CS 6463 Cloud and Big Data

Assignment 2: Setting up a Hadoop Cluster and Running Benchmarks

Due Midnight Monday, November 3, 2014

1. Download and execute the *hadoop-install.sh* file on each of your VMs.

   ```
   wget http://cs.utsa.edu/~plama/CS6463/hadoop-install.sh

   chmod +x hadoop-install.sh

   ./hadoop-install.sh groupX
   ```

2. Password less Authentication: Select one of your VM as the master node, and make sure that the master node is able make ssh connection with all the other worker nodes without providing any password. Also, make sure that each worker node is able to make ssh connection with the master node without providing any password.

3. Setup the Hadoop cluster by following the instructions given at

   http://cs.utsa.edu/~plama/CS6463/hadoop-configuration.txt

4. Running Wordcount benchmark from the MASTER node:
   (a) Download plain text file from Wikimedia dump and extract it as follows:

   ```
   wget -P /tmp/wc-input/ http://dumps.wikimedia.org/enwiki/20140102/enwiki-20140102-pages-articles-multistream-index.txt.bz2

   bunzip2 http://dumps.wikimedia.org/enwiki/20140102/enwiki-20140102-pages-articles-multistream-index.txt.bz2
   ```

   (b) Copy the files from the local filesystem to HDFS as follows:

   ```
   cd $HADOOP_PREFIX

   bin/hadoop fs -copyFromLocal /tmp/wc-input /user/groupX/wc-input
   ```

   (c) List the files from the HDFS

   ```
   bin/hadoop fs -ls /user/groupX/wc-input
   ```

   (d) Run the *wordcount* benchmark with different number of *reduce* tasks, and observe how the job execution time is affected.

   For example:

```
bin/hadoop jar hadoop*examples*.jar wordcount /user/groupX/wc-input
/user/groupX/wc-output-r1

bin/hadoop jar hadoop*examples*.jar wordcount –D mapred.reduce.tasks=3
/user/groupX/wc-input /user/groupX/wc-output-r3
```

(e) Find one more configuration parameter that can improve the job execution time, and collect the job execution time results by changing the value of that parameter.

Important Note:

Job execution results can be obtained by using the following command

```
bin/hadoop job –history wc-output-r1
```

Here, the job output directory is used to identify the job, whose execution history is being fetched. Alternatively, job history can also be obtained from the jobtracker web interface:

http://10.242.144.xx:50030/jobhistoryhome.jsp

Troubleshooting Tips:

(a) If a job is long running, you can let it run in the background, and free the shell to do other stuff by using:

```
Ctrl-C
```

(b) Job progress can be monitored from the Jobtracker Web Interface. You can find out which task is taking too long or which failed.

(c) If a job hangs (making no progress), you can kill the job as follows:

```
bin/hadoop job –list

bin/hadoop job –kill job_2014----
```

(d) To troubleshoot the task that took too long, check the corresponding log file under the directory,

```
/usr/local/hadoop-1.2.1/logs/userlogs
```

Or,

Check the Log files from the jobtracker web interface.

```
http://10.242.144.xx:50030/logs/userlogs
```

Submission Policy and Deliverables

Please submit a PDF report named as “Assign2GroupX.pdf” to the blackboard by the due date. Only one submission per group is required. The report **must** include the following:

1. A graph showing the impact of increasing the **number of reduce tasks** on the execution time of wordcount benchmark.

2. A graph showing the impact of changing another configuration parameter on the execution time of wordcount benchmark.

3. One representative Screenshot of the console output when you execute the benchmark.

4. One representative Screenshot of the Jobtracker’s web interface which shows the status of running/completed jobs. (http://10.242.144.xx:50030)

5. Describe how the work was divided among your group members.