# CIND719-DK0T
# Assignment 1:

Ramello Peralta 500519802

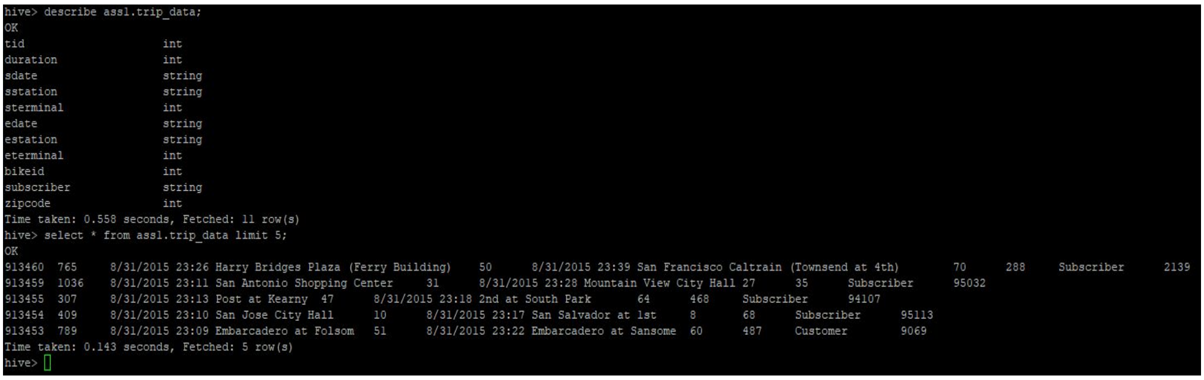
Creating the database + tables:

**Create database ass1;**

**Use ass1;**

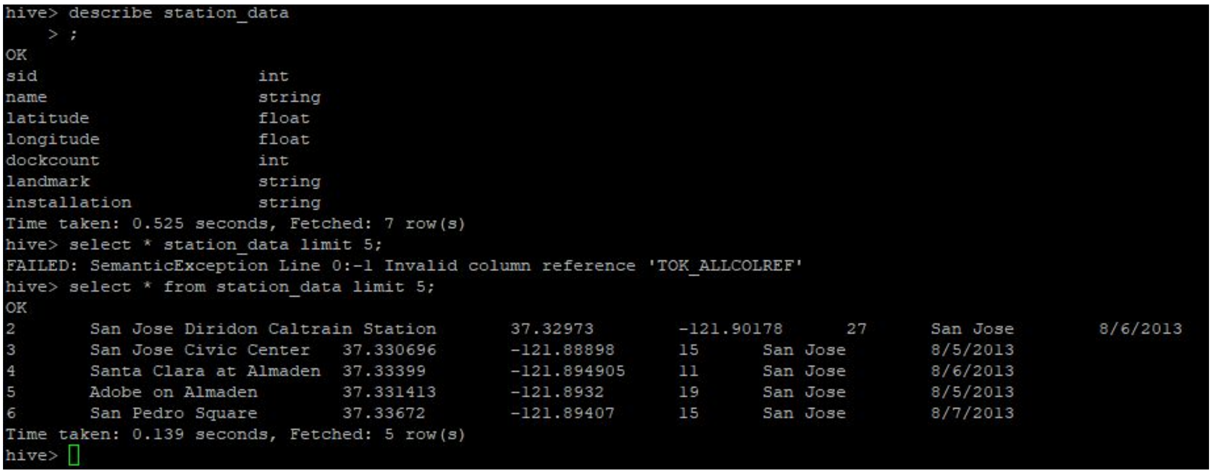
**Create table trip_data (tid int, duration int, date string, station string, terminal int, edate string, estation string, eterminal int, bikeid int, subscriber string, zipcode int) row format delimited fields terminated by '\t';**

**Load data inpath '/user/CIND719/trip_data.csv' overwrite into table ass1.trip_data;**



```
hive> describe ass1.trip_data;
OK
tid                     int
duration                int
sdate                   string
sstation                string
sterminal               int
edate                   string
estation                string
eterminal               int
bikeid                  int
subscriber              string
zipcode                 int
Time taken: 0.558 seconds, Fetched: 11 row(s)
hive> select * from ass1.trip_data limit 5;
OK
913460  765     8/31/2015 23:26 Harry Bridges Plaza (Ferry Building)    50      8/31/2015 23:39 San Francisco Caltrain (Townsend at 4th)        70      288     Subscriber      2139
913459  1036    8/31/2015 23:11 San Antonio Shopping Center     31      8/31/2015 23:28 Mountain View City Hall 27      35      Subscriber      95032
913455  307     8/31/2015 23:13 Post at Kearny  47      8/31/2015 23:18 2nd at South Park       64      468     Subscriber      94107
913454  409     8/31/2015 23:10 San Jose City Hall      10      8/31/2015 23:17 San Salvador at 1st     8       68      Subscriber      95113
913453  789     8/31/2015 23:09 Embarcadero at Folsom   51      8/31/2015 23:22 Embarcadero at Sansome  60      487     Customer        9069
Time taken: 0.143 seconds, Fetched: 5 row(s)
hive> 
```

