DSCI 551 – HW5

(Hadoop and Spark)

(Fall 2022)

100 points, Due 11/30, Wednesday

1. [40 points] In this homework, your task is to write a Hadoop MapReduce program named SQL2MR.java that finds the answer to the following SQL query on the cars.csv data set. Recall that cars.csv was provided to you in homework 1 and also as an example for your project. The file has a header followed by rows, one for each car with the information on car's body type, highway MPG, and price.

   ```
   select carbody, max(highwaympg)
   from cars
   where price >= 10000
   group by carbody
   having count(highwaympg) >= 5;
   ```

   You are provided with a template for SQL2MR.java where you can provide the missing code to produce a complete program. The template also has some hints that may help you.

   You are reminded of the following steps to compile and run the program. **The steps assume that you have removed the header from cars.csv file and save it under a directory called cars-input. You do not need to create the cars-output directory ahead of time.**

   - hadoop com.sun.tools.javac.Main SQL2MR.java
   - jar cf sql2mr.jar SQL2MR*.class
   - hadoop jar sql2mr.jar SQL2MR cars-input cars-output

2. [Spark DataFrame, 30 points] Using the same country data set as in lab3 (country.json, city.json, and countrylanguage.json), write a Spark DataFrame script for each of the following questions. You use "import pyspark.sql.functions as fc". You need to display the full content of the columns (e.g., using show(truncate=False)).
   a. Find top-10 most popular official language, ranked by the number of countries where the language is official. Return the language and count in the descending order of the count.
   b. Find names of countries and their capital cities, for all countries in North America and having a GNP of at least 100,000. Output country and capital city names only.
   c. Find names of countries in North America continent where English is an official language.
   d. Find the maximum population over all cities in USA.

e. Find country codes of the countries where both English and French are official languages.

3. [Spark RDD, 30 points] Using the country data set, write a Spark RDD script for each of the following questions. You use “import pyspark.sql.functions as fc”.
   a. Find out how many countries have a GNP between 10,000 and 20,000 inclusive.
   b. For each continent, find the maximum GNP of countries in the continent.
   c. Find the first 20 countries and names of their capital cities, ordered by the names of countries, descending.
   d. Find the maximum population of cities in USA.
   e. Find country codes of the countries where both English and French are official languages.

   Note: you should only use collect function to produce the final result. That is, when you use rdd.collect(), the rdd should contain the final result for the question.

Submission:

Q1: Submit 3 files - SQL2MR.java, sql2mr.jar, and part-r-00000 file under cars-output directory.

Q2+Q3: Submit one txt file in total for Q2 and Q3, which including both code and output for each question. Please run all your code on ec2 for each question to get the output and paste them into the txt file.

Please Note: Please do not zip your files, do not submit other files. You do not need to submit any data file.