SQL Queries in a Hadoop Cluster (HiveQL)

Data Preparation

1. Add files to Hadoop Cluster (Linux)

```
hadoop fs -put /root/lab/station_data.csv /user/lab/station_dat.csv
hadoop fs -put /root/lab/trip_data.csv /user/lab/trip_dat.csv
```

2. Create database (SQL)

```
create database bikes;
use bikes;
```

4. Create Tables

```
create table bikes.stationtemp (
station_id int, name string, lat float, lon float, dockcount int, landmark string, install string)
row format delimited
fields terminated by ',';

load data inpath '/user/lab/station_dat'
overwrite into table bikes.stationtemp;

create table bikes.station as
select station_id, name, lat, lon, dockcount, landmark,
 from_unixtime(unix_timestamp(install , 'M/d/yyyy')) as install_date
from bikes.stationtemp;
```

5. Test if bikes.station exists, then drop temporary table.

```
select * from bikes.station limit 5;
drop table bikes.stationtemp;

create table bikes.triptemp (
trip_id int, duration float, start_date string, start_station string, start_terminal int, end_date string,
end_station string, end_terminal int, bike_number int, sub_type string, zip int)
row format delimited
fields terminated by ',';

load data inpath '/user/lab/trip_dat.csv'
overwrite into table bikes.triptemp;



create table bikes.trip as
```

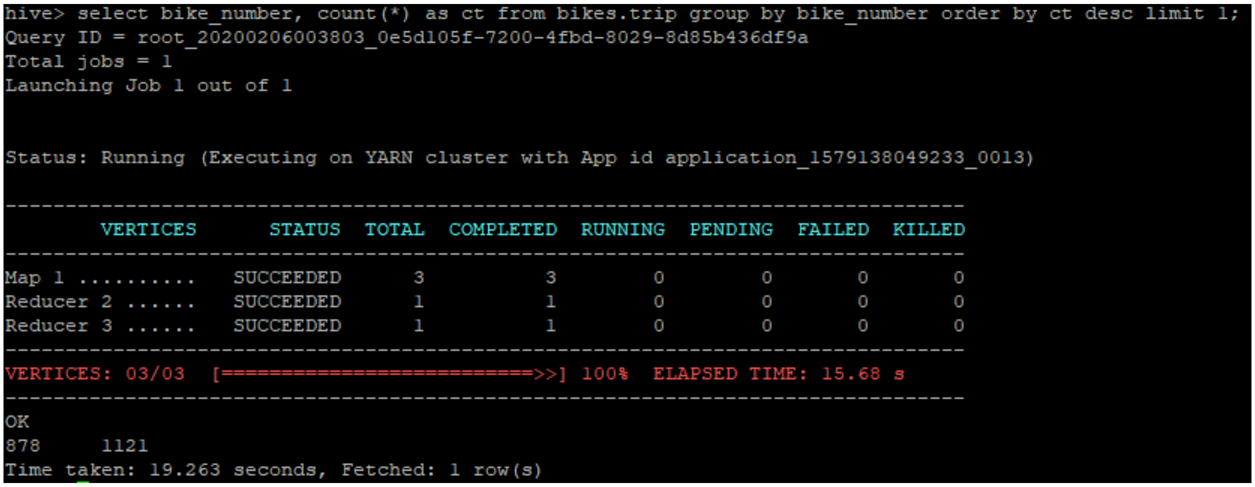
select trip_id, duration, from_unixtime(unix_timestamp(start_date , 'M/d/yyyy H:m')) as start_date, start_station, start_terminal, from_unixtime(unix_timestamp(end_date , 'M/d/yyyy H:m')) as end_date, end_station, end_terminal, bike_number, sub_type, zip
from bikes.triptemp;

6. Test if bikes.trip exists, then drop temp table.

select * from bikes.trip limit 5;
drop table bikes.triptemp;

Find the 'most popular' bike (the bike that has made the highest number of trips)

select bike_number, count(*) as ct from bikes.trip group by bike_number order by ct desc limit 1;