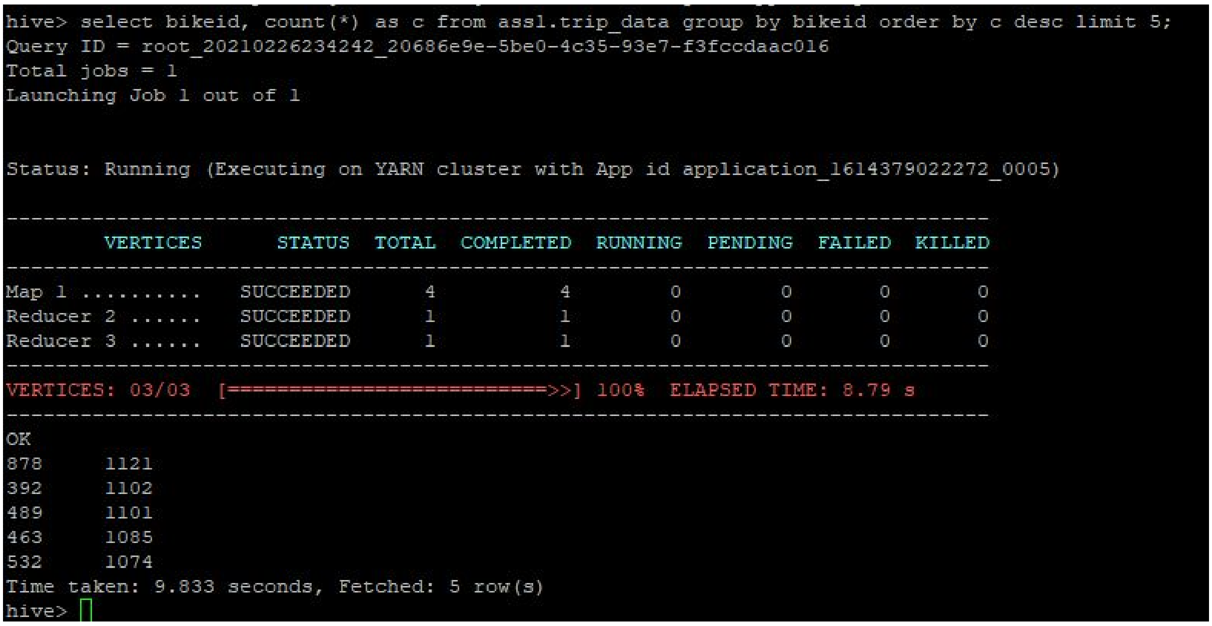
- same was done with station_data table



```
hive> describe station_data
    > ;
OK
sid                     int
name                    string
latitude                float
longitude               float
dockcount               int
landmark                string
installation            string
Time taken: 0.525 seconds, Fetched: 7 row(s)
hive> select * station_data limit 5;
FAILED: SemanticException Line 0:-1 Invalid column reference 'TOK_ALLCOLREF'
hive> select * from station_data limit 5;
OK
2       San Jose Diridon Caltrain Station       37.32973        -121.90178      27      San Jose        8/6/2013
3       San Jose Civic Center   37.330696       -121.88898      15      San Jose        8/5/2013
4       Santa Clara at Almaden  37.33399        -121.894905     11      San Jose        8/6/2013
5       Adobe on Almaden        37.331413       -121.8932       19      San Jose        8/5/2013
6       San Pedro Square        37.33672        -121.89407      15      San Jose        8/7/2013
Time taken: 0.139 seconds, Fetched: 5 row(s)
hive> 
```

1. Find the 'most popular' bike, i.e. the bike that has made the highest number of trips (1.5 pts)

**Select bikeid, count(*) as c from ass1.trip_data group by bikeid order by c desc limit 5;**

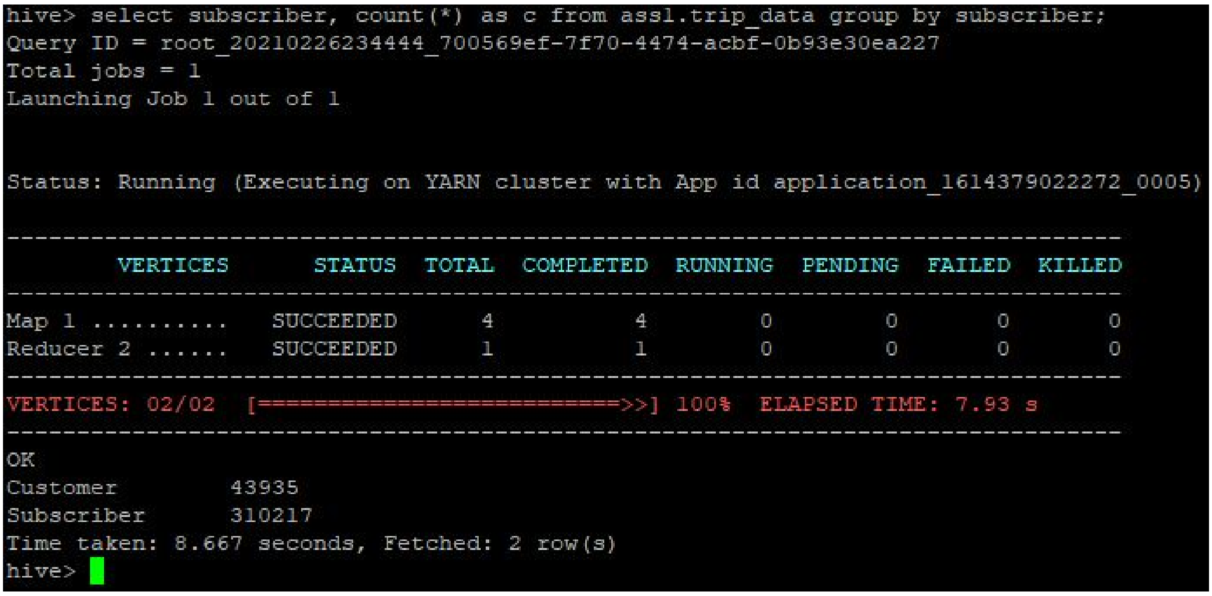


```
hive> select bikeid, count(*) as c from ass1.trip_data group by bikeid order by c desc limit 5;
Query ID = root_20210226234242_20686e9e-5be0-4c35-93e7-f3fccdaac016
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1614379022272_0005)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 8.79 s
----------------------------------------------------------------------------------------------
OK
878     1121
392     1102
489     1101
463     1085
532     1074
Time taken: 9.833 seconds, Fetched: 5 row(s)
hive>
```

