Session 8

HIVE BASICS

Assignment 1

PROBLEM STAEMENT-

Task 1

Create a database named 'custom'.

Create a table named temperature_data inside custom having below fields:

1. date (mm-dd-yyyy) format

2. zip code

3. temperature

The table will be loaded from comma-delimited file.

Load the dataset.txt (which is ',' delimited) in the table.

SOLUTION-

```
CREATE DATABASE custom;
CREATE TABLE temperature_data
(
full_date STRING,
zip INT,
temperature INT
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';
LOAD DATA LOCAL INPATH '/home/acadgild/hadoop/temp.txt'
INTO TABLE custom.temperature_data;
```

OUTPUT-

```
hive> select * from temperature_data;
OK
10-01-1990      123112  10
14-02-1991      283901  11
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902  9
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
10-01-1993      123112  11
14-02-1994      283901  12
10-03-1993      381920  16
10-01-1994      302918  23
12-02-1991      384902  10
10-01-1991      123112  11
14-02-1990      283901  12
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 9.448 seconds, Fetched: 20 row(s)
hive>
```

Task 2

1. Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.
2. Calculate maximum temperature corresponding to every year from temperature_data table.
3. Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.
4. Create a view on the top of last query, name it temperature_data_vw.
5. Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

SOLUTION –

1. select * from temperature_data where zip BETWEEN 300000 AND 399999;

```
hive> select * from temperature_data where zip BETWEEN 300000 AND 399999;
OK
10-03-1990      381920  15
10-01-1991      302918  22
12-02-1990      384902  9
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
10-03-1993      381920  16
10-01-1994      302918  23
12-02-1991      384902  10
10-03-1991      381920  16
10-01-1990      302918  23
12-02-1991      384902  10
Time taken: 2.257 seconds, Fetched: 12 row(s)
hive>
```

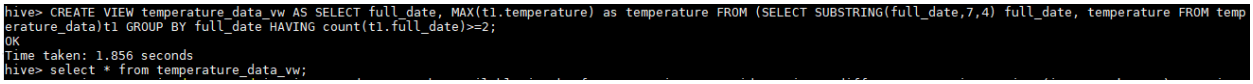
2. SELECT SUBSTRING(full_date,7,4), MAX(temperature) FROM temperature_data GROUP BY SUBSTRING(full_date,7,4);

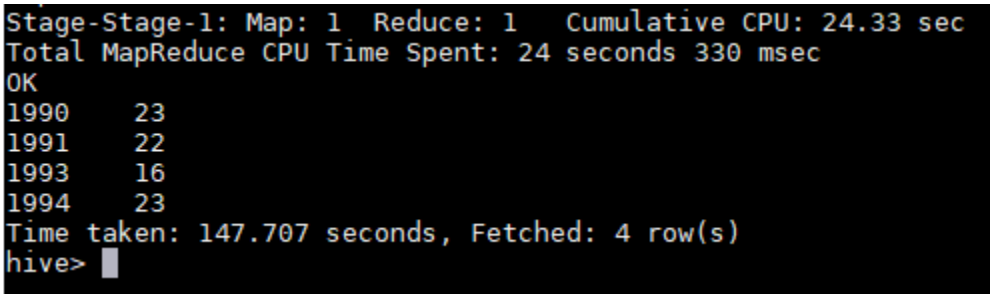
```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 19.74 sec
Total MapReduce CPU Time Spent: 19 seconds 740 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 138.164 seconds, Fetched: 4 row(s)
hive>
```

3. SELECT full_date, MAX(t1.temperature) as temperature FROM (SELECT SUBSTRING(full_date,7,4) full_date, temperature FROM temperature_data)t1 GROUP BY full_date HAVING count(t1.full_date)>=2;

```
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 24.58 sec
Total MapReduce CPU Time Spent: 24 seconds 580 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 107.45 seconds, Fetched: 4 row(s)
hive>
```

4. CREATE VIEW temperature_data_vw AS SELECT full_date, MAX(t1.temperature) as temperature FROM (SELECT SUBSTRING(full_date,7,4) full_date, temperature FROM temperature_data)t1 GROUP BY full_date HAVING count(t1.full_date)>=2;





5. INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hadoop/temperature_data_vw.txt' ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' SELECT * FROM temperature_data_vw;