CSE 482: Big Data Analysis (Spring 2019) Homework 5

Due date: Monday, April 17, 2019 (before midnight)

Submit a PDF version of your homework via D2L along with a snapshot image of the AWS cluster usage and the source code and results for questions 2 and 3.

1. Consider the following two data files (`student.txt` and `transcript.txt`).

```
hadoop> cat student.txt
john,senior,cse
mary,senior,cse
bob,junior,ece
lee,sophomore,ece

hadoop> cat transcript.txt
john,cse482,3.5
john,cse335,3.5
mary,cse482,3.5
mary,cse335,4.0
bob,cse335,2.5
bob,cse232,3.0
lee,cse335,3.5
```

   State each of the following Pig Latin queries in plain English (i.e., explain what it is trying to do) and show the query result.

   (a)
```
data = LOAD 'transcript.txt' USING PigStorage(',')
            AS (Sname, Course, GPA:float);
tmp = FILTER data BY Course == 'cse482';
result = FOREACH tmp GENERATE Sname, GPA;
DUMP result;
```

   (b)
```
data = LOAD 'transcript.txt' USING PigStorage(',')
            AS (Sname, Course, GPA:float);
grp = GROUP data BY Course;
tmp = FOREACH grp GENERATE group AS Course, AVG(data.$2) AS avgGPA;
result = ORDER tmp BY avgGPA desc;
dump result;
```

   (c)
```
std = LOAD 'student.txt' USING PigStorage(',')
            AS (Name, Status, Dept);
cs = FILTER std BY Dept == 'cse';
data = LOAD 'transcript.txt' USING PigStorage(',')
            AS (Sname, Course, GPA:float);
tmp = JOIN cs by Name, data by Sname;
tmp2 = FOREACH tmp GENERATE Sname, Course;
result = DISTINCT tmp2;
dump result;
```

(d)
```
std = LOAD ‘student.txt’ USING PigStorage(‘,’);
data = LOAD ‘transcript.txt’ USING PigStorage(‘,’);
enrollment = JOIN std by $0, data by $0;
tmp = FOREACH enrollment GENERATE $0, $2, $4, $5;
tmp2 = GROUP tmp BY $2;
result = FOREACH tmp2 GENERATE group, MIN(tmp.$3);
dump result;
```

(e)
```
data = LOAD 'transcript.txt' USING PigStorage(',');
grp1 = GROUP data BY $1;
tmp = FOREACH grp1 GENERATE group, 'all', AVG($1.$2);
grp2 = GROUP data all;
tmp2 = FOREACH grp2 GENERATE $0, AVG($1.$2);
tmp3 = JOIN tmp BY $1, tmp2 BY $0;
tmp4 = FILTER tmp3 BY $2 > $4;
result = FOREACH tmp4 GENERATE $0;
dump result;
```

2. For this question, you need to write the Pig Latin scripts for processing the lastFM streaming data set. First, you should download the lastFM sample dataset from `http://www.cse.msu.edu/~cse482/lastFM.tar`. After unzipping the file, you should obtain 2 files, `lastFM-users.csv` and `lastFM-ratings.csv`. The format in each line of the data files are as follows:

```
lastFM-users.csv:   userID,gender,age,country
lastFM-ratings.csv: userID,artist,number_of_plays
```

For each question below, the Pig Latin code should be written in a script file named q2*.pig. For example, the script for the first question should be named q2a.pig, the second question q2b.pig, and so on. The query results should also be stored in its corresponding directory named q2* using the `store` command. Create a zip/tar file to compress/archive all the script and result files into a single file, e.g., `question2.tar` or `question2.zip` for submission. To simplify the process, you should run the pig program in local mode instead of distributed mode.

(a) Write a Pig Latin script that finds the IDs of all users who have streamed a song performed by “the beatles”. Save the output in a directory named `q2a`.

(b) Write a Pig Latin script that returns the top 10 most popular artists (where popularity is measured in terms of the number of users who have streamed a song performed by the artist). The query result should contain only 2 columns (artist and number of users). Save the output in a directory named `q2b`.

(c) Write a Pig Latin script that returns the top 10 most popular artists (where popularity is measured in terms of the total number of plays a song performed by the artist was streamed). The query result should contain only 2 columns (artist and number of plays). Save the output in a directory named `q2c`.

(d) Write a Pig Latin script that returns the number of male and female listeners of songs performed by an artist named "the beatles". The query result should contain 1 row and 2 columns (number of male listeners, number of female listeners). Save the output in a directory named `q2d`.

3. For this question, you will use the same lastFM dataset from question 2. You should save the source code into a script file named question3*.sql (e.g., question3a.sql, question3b.sql, etc) as well as the output tables `question3b.txt`, `question3c.txt`, and `question3d.txt` (except for question 3a) and submit them to D2L. To create the output tables for question 3(b), (c), and (d) below, you need to explicitly use the `create table` command as shown in the example below (see lecture 24):

```
CREATE TABLE question3b
    ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '\t'
    LINES TERMINATED BY '\n'
    STORED AS TEXTFILE
AS
    -- Enter your query here
```

The preceding example will create a table named question3b for the query results. The table is physically stored on HDFS in a directory named `/user/hive/warehouse/question3b`. You can use `hadoop fs -getmerge /user/hive/warehouse/question3b question3b.txt` to merge the query result files in the HDFS directory into a file on the local filesystem named `question3b.txt`. You can then use `sftp` to transfer the result file back to your own machine or to one of the CSE machines (such as arctic).

(a) Write the corresponding HiveQL queries for creating the following two internal tables and load their corresponding data from HDFS: *Users* from `lastFM-users.csv` and *Streaming* from `lastFM-ratings.csv`. The schema for the tables are as follows:

Users(UserID: string, Gender: char(1), Age: int, Country: string)

Streaming(UserID: string, Artist: string, NumPlays: int)

(b) Write the HiveQL query to find the average number of times songs performed by each artist were streamed by users. The output must have 2 columns: artist and average number of plays.

(c) Write the corresponding HiveQL query to find the names of popular artists, i.e., artists whose songs have been streamed by more than 25,000 users. The output must have 2 columns: artist and number of users who streamed songs performed by the artist.

(d) Write the corresponding HiveQL query to count the number of users from each country who had streamed songs performed by the artist named 'the beatles'. The output must have 2 columns: country and number of users from the country who had streamed a song performed by 'the beatles'.