CS4545/CS6545 – Hands-on 4
Winter 2021, UNB, Fredericton
Due: March 16, 2020 at 11 am

In this hands-on, you'll become familiar with Hadoop-based data warehouse system Hive.

INSTRUCTIONS: Launching

1. Remotely connect to the FCS lab (see *RemoteDesktopToALabMachine.pdf*)
   Launch and login to the BigDataSystems Master VM and the 3 Slave VMs.

2. On BigDataSystemsMaster VM type the command:
$ start-all.sh

This will start the Hadoop cluster (on all 4 VMs).

3. On BigDataSystemsMaster VM, go to *Datasets* folder
$ cd Datasets/

4. Run the command: hive

$ hive
hive>

5. Show the existing tables in Hive
hive> show tables;

INSTRUCTIONS: DEMO1

6. Create table *emp* and load data from file.

```
hive> CREATE TABLE emp (name STRING, age INT, eid INT)
        ROW FORMAT DELIMITED
        FIELDS TERMINATED BY '|' ;

hive> LOAD DATA LOCAL INPATH '/home/bigdata/hive/apache-hive-1.2.1-bin/examples/files/emp.txt'
OVERWRITE INTO TABLE emp;
```

7. Run a HiveQL query on *emp*. See that a map-reduce task is launched
hive> select name, age, eid from emp;

INSTRUCTIONS: DEMO2

8. Create a partitioned table *invites* and load data from files.

```
hive> CREATE TABLE invites (foo INT, bar STRING) PARTITIONED BY (ds STRING);

hive> LOAD DATA LOCAL INPATH '/home/bigdata/hive/apache-hive-1.2.1-bin/examples/files/kv2.txt'
OVERWRITE INTO TABLE invites PARTITION (ds='2008-08-15');
```

hive> LOAD DATA LOCAL INPATH '/home/bigdata/hive/apache-hive-1.2.1-bin/examples/files/kv3.txt' OVERWRITE INTO TABLE invites PARTITION (ds='2008-08-08');

hive> select count(*) from invites;

INSTRUCTIONS: Task and deliverables

1. Download the zip file and unzip it:
wget https://www.cs.unb.ca/~sray/teaching/bds/handson/h4datasets.zip

It contains two files. File *countries.csv* contains information about countries. Whereas, *covid19-detail.csv* contains information about daily Covid 19 statistics for each country.

2. Create two Hive-based tables *covid* and *countries*. They should have the following schema. Inspect the data files (that you downloaded above) to determine the datatype of the columns.

covid (icode, continent, location, rdate, new_cases, new_deaths, icu_patients, hospital_patients, new_vaccinations)

countries (country, population, area, currency)

3. Load data into *covid* from *covid19-detail.csv*.
Then, load data into *countries* from *countries.csv*.

4. Then answer the queries below with HiveQL:

Q4.1: Show the total number of recoveries in North America (continent).
Note: the number of recoveries can be calculated by subtracting the total deaths from total cases.

Q4.2: Show the country name and the number of cases for each country where the total number of cases is more than 1 million.

Q4.3: Show the country name of the top 10 countries by the number of ICU patients.

Q4.4: Show the country name and the mortality rate (i.e. total number of deaths per 100,000 population) of the top 10 countries by mortality rate.

Q4.5: Show the name of the top 3 countries with the highest vaccination rate (i.e. % of the population vaccinated) where the country population is larger than 1 million.

INSTRUCTIONS: D2L Submission

Submit via Desire To Learn (D2L). In a single file: h4_<your_student_id>.txt include the following:

1. The create table statement for *covid* and *countries*
2. Table data loading statements for loading *covid* and *countries*
3. Queries Q4.1 to Q4.5
4. Output of the queries Q4.1 to Q4.5

NOTES ABOUT PLAGIARISM

Please note that the handsons are meant to be done individually. Any submission that appears to be in violation of an academic offence (plagiarism) may be reported to the Registrar's Office as per UNB regulations (See section VII of UNB Undergraduate Calendar).