Hadoop – 0.20.2 Installation

1. Copy from ~ css534 and installing the Hadoop executable:

1. cp ~css534/hadoop.0.20.2.tar.gz ~’
2. gzip –d hadoop\_0.20.2.tar.gz
3. tar –xvf - < hadoop\_0.20.2.tar

2. Download the Hadoop 0.20.2 user manual:

Let’s assume that you have installed your hadoop at /home/mickey/. Then,

1. Download files from cssmpi1h.uwb.edu:/home/mickey/hadoop-0.20.2/docs to your lapotop/home computer, using WinSCP, Filezilla, or scp.

Scp example:

**scp -r mfukuda@cssmpi1h.uwb.edu:/home/NETID/css534/hadoop-0.20.2/docs .**

1. Open your local docs/ folder and click index.html, or specify this index.html file in your browser’s URL:

Browser example:

**file:///Users/munehiro/Desktop/docs/index.html**

3. Update your .bash\_profile.

**export HADOOP\_HOME=$HOME/hadoop-0.20.2**

**export HADOOP\_VERSION=0.20.2**

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

4. Make directories.

1. Under ~/hadoop-0.20.2/

mkdir input

mkdir dfs

mkdir dfs/namedir

mkdir dfs/namesecondary

mkdir mapred

mkdir mapred/system

5. Sett up configuration files under ~/hadoop\_0.20.2/conf/.

1. hadoop-env.sh

**export JAVA\_HOME=/usr/**

Or, you can simply copy ~css534/hadoop\_0.20.2/conf/hadoop-env.h to your directory:

**cp /home/NETID/css534/hadoop\_0.20.2/conf/hadoop-env.sh ~/hadoop\_0.20.2/conf/hadoop-env.sh**

1. core-site.xml

**<configuration>**

**<property>**

**<name>fs.default.name</name>**

**<value>hdfs://cssmpi1h.uwb.edu:xxxx0</value>**

**<final>true</final>**

**</property>**

**</configuration>**

Please change xxxx0 into the first port # of your assigned port range. The professor’s port range is 28000 – 28010, so that it should be 28000.

Or, you can simply copy ~css534/hadoop-0.20.2’s file to your directory:

cp /home/NETID/css534/hadoop\_0.20.2/conf/core-site.xml ~/hadoop\_0.20.2/conf/core-site.xml

However, please don’t forget to change the port number 28000!

1. hdfs-site.xml

**<configuration>**

**<property>**

**<name>hadoop.tmp.dir</name>**

**<value>/tmp/hadoop-YOURNETID</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>dfs.name.dir</name>**

**<value>/home/NETID/css534/hadoop-0.20.2/dfs/name</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>dfs.data.dir</name>**

**<value>${hadoop.tmp.dir}/dfs/data</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>fs.checkpoint.dir</name>**

**<value>/home/NETID/css534/hadoop-0.20.2/dfs/namesecondary</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>dfs.secondary.http.address</name>**

**<value>cssmpi1h:xxxx1</value>**

**</property>**

**<property>**

**<name>dfs.datanode.address</name>**

**<value>0.0.0.0:xxxx2</value>**

**</property>**

**<property>**

**<name>dfs.datanode.http.address</name>**

**<value>0.0.0.0:xxxx3</value>**

**</property>**

**<property>**

**<name>dfs.datanode.ipc.address</name>**

**<value>0.0.0.0:xxxx4</value>**

**</property>**

**<property>**

**<name>dfs.http.address</name>**

**<value>0.0.0.0:xxxx5</value>**

**</property>**

**</configuration>**

Please note that your hadoop temporary directory will be created under /tmp as hadoop-YourUnetID. For instance, if your UnetID is mickey, it should be /tmp/hadoop-mickey\_css534.

