CSE 482 Exercise 11 (Date: April 5, 2019)

The goal of this exercise is to help you get familiarize with Pig Latin. Follow the instructions below to complete the exercise. Read carefully the slides from lectures 21 and 22 before attempting this exercise.

1. Use wget to download the source files from http://www.cse.msu.edu/~cse482/exercise11.tar. Unarchive the file.
2. Launch pig in the local mode by typing

   hadoop> pig –x local (this will launch pig in local mode)

   To turn off pig messages, you can also use launch pig as follows:

   hadoop> pig –x local -4 nolog.conf (make sure nolog.conf is in the working directory)

   Now, load the wiki_edit.txt file into an alias called data by typing

   grunt> data = LOAD 'wiki_edit.txt' USING PigStorage(' ')
   AS (revID, article, ts, username);
3. Check to make sure the data is loaded correctly by typing

   grunt> describe data;
4. In this exercise, you need to write a Pig program to count the number of distinct users who edited each article. Output only those articles edited by more than 15 distinct users. Hint:

   grunt> edits = FOREACH data GENERATE article, username
   grunt> uniq_edits = DISTINCT ????
   grunt> articles = GROUP uniq_edits BY ????
   grunt> counts = FOREACH articles GENERATE ???? AS article, COUNT(????) AS numUsers;
   grunt> results = FILTER counts BY (int) numUsers > ????
   grunt> STORE results INTO 'output'
5. Store all the Pig Latin commands above (from loading the data to storing the output) into a script file named exercise11.pig.

Deliverables: Submit the following files: (1) the Pig latin script file named exercise11.pig, (2) the query results file located in output directory, and (3) a snapshot image of the AWS EMR cluster usage (you should indicate which cluster was used to run the job).