Programing Map-Reduce

(Hadoop)

with Eclipse

+

Wei-Yu Chen

NCHC

2008/05/27

see more : http://trac.nchc.org.tw/cloud/

1. Prepare :

- **System :**

  - Ubuntu 7.10
  - Hadoop 0.16

- **Requirement :**

  - Eclipse (3.2.2)

```
$ apt-get install eclipse
```

  - java 6

```
$ apt-get install sun-java6-bin sun-java6-jdk sun-java6-jre sun-java6-plugin
```

  - suggest to remove the default java compiler gcj

```
$ apt-get purge java-gcj-compat
```

  - Append two codes to /etc/bash.bashrc to setup Java Class path

```
export JAVA_HOME=/usr/lib/jvm/java-6-sun
export HADOOP_HOME=/home/waue/workspace/hadoop/
export CLASSPATH=.:$JAVA_HOME/lib/dt.jar:$JAVA_HOME/lib/tools.jar
```

- **Building UP Path**

Name	Path
Hadoop Home	/home/waue/workspace/hadoop/
Java Home	/usr/lib/jvm/java-6-sun

2. Hadoop Setup

1. Generate an SSH key for the user.

```
$ ssh-keygen -t rsa -P ""
$ cat ~/.ssh/id_rsa.pub >> ~.ssh/authorized_keys
$ ssh localhost
$ exit
```

2. Installation Hadoop

```
$ cd /home/waue/workspace
$ sudo tar xzf hadoop-0.16.0.tar.gz
$ sudo mv hadoop-0.16.0 hadoop
$ sudo chown -R waue:waue hadoop
$ cd hadoop
```

3. Configuration

   1. hadoop-env.sh ($HADOOP_HOME/conf/)

   - Change

```
# The java implementation to use.  Required.
# export JAVA_HOME=/usr/lib/j2sdk1.5-sun
```

to

```
# The java implementation to use.  Required.
export JAVA_HOME=/usr/lib/jvm/java-6-sun
export HADOOP_HOME=/home/waue/workspace/hadoop
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export HADOOP_SLAVES=$HADOOP_HOME/conf/slaves
```

   2. hadoop-site.xml ($HADOOP_HOME/conf/)

   - modify the contents of conf/hadoop-site.xml as below

```
<configuration>
<property>
  <name>fs.default.name</name>
  <value>localhost:9000</value>
  <description>
  </description>
</property>
<property>
  <name>mapred.job.tracker</name>
  <value>localhost:9001</value>
  <description>
  </description>
<property>
```

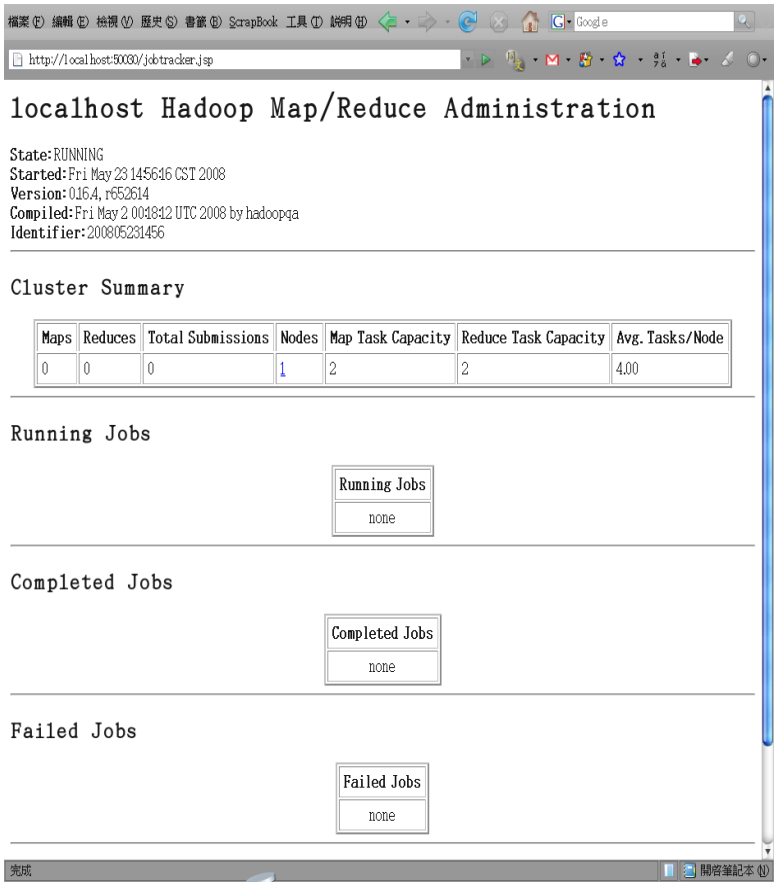
```
  <name>mapred.map.tasks</name>
  <value>1</value>
  <description>
    define mapred.map tasks to be number of slave hosts
  </description>
</property>
<property>
  <name>mapred.reduce.tasks</name>
  <value>1</value>
  <description>
    define mapred.reduce tasks to be number of slave hosts
  </description>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

4. Start Up Hadoop