Please change all port numbers:

xxxx1: your first assigned port + 1

xxxx2: your first assigned port + 2

xxxx3: your first assigned port + 3

xxxx4: your first assigned port + 4

xxxx5: your first assigned port + 5

Or, you can simply copy /home/NETID/css534’s file to your directory:

cp /home/NETID/css534/hadoop\_0.20.2/conf/hdfs-site.xml ~/hadoop\_0.20.2/conf/hdfs-site.xml

However, please don’t forget to change /tmp/hadoop-css534, /home/NETID/css534/hadoop-0.20.2 as well as all port numbers such as 28001, 28002, 28003, 28004, and 28005!

1. mapred-site.xml

**<configuration>**

**<property>**

**<name>mapred.job.tracker</name>**

**<value>cssmpi1h:xxxx6</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>mapred.job.tracker.http.address</name>**

**<value>cssmpi1h:xxxx7</value>**

**</property>**

**<property>**

**<name>mapred.task.tracker.http.address</name>**

**<value>0.0.0.0:xxxx8</value>**

**</property>**

**<property>**

**<name>mapred.local.dir</name>**

**<value>/tmp/hadoop-css534/mapred/local</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>mapred.system.dir</name>**

**<value>/home/NETID/css534/hadoop-0.20.2/mapred/system</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>mapred.tasktracker.map.tasks.maximum</name>**

**<value>4</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>mapred.tasktracker.reduce.tasks.maximum</name>**

**<value>4</value>**

**<final>true</final>**

**</property>**

**<property>**

**<name>mapred.child.java.opts</name>**

**<value>-Xmx1g</value>**

**<final>true</final>**

**</property>**

**</configuration>**

Please set your mapred local directory to /tmp/hadopp-YourUnetID/mapred/local

For instance, if your UnetID is mickey, it should be /tmp/hadoop-mickey\_css534/mapred/local.

Please change the job tracker port:

xxxx6: your first assigned port + 6

Or, you can simply copy /home/NETID/css534’s file to your directory:

cp /home/NETID/css534/hadoop\_0.20.2/conf/mapred-site.xml ~/hadoop\_0.20.2/conf/map-site.xml

However, please don’t forget to change /tmp/hadoop-css534, /home/NETID/css534/hadoop-0.20.2 as well as the port number 28006, 28007, and 28008!

1. slaves

cssmpi1h

cssmpi2h

cssmpi3h

cssmpi4h

Choose 4 computing nodes. The first node should be the same as the node that you specified in core-site.xml.

Or, you can simply copy /home/NETID/css534’s file to your directory:

cp /home/NETID/css534/hadoop\_0.20.0/conf/slave ~/hadoop\_0.20.2/conf/slave

6. Format a new file system (just one time).

Make sure that you logged in cssmpi1h.uwb.edu or the master node you decided.

cd ~/hadoop-0.20.2

./bin/hadoop namenode –format

7. Start HDFS.

./bin/start-all.sh

8. Create your hadoop account by yourself.

In the following example, you need to replace css534 with YourUNetID.

./bin/hadoop fs –mkdir /user

./bin/hadoop fs –mkdir /user/css534

./bin/hadoop fs –chown css534:css534 /user/css534

./bin/hadoop dfsadmin –setSpaceQuota 10g /user/css534

./bin/hadoop fs –ls /user/

9. Compile a MapReduce application, (i.e., WordCount.java)

Copy ~css534/programming/MapReduce/worcount\_2.0 to your directory.

cp –r /home/NETID/css534/programming/MapReduce/wordcount\_2.0 .

cd wordcount\_2.0

javac -classpath ${HADOOP\_HOME}/hadoop-${HADOOP\_VERSION}-core.jar WordCount.java

jar –cvf wordcount.jar \*.class

10. Set up input/ and output/ directories under your account, (i.e., /user/YourUNetID/input and /user/YourUNetID/input) and copy your Unix wordcount\_2.0/input/’s files to Hadoop’s /user/YouUNeetID/input.