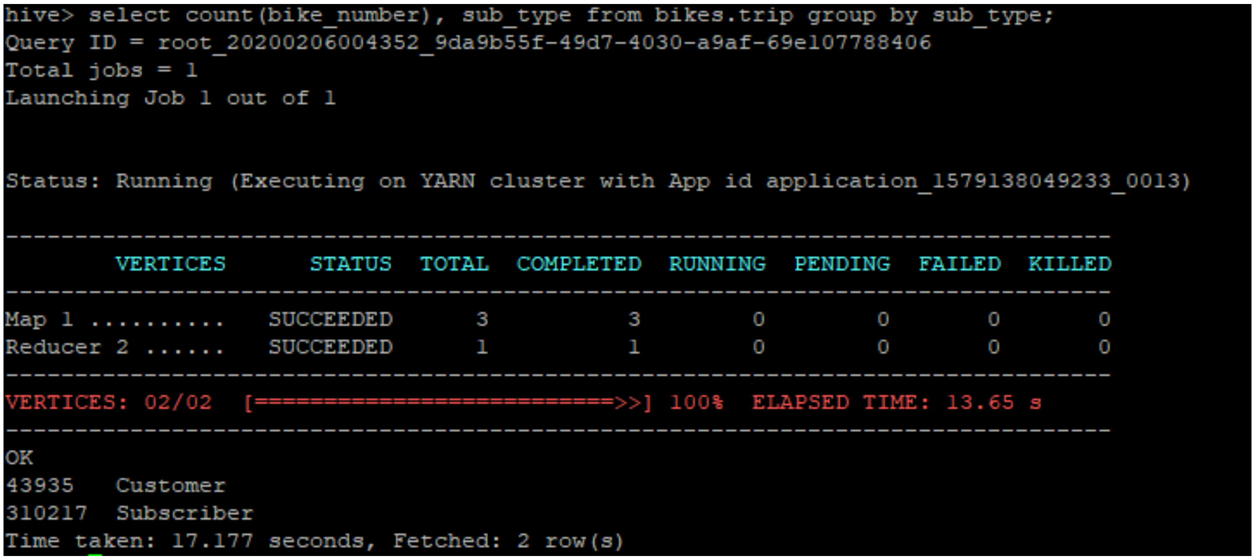
```
hive> select bike_number, count(*) as ct from bikes.trip group by bike_number order by ct desc limit 1;
Query ID = root_20200206003803_0e5d105f-7200-4fbd-8029-8d85b436df9a
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1579138049233_0013)

----------------------------------------------------------------------------------------------
        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
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 15.68 s
----------------------------------------------------------------------------------------------
OK
878     1121
Time taken: 19.263 seconds, Fetched: 1 row(s)
```

Find the number of trips made by each subscription type

select count(bike_number), sub_type from bikes.trip group by sub_type;



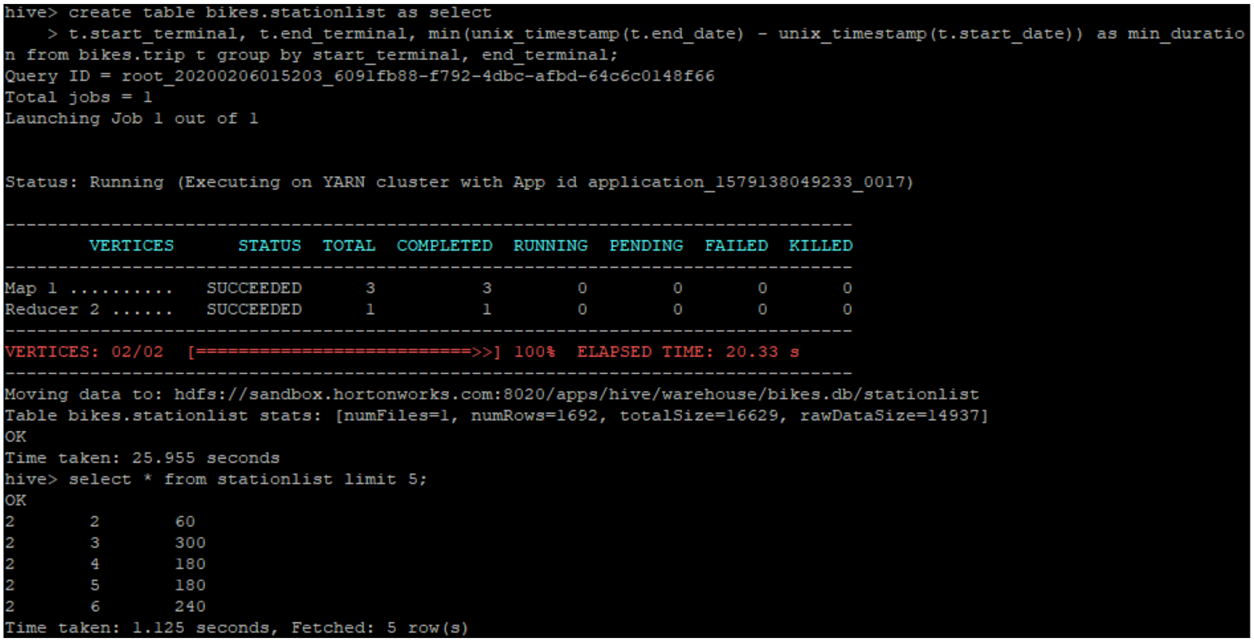
```
hive> select count(bike_number), sub_type from bikes.trip group by sub_type;
Query ID = root_20200206004352_9da9b55f-49d7-4030-a9af-69e107788406
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1579138049233_0013)

----------------------------------------------------------------------------------------------
        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: 13.65 s
----------------------------------------------------------------------------------------------
OK
43935   Customer
310217  Subscriber
Time taken: 17.177 seconds, Fetched: 2 row(s)
```

