Assignment Day 8

Input Dataset:

https://drive.google.com/file/d/0Bxr27gVaXO5sa0JBamZXdkpYUFk/view?usp=sharing

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.

Ans:

create database if not exists custom**;**
(Creates a database named 'custom' if it is not existing already)

show databases;
(Lists the available databases onto console)

```
CREATE TABLE temperature_data
(
  full_date STRING,
  zip INT,
  temperature INT
)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
```

Explanation:
(Creates a table temperature_data with fields full_date, zip, temperature**).**

LOAD DATA LOCAL INPATH
'/home/acadgild/Desktop/TestHadoop/hive/**temperature_dataset.csv' INTO TABLE** temperature_data**;**
Explanation:
(Loads data from input file on local into specified table name**)**
ScreenShot:

```
hive> create database custom;
OK
Time taken: 16.113 seconds
hive> show databases;
OK
custom
default
simplidb
test
Time taken: 0.442 seconds, Fetched: 4 row(s)
hive> use custom;
OK
Time taken: 0.067 seconds
hive> show tables;
OK
Time taken: 0.222 seconds
hive> create database if not exists custom;
OK
Time taken: 0.026 seconds
```

```
hive> CREATE TABLE temperature_data
    > (
    > full_date  STRING,
    > zip INT,
    > temperature INT
    > )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 1.638 seconds
hive> show tables;
OK
temperature_data
Time taken: 0.081 seconds, Fetched: 1 row(s)
hive> desc temperature_data;
OK
full_date               string
zip                     int
temperature             int
Time taken: 0.43 seconds, Fetched: 3 row(s)
hive> LOAD DATA LOCAL INPATH '/home/acadgild/Desktop/TestHadoop/hive/temperature_dataset.csv' INTO TABLE temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 2.501 seconds
```

```
hive> select * from temperature_data limit 5;
OK
temperature_data.full_date      temperature_data.zip    temperature_data.temperature
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
Time taken: 0.433 seconds, Fetched: 5 row(s)
hive>
```

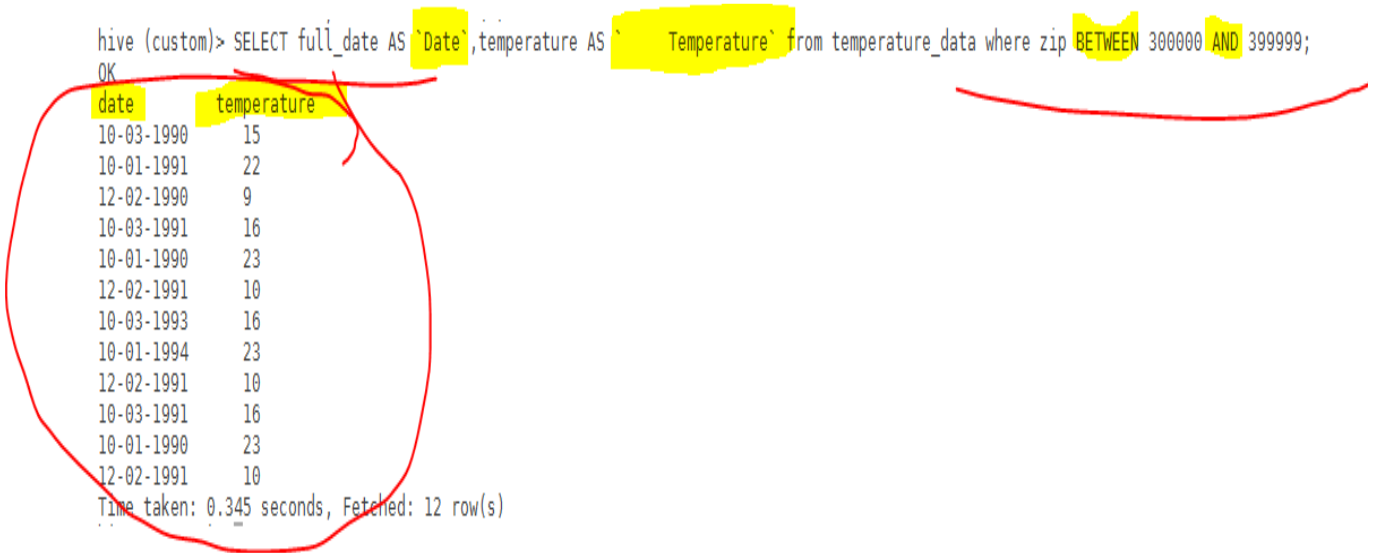
Task 2:
a) Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.
b) Calculate maximum temperature corresponding to every year from temperature_data table.
c) Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.
d) Create a view on the top of last query, name it temperature_data_vw.
e) Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

Ans(a):

SELECT full_date **AS** `Date`,temperature **AS** ` Temperature` **from temperature_data where** zip **BETWEEN** 300000 **AND** 399999**;**