```
$ cd $HADOOP_HOME
$ bin/hadoop namenode -format
08/05/23 14:52:16 INFO dfs.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = Dx7200/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 0.16.4
STARTUP_MSG:   build = http://svn.apache.org/repos/asf/hadoop/core/branches/branch-0.16 -r 652614; compiled by 'hadoopqa' on Fri May  2 00:18:12 UTC 2008
************************************************************/
08/05/23 14:52:17 INFO fs.FSNamesystem: fsOwner=waue,waue,adm,dialout,cdrom,floppy,audio,dip,video,plugdev,staff,scanner,lpadmin,admin,netdev,powerdev,vboxusers
08/05/23 14:52:17 INFO fs.FSNamesystem: supergroup=supergroup
08/05/23 14:52:17 INFO fs.FSNamesystem: isPermissionEnabled=true
08/05/23 14:52:17 INFO dfs.Storage: Storage directory /tmp/hadoop-waue/dfs/name has been successfully formatted.
08/05/23 14:52:17 INFO dfs.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at Dx7200/127.0.1.1
************************************************************/
```

```
$ /bin/start-all.sh
starting namenode, logging to /home/waue/workspace/hadoop/logs/hadoop-waue-namenode-Dx7200.out
localhost: starting datanode, logging to /home/waue/workspace/hadoop/logs/hadoop-waue-datanode-Dx7200.out
localhost: starting secondarynamenode, logging to /home/waue/workspace/hadoop/logs/hadoop-waue-secondarynamenode-Dx7200.out
starting jobtracker, logging to /home/waue/workspace/hadoop/logs/hadoop-waue-jobtracker-Dx7200.out
localhost: starting tasktracker, logging to /home/waue/workspace/hadoop/logs/hadoop-waue-tasktracker-Dx7200.out
```

Then make sure http://localhost:50030/ by your explorer is on going.



Ps : if your system had error after restart, you could do there for resolving and renewing one.

```
$ cd $HADOOP_HOME
$ bin/stop-all.sh
$ rm -rf /tmp/*
$ rm -rf  logs/*
```

And repeat to 4. start up Hadoop

3. Eclipse Setup

3.1 install IBM mapReduce tool

1. Download the IBM MapReduce Tools zip file and extract to /tmp/.

2. Make sure Eclipse is closed and ...