Bike ID 878 has the highest number of trips made at 1121 trips.

2. Find the number of trips made by each subscription type. (1.5 pts)

**Select subscriber, count(*) as c from ass1.trip_data group by subscriber;**

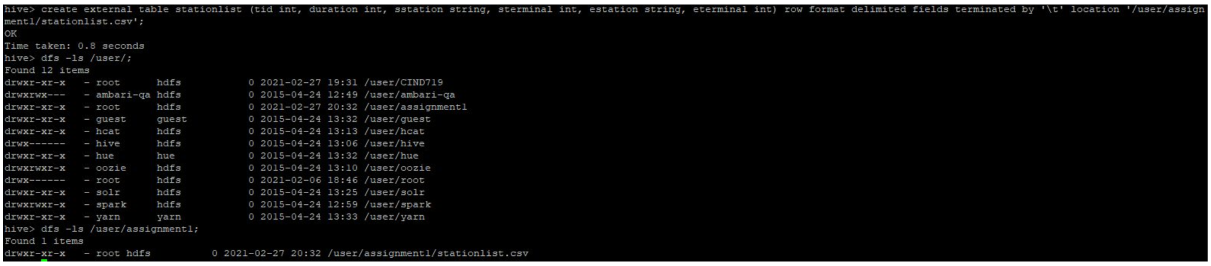


```
hive> select subscriber, count(*) as c from ass1.trip_data group by subscriber;
Query ID = root_20210226234444_700569ef-7f70-4474-acbf-0b93e30ea227
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1614379022272_0005)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 7.93 s
----------------------------------------------------------------------------------------------
OK
Customer        43935
Subscriber      310217
Time taken: 8.667 seconds, Fetched: 2 row(s)
hive>
```

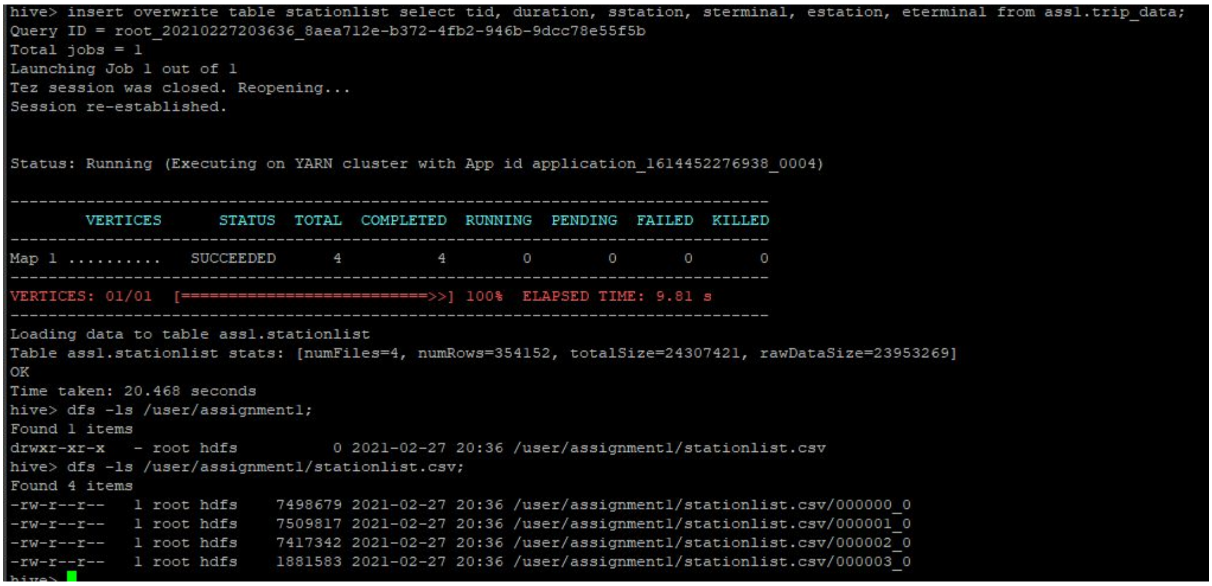
Subscribers have made 310217 trips while Customers have made 43935.