Build a table that shows which stations are connected, and the minimum duration between them.

create table bikes.stationlist as select
t.start_terminal, t.end_terminal,
min(unix_timestamp(t.end_date) - unix_timestamp(t.start_date)) as min_duration
from bikes.trip t
group by start_terminal, end_terminal;



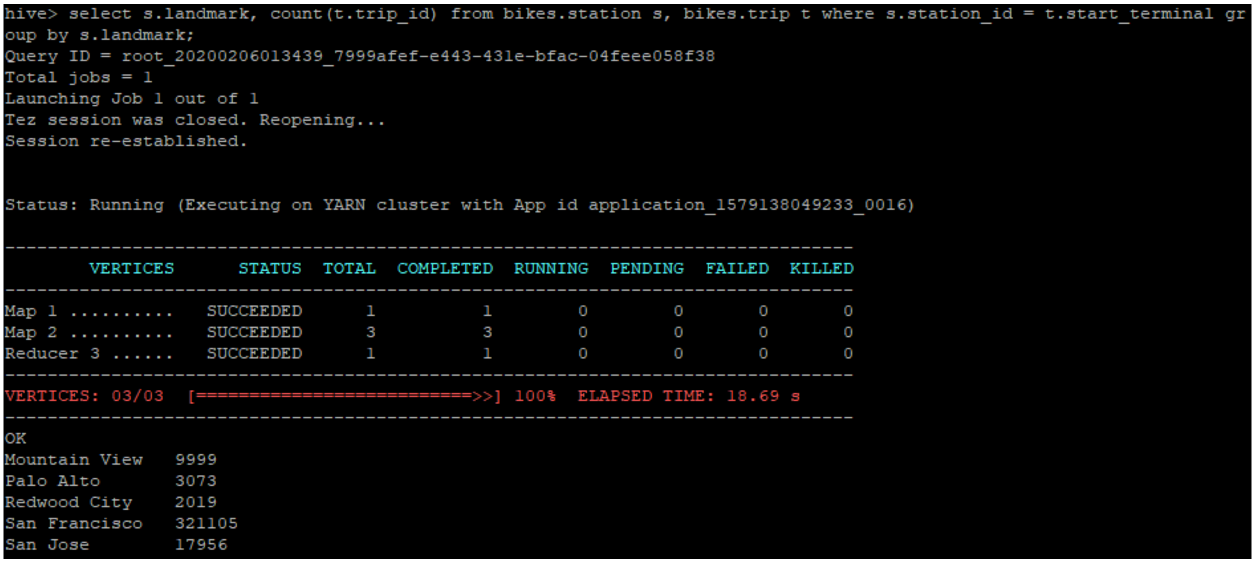
```
hive> create table bikes.stationlist as select
    > t.start_terminal, t.end_terminal, min(unix_timestamp(t.end_date) - unix_timestamp(t.start_date)) as min_duratio
n from bikes.trip t group by start_terminal, end_terminal;
Query ID = root_20200206015203_6091fb88-f792-4dbc-afbd-64c6c0148f66
Total jobs = 1
Launching Job 1 out of 1


Status: Running (Executing on YARN cluster with App id application_1579138049233_0017)

----------------------------------------------------------------------------------------------
        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: 20.33 s
----------------------------------------------------------------------------------------------
Moving data to: hdfs://sandbox.hortonworks.com:8020/apps/hive/warehouse/bikes.db/stationlist
Table bikes.stationlist stats: [numFiles=1, numRows=1692, totalSize=16629, rawDataSize=14937]
OK
Time taken: 25.955 seconds
hive> select * from stationlist limit 5;
OK
2       2       60
2       3       300
2       4       180
2       5       180
2       6       240
Time taken: 1.125 seconds, Fetched: 5 row(s)
```

4. Find the number of trips originating from each landmark.

select s.landmark, count(t.trip_id) from bikes.station s, bikes.trip t
where s.station_id = t.start_terminal group by s.landmark