~/hadoop-0.20.2/bin/hadoop fs –mkdir /user/css534/input

~/hadoop-0.20.2/bin/hadoop fs –put input/\* /user/css534/input

~/hadoop-0.20.2/bin/hadoop fs -ls /user/css534/input

11. Run WordCount and check the results

~/hadoop-0.20.2/bin/hadoop jar wordcount.jar WordCount input output

~/hadoop-0.20.2/bin/hadoop fs –ls /user/css534/output

~/hadoop-0.20.2/bin/hadoop fs –get /user/css534/output/part-00000 output

cd output/; ls

cat part-0000

12. Delete /user/css534/output for a next MapReduce run

~/hadoop-0.20.2/bin/hadoop fs –rmr /user/css534/output

~/hadoop-0.20.2/bin/hadoop fs –ls /user/css534

13. Stop HDFS

~/hadoop-0.20.2/bin/stop-all.sh

NOTE 1:

If your hadoop could not configure correctly, please read the following log files under hadoop-0.20.2/logs/ to make sure that no port collisions happened. Note that these files assume: you are using cssmpi1 as the master (name) node and cssmpi1-4 as data nodes.

|  |  |
| --- | --- |
| Files | To check ports in use |
| hadoop-YOURACCOUNT-namenode-cssmpi1.log | 28xx0, 28xx5 |
| hadoop-YOURACCOUNT-secondarynamenode-cssmpi1.log | 28xx1 |
| hadoop-0.20.2/logs/hadoop-YOURACCOUNT-datanode-cssmpi4.log | 28xx2, 28xx3, 28xx4 |
| hadoop-0.20.2/logs/hadoop-YOURACCOUNT-jobtracker-cssmpi1.log | 28xx6, 28007 |
| hadoop-0.20.2/logs/hadoop-YOURACCOUNT-tasktracker-cssmpi4.log | 28xx8 |

If all configurations sound fine but you still have problems, please refer to

**https://hadoop.apache.org/docs/r2.5.2/hadoop-mapreduce-client/hadoop-mapreduce-client-core/mapred-default.xml**

NOTE 2:

If your hadoop could not configure correctly due to port conflicts or errors in your configuration files or if you want to change the number of computing nodes,

1. Make sure you stop hadoop: ~/hadoop-0.20.2/bin/stop-all.sh
2. Delete all your files under /tmp at each machine, (i.e., cssmpi1 – cssmpi4 in this example).

For the css534 example, delete all the following files:

**hadoop-css534**

**hadoop-css534-datanode.pid**

**hadoop-css534-jobtracker.pid**

**hadoop-css534-namenode.pid**

**hadoop-css534-secondarynamenode.pid**

**hadoop-css534-tasktracker.pid**

**hsperfdata\_css534**

**[11:32:18] css534@cssmpi1h: ~ $ cd /tmp**

**[11:33:05] css534@cssmpi1h: /tmp $ rm -rf hadoop-css534**

**[11:35:11] css534@cssmpi1h: /tmp $ rm -rf hsperfdata\_css534**

**[11:35:19] css534@cssmpi1h: /tmp $ ssh cssmpi2h**

**Last login: Sun Oct 20 20:23:18 2019 from cssmpi1h.uwb.edu**

**[11:35:30] css534@cssmpi2h: ~ $ cd /tmp**

**[11:35:33] css534@cssmpi2h: /tmp $ rm -rf hadoop-css534**

**[11:35:39] css534@cssmpi2h: /tmp $ rm -rf hsperfdata\_css534**

**[11:35:44] css534@cssmpi2h: /tmp $ ssh cssmpi3h**

**Last login: Sun Oct 20 20:23:19 2019 from cssmpi1h.uwb.edu**

**[11:35:49] css534@cssmpi3h: ~ $ cd /tmp**

**[11:35:51] css534@cssmpi3h: /tmp $ rm -rf hadoop-css534**

**[11:36:00] css534@cssmpi3h: /tmp $ rm -rf hsperfdata\_css534**

**[11:36:06] css534@cssmpi3h: /tmp $ ssh cssmpi4h**

**Last login: Sun Oct 20 20:23:25 2019 from cssmpi1h.uwb.edu**

**[11:36:12] css534@cssmpi4h: ~ $ cd /tmp**

**[11:36:15] css534@cssmpi4h: /tmp $ rm -rf hadoop-css534**

**[11:36:20] css534@cssmpi4h: /tmp $ rm -rf hsperfdata\_css534**

**[11:36:25] css534@cssmpi4h: /tmp $**

1. Make sure no zombies are hogging ports:

The professor's ports are 28000 - 28009.