3. Build a table that shows which stations are connected, and the minimum duration between them. You can use either station id or station name. Save this table as a comma separated text file in ‘/user/assignment1/stationlist.csv’ in HDFS. Include the directory listing of the output directory and first five lines of the output file in your submission. (3 pts)

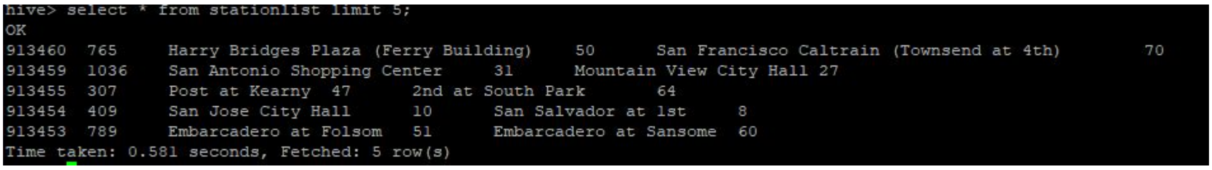
**Create external table stationlist (tid int, duration int, station string, terminal int, estation string, eterminal int) row format delimited fields terminated by ‘,’ location ‘/user/assignment1/stationlist.csv’;**



```
hive> create external table stationlist (tid int, duration int, sstation string, sterminal int, estation string, eterminal int) row format delimited fields terminated by '\t' location '/user/assign
ment1/stationlist.csv';
OK
Time taken: 0.8 seconds
hive> dfs -ls /user/;
Found 12 items
drwxr-xr-x   - root      hdfs          0 2021-02-27 19:31 /user/CIND719
drwxrwx---   - ambari-qa hdfs          0 2015-04-24 12:49 /user/ambari-qa
drwxr-xr-x   - root      hdfs          0 2021-02-27 20:32 /user/assignment1
drwxr-xr-x   - guest     guest         0 2015-04-24 13:32 /user/guest
drwxr-xr-x   - hcat      hdfs          0 2015-04-24 13:13 /user/hcat
drwx------   - hive      hdfs          0 2015-04-24 13:06 /user/hive
drwxr-xr-x   - hue       hue           0 2015-04-24 13:32 /user/hue
drwxrwxr-x   - oozie     hdfs          0 2015-04-24 13:10 /user/oozie
drwx------   - root      hdfs          0 2021-02-06 18:46 /user/root
drwxr-xr-x   - solr      hdfs          0 2015-04-24 13:25 /user/solr
drwxrwxr-x   - spark     hdfs          0 2015-04-24 12:59 /user/spark
drwxr-xr-x   - yarn      yarn          0 2015-04-24 13:33 /user/yarn
hive> dfs -ls /user/assignment1;
Found 1 items
drwxr-xr-x   - root hdfs          0 2021-02-27 20:32 /user/assignment1/stationlist.csv
```



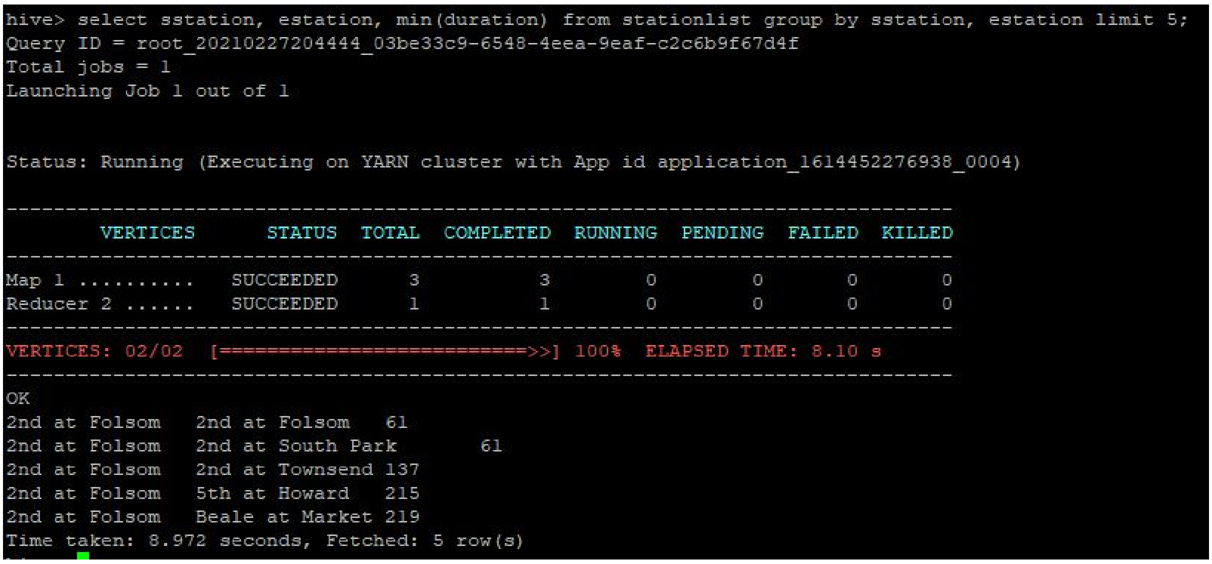
```
hive> insert overwrite table stationlist select tid, duration, sstation, sterminal, estation, eterminal from ass1.trip_data;
Query ID = root_20210227203636_8aea712e-b372-4fb2-946b-9dcc78e55f5b
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1614452276938_0004)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 9.81 s
--------------------------------------------------------------------------------
Loading data to table ass1.stationlist
Table ass1.stationlist stats: [numFiles=4, numRows=354152, totalSize=24307421, rawDataSize=23953269]
OK
Time taken: 20.468 seconds
hive> dfs -ls /user/assignment1;
Found 1 items
drwxr-xr-x   - root hdfs          0 2021-02-27 20:36 /user/assignment1/stationlist.csv
hive> dfs -ls /user/assignment1/stationlist.csv;
Found 4 items
-rw-r--r--   1 root hdfs    7498679 2021-02-27 20:36 /user/assignment1/stationlist.csv/000000_0
-rw-r--r--   1 root hdfs    7509817 2021-02-27 20:36 /user/assignment1/stationlist.csv/000001_0
-rw-r--r--   1 root hdfs    7417342 2021-02-27 20:36 /user/assignment1/stationlist.csv/000002_0
-rw-r--r--   1 root hdfs    1881583 2021-02-27 20:36 /user/assignment1/stationlist.csv/000003_0
hive>
```



```
hive> select * from stationlist limit 5;
OK
913460  765     Harry Bridges Plaza (Ferry Building)    50      San Francisco Caltrain (Townsend at 4th)        70
913459  1036    San Antonio Shopping Center     31      Mountain View City Hall 27
913455  307     Post at Kearny  47      2nd at South Park       64
913454  409     San Jose City Hall      10      San Salvador at 1st     8
913453  789     Embarcadero at Folsom   51      Embarcadero at Sansome  60
Time taken: 0.581 seconds, Fetched: 5 row(s)
```

