Assignment 2.8

HIVE

Creating temporary table

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS temp_table (the_date String,zipcode int,temperature int) ROW FORMAT DELIMIT
ED FIELDS TERMINATED BY ',';
OK
Time taken: 0.69 seconds
```

```
hive> LOAD data local inpath 'dataset.txt' overwrite into table temp_table;
Loading data to table custom.temp_table
OK
Time taken: 1.624 seconds
hive> select * from temp_table;
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: 2.607 seconds, Fetched: 20 row(s)
hive>
```

Creating the actual table

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS temparature_data (the_date timestamp ,zipcode int,temperature int);
OK
Time taken: 0.089 seconds
```

Inserting data after converting to correct date format

```
hive> INSERT OVERWRITE TABLE temparature_data SELECT from_unixtime(unix_timestamp(the_date,'dd-MM-yyyy')),zipcode,tem
perature from temp_table;
```

```
hive> select * from temparature_data;
OK
1990-01-10 00:00:00     123112  10
1991-02-14 00:00:00     283901  11
1990-03-10 00:00:00     381920  15
1991-01-10 00:00:00     302918  22
1990-02-12 00:00:00     384902  9
1991-01-10 00:00:00     123112  11
1990-02-14 00:00:00     283901  12
1991-03-10 00:00:00     381920  16
1990-01-10 00:00:00     302918  23
1991-02-12 00:00:00     384902  10
1993-01-10 00:00:00     123112  11
1994-02-14 00:00:00     283901  12
1993-03-10 00:00:00     381920  16
1994-01-10 00:00:00     302918  23
1991-02-12 00:00:00     384902  10
1991-01-10 00:00:00     123112  11
1990-02-14 00:00:00     283901  12
1991-03-10 00:00:00     381920  16
1990-01-10 00:00:00     302918  23
1991-02-12 00:00:00     384902  10
Time taken: 0.172 seconds, Fetched: 20 row(s)
```

Problem Statement

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

SELECT date_format(the_date,"MM-dd-YYYY"), temperature from temparature_data WHERE zipcode > 300000 AND zipcode < 399999;

```
hive> SELECT date_format(the_date,"MM-dd-YYYY"), temperature from temparature_data WHERE zipcode > 300000 AND zipcode
 < 399999;
OK
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: 0.237 seconds, Fetched: 12 row(s)
hive>
```

2. **Calculate maximum temperature corresponding to every year from temperature_data table.**

SELECT year(the_date) as the_year, MAX(temperature) from temparature_data GROUP BY year(the_date);

```
hive> SELECT year(the_date) as the_year, MAX(temperature) from temparature_data GROUP BY year(the_date);
Total MapReduce CPU Time Spent: 4 seconds 210 msec
OK
1990    23
1991    22
1993    16
1994    23
```

3. **Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.**

SELECT year(the_date) as the_year,MAX(temperature) as max_temp FROM temparature_data GROUP BY year(the_date) HAVING COUNT(year(the_date)) >= 2;

```
hive> SELECT year(the_date) as the_year,MAX(temperature) as max_temp FROM temparature_data GROUP BY year(the_date) HA
VING COUNT(year(the_date)) >= 2;
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 30.114 seconds, Fetched: 4 row(s)
```

4. **Create a view on the top of last query, name it temperature_data_vw.**

CREATE VIEW temperature_data_vw AS SELECT year(the_date) as the_year,MAX(temperature) as max_temp FROM temparature_data GROUP BY year(the_date) HAVING COUNT(year(the_date)) >= 2;

```
hive> CREATE VIEW temperature_data_vw AS SELECT year(the_date) as the_year,MAX(temperature) as max_temp FROM temparat
ure_data GROUP BY year(the_date) HAVING COUNT(year(the_date)) >= 2;
OK
Time taken: 0.255 seconds
```

```
hive> select * FROM temperature_data_vw;
OK
1990    23
1991    22
1993    16
1994    23
```

5. **Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.**

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

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/temperature_data_vw.txt' ROW FORMAT DELIMITED FIELDS TERMINATE
D BY '|' SELECT * FROM temperature_data_vw;
```

```
[acadgild@localhost ~]$ cat temperature_data_vw.txt/000000_0
1990|23
1991|22
1993|16
1994|23
```