Hive

Q1) Fetch date and temperature from temperature_data where zip code is greater than

300000 and less than 399999

We are going to display only the date and temperature data, from the temperature_data items for which the zip code is greater than 30000 and less than 399999. The query for this is given below.

Query:

SELECT full_date,temperature from temperature_data where zip>300000 and zip<399999;

Result:

10-03-1990	15
10-01-1991	22
12-02-1990	9
10-03-1991	16
10-01-1990	23
12-02-1991	10
10-03-1993	16
10-01-1994	23
12-02-1991	10
10-03-1991	16
10-01-1990	23
12-02-1991	10

Screenshot of Mobaxterm for Q1:

```
hive> SELECT full_date,temperature from temperature_data where zip>300000 and zip<399999;
OK
10-03-1990      15
10-01-1991      22
12-02-1990      9
10-03-1991      16
10-01-1990      23
12-02-1991      10
10-03-1993      16
10-01-1994      23
12-02-1991      10
10-03-1991      16
10-01-1990      23
12-02-1991      10
Time taken: 0.287 seconds, Fetched: 12 row(s)
```

Q2) Calculate maximum temperature corresponding to every year from temperature_data

table.

We are having the date in DD-MM-YYYY format, but we need to select and diaplay YYYY and maximum temperature recorded for each YYYY. The query for this is given below.

Query:

SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBSTR(full_date,7,4);

Result:

1990 23
1991 22
1993 16
1994 23

Screenshot of Mobaxterm for Q2:

```
hive> SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBSTR(full_date,7,4);
Query ID = acadgild_20171029211212_dd17c8c6-0055-4ecd-816a-011ea17971a7
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_1509280519342_0007, Tracking URL = http://localhost:8088/proxy/application_1509280519342_0007/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509280519342_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-29 21:13:08,683 Stage-1 map = 0%,  reduce = 0%
2017-10-29 21:13:18,857 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.9 sec
2017-10-29 21:13:27,842 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 3.52 sec
MapReduce Total cumulative CPU time: 3 seconds 520 msec
Ended Job = job_1509280519342_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 3.52 sec   HDFS Read: 676 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 520 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 33.973 seconds, Fetched: 4 row(s)
```

Q3) Calculate maximum temperature from temperature_data table corresponding to those

years which have at least 2 entries in the table..

We are drilling down further from the previous query and we want to display the YYYY and max temperature for the records that has minimum 2 entries for the same year, YYYY in temperature_data table. The query for this is given below.

Query:

SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBSTR(full_date,7,4) HAVING COUNT(SUBSTR(full_date,7,4))>2;

Result:

1990 23
1991 22

Screenshot of Mobaxterm for Q3:

```
hive> SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBSTR(full_date,7,4) HAVING COUNT(SUBST
R(full_date,7,4))>2;
Query ID = acadgild_20171029182727_a773f838-03ab-488f-8822-73266561d628
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_1509280519342_0005, Tracking URL = http://localhost:8088/proxy/application_1509280519342_0005/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509280519342_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-29 18:27:55,494 Stage-1 map = 0%,  reduce = 0%
2017-10-29 18:28:03,455 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.84 sec
2017-10-29 18:28:13,048 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.22 sec
MapReduce Total cumulative CPU time: 4 seconds 220 msec
Ended Job = job_1509280519342_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.22 sec   HDFS Read: 676 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 220 msec
OK
1990    23
1991    22
Time taken: 28.998 seconds, Fetched: 2 row(s)
```

Q4) Create a view on the top of last query, name it temperature_data_vw.

We are creating a view for the last query The query to create a view is given below.

Query:

CREATE VIEW temperature_data_vw AS SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBSTR(full_date,7,4) HAVING COUNT(SUBSTR(full_date,7,4))>2;

Screenshot of Mobaxterm for Q4:

```
hive> CREATE VIEW temperature_data_vw AS SELECT SUBSTR(full_date,7,4) AS Year,MAX(temperature) from temperature_data GROUP BY SUBST
R(full_date,7,4) HAVING COUNT(SUBSTR(full_date,7,4))>2;
OK
Time taken: 0.175 seconds
```

Q5) Export contents from temperature_data_vw to a file in local file system, such that each

file is '|' delimited.

We are storing the above view inside a local file. The directory path to store the view is given in the below query.

Query:

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

Result (contents of temp_export):

```
1990|23
1991|22
```

Screenshot of Mobaxterm for Q5

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hive/temp_export'
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY '|'
    > SELECT * FROM temperature_data_vw;
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_1509280519342_0006, Tracking URL = http://localhost:8088/proxy/application_1509280519342_0006/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job  -kill job_1509280519342_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-10-29 18:30:42,135 Stage-1 map = 0%,  reduce = 0%
2017-10-29 18:30:50,723 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.86 sec
2017-10-29 18:30:59,453 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.15 sec
MapReduce Total cumulative CPU time: 4 seconds 150 msec
Ended Job = job_1509280519342_0006
Copying data to local directory /home/acadgild/hive/temp_export
Copying data to local directory /home/acadgild/hive/temp_export
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.15 sec   HDFS Read: 676 HDFS Write: 16 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 150 msec
OK
Time taken: 28.974 seconds
```

Contents of temp_export from Mobaxterm for Q5