**Select sstation, estation, min(duration) from stationlist group by sstation, estation limit 5;**

- Listing the first 5 trip combinations between start and end terminal with lowest duration



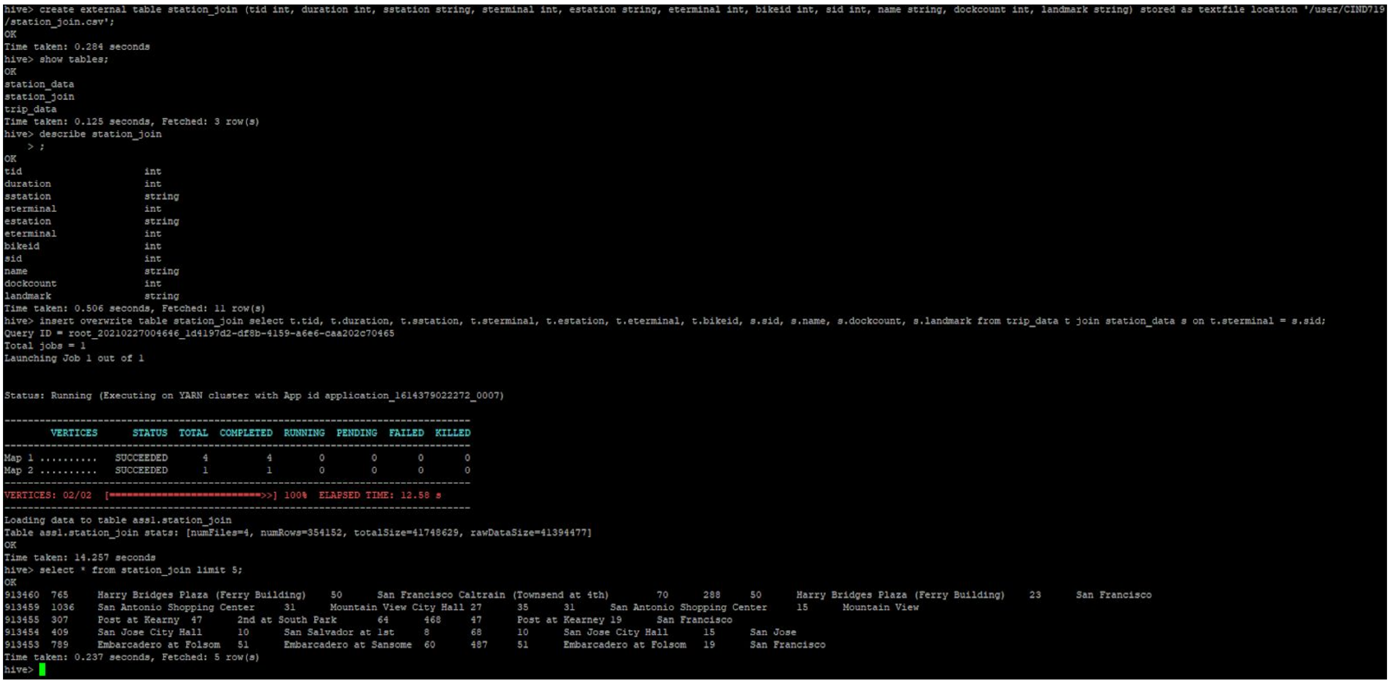
```
hive> select sstation, estation, min(duration) from stationlist group by sstation, estation limit 5;
Query ID = root_20210227204444_03be33c9-6548-4eea-9eaf-c2c6b9f67d4f
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1614452276938_0004)

--------------------------------------------------------------------------------
        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: 8.10 s
--------------------------------------------------------------------------------
OK
2nd at Folsom   2nd at Folsom   61
2nd at Folsom   2nd at South Park       61
2nd at Folsom   2nd at Townsend 137
2nd at Folsom   5th at Howard   215
2nd at Folsom   Beale at Market 219
Time taken: 8.972 seconds, Fetched: 5 row(s)
```

