

Using Hadoop on the neighbor cluster

Chris Hlady

February 26, 2009

1 Introduction

- ssh studentXX@neighbor.cs.uiowa.edu to login to the neighbor cluster.
- haddop is located in /usr/local/hadoop – You may want to add the bin folder to your PATH. To do that:
Add export PATH=\$PATH:/usr/local/hadoop/bin to the botoom of /home/studentXX/.bashrc
- Add quotes to your arguments to hadoop dfs if they contain *
hadoop dfs -rmr "*.out"
- You might want to do a hadoop dfs -chmod 600 on your inputs

2 More information

Here are some websites for finding out more information about hadoop 0.18.3

- The API: <http://hadoop.apache.org/core/docs/r0.18.3/api/index.html>
- A tutorial: http://hadoop.apache.org/core/docs/current/mapred_tutorial.html
- Streaming (setting up hadoop jobs in arbitrary languages): http://www.michael-noll.com/wiki/Writing_An_Hadoop_Ma

3 Doing a wordcount on Shakespeare

1. Fortunately Hadoop comes with some example MapReduce programs, one of which implements word-count
2. Visit the Gutenberg Project <http://www.gutenberg.org/etext/100> to read the description of the Complete Works of William Shakespeare
3. Download the zip file
 - cd /home/studentXX
 - wget <http://www.gutenberg.org/dirs/etext94/shaks12.zip>
 - unzip shaks12.txt
 - rm shaks12.zip
4. Copy the file to the HDFS
 - hadoop dfs -copyFromLocal /home/studentXX/shaks12.txt shaks12.txt

- `hadoop dfs -ls`
 - `hadoop dfs -chmod 600 shaks12.txt`
5. Start the wordcount example
 - Check <http://neighbor.cs.uiowa.edu:50030/jobtracker.jsp> and make sure there are no other jobs running
 - `hadoop jar /usr/local/hadoop/hadoop-0.18.3-examples.jar wordcount shaks12.txt wordcount.out`
 6. Follow the progress on the jobtracker. If the job is going to take a long time, you can log out while its running. (Ctrl-C; exit;)
 7. Once the job is done, it's time to read the output off of the HDFS to your home directory
 - `hadoop dfs -copyToLocal wordcount.out /home/studentXX/wordcount.out`
 - `cd /home/studentXX/wordcount.out`
 - `cat part-* > combined.txt`
 - Move the results to your home machine as soon as possible. There is no backup strategy for the cluster.
 - When you're done, clean up after yourself on the HDFS
 - `hadoop dfs -rm shaks12.txt`
 - `hadoop dfs -rmr wordcount.out`

4 Getting distribution of UIHC EMR logins by hour

4.1 Writing the application

4.1.1 mapper.py

On neighbor: `/home/chlady/compepi-hadoop/mapper.py`

4.1.2 reducer.py

http://www.michael-noll.com/wiki/Writing_An_Hadoop_MapReduce_Program_In_Python

4.2 Writing the input data to HDFS

`hadoop dfs -copyFromLocal /home/chlady/logins.csv logins.csv`

4.3 Running the job

- See what options `streaming.jar` takes:
- `hadoop jar /usr/local/hadoop/contrib/streaming/hadoop-0.18.3-streaming.jar`
- I've written a custom mapper and reducer, and I want to set `-input`, `-output` obviously. Also, instead of the default of 40 reducers, I'd like to just run one so I don't have to sort (this comes with a performance penalty)
- `hadoop jar /usr/local/hadoop/contrib/streaming/hadoop-0.18.3-streaming.jar -mapper /home/chlady/compepi-hadoop/mapper.py -reducer /home/chlady/compepi-hadoop/reducer.py -input logins.csv -output loginsByHourOfWeek -numReduce 1`

4.4 Getting the data off the HDFS

Nothing new here:

```
hadoop dfs -copyToLocal loginsByHourOfWeek /home/chlady/loginsByHourOfWeek
```

4.5 Combining the results

This isn't necessary since there was only one reducer.

5 Limitations

- There is 36G disk space on each node, so theoretically you could have up to 100GB for your input and output files.
- Everybody gets their own HDFS directory. Please remove data you put on the HDFS as soon as you are done. You can immediately move the output to your home directory, and then from there scp it to your home machine.
- Your mappers and reducers can probably use about 200-400MB RAM above what the basic examples require
- Running multiple jobs at the same time is possible, but not advisable.