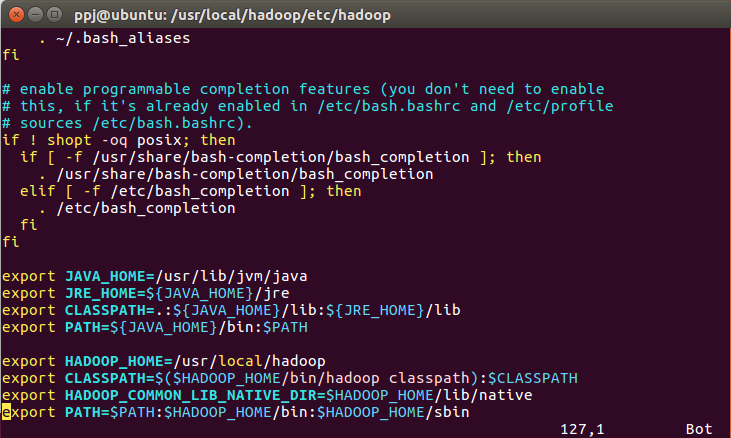
给hadoop配置环境变量，将下面代码添加到.bashrc文件:

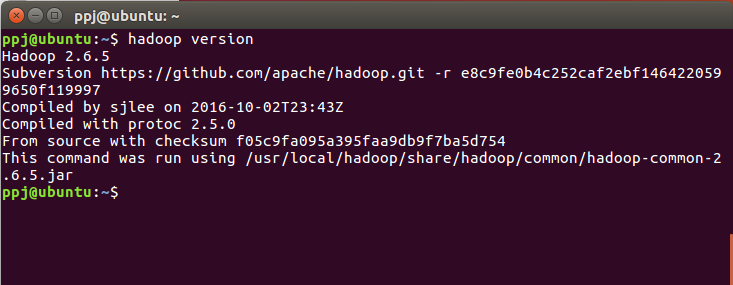
export HADOOP\_HOME=/usr/local/hadoop

export CLASSPATH=$($HADOOP\_HOME/bin/hadoop classpath):$CLASSPATH

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

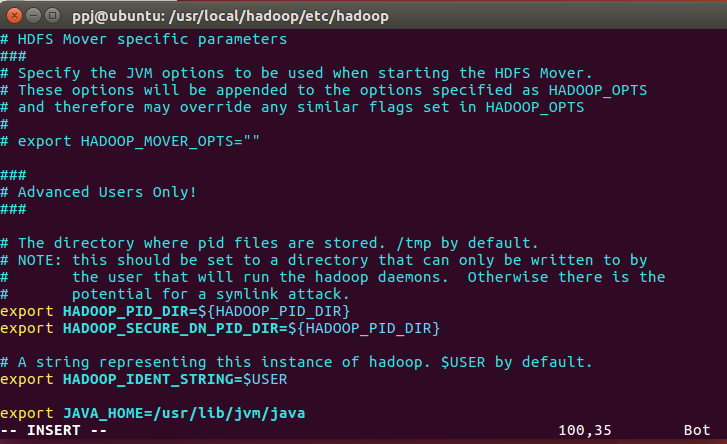
export PATH=$PATH:$HADOOP\_HOME/bin:$HADOOP\_HOME/sbin



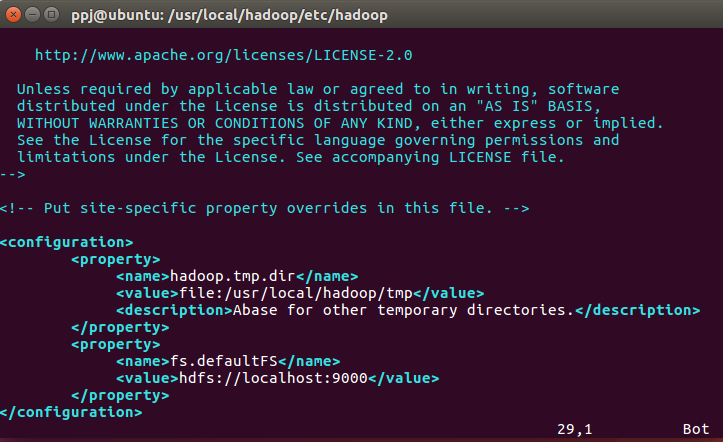


5、伪分布式配置

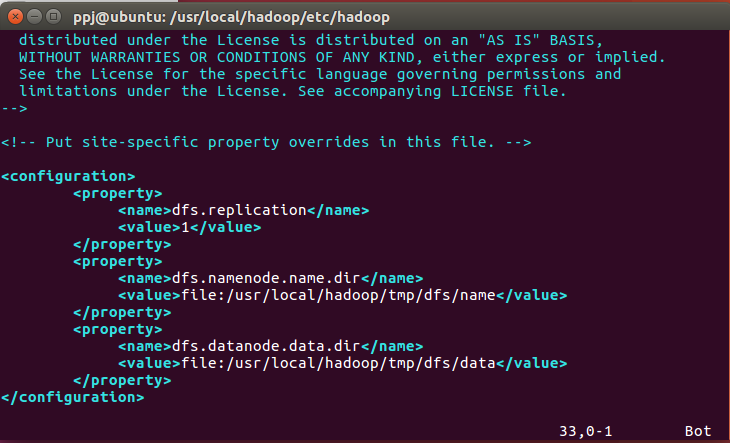
（1）将jdk的路径添（export JAVA\_HOME=/usr/lib/jvm/java ）加到hadoop-env.sh文件



（2）修改配置文件 core-site.xml



（3）修改配置文件hdfs-site.xml



（4）执行 NameNode 的格式化

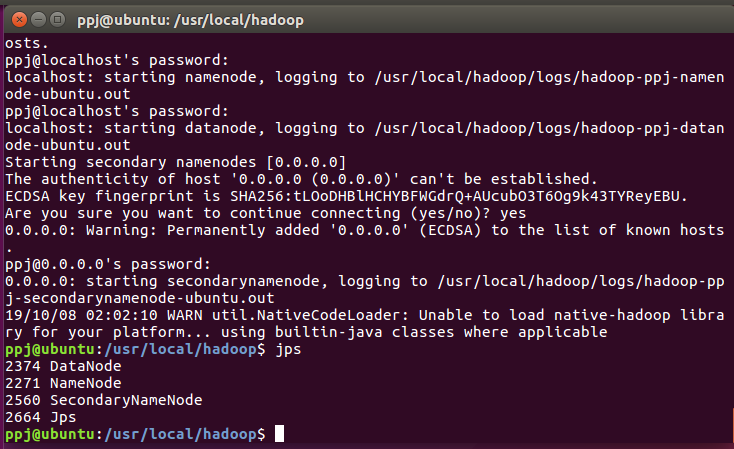
$ ./bin/hdfs namenode -format

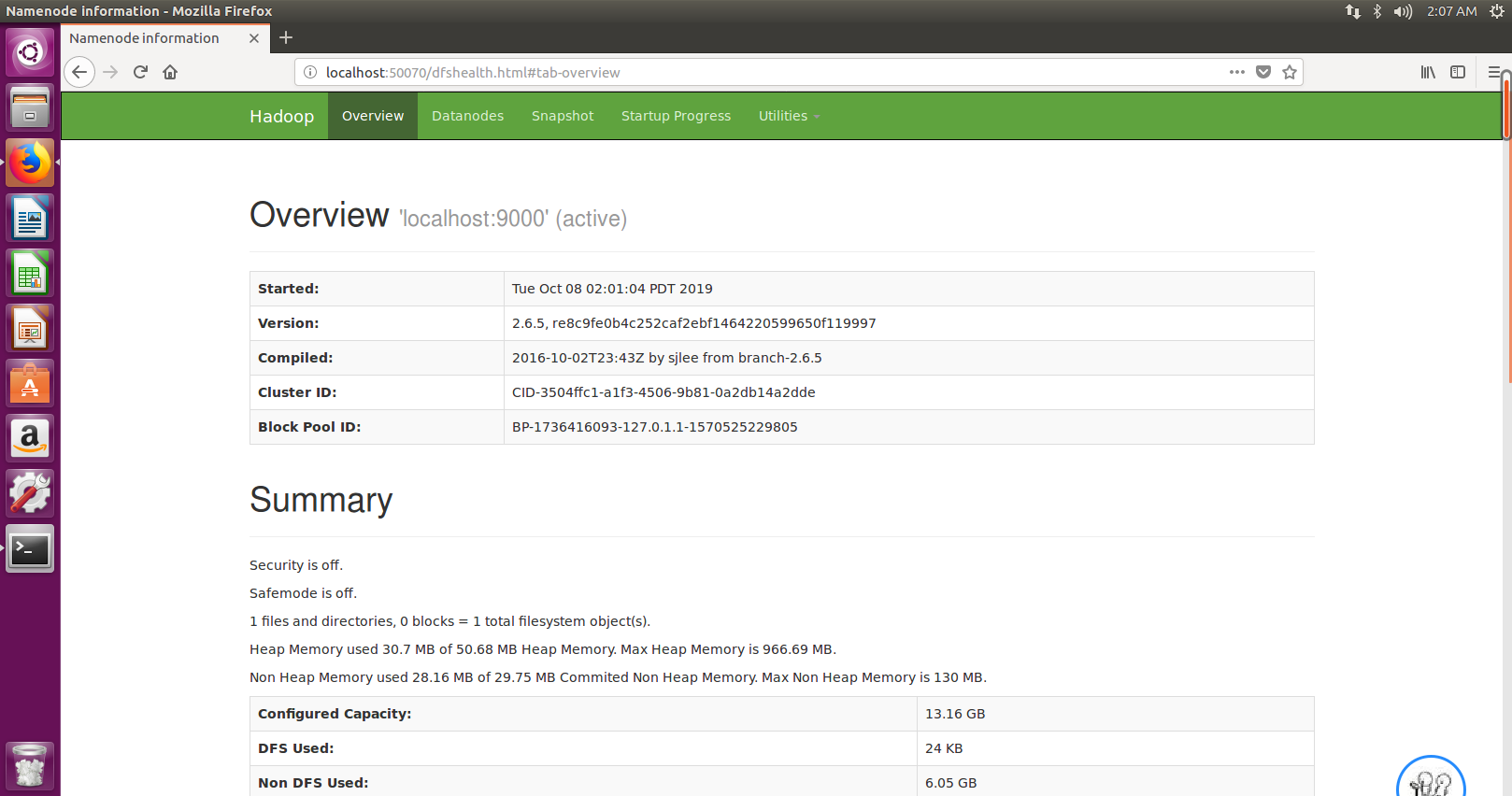
（5）启动namenode和datanode进程

$ ./sbin/start-dfs.sh

（6）查看启动结果

$ jps



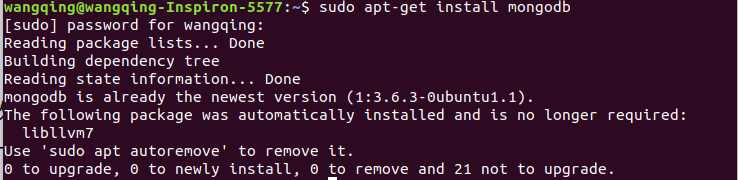


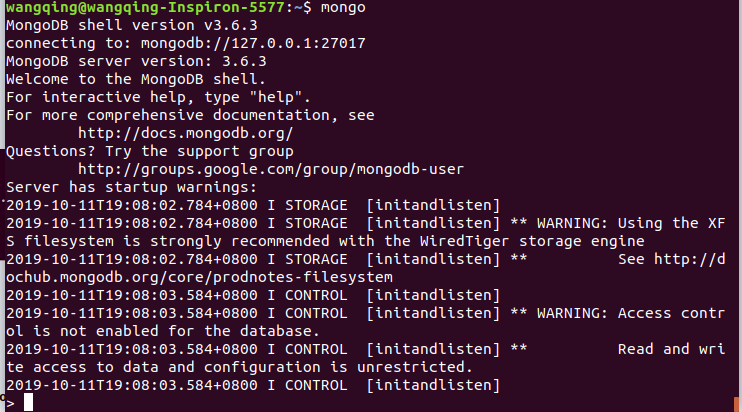
### 2.3 nosql 数据库的应用

**mongodb数据库：**

1. mongodb数据库的安装

在ubuntu系统上，直接使用命令sudo apt-get install mongodb进行安装安装完成之后就可以直接使用，默认端口号为：27017。下图即为安装之后的截图：

