Lab 7
Big Data Spring 2016

In this lab we will understand and run Pig scripts on an EMR cluster. The Pig scripts process a search query log file from the Excite search engine to find search phrases that occur with particular high frequency during certain times of the day. These scripts come from the Apache Pig tutorial[1].

Part I: Setup

1. Start an EMR cluster using your key pair. While your cluster is starting, we will go over the Pig scripts and look at what they do.
2. Once your master node is ready, SSH into the master node.
3. Download the file pigtutorial.tar.gz from NYU Classes and SCP this file to the master node.
4. Unpack the tutorial files by typing `tar -xvf pigtutorial.tar.gz`

Part II: Pig in Local Mode

1. Move to the directory with the tutorial files by typing `cd pigtmp`
2. To run the Pig script locally, type `sudo pig -x local script1-local.pig`

Part III: Pig in Mapreduce Mode

1. Move the data files to HDFS by typing `hadoop fs -copyFromLocal excite.log.bz2`
2. Run the Pig script in Mapreduce mode by typing `pig script1-hadoop.pig`
3. Get the output files from HDFS by typing `hadoop fs -get script1-hadoop-results`

Part IV: Deliverable

You must submit answers to the following questions on NYU Classes (either type into the text field or attach a text file). The answers to these questions can be found by looking through the job monitoring statistics output by running the Mapreduce job. **Due date: Monday, March 28, 12:00pm (noon).**

1. How long did the Pig script take to run?
2. How many Mapreduce jobs were generated by the Pig script? Which Mapreduce job was the slowest?
3. For each Mapreduce job that was generated, list the relations (aliases) that were mapped to it and the Pig operations (features) that were used.

[1] http://pig.apache.org/docs/r0.7.0/tutorial.html