CSE 482 Exercise 12 (Date: April 12, 2019)

The purpose of this exercise is to help you get started using Hive. Follow the instructions below to complete the exercise.

1. Launch an AWS cluster. You need to wait for at least 5 minutes to ensure hive is fully installed on AWS. Use wget to download the data from http://www.cse.msu.edu/~cse482/exercise12.tar. After unarchiving the tar file, you will see 2 data files: grade.txt and major.txt.
2. Create the directories grade and major on HDFS. Upload the data files grade.txt and major.txt to the following paths on HDFS: /user/hadoop/grade/grade.txt and /user/hadoop/major/major.txt.
3. Create a script file named exercise12.sql to load the raw data into two external tables, named grade and major, respectively. Hint: your script should contain the following code:

```
DROP TABLE IF EXISTS grade;
CREATE EXTERNAL TABLE IF NOT EXISTS grade (
        name  STRING,
        hw1  INT,
        …                       -- fill in the rest of the schema for hw2 and hw3 grades
 ) ROW FORMAT DELIMITED
FIELDS TERMINATED BY ‘,’
STORED AS TEXTFILE
LOCATION ‘/user/hadoop/grade‘;    -- path to the directory that contains grade.txt.

DROP TABLE IF EXISTS major;
CREATE EXTERNAL TABLE IF NOT EXISTS major (
        name  STRING,
        status STRING,
        dept STRING
) ROW FORMAT DELIMITED
FIELDS TERMINATED BY …          -- fill in the blanks
STORED AS TEXTFILE
LOCATION ‘…’;                   -- file in the blank by specifying the directory name
```

4. Add the following statement to the script file to create a table named transcript:

```
CREATE TABLE transcript
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
LINES TERMINATED BY '\n' AS
        SELECT grade.name AS name, status, dept, hw1+hw2+hw3 as hwgrade
        FROM grade, major
        WHERE grade.name = major.name;
```

5. Launch beeline by typing the following:

   ```
   hadoop@ip-xx-xx >  beeline  -u “jdbc:hive2://localhost:10000/default”  -n hadoop
   ```

6. Execute the script file in beeline by typing the following statement on beeline:

   ```
   jdbc:hdbc://localhost:10000/default>  source exercise12.sql;
   ```

   By executing the script, this will create 3 tables, grade, major, and transcript. To check that the tables exist, type the following:

   ```
   jdbc:hdbc://localhost:10000/default>  show tables;
   jdbc:hdbc://localhost:10000/default>  select * from transcript;
   ```

7. Download the transcript table from HDFS as follows:

   ```
   hadoop@ip-xx-xx > hadoop fs –getmerge /user/hive/warehouse/transcript  transcript.txt
   ```

   This will create an output file named transcript.txt on the local filesystem of AWS master node.

Deliverables: Submit the script file exercise12.sql and the output file transcript.txt. You should also submit a proof of your AWS cluster usage.