CSE 482 Exercise 10 (Date: March 29, 2019)

The purpose of this exercise is to help you familiarize with compiling and executing a Hadoop program. Read carefully lectures 18 and 19 before doing the exercise.

1. Launch an AWS EMR cluster. Use SSH to connect to the master node of the EMR cluster
2. Once you've connected to the master node, download the data and source code from http://www.cse.msu.edu/~cse482/lecture19.tar using wget (see Exercise 9).
    a. Unarchive the tar file.
    b. Set the environment variables for JAVA_HOME and HADOOP_CLASSPATH appropriately (refer to the env.sh file from exercise 9 on how to set the environment variables). Check to make sure that the paths to JAVA_HOME and HADOOP_CLASSPATH exist on the AWS master node you have connected to via SSH. Otherwise, you need to modify the env.sh file to point to the correct location (or appropriate version of Java).
    c. Upload the sentiment.data file to HDFS.
3. Modify the getSentiment.java program so that the reducer output key-value pair is as follows:

    Key: word,   Value: sentiment (ratioVal)

    For example, instead of the following key-value pair:

    sad   negative,1109,21470

    the reducer output should look like this:

    sad   negative (0.052)
4. Compile and execute the modified getSentiment.java program by following the procedure described in lecture 19. Set the number of reducers to 4, threshold to 800, and ratio to 10. Use hadoop fs -getmerge command to concatenate the reducer outputs into a single file named results.txt. Submit the modified source file (getSentiment.java) and results.txt file to D2L along with a screen shot of the AWS EMR cluster that you've run (see http://www.cse.msu.edu/~ptan/CSE482/class/exercises/ex9/EMR.jpg).
5. Terminate your AWS EMR cluster (VERY IMPORTANT) to avoid incurring further charges.