```
$ cd /tmp/
$ unzip mapreduce_tools.zip
$ mv plugins/com.ibm.hipods.mapreduce* /usr/lib/eclipse/plugins/
```

3. Restart Eclipse

Check IBM MapReduce Tools plugin installing well

Eclipse
File > New > Project ● see **MapReduce category**

3.2 Eclipse configure

Eclipse
Window > Preferences > java> compiler ● set **compiler compliance level** to **5.0**

- Some eclipse-plugin may exhaust much resource, you may happen to out of memory error . We suggest to execute eclipse with some parameters as that :

```
$ eclipse -vmargs -Xmx 512m
```

4. Run on Eclipse

4.1 map-reduce sample code

Eclipse
File > new > project > map-reduce project > next > • **project name : sample** • **use default location : V** • **use default Hadoop : V** **> Finish**

at Project explorer , you will see sample tree. Now, you should create a sample code.

Eclipse
right click **sample > new > file >** • **file name : WordCount.java**

the sample code is here

http://trac.nchc.org.tw/cloud/attachment/wiki/hadoop-sample-code/WordCount.java

paste the contents to your new adding file WordCount.java

4.2. Connect to Hadoop File System

Enable the MapReduce servers window

Eclipse
Window > Show View > Other... > MapReduce Tools > MapReduce Servers

At the bottom of your window, you should have a "**MapReduce Servers**" tab. If not, see second bullet above. Switch to that tab.

At the top right edge of the tab, you should see a little blue elephant icons.

Eclipse
Click blue elephant to add a new MapReduce server location. • **Server name : any_you_want** • **Hostname : localhost** • **Installation directory: /home/waue/workspace/nutch/**

- **Username : waue**

If any password prompt, please input the **password** which you login to local

It should show up under a little elephant icon in the Project Explorer (on the left side of Eclipse).

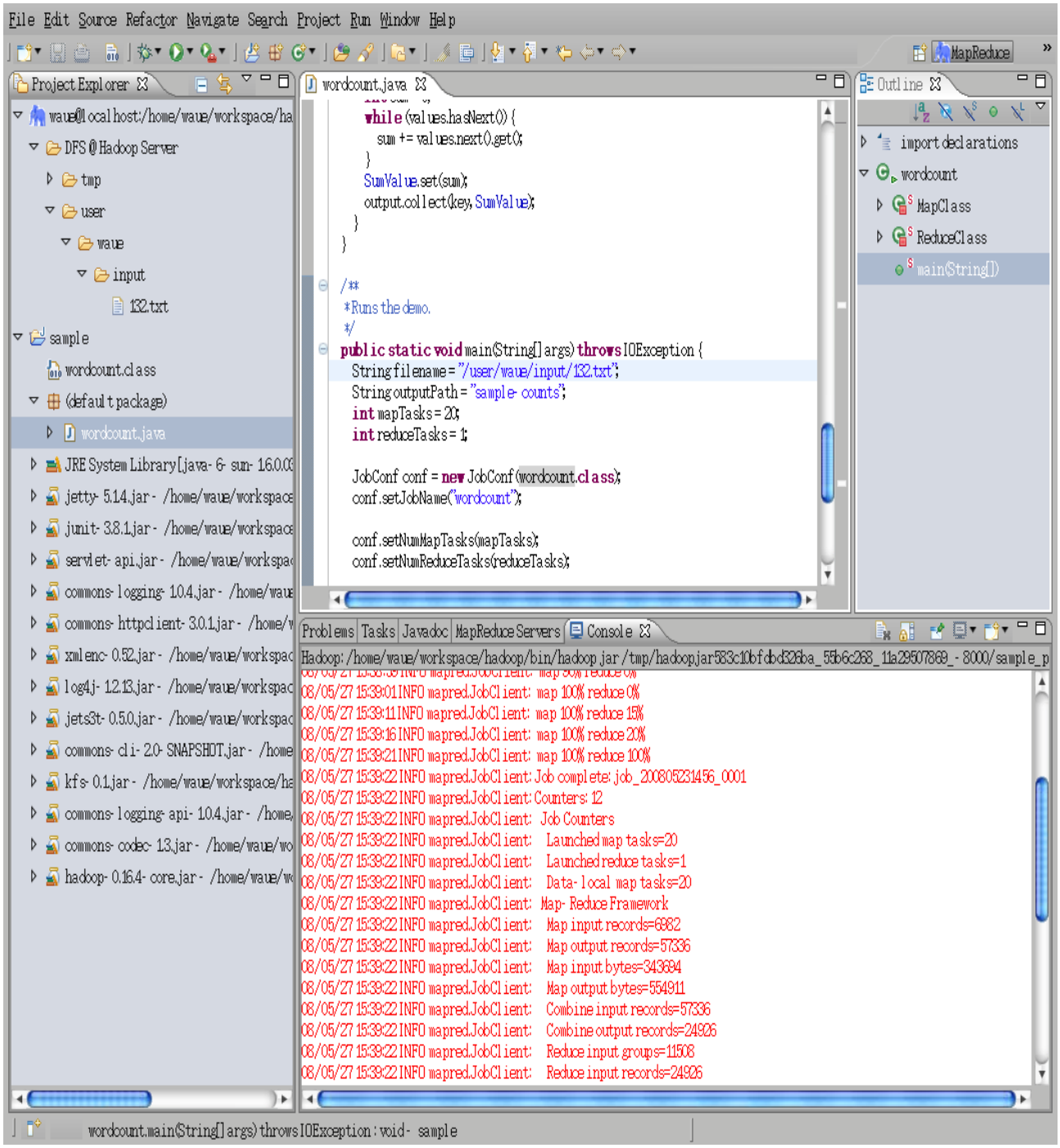
ps : Pleast make sure your Hadoop is working on local system. If not, please refer session 2 Hadoop Setup for debuging, or you can not pass through.