4. Find the number of trips originating from each landmark. Your output should include the landmark name and the number of trips originating from it. (3 pts)

**Create external table station_join (tid int, duration int, sstation string, sterminal int, estation string, eterminal int, bikeid int, sid int, name string, dockcount int, landmark string) stored as textfile location ‘/user/CIND719/station_join.csv’;**

- Joining station_data and trip_data on start terminal to match landmark name and start terminal

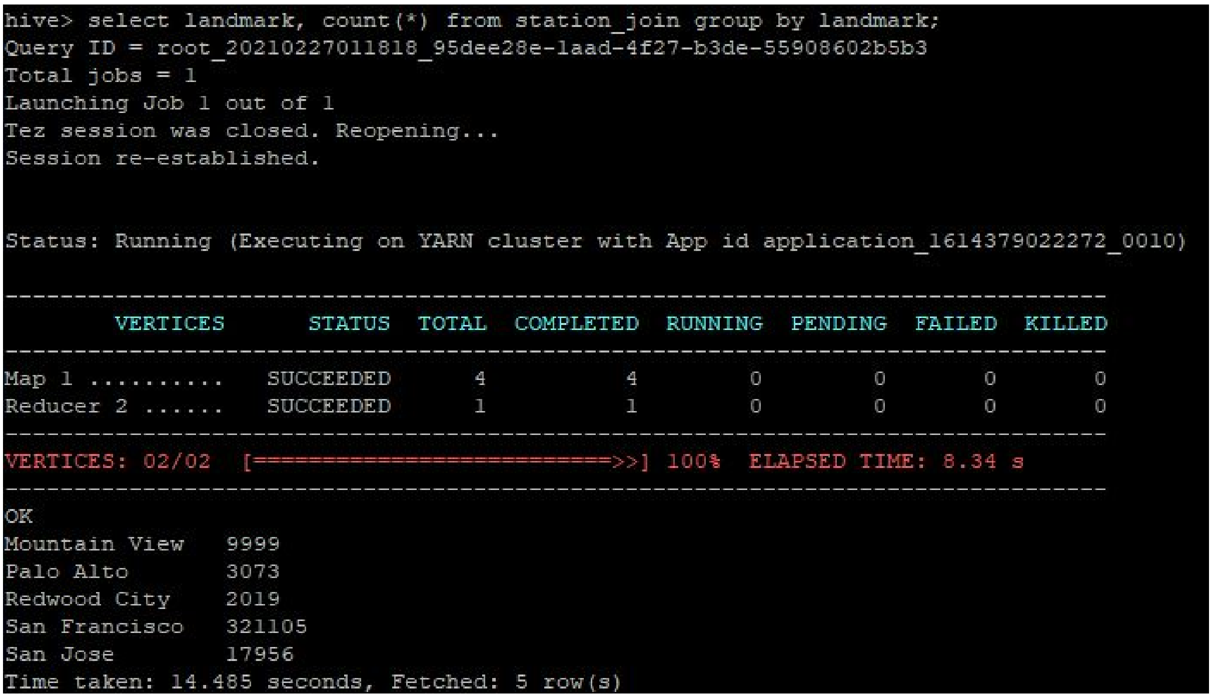
**Insert overwrite table station_join select t.tid, t.duration, t.sstation, tsterminal, t.estation, t.eterminal, t.bikeid, s.sid, s.name, s.dockcount, s.landmark from trip_data t join station_data s on t.sterminal = s.sid;**



```
hive> create external table station_join (tid int, duration int, sstation string, sterminal int, estation string, eterminal int, bikeid int, sid int, name string, dockcount int, landmark string) stored as textfile location '/user/CIND719
/station_join.csv';
OK
Time taken: 0.284 seconds
hive> show tables;
OK
station_data
station_join
trip_data
Time taken: 0.125 seconds, Fetched: 3 row(s)
hive> describe station_join
    > ;
OK
tid                     int
duration                int
sstation                string
sterminal               int
estation                string
eterminal               int
bikeid                  int
sid                     int
name                    string
dockcount               int
landmark                string
Time taken: 0.506 seconds, Fetched: 11 row(s)
hive> insert overwrite table station_join select t.tid, t.duration, t.sstation, t.sterminal, t.estation, t.eterminal, t.bikeid, s.sid, s.name, s.dockcount, s.landmark from trip_data t join station_data s on t.sterminal = s.sid;
Query ID = root_20210227004646_1d4197d2-df8b-4159-a6e6-caa202c70465
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1614379022272_0007)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Map 2 ..........   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 12.58 s
--------------------------------------------------------------------------------
Loading data to table assl.station_join
Table assl.station_join stats: [numFiles=4, numRows=354152, totalSize=41748629, rawDataSize=41394477]
OK
Time taken: 14.257 seconds
hive> select * from station_join limit 5;
OK
913460  765     Harry Bridges Plaza (Ferry Building)    50      San Francisco Caltrain (Townsend at 4th)        70      288     50      Harry Bridges Plaza (Ferry Building)    23      San Francisco
913459  1036    San Antonio Shopping Center     31      Mountain View City Hall 27      35      31      San Antonio Shopping Center     15      Mountain View
913455  307     Post at Kearny  47      2nd at South Park       64      468     47      Post at Kearney 19      San Francisco
913454  409     San Jose City Hall      10      San Salvador at 1st     8       68      10      San Jose City Hall      15      San Jose
913453  789     Embarcadero at Folsom   51      Embarcadero at Sansome  60      487     51      Embarcadero at Folsom   19      San Francisco
Time taken: 0.237 seconds, Fetched: 5 row(s)
hive> 
```

