Session 8: HIVE BASICS

Assignment 1

Task 1:

a) **Create a database named 'custom'.**

Solution:

```
hive>
    > create database custom;
OK
Time taken: 0.754 seconds
hive> show databases
    > ;
OK
custom
default
Time taken: 0.316 seconds, Fetched: 2 row(s)
hive>
```

b) **Create a table named temperature_data inside custom having below fields:**
   1. **date (mm-dd-yyyy) format**
   2. **zip code**
   3. **temperature**

Solution:

```
hive> create table custom.temperature_data(recordeddate String,zipcode int,temp
int);
OK
Time taken: 1.034 seconds
```

```
hive> use custom;
OK
Time taken: 0.056 seconds
hive> show tables;
OK
temperature_data
Time taken: 0.087 seconds, Fetched: 1 row(s)
hive>
```

c) **The table will be loaded from comma-delimited file.**
Load the dataset.txt (which is ',' delimited) in the table.

Solution:

```
hive> create table custom.temperature_data_temp(recordeddate String, zipcode int
, temp int)
    > row format delimited
    > fields terminated by ','
    > lines terminated by '\n'
    > stored as textfile;
OK
Time tload data local inpath '/home/acadgild/Downloads/dataset_Session 14.txt' o
verwrite into table custom.temperature_data_temp;
Loading data to table custom.temperature_data_temp
OK
Time taken: 0.949 seconds
hive> select * from custom.temperature_data_temp;
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
```

```
OK
Time taken: 0.225 seconds
hive> CREATE TABLE Temperature_data AS SELECT CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(
RecordedDate,'dd-MM-yyyy'),'MM-dd-yyyy') AS STRING) AS RecordedDate ,ZipCode,tem
p FROM temperature_data_temp;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = acadgild_20181226235410_fa508b73-a09d-45d8-bb22-395196eec6cb
Total jobs = 3
```

Output:

Column in MM-dd-yyyy format as asked

```
hive> SELECT * FROM temperature_Data;
OK
temperature_data.recordeddate   temperature_data.zipcode        temperature_data
.temp
01-10-1990      123112  10
02-14-1991      283901  11
03-10-1990      381920  15
01-10-1991      302918  22
02-12-1990      384902  9
01-10-1991      123112  11
02-14-1990      283901  12
03-10-1991      381920  16
01-10-1990      302918  23
02-12-1991      384902  10
01-10-1993      123112  11
02-14-1994      283901  12
03-10-1993      381920  16
01-10-1994      302918  23
02-12-1991      384902  10
01-10-1991      123112  11
02-14-1990      283901  12
03-10-1991      381920  16
01-10-1990      302918  23
02-12-1991      384902  10
Time taken: 0.206 seconds, Fetched: 20 row(s)
```

Task 2:

(a)Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.

Solution/Output:

```
hive> SELECT recordeddate,temp  FROM temperature_Data WHERE zipcode >300000 and zipco
de<399999;
OK
recordeddate    temp
03-10-1990      15
01-10-1991      22
02-12-1990      9
03-10-1991      16
01-10-1990      23
02-12-1991      10
03-10-1993      16
01-10-1994      23
02-12-1991      10
03-10-1991      16
01-10-1990      23
02-12-1991      10
Time taken: 0.289 seconds, Fetched: 12 row(s)
```

(b) Calculate maximum temperature corresponding to every year from temperature_data table.

Solution:

```
hive> select MAX(temp),CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(recordeddate,'MM-dd-yyyy'),'
yyyy')AS String) AS yeardata FROM temperature_data GROUP BY CAST(FROM_UNIXTIME(UNIX_T
IMESTAMP(recordeddate,'MM-dd-yyyy'),'yyyy')AS String);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future ve
```

Output:

```
Total MapReduce CPU Time Spent: 4 seconds 620 ms
OK
_c0     yeardata
23      1990
22      1991
16      1993
23      1994
Time taken: 30.081 seconds, Fetched: 4 row(s)
hive>
```

(c) Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

Solution:

```
hive> SELECT MaxTemp,yeardata FROM (select MAX(temp)as MaxTemp, COUNT(*) AS cnt ,CAST
(FROM_UNIXTIME(UNIX_TIMESTAMP(recordeddate,'MM-dd-yyyy'),'yyyy')AS String) AS yeardat
a FROM temperature_data GROUP BY CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(recordeddate,'MM-d
d-yyyy'),'yyyy')AS String)) T1 WHERE cnt>=2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future ve
```

Output:

```
Total MapReduce CPU Time Spent: 5 seconds 230 mse
OK
maxtemp yeardata
23      1990
22      1991
16      1993
23      1994
Time taken: 31.234 seconds, Fetched: 4 row(s)
```

(d) Create a view on the top of last query, name it temperature_data_vw.

Solution:

```
Time taken: 31.234 seconds, Fetched: 4 row(s)
hive> CREATE VIEW temperature_data_vw AS SELECT MaxTemp,yeardata FROM (select MAX(tem
p)as MaxTemp, COUNT(*) AS cnt ,CAST(FROM_UNIXTIME(UNIX_TIMESTAMP(recordeddate,'MM-dd-
yyyy'),'yyyy')AS String) AS yeardata FROM temperature_data GROUP BY CAST(FROM_UNIXTIM
E(UNIX_TIMESTAMP(recordeddate,'MM-dd-yyyy'),'yyyy')AS String)) T1 WHERE cnt>=2;
OK
```

Output:

```
rite: 167 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 440 msec
OK
temperature_data_vw.maxtemp     temperature_data_vw.yeardata
23      1990
22      1991
16      1993
23      1994
Time taken: 39.754 seconds, Fetched: 4 row(s)
```

(e) Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

Solution:

Output: