McGill | School of Continuing Studies | École d'éducation permanente

Assessment
Data at Scale
YCBS-257

Assignment 1

15% - Part 2 of 2 due on 23:59 Sunday Feb 10 2019

Hadoop Data Analysis using Apache Hive / Impala

Provided files:

9 Bixi Data files (one file for Stations, 8 files for rides)
Download this files to your machine in order to perform the assignment

1. Hive 75%

Like Pig, Apache Hive makes it possible to write complex queries over big datasets, but it has the advantage of using HiveQL, a language that is very easy to learn for people who are used to work with databases since it is very similar to SQL.

While doing this exercise, you are encouraged to look up the online documentation https://cwiki.apache.org/confluence/display/Hive/LanguageManual

https://www.cloudera.com/documentation/enterprise/latest/topics/impala_functions.html

Question 1

Write a HiveQL script that create two non-managed tables, named **data** and **stations**, that contain the data from files OD_2018*.csv and Stations_2018.csv, respectively.

Question 2:

Write a HiveQL script to compute and print on the screen the number of rows for each table (data, stations).

Question 3:

Write a HiveQL script to split the data table into two tables **members** and **notmembers**. The members table should contains only peoples that are members and the notmembers table should contains all non-member people.

Question 4:

Write a HiveQL script to compute and print on the screen the number of rows for each table (members, notmembers).

Question 5:

Write a HiveQL script that given some station's code (hard-coded constant) will return the stations record if found and all the ride for this station.

Question 6:

Write a HiveQL script that will calculate the number of rides departures per station. The output has to be sorted descending.

Question 7:

Write a HiveQL script that will calculate the count of rides per station (start station) and the MIN, MAX, AVG of the ride's duration for members and non-members.

Question 8:

Write a HiveQL script that will list the Top 5 names of the start station for members and for non-members

Question 9:

Write a HiveQL script that will list the longest and shortest ride done by a member and by a non-member.

Question 10:

Write a HiveQL script that will:

- Create a new partitioned table by the (is_member) column
- Populate this table with rows from the data table
- Dynamically create partition for inserted rows

Question 11:

Write a HiveQL script that will

- Create a new partitioned table by the (is_member) column
- With **8** buckets clustered by (start_station_code) and sorted by (duration_sec)
- Dynamically organize data into **8** buckets

Question 12:

Write a HiveQL script that will return rows count for all bucket **4** of the bucketed table you just created it in the previous question.

2. Impala 25%

Like Apache Hive, Impala makes it possible to write complex queries over big datasets, but it has the advantage of using interactive HiveQL.

Question 1:

Write a HiveQL script that will calculate:

- The percentage of the top 5 start stations
- Dhe percentage of the top 5 end stations.
- Display the name of these stations with the persentage rather than the station code.

Question 2:

Write a HiveQL script that will calculate the distance in KM for the shortest and the longest ride

To calculate a distance between two points based on their longitude and latitude coordinates you can use the Haversine formula.

https://en.wikipedia.org/wiki/Haversine_formula

$$d = 2r \arcsin\left(\sqrt{\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

What to submit:

One Microsoft Word file

- ○ <your name> - W2019A1P2.docx

✓ Sections inside should be separated and named

✓ Comments can be submitted in French or in English in the text file