**Problem Objective**

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.

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.

Code/Script

1. Fetch date and temperature from temperature\_data where zip code is greater than 300000 and less than 399999.

-- create a database custom

create database custom;

-- entering to database custom

use custom;

-- create table temperature\_data

create table temperature\_data

(date string,zipcode int,temperature float)

row format delimited

fields terminated by ',';

--create temp table for conversion date to expected format (MM-dd-yyyy)

create table temp

(date string,zipcode int,temperature float)

row format delimited

fields terminated by ',';

-- load data into temperature\_data

load data local inpath '/home/acadgild/Downloads/dataset\_Session\_14.txt' into table temp;

-- move data from temp to temperature\_data

insert into table temperature\_data select to\_date(from\_unixtime(unix\_timestamp(date,'dd-MM-yyyy'),'yyyy-MM-dd')),zipcode,temperature from temp;

-- select date and temperature b/w 300000 to 399999

select date,temperature,zipcode from temperature\_data where zipcode between 300000 and 399999;

Output

hive> select date,temperature,zipcode from temperature\_data where zipcode between 300000 and 399999;

OK

1990-03-10 15.0 381920

1991-01-10 22.0 302918

1990-02-12 9.0 384902

1991-03-10 16.0 381920

1990-01-10 23.0 302918

1991-02-12 10.0 384902

1993-03-10 16.0 381920

1994-01-10 23.0 302918

1991-02-12 10.0 384902

1991-03-10 16.0 381920

1990-01-10 23.0 302918

1991-02-12 10.0 384902

Time taken: 0.613 seconds, Fetched: 12 row(s)

Code/Script

1. Calculate maximum temperature corresponding to every year from temperature\_data table.

-- find out maximum temperature by year

select year(date),max(temperature) from temperature\_data group by year(date);

Output

1990 23.0

1991 22.0

1993 16.0

1994 23.0

Time taken: 499.925 seconds, Fetched: 4 row(s)

Code/Script  
Calculate maximum temperature from temperature\_data table corresponding to those years which have at least 2 entries in the table.

-- find out maximum temperature by year at least 2 entries

select year(date),max(temperature) from temperature\_data group by year(date) having count(1) > 1 ;

Output

1990 23.0

1991 22.0

1993 16.0

1994 23.0

Code/Script

Create a view on the top of last query, name it temperature\_data\_vw.

-- create view

create view temperature\_data\_vw as select year(date),max(temperature) from temperature\_data group by year(date) having count(1) > 1 ;

Output

hive> create view temperature\_data\_view as select year(date),max(temperature) from temperature\_data group by year(date) having count(1) > 1 ;

OK

Time taken: 2.243 seconds

hive> select \* from temperature\_data\_view;

Query ID = acadgild\_20170401233737\_6dbab3aa-1afd-4b10-a9fe-2d3610d79c99

Total jobs = 1

Launching Job 1 out of 1

Number of reduce tasks not specified. Estimated from input data size: 1

In order to change the average load for a reducer (in bytes):

set hive.exec.reducers.bytes.per.reducer=<number>

In order to limit the maximum number of reducers:

set hive.exec.reducers.max=<number>

In order to set a constant number of reducers:

set mapreduce.job.reduces=<number>

Starting Job = job\_1491017400247\_0016, Tracking URL = http://localhost:8088/proxy/application\_1491017400247\_0016/

Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job\_1491017400247\_0016

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2017-04-01 23:39:10,763 Stage-1 map = 0%, reduce = 0%

2017-04-01 23:40:11,072 Stage-1 map = 0%, reduce = 0%

2017-04-01 23:41:11,248 Stage-1 map = 0%, reduce = 0%

2017-04-01 23:41:37,247 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 15.82 sec

2017-04-01 23:42:37,754 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 15.82 sec

2017-04-01 23:43:27,252 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 20.25 sec

2017-04-01 23:43:51,438 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 31.62 sec

MapReduce Total cumulative CPU time: 31 seconds 620 msec

Ended Job = job\_1491017400247\_0016

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 33.05 sec HDFS Read: 690 HDFS Write: 40 SUCCESS

Total MapReduce CPU Time Spent: 33 seconds 50 msec

OK

1990 23.0

1991 22.0

1993 16.0

1994 23.0

Time taken: 407.474 seconds, Fetched: 4 row(s)

Code/Script

Export contents from temperature\_data\_vw to a file in local file system, such that each file is '|' delimited.

-- write view to local file sytem with fields terminated by '|'

insert overwrite local directory '/home/acadgild/Downloads/temp\_output' row format delimited

fields terminated by '|' select \* from temperature\_data\_view;

Output

1990|23.0

1991|22.0

1993|16.0

1994|23.0