安装单机 Hadoop 系统和 WordCount 程序实验

姓名 查鹏　邮箱:1613541957@qq.com　联系方式：15895987843

(南京大学 计算机科学与技术系, 南京　210046)

1　实验要求

1.1　在本地电脑上正确安装和运行伪分布式Hadoop系统；

1.2　从网页中获取一组英文网页数据，在本机自带的WordCount可执行程序文件并产生结果。

2　实验环境和数据说明

2.1　Ubuntu环境：Ubuntu12.04.5

2.2　Java环境：JDK1.8.0_77

2.3　Hadoop版本：Hadoop1.0.4

2.4　WordCount实验网页数据：

http://stackoverflow.com/questions/36598111/not-able-to-stop-dbms-scheduler-job

总共一个网页。

3　实验过程

3.1　安装和配置JDK

3.1.1　从网页上下载 Linux 版本的 JDK，我下载的是 JDK1.8.0_77，将其拷入/usr 目录下；

3.1.2　使用指令 sudo vim /etc/profile 配置 JAVA_HOME 和 CLASS_PATH，截图如下：

```
export JAVA_HOME=/usr/jdk1.8.0_77
export PATH=$JAVA_HOME/bin:$PATH
export CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar
export JAVA_HOME PATH CLASSPATH
```

3.1.3　使用指令 source /etc/profile 使得刚刚的配置文件成效，然后使用指令 java –version 查看当前 Java 版本：

```
user@ubuntu:~$ java -version
java version "1.8.0_77"
Java(TM) SE Runtime Environment (build 1.8.0_77-b03)
Java HotSpot(TM) Client VM (build 25.77-b03, mixed mode)
```

3.2　下载安装Hadoop

从官网下载 Hadoop 压缩包，我下载的是 1.0.4 版本。移入 Ubuntu 中，然后使用指令 tar –xvf hadoop-1.0.4.tar.gz 进行解压。

3.3 配置SSH

3.3.1 使用指令 ssh-key –t rea 生成密钥对，然后一直按回车键，按照默认方式运行到结束；

```
user@ubuntu:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/user/.ssh/id_rsa):
/home/user/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/user/.ssh/id_rsa.
Your public key has been saved in /home/user/.ssh/id_rsa.pub.
The key fingerprint is:
b1:a2:86:67:dc:68:aa:5d:35:47:a5:fb:5b:bb:91:cc user@ubuntu
The key's randomart image is:
+--[ RSA 2048]----+
|          .      |
|         o       |
|        +        |
|       . +       |
|      + S        |
|   o = + . o .   |
|  . X .   . E    |
| . B       o o   |
|..o        . o.  |
+-----------------+
```

3.3.2 使用指令 cd .ssh 进入.ssh 目录，然后使用指令 cp id_rsa.pub authorized_keys；

3.3.3 然后执行命令 ssh localhost，测试一下是否可以实现用 SSH 进行连接并且不需要输入密码。在执行这一步之前，需要先用指令 sudo apt-get install openssh-server 按照 openssh-server。最后结果如下：

```
user@ubuntu:~/.ssh$ ssh localhost
Welcome to Ubuntu 12.04.5 LTS (GNU/Linux 3.5.0-23-generic i686)
```

3.4 Hadoop环境配置和运行

3.4.1 使用指令 cd hadoop-1.0.4 进入 hadoop 文件目录；

3.4.2 使用指令 vim core-site.xml 进入 core-site.xml 文件，进行配置

```
<configuration>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
<description>The name of the default file system. A URI whose
scheme and authority determine the FileSystem implementation.
</description>
</property>
</configuration>
```

3.4.3 使用指令 vim hdfs-site.xml 进入 hdfs-site.xml 文件进行 hdfs 配置

```
<configuration>
<property>
<name>dfs.republication</name>
<value>1</value>
<description>The actual number of republication can be specified when the
file is created.</description>
</property>
</configuration>
```

3.4.4 使用指令 vim mapred-site.xml 进入 mapred-site.xml 文件进行配置

```
<configuration>
<property>
<name>mapred.job.tracker</name>
<value>localhost:9001</value>
<description> The host and port that the MapReduce job tracker runs at.
</description>
</property>
</configuration>
```

