Assignment 1: Introduction to Hadoop
ITCS 6190/8190: Cloud Computing for Data Analysis

The goal of this assignment is for you to have:
1. A working Hadoop installation.
2. Your Amazon Web Services (AWS) account setup and running.

Please submit your exercise output on Canvas before midnight on January 22, 2017.

WordCount on Hadoop VM:

1. Install Hadoop in a virtual machine using the Hadoop VM installation instructions handout, and verify the WordCount program is working fine.

   Look at the source code and description of the WordCount program to understand how it works:
   https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount1_source.html

   You may also look at how to build and run the WordCount program:
   https://www.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_usage.html

2. Download the Canterbury Corpus (http://corpus.canterbury.ac.nz/descriptions/) from http://corpus.canterbury.ac.nz/resources/cantrbry.zip .
3. Run WordCount on the alice29.txt input file and then on the asyoulik.txt input file.
4. Save the last 20 lines of output from WordCount for each of the two input files. The output files (each consisting of the last 20 lines of output) corresponding to the input files alice29.txt and asyoulik.txt should be named alice20-hadoop.txt and asyoulik20-hadoop.txt respectively.

   Additional information on shell commands for the filesystem is at:
   http://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html

   For example, if you want to save the output file in HDFS to the local filesystem, you can use the command:
   hadoop fs –get /user/cloudera/wordcount/output/part-00000 MY_LOCAL_DIR

   Here MY_LOCAL_DIR is the path to the directory where you want to save the file.

5. If you are unfamiliar with UNIX, it is **strongly recommended** that you work through the UNIX tutorial at http://www.ee.surrey.ac.uk/Teaching/Unix/

Setup and Test AWS Account:

1. Follow the instructions in AWSLab-Hadoop to create an AWS account. The instructions will also guide you to creating an EMR cluster and connecting to the master node.
2. After you connect to the master node, take a screenshot of your shell screen and save it to your local disk.

Submissions

Upload to Canvas a zip file named your_uncc_id.zip (e.g. wshalaby.zip if your email address is wshalaby@uncc.edu). The file should contain:

1. The two output files from WordCount on Hadoop VM (alice20-hadoop.txt and asyoulik20-hadoop.txt)
2. A screenshot of your shell screen after you perform step 5 of the AWSLab-Hadoop "Connect to the Master Node". Name your screenshot file aws-shell.jpg

Grading Rubric

Total 50 points.

1. 10 points for submitting your_uncc_id.zip with required files and proper naming.
2. 15+15 points for correct output of alice20-hadoop.txt and asyoulik20-hadoop.txt.
3. 10 points for aws-shell.jpg
4. Incorrect output (e.g., did not Consider running WordCount on each file individually) = -5 Points