Basic Hive Assignment :

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:
1.Create database custom;
2.use custom;

```
hive> show databases;
OK
default
Time taken: 0.019 seconds, Fetched: 1 row(s)
hive> create database custom;
OK
Time taken: 2.292 seconds
hive> show databases;
OK
custom
default
Time taken: 0.022 seconds, Fetched: 2 row(s)
hive>
```

```
hive> create table temperature_data(
    > temp_date String,
    > zipcode int,
    > temp int) ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LINES TERMINATED BY '\n'
    > STORED AS TEXTFILE;
OK
Time taken: 0.184 seconds
hive>
```

3. create table temperature_data(
> temp_date String,
> zipcode int,
> temp int) ROW FORMAT DELIMITED
> FIELDS TERMINATED BY ','
> LINES TERMINATED BY '\n'
> STORED AS TEXTFILE;

4. load data local inpath '/home/admin/Documents/hive/dataset_hive.txt' overwrite into table temperature_data;

```
hive> load data local inpath '/home/admin/Documents/hive/dataset_hive.txt' overwrite into table temperature_data;
Loading data to table custom.temperature_data
Table custom.temperature_data stats: [numFiles=1, numRows=0, totalSize=437, rawDataSize=0]
OK
Time taken: 0.805 seconds
hive>
```

5. select from_unixtime(unix_timestamp(temp_date,'dd-MM-yyyy'),'MM-dd-yyyy') , zipcode , temp from temperature_data;

```
hive> select from_unixtime(unix_timestamp(temp_date,'dd-MM-yyyy'),'MM-dd-yyyy') , zipcode , temp  from temperature_data;
OK
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: 0.058 seconds, Fetched: 20 row(s)
```

Task 2

2.1

Problem Statement:

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

Solution:

hive> select zipcode , temp from temperature_data where zipcode > 300000 and zipcode < 399999;

```
hive> select  zipcode , temp  from temperature_data where zipcode > 300000 and zipcode < 399999;
OK
381920  15
302918  22
384902  9
381920  16
302918  23
384902  10
381920  16
302918  23
384902  10
381920  16
302918  23
384902  10
Time taken: 0.232 seconds, Fetched: 12 row(s)
```

2.2

Calculate maximum temperature corresponding to every year from temperature_data table

Solution:

create view temperature_data_vw as select from_unixtime(unix_timestamp(temp_date,'dd-MM-yyyy'),'yyyy') as year, max(temp) as tmp from temperature_data group by temp_date;

select year, max(tmp) from temperature_data_vw group by year ;

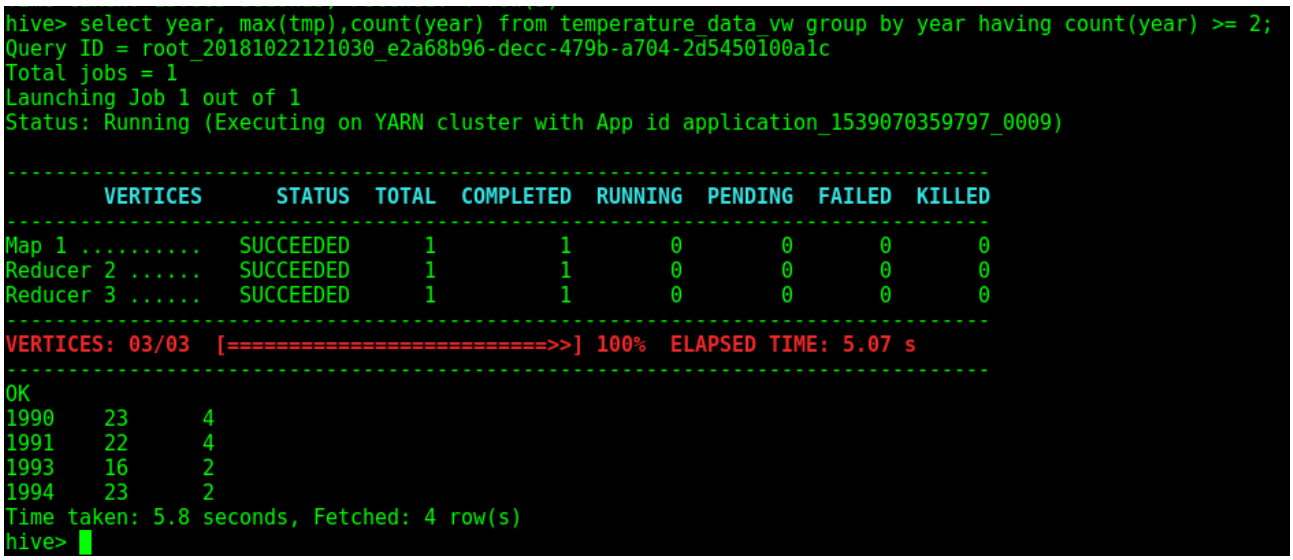
```
hive> select year, max(tmp) from temperature_data_vw group by year ;
Query ID = root_20181022115329_980eeb2d-195b-4ab4-80ca-62e067b7e940
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1539070359797_0008)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
Reducer 3 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 6.51 s
--------------------------------------------------------------------------------
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 7.179 seconds, Fetched: 4 row(s)
```

2.3

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

Solution:

select year, max(tmp),count(year) from temperature_data_vw group by year having count(year) >= 2;



2.4

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

Solution:

create view temperature_data_vw as select from_unixtime(unix_timestamp(temp_date,'dd-MM-yyyy'),'yyyy') as year, max(temp) as tmp from temperature_data group by temp_date;

2.5

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

Solution:

hive> insert overwrite local directory '/home/admin/Documents/hive/' row format delimited fields terminated by '|' stored as TextFile select * from temperature_data_vw;

```
hive> insert overwrite local directory '/home/admin/Documents/hive/' row format delimited fields terminated by '|' stored as TextFile select * from
mperature_data_vw;
Query ID = root_20181022123025_7712ecb5-5e22-4bbf-b42a-4b54c4cde256
Total jobs = 1
Launching Job 1 out of 1
Status: Running (Executing on YARN cluster with App id application_1539070359797_0011)

--------------------------------------------------------------------------------
        VERTICES      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
--------------------------------------------------------------------------------
Map 1 ..........   SUCCEEDED      1          1        0        0       0       0
Reducer 2 ......   SUCCEEDED      1          1        0        0       0       0
--------------------------------------------------------------------------------
VERTICES: 02/02  [==========================>>] 100%  ELAPSED TIME: 5.70 s
--------------------------------------------------------------------------------
Moving data to local directory /home/admin/Documents/hive
OK
Time taken: 8.581 seconds
hive> exit
    > ;
[root@hdpmaster hive]# ll
total 4
-rw-r--r-- 1 root root 95 Oct 22 12:30 000000_0
[root@hdpmaster hive]# cat 000000_0
1990|23
1991|22
1993|11
1994|23
1990|15
1991|16
1993|16
1990|9
1991|10
1990|12
```