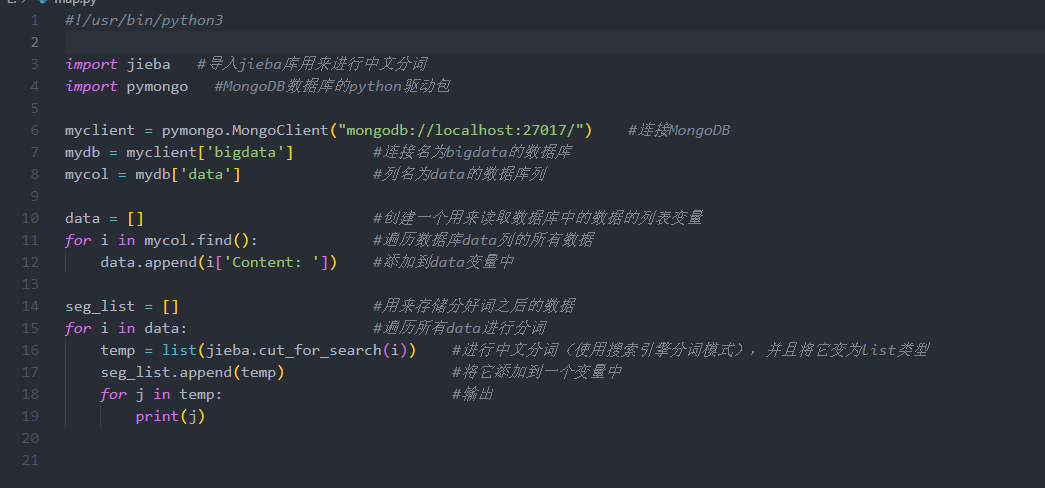


1. 运行截图

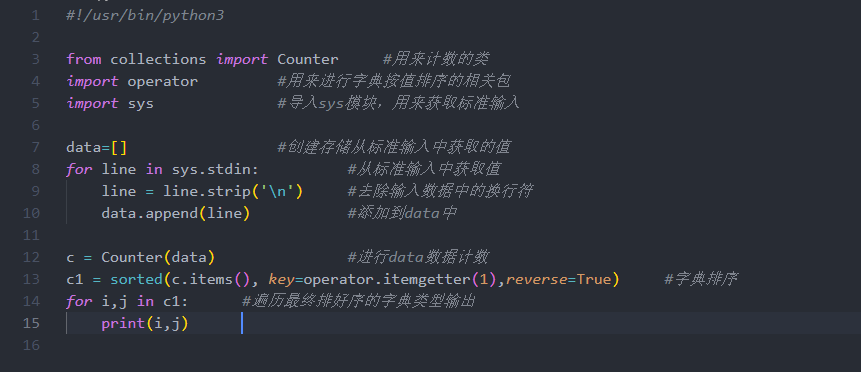
### 2.4、MapReduce原理详解（给出一个运行示例）

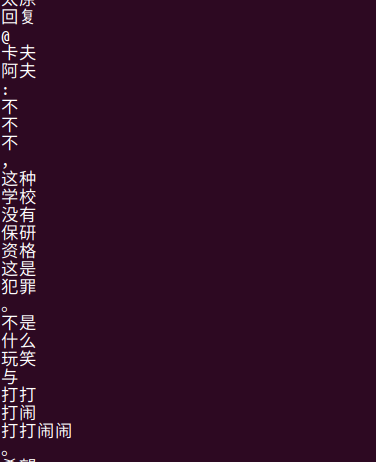
**通过Mongodb和python3.6进行MapReduce：**

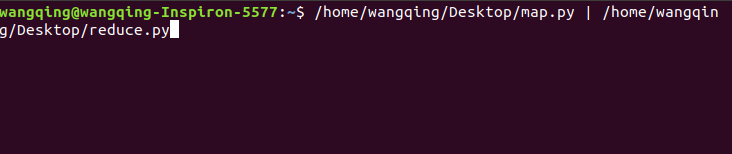
1. mapper代码截图

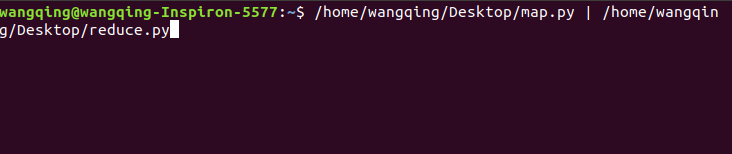
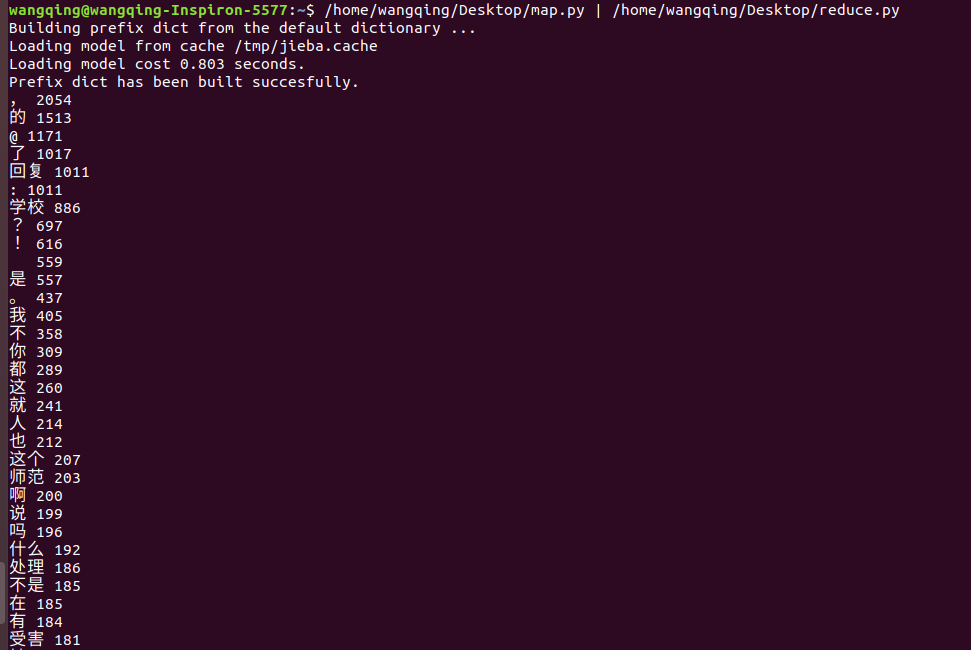


1. reducer代码截图



1. ****mapper代码的运行

****

1. ****MapReduce的执行结果

## 三、 实验总结

通过本次实验了解了Hadoop伪分布模式是在单机上模拟Hadoop分布式。是单机上的分布式并不是真正的分布式，而是使用线程模拟分布式。

对于hadoop有了初步的认识。Hadoop包括并行计算模型Map/Reduce、分布式文件系统HDFS，以及分布式数据库Hbase。其中MapReduce是hadoop的核心组件之一，hadoop要实现分布式需要包括两部分，一部分是分布式文件系统hdfs，一部分是分布式计算框架mapreduce，缺一不可，可以通过mapreduce很容易在hadoop平台上进行分布式的计算编程。

学习了nosql的相关知识，掌握了MongoDB的简单使用方法。Nosql满足CAP定理：

Consistency(一致性), 数据一致更新，所有数据变动都是同步的

Availability(可用性), 好的响应性能

Partition tolerance(分区容错性) 可靠性

P: 系统中任意信息的丢失或失败不会影响系统的继续运作