Explanation:
(
SELECT - to select columns from table
AS – to provide an alias to output
where - The condition to filter
BETWEEN - to check for lower & upper range for two boundary values
AND – Boolean operator , true only if left & write operands are true.
)

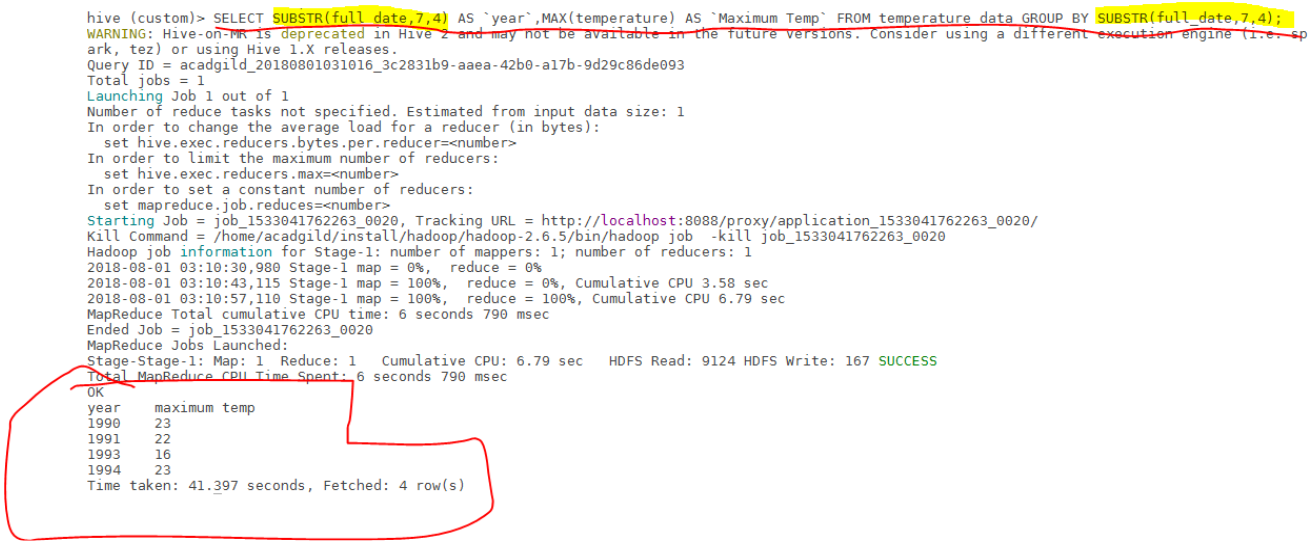
ScreenShot:



Ans(b):
SELECT SUBSTR(full_date,7,4**) AS** `year`**,MAX(**temperature**) AS** `Maximum Temp` **FROM** temperature_data **GROUP BY SUBSTR(**full_date,7,4**);**

Explanation:
(
SUBSTR – Truncating string date & selecting year field starting from 7th index & picking next four indices.
MAX – Selecting Maximum from all the available temperature.
GROUP BY SUBSTR(full_date,7,4**) –** Grouping records based on Year.
)

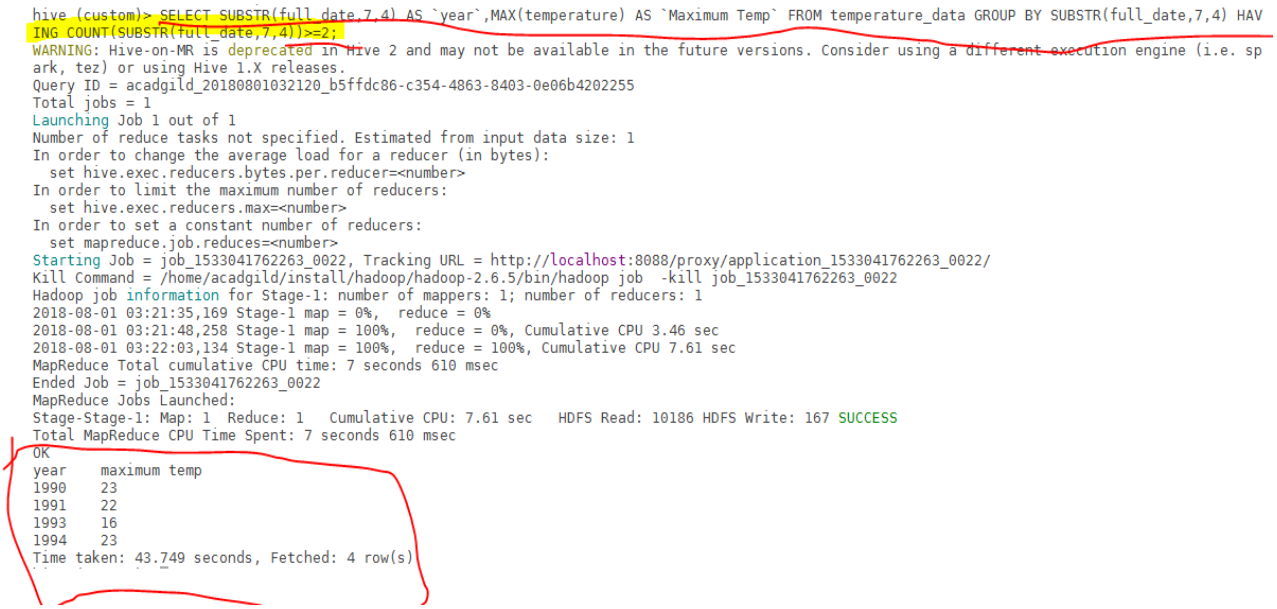
ScreenShot:



