CSE 482 Exercise 9 (Date: March 22, 2019)

The purpose of this exercise is to help you get started compiling and running a Hadoop program on Amazon Web Services.

1. Download and read the supplementary Powerpoint slides entitled “Instructions on Accessing AWS” (lecture17b.pptx) from the class Web page.

   a. Create an account on AWS and sign up for the AWS Educate grant.

   b. Launch an AWS EMR cluster and follow the steps given in lecture17b.pptx.

2. Use SSH to connect to the master node of the EMR cluster

   a. Once you’ve connected to the master node, download the data and Hadoop source code from http://www.cse.msu.edu/~ptan/CSE482/exercises/ex9/ex9.tar. For example, you can use wget to do this:

   ```
   hadoop> wget  http://www.cse.msu.edu/~ptan/CSE482/exercises/ex9/ex9.tar
   ```

   b. Unarchive the tar file to obtain the following three files: WordCount.java, document.txt, and env.sh.

   c. Run env.sh to set the environment variables for JAVA_HOME and HADOOP_CLASSPATH.

   d. Compile the Java code WordCount.java.

   e. Create a Java archive (jar) file named wc.jar that contains all the *.class files.

   f. Upload the data file document.txt to HDFS.

   g. Run the Hadoop program WordCount from the wc.jar file by typing the following:

   ```
   hadoop jar  wc.jar  WordCount document.txt  output
   ```

   h. After the program has been successfully executed, download the result file from the output directory on HDFS to the local directory by typing the following command:

   ```
   hadoop  fs  –getmerge output ./result.txt
   ```

   i. Run the sftp program on the AWS host machine to transfer the results.txt file to your CSE account:

   ```
   sftp <yourMSUID>@arctic.cse.msu.edu
   sftp> put  result.txt
   sftp> quit
   ```

   j. Terminate your AWS EMR cluster (VERY IMPORTANT) to avoid incurring further charges.

Deliverables: Submit (via D2L) the result.txt file