CSE 482 Exercise 7 ((tentative) due date : Nov 27, 2020)

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

1. Launch an AWS EMR cluster. Try to use a small cluster so that the cost for running your program is not too expensive. Follow the steps given in lecture16.pptx and lecture16b.pptx.
2. Use SSH to connect to the master node of the EMR cluster
    a. Once you've connected to the master node, use wget to download the data and source code from http://cse.msu.edu/~karimiha/cse482/exercises/exercise7/materials.tar
    b. Unarchive the tar file to obtain the following three files: countEdits.java, wiki_edit.txt, and env.sh.
    c. Run env.sh to set the environment variables for JAVA_HOME and HADOOP_CLASSPATH.
    d. Compile the Java code countEdits.java.
    e. Create a Java archive (jar) file named wiki.jar that contains all the *.class files.
    f. Upload the data file wiki_edit.txt to HDFS.
    g. Run the Hadoop program countEdits from the wiki.jar file by typing the following:

    ```
    hadoop jar  wiki.jar  countEdits  wiki_edit.txt  output
    ```

    h. After the program has been successfully executed, download the result file by typing the following command:

    ```
    hadoop  fs  –copyToLocal  output/part-r-00000  ./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
    ```

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

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