```
$ cd /home/waue/workspace/hadoop/
$ wget http://www.gutenberg.org/etext/132/132.txt
$ bin/hadoop dfs -mkdir input
$ bin/hadoop dfs -ls
Found 1 items
/user/waue/input    <dir>    2008-05-23 15:15    rwxr-xr-x    waue    supergroup
$ bin/hadoop dfs -put 132.txt input
```

4.3 Run

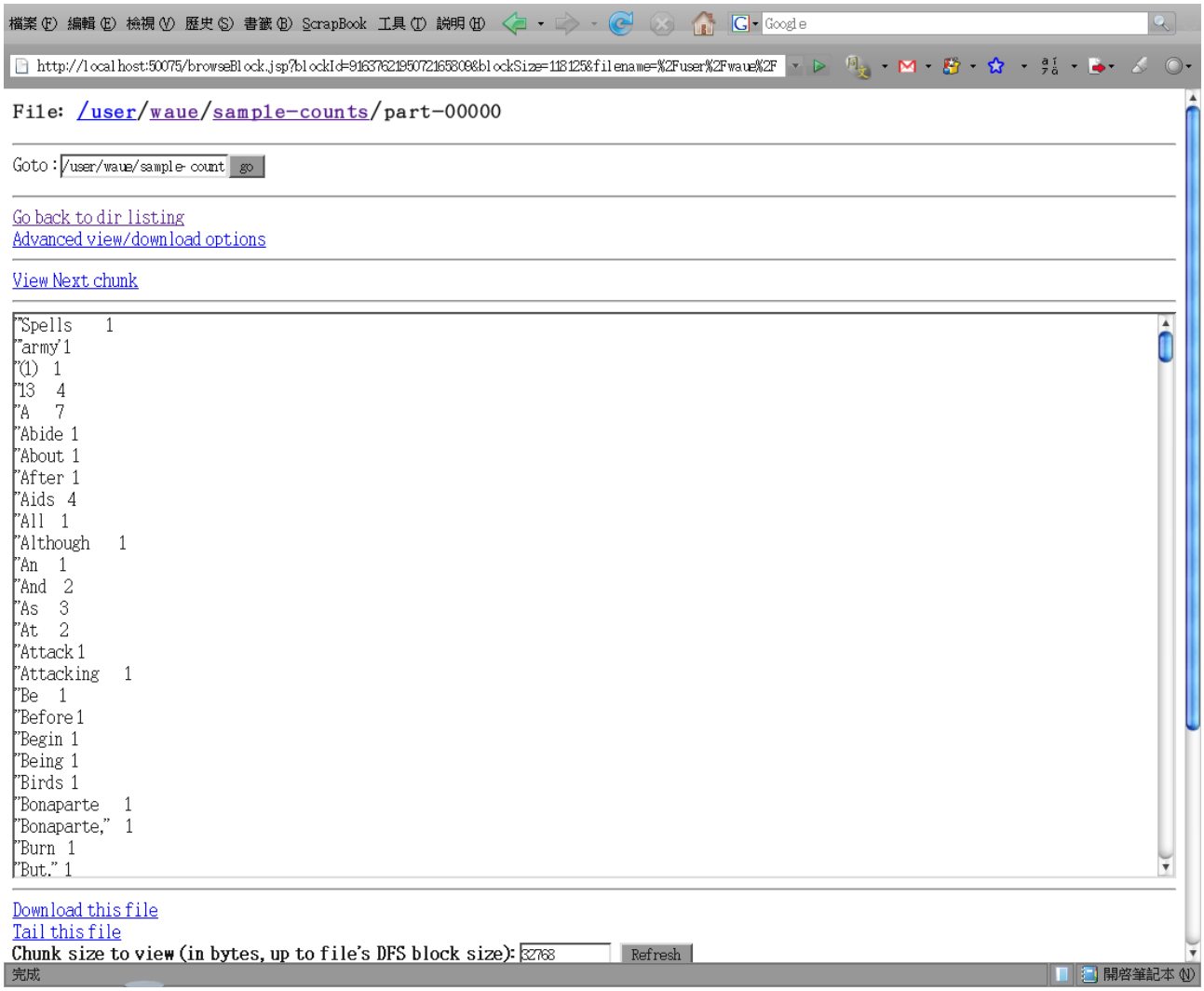
Eclipse
sample > right click **WordCount.java > run as ... > run on Hadoop > choose an existing server from the list below > finish**

A console tag will show beside MapReduce Server tag.



While Map Reduce is running, you can visit http://localhost:50030/ to view that Hadoop is dispatching jobs by Map Reduce.

After finish, you can go to http://localhost:50060/ to see the result.


檔案(F) 編輯(E) 檢視(V) 歷史(S) 書籤(B) ScrapBook 工具(T) 說明(H)
Google
http://localhost:50075/browseBlock.jsp?blockId=9163762195072165808&blockSize=118125&filename=%2Fuser%2Fwaue%2F
File: /user/waue/sample-counts/part-00000
Goto : /user/waue/sample-count go
Go back to dir listing
Advanced view/download options
View Next chunk
"Spells 1
"army 1
"(1) 1
"13 4
"A 7
"Abide 1
"About 1
"After 1
"Aids 4
"All 1
"Although 1
"An 1
"And 2
"As 3
"At 2
"Attack 1
"Attacking 1
"Be 1
"Before 1
"Begin 1
"Being 1
"Birds 1
"Bonaparte 1
"Bonaparte," 1
"Burn 1
"But." 1
Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768 Refresh
完成
開啓筆記本(N)

5. Reference

- NCHC Cloud Technique Develop Group http://trac.nchc.org.tw/cloud/
- IBM Map-Reduce http://www.alphaworks.ibm.com/tech/mapreducetools
- Cloud9 http://www.umiacs.umd.edu/~jimmylin/cloud9/umd-hadoop-dist/cloud9-docs/howto/start.html
- Runing Hadoop http://www.michael-noll.com/wiki/Running_Hadoop_On_Ubuntu_Linux_%28Single-Node_Cluster%29

- **Related Files :**
  - Hadoop
    http://apache.ntu.edu.tw/hadoop/core/
  - IBM map reduce tool :
    http://www.alphaworks.ibm.com/tech/mapreducetools
  - word sample 1 : The Art of War by 6th cent. B.C. Sunzi
    http://www.gutenberg.org/etext/132
  - word sample 2 : The Adventures of Sherlock Holmes by Sir Arthur Conan Doyle http://www.gutenberg.org/etext/1661