3.4.5 使用指令 bin/hadoop namenode –format 格式化分布式文件系统 HDFS，然后使用指令 bin/start-all.sh 启动 Hadoop 守护进程，然后使用是在 jps 查看是否正常启动。

```
user@ubuntu:~/hadoop-1.0.4$ jps
3475 SecondaryNameNode
3876 Jps
3253 DataNode
3574 JobTracker
3000 NameNode
3816 TaskTracker
```

3.5 WordCount

3.5.1 从网页中下载一份网页数据，我是从下面这个网页下载的网页数据：

http://stackoverflow.com/questions/36598111/not-able-to-stop-dbms-scheduler-job

将下载的网页数据存入 file 文件夹中，并将 file 文件夹移入 Ubuntu 的/home/user 目录下；

3.5.2 使用指令 bin/hadoop dfs -copyFromLocal /home/user/file test-in，将文件复制到 HDFS 文件系统中；

3.5.3 使用指令 bin/hadoop jar hadoop-sss.examples.jar wordcount test-in test out，进行词频统计实验，代码运行的过程和运行经过在下面给出。

3.5.4 代码运行结束后，使用指令 binhadoop dfs –copyToLocal test-in test –out，将得到的 test-out 文件移到本地目录下，方便查看得到的统计数据。

4 实验结果

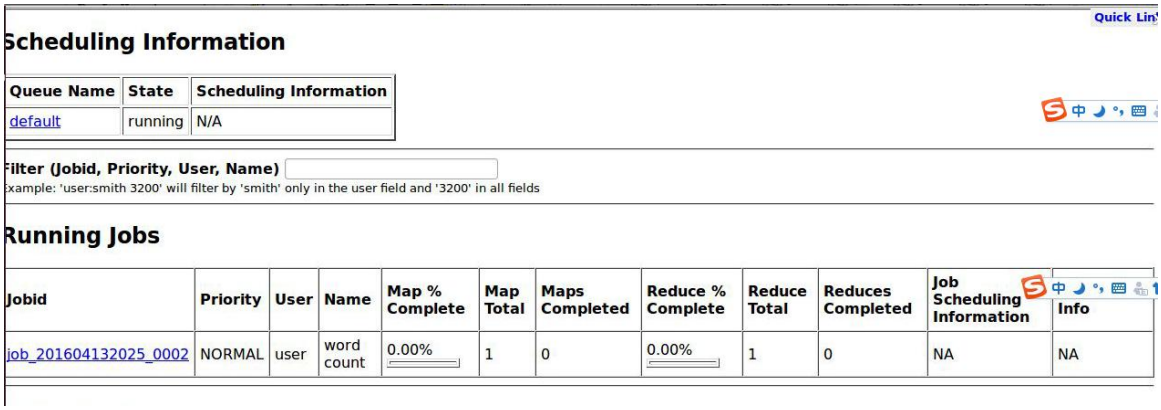
4.1 Java版本查看

```
user@ubuntu:~$ java -version
java version "1.8.0_77"
Java(TM) SE Runtime Environment (build 1.8.0_77-b03)
Java HotSpot(TM) Client VM (build 25.77-b03, mixed mode)
```

4.2 Hadoop安装运行结果

```
user@ubuntu:~/hadoop-1.0.4$ jps
3475 SecondaryNameNode
3876 Jps
3253 DataNode
3574 JobTracker
3000 NameNode
3816 TaskTracker
```