**[13:20:05] css534@cssmpi1h: /tmp $ netstat | grep 280**

**[13:20:15] css534@cssmpi1h: /tmp $ ssh cssmpi2h**

**Last login: Mon Nov 4 13:20:32 2019 from cssmpi4h.uwb.edu**

**[13:21:14] css534@cssmpi2h: ~ $ netstat | grep 280**

**[13:21:20] css534@cssmpi2h: ~ $ ssh cssmpi3h**

**Last login: Mon Nov 4 13:20:35 2019 from cssmpi4h.uwb.edu**

**[13:21:24] css534@cssmpi3h: ~ $ netstat | grep 280**

**unix 3 [ ] STREAM CONNECTED 28058**

**[13:21:31] css534@cssmpi3h: ~ $ ssh cssmpi4h**

**Last login: Mon Nov 4 13:20:42 2019 from cssmpi4h.uwb.edu**

**[13:21:36] css534@cssmpi4h: ~ $ netstat | grep 280**

**unix 2 [ ] DGRAM 28098**

**unix 3 [ ] STREAM CONNECTED 28097**

1. Make sure no zombies are running:

**[13:23:06] css534@cssmpi1h: /tmp $ ps -u css534**

**PID TTY TIME CMD**

**9530 ? 00:00:00 sshd**

**9531 pts/9 00:00:00 bash**

**10455 pts/9 00:00:00 ps**

**[13:23:13] css534@cssmpi1h: /tmp $ ssh cssmpi2h**

**Last login: Mon Nov 4 13:21:14 2019 from cssmpi4h.uwb.edu**

**[13:23:17] css534@cssmpi2h: ~ $ ps -u css534**

**PID TTY TIME CMD**

**10601 ? 00:00:00 sshd**

**10602 pts/8 00:00:00 bash**

**10661 pts/8 00:00:00 ps**

**[13:23:21] css534@cssmpi2h: ~ $ ssh cssmpi3h**

**Last login: Mon Nov 4 13:21:24 2019 from cssmpi2h.uwb.edu**

**[13:23:24] css534@cssmpi3h: ~ $ ps -u css534**

**PID TTY TIME CMD**

**29852 ? 00:00:00 sshd**

**29853 pts/11 00:00:00 bash**

**29912 pts/11 00:00:00 ps**

**[13:23:29] css534@cssmpi3h: ~ $ ssh cssmpi4h**

**Last login: Mon Nov 4 13:21:35 2019 from cssmpi3h.uwb.edu**

**[13:23:35] css534@cssmpi4h: ~ $ ps -u css534**

**PID TTY TIME CMD**

**29684 ? 00:00:00 sshd**

**29685 pts/6 00:00:00 bash**

**29790 pts/6 00:00:00 ps**

1. Thereafter, restart your installation with a hadoop reformatting.

cd ~/hadoop-0.20.2/

./bin/hadoop namenode –format

Note 3:

To check the status of Hadoop and MapReduce tasks, you can access their http servers:

|  |  |
| --- | --- |
| Master node (assuming you are using cssmpi1h)  HDFS name node  MapReduce job tracker | http://cssmpi1h.uwb.edu:xxxx5  http://cssmpi1h.uwb.edu:xxxx7 |
| Slave nodes (assuming you are using cssmpi1h-4h)  MapReduce task tracker | http://cssmpiNh.uwb.edu:xxxx8 |