Systems for Big Data Processing 2019/20 – 1st semester

Project nº 2

Consider you want to process information about taxi rides in some city to better understand the behavior of communities and help taxi drivers maximize their profit.

To this end, you have information about taxi drivers, including information for pickup and drop-off location.

The data set is available at

https://drive.google.com/file/d/1nNIPI12Vj5arlPhdntR1K7QMFYKi2kwo/view?usp=sharing

and has the following columns:

Col#	Name	Type	Description
0	row#	INT	Row number
1	id	CHAR(9)	Service's ID
2	vendor_id	INT	Taxi company's ID
3	pickup_datetime	CHAR(19)	Date and time of pickup e.g., 2019-01-01 00:46:40
4	dropoff_datetime	CHAR(19)	Date and time of drop off e.g., 2019-01-01 00:53:20
5	passenger_count	INT	# passengers in taxi
6	pickup_longitude	FLOAT	—
7	pickup_latitude	FLOAT	—
8	dropoff_longitude	FLOAT	—
9	dropoff_latitude	FLOAT	—
10	haversine_distance	FLOAT	Distance between pickup and dropoff locations
11	maximum_temperature	INT	—
12	minimum_temperature	INT	—
13	average_temperature	FLOAT	—
14	precipitation	FLOAT	—
15	snow_fall	FLOAT	—
16	snow_depth	FLOAT	—

Because locations are represented as floats, you are advised to divide the world into square areas from latitude/longitude, by using only two decimal digits, i.e.,

-73.9821548462, 40.7679367065 → -73.98, 40.76

You are asked to use Apache Hive to create indexes for answering the following queries:

1. *What are the 10 most frequent routes (pickup-area, dropoff-area), given a weekday (0=sun, 6=sat) and hour (no minutes)?*
2. *What is the expected duration and distance of a taxi ride, given the pickup area, the weekday (0=sun, 6=sat) and time?* Consider time intervals of 15 minutes.
3. *What is the factor of additional clients when it is raining or snowing (merge these conditions into one) versus dry weather, given the pickup-area?*
4. An **optional 4th index** that will answer a non-trivial and interesting question over the given data set. If you add this to your project you will get some extra points.

To solve this project, you should use a new/different container image.

Download the container image:

```
docker pull prasanthj/docker-hive-on-tez
```

Download execute the container image (just once):

```
docker run -v path_to_local_folder:/root/work --name docker-hive-on-tez -it
          -P prasanthj/docker-hive-on-tez /etc/hive-bootstrap.sh -bash
```

please note that you must replace the "`path_to_local_folder`" with the path to the shared folder in your laptop. In Windows Home with Docker Toolbox, be sure to replace all the "\" with "/", and replace the initial "`c:\Users\…`" with "`/c/Users/…`"

Download the container image (stop the container image):

```
docker stop docker-hive-on-tez
```

Download the container image (resume the previously stopped container image):

```
docker restart docker-hive-on-tez; docker exec -it docker-hive-on-tez bash
```

Then you can access the host local folder at "/root/work". Some more info at

https://github.com/prasanthj/docker-hive-on-tez

To run a HQL script inside the container, inside the container execute

```
hive -f script.sql
```

The project team has two members and must deliver:

1. One file with the HQL (hive query language) script to calculate each of the required indexes (3 or 4 files in total) grouped in a ZIP file.
2. A small report (two pages maximum) with a description of your approach/strategy/results/…

The project files (the HQL scripts and the report in PDF) shall be uploaded as a single ZIP file with the name **`Gnn_AAAAA_BBBBB.zip`** where:

`Gnn` -> group number, e.g., G04

`AAAAA` -> student 1 number, e.g., 45454

`BBBBB` -> student 2 number, e.g., 54321

(the numbers AAAAA and BBBBB must be in increasing order),

using the form at the address:

https://forms.gle/fXkwao27KAMwuZw7A

no later than Sunday, December 6, 2019 @ 23:59.