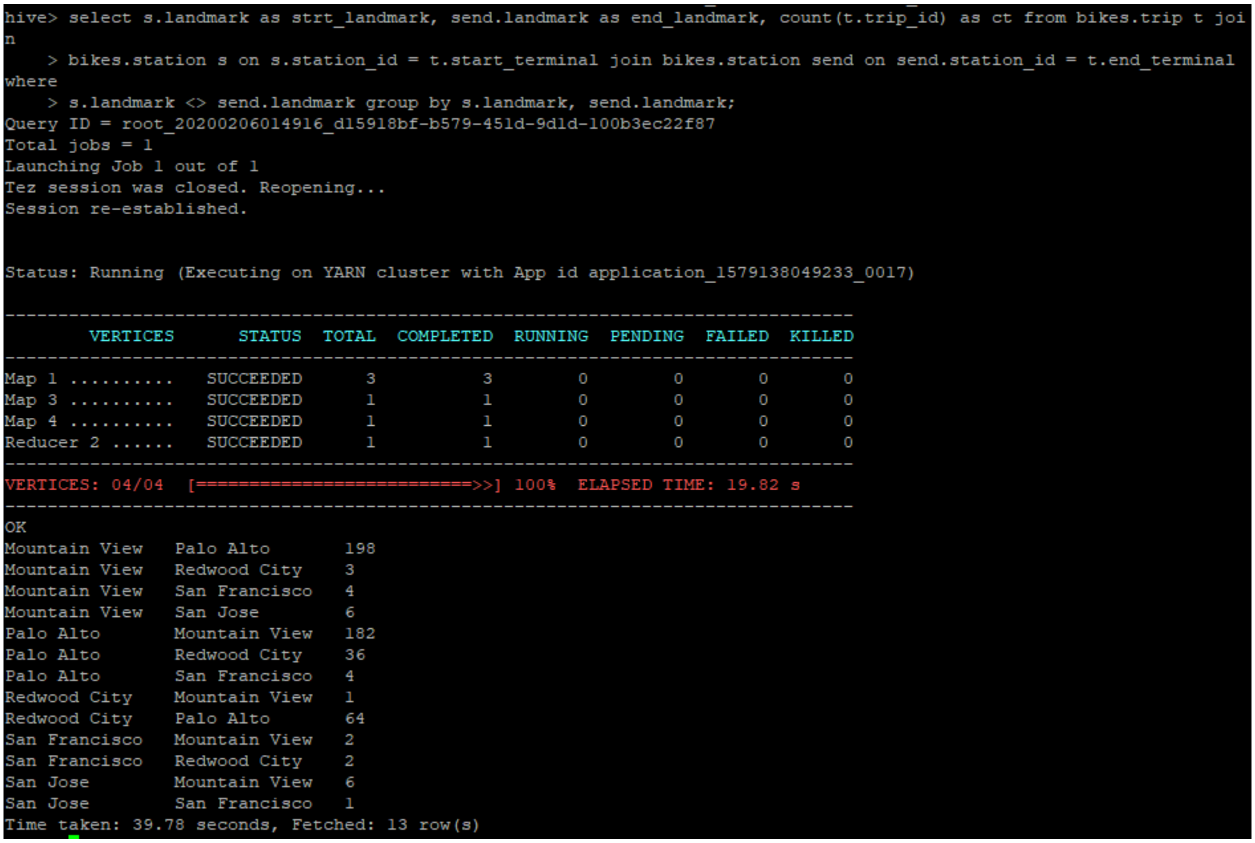
```
hive> select s.landmark, count(t.trip_id) from bikes.station s, bikes.trip t where s.station_id = t.start_terminal gr
oup by s.landmark;
Query ID = root_20200206013439_7999afef-e443-431e-bfac-04feee058f38
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.


Status: Running (Executing on YARN cluster with App id application_1579138049233_0016)

----------------------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Map 2 ..........   SUCCEEDED      3          3        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 18.69 s
----------------------------------------------------------------------------------------------
OK
Mountain View   9999
Palo Alto       3073
Redwood City    2019
San Francisco   321105
San Jose        17956
```

Find the number of trips crossing landmarks.

select s.landmark as strt_landmark, send.landmark as end_landmark, count(t.trip_id) as ct
from bikes.trip t join bikes.station s on s.station_id = t.start_terminal join bikes.station send
on send.station_id = t.end_terminal where s.landmark <> send.landmark
group by s.landmark, send.landmark;



```
hive> select s.landmark as strt_landmark, send.landmark as end_landmark, count(t.trip_id) as ct from bikes.trip t joi
n
    > bikes.station s on s.station_id = t.start_terminal join bikes.station send on send.station_id = t.end_terminal
where
    > s.landmark <> send.landmark group by s.landmark, send.landmark;
Query ID = root_20200206014916_d15918bf-b579-451d-9d1d-100b3ec22f87
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_1579138049233_0017)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      3          3        0        0       0       0
Map 3 ..........   SUCCEEDED      1          1        0        0       0       0
Map 4 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 04/04  [==========================>>] 100%  ELAPSED TIME: 19.82 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: 39.78 seconds, Fetched: 13 row(s)
```