4.3 词频统计过程中的网页截图



Scheduling Information

Queue Name	State	Scheduling Information
default	running	N/A

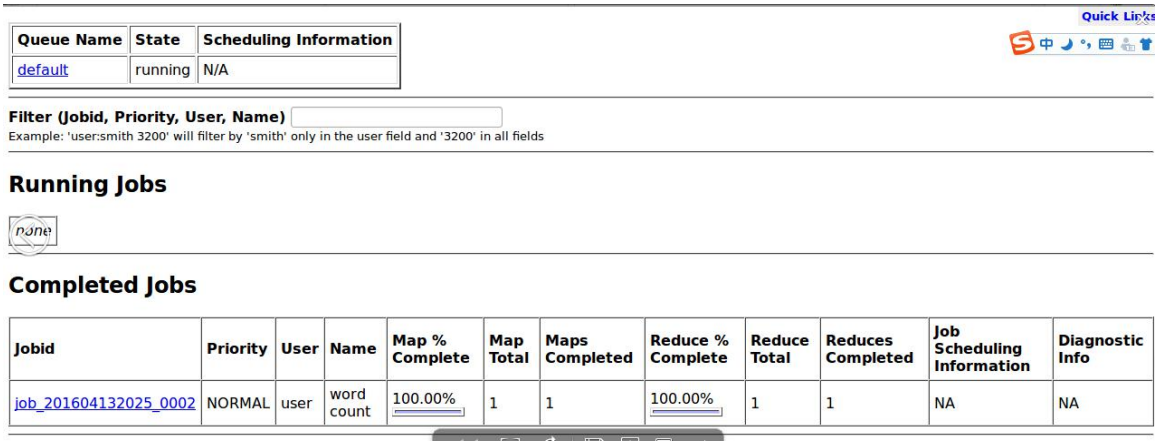
Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Info
job_201604132025_0002	NORMAL	user	word count	0.00%	1	0	0.00%	1	0	NA	NA

4.4 实验完成后的网页截图



Queue Name	State	Scheduling Information
default	running	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

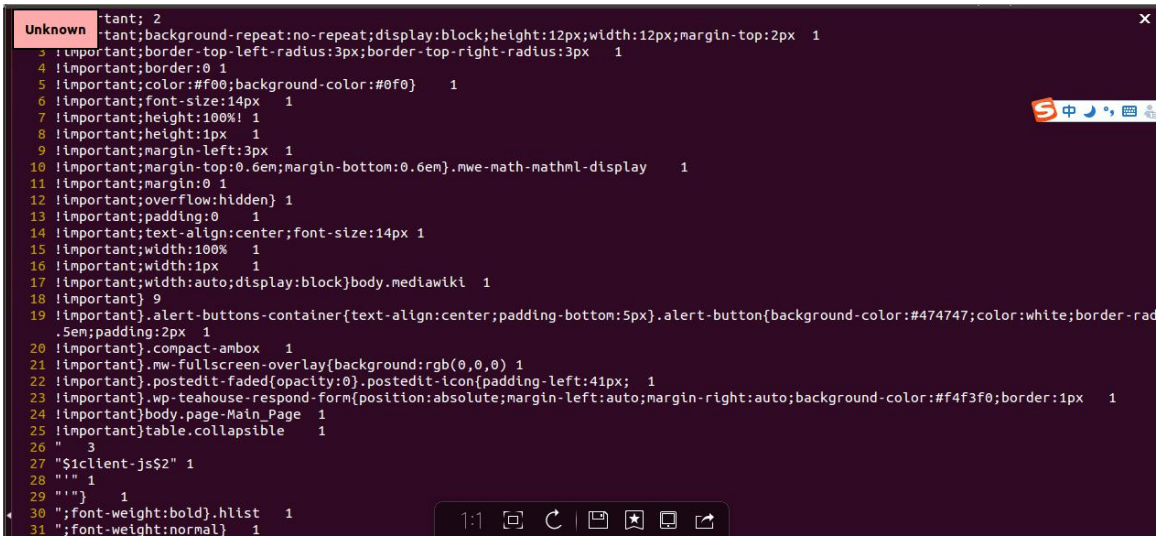
Running Jobs

none

Completed Jobs

Jobid	Priority	User	Name	Map % Complete	Map Total	Maps Completed	Reduce % Complete	Reduce Total	Reduces Completed	Job Scheduling Information	Diagnostic Info
job_201604132025_0002	NORMAL	user	word count	100.00%	1	1	100.00%	1	1	NA	NA

4.5 词频统计结果



5 实验体会

这次实验按照讲义和课本上的步骤做就做好了，基本没有什么难度，只是有的地方需要用 bin/hadoop，而且在 SSH 配置时需要事先按照 openssh-server。

这次 Hadoop 安装实验，我了解了 hadoop 编程和运行的基本思路，为以后的实验做准备。