**Select landmark, count(*) from station_join group by landmark;**

- Showing start terminal landmark name and number of trips originating from this landmark

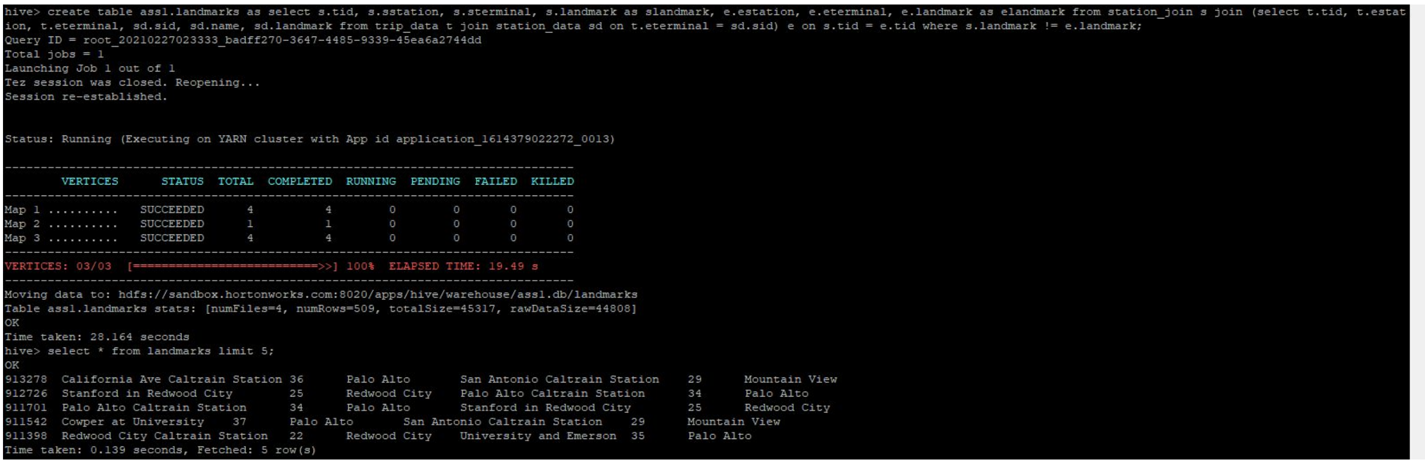


```
hive> select landmark, count(*) from station_join group by landmark;
Query ID = root_20210227011818_95dee28e-1aad-4f27-b3de-55908602b5b3
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1614379022272_0010)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 8.34 s
--------------------------------------------------------------------------------
OK
Mountain View   9999
Palo Alto       3073
Redwood City    2019
San Francisco   321105
San Jose        17956
Time taken: 14.485 seconds, Fetched: 5 row(s)
```

5. Find the number of trips crossing landmarks, i.e. trips that originate in one landmark and end in another. Your output should include the originating and ending landmark names and the number of trips between them. (6 pts)

