**Paul Ycay**

**500709618**

**CIND719 Big Data Analytics Tools**

**Fall 2019, Wednesday 6:30-9:30 pm section**

**Wednesday, October 2, 2019**

## Assignment 1

hadoop fs –put /root/lab/station_data.csv /user/lab
hadoop fs –put /root/lab/trip_data.csv /user/lab
hadoop fs –mkdir /root/lab

hadoop fs -put /root/station_data.csv /user/lab

hadoop fs -put /root/trip_data.csv /user/lab

hadoop fs -ls /user/lab

create table trip_data ( trip_id int, duration string, start_date string, start_station string, start_terminal int, end_date date, end_station string, end_terminal int, bike_id int, subscription_type string, zip_code int) row format delimited fields terminated by ',';

create table station_data ( station_id int, station_name string, lat string, lon string, dockcount int, landmark string, installation string ) row format delimited fields terminated by ',';

load data inpath '/user/lab/trip_data.csv' overwrite into table trip_data;

load data inpath '/user/lab/station_data.csv' overwrite into table station_data;

Question 1

```
hive> select max (bike_id) from trip_data as popularbike;
Query ID = root_20191003002258_130d31c4-b3f0-45d6-808a-8bb70166d3b4
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1569455041304_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 16.23 s
--------------------------------------------------------------------------------
OK
878
Time taken: 23.38 seconds, Fetched: 1 row(s)
```

select max (bike_id) from trip_data as popularbike;

*or*

SELECT bike_id, COUNT(*)

FROM trip_data

GROUP BY bike_id

ORDER BY COUNT(*) DESC

LIMIT 1;

Question 2

```
hive> select subscription_type, count (subscription_type) from trip_data group by subscription_type;
Query ID = root_20191003004213_8f8c1a73-2e66-414b-ba65-39404fc16298
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1569455041304_0011)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 13.19 s
--------------------------------------------------------------------------------
OK
Customer        43935
Subscriber      310217
Time taken: 30.307 seconds, Fetched: 2 row(s)
```

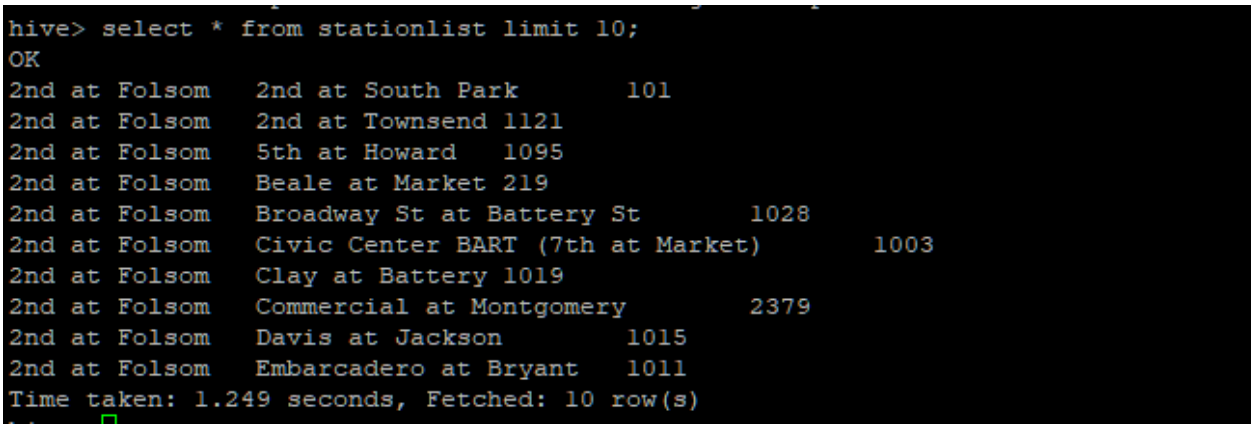
Select subscription_type, count (subscription_type) from trip_data group by subscription_type;

Question 3

```
hive> create table stationlist as select start_station, end_station, min(duration) from trip_data where start_station
 != end_station group by start_station, end_station;
Query ID = root_20191003062100_c7b627dc-975f-4c97-8ab0-1b5b3df2f84c
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1570081996416_0006)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 9.58 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/stationlist
Table default.stationlist stats: [numFiles=1, numRows=1622, totalSize=84157, rawDataSize=82535]
OK
Time taken: 23.753 seconds
```




create table stationlist as select start_station, end_station, min(duration) from trip_data where start_station != end_station group by start_station, end_station;

select * from stationlist limit 10;