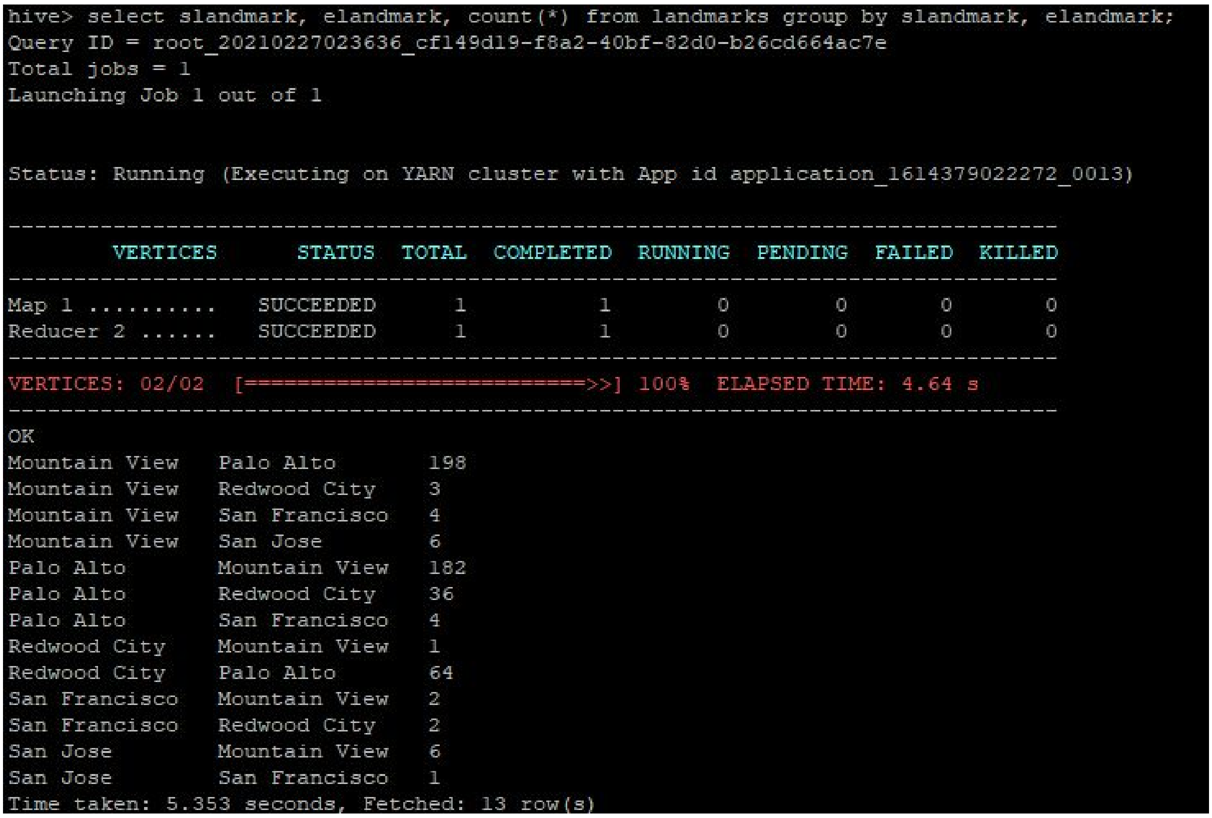
```
Create table ass1.landmarks as
        select s.tid, s.sstation, s.sterminal, s.landmark as slandmark, e.estation,
        e.eterminal, e.landmark as elandmark from station_join s
                join (select t.tid, t.estation, t.eterminal, sd.sid, sd.name, sd.landmark
                from trip_data t join station_data sd on t.eterminal = sd.sid) e
                on s.tid = e.tid
                where s.landmark != e.landmark;
```

- Creating a landmarks table from nested query
- **(select t.tid, t.estation, t.eterminal, sd.sid, sd.name, sd.landmark from trip_data t join station_data sd on t.eterminal = sd.sid)**
    - Joining trip_data and station_data tables on station id to find landmark name for END terminal(t.eterminal) this time.
- Outer query includes the unique trip id and start/end terminal names by joining the inner query (joined on end terminal) with station_join (station_join table already is joined on start terminal) filtered by trips that have different start/end landmarks



```
hive> create table ass1.landmarks as select s.tid, s.sstation, s.sterminal, s.landmark as slandmark, e.estation, e.eterminal, e.landmark as elandmark from station_join s join (select t.tid, t.estat
ion, t.eterminal, sd.sid, sd.name, sd.landmark from trip_data t join station_data sd on t.eterminal = sd.sid) e on s.tid = e.tid where s.landmark != e.landmark;
Query ID = root_20210227023333_badff270-3647-4485-9339-45ea6a2744dd
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_1614379022272_0013)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      4          4        0        0       0       0
Map 2 ..........   SUCCEEDED      1          1        0        0       0       0
Map 3 ..........   SUCCEEDED      4          4        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 19.49 s
--------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/ass1.db/landmarks
Table ass1.landmarks stats: [numFiles=4, numRows=509, totalSize=45317, rawDataSize=44808]
OK
Time taken: 28.164 seconds
hive> select * from landmarks limit 5;
OK
913278  California Ave Caltrain Station 36      Palo Alto       San Antonio Caltrain Station    29      Mountain View
912726  Stanford in Redwood City        25      Redwood City    Palo Alto Caltrain Station      34      Palo Alto
911701  Palo Alto Caltrain Station      34      Palo Alto       Stanford in Redwood City        25      Redwood City
911542  Cowper at University    37      Palo Alto       San Antonio Caltrain Station    29      Mountain View
911398  Redwood City Caltrain Station   22      Redwood City    University and Emerson  35      Palo Alto
Time taken: 0.139 seconds, Fetched: 5 row(s)
```

**Select slandmark, elandmark, count(*) from landmarks group by slandmark, elandmark;**



```
hive> select slandmark, elandmark, count(*) from landmarks group by slandmark, elandmark;
Query ID = root_20210227023636_cf149d19-f8a2-40bf-82d0-b26cd664ac7e
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1614379022272_0013)

----------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 4.64 s
----------------------------------------------------------------------------------
OK
Mountain View   Palo Alto       198
Mountain View   Redwood City    3
Mountain View   San Francisco   4
Mountain View   San Jose        6
Palo Alto       Mountain View   182
Palo Alto       Redwood City    36
Palo Alto       San Francisco   4
Redwood City    Mountain View   1
Redwood City    Palo Alto       64
San Francisco   Mountain View   2
San Francisco   Redwood City    2
San Jose        Mountain View   6
San Jose        San Francisco   1
Time taken: 5.353 seconds, Fetched: 13 row(s)
```

End of Assignment 1.