Ans(c):

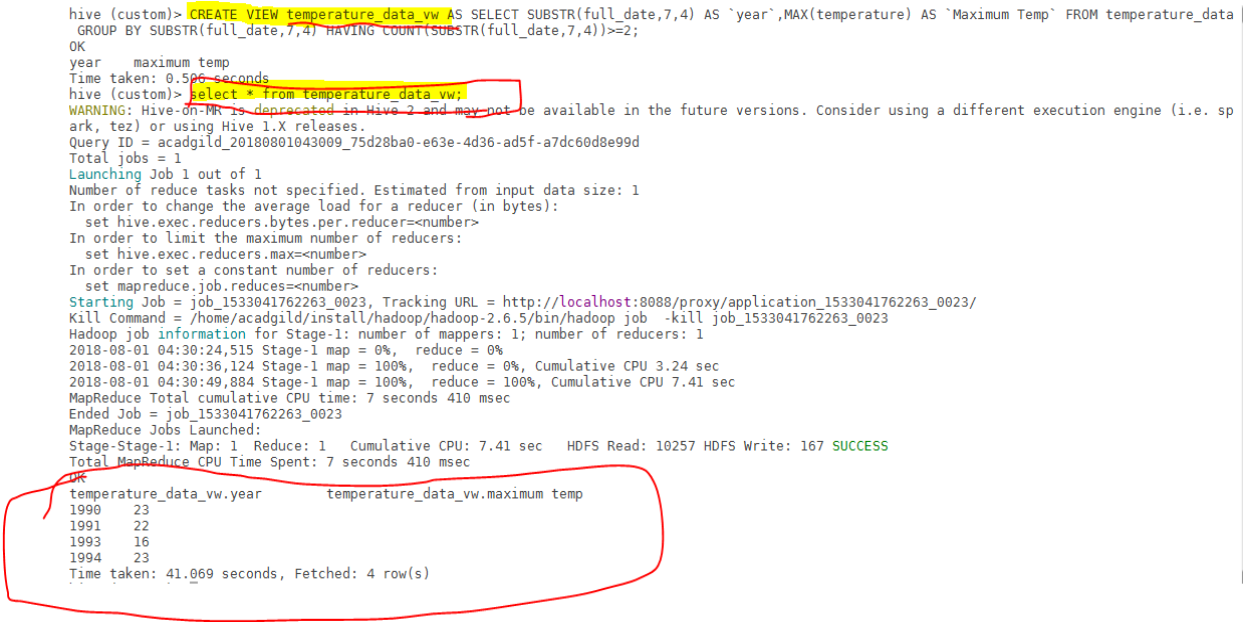
SELECT SUBSTR(full_date,7,4**) AS `year`,MAX(**temperature**) AS** `Maximum Temp` **FROM** temperature_data **GROUP BY SUBSTR(**full_date,7,4**) HAVING COUNT(SUBSTR(**full_date,7,4**))>=2;**

Explanation:
(
HAVING COUNT – Counts the occurance of each year to be atleast two times within the Dataset.
)
ScreenShot:



Ans(d):
CREATE VIEW temperature_data_vw **AS SELECT SUBSTR(**full_date,7,4**) AS** `year`,**MAX(**temperature**) AS** `Maximum Temp` **FROM** temperature_data **GROUP BY SUBSTR(**full_date,7,4**) HAVING COUNT(SUBSTR(**full_date,7,4))>=2;
Explanation:
Creating View as temperature_data_vw for the query in task 2d.

ScreenShot:



```
hive (custom)> CREATE VIEW temperature_data_vw AS SELECT SUBSTR(full_date,7,4) AS `year`,MAX(temperature) AS `Maximum Temp` FROM temperature_data
 GROUP BY SUBSTR(full_date,7,4) HAVING COUNT(SUBSTR(full_date,7,4))>=2;
OK
year    maximum temp
Time taken: 0.526 seconds
hive (custom)> select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. sp
ark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180801043009_75d28ba0-e63e-4d36-ad5f-a7dc60d8e99d
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_1533041762263_0023, Tracking URL = http://localhost:8088/proxy/application_1533041762263_0023/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533041762263_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-01 04:30:24,515 Stage-1 map = 0%,  reduce = 0%
2018-08-01 04:30:36,124 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.24 sec
2018-08-01 04:30:49,884 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.41 sec
MapReduce Total cumulative CPU time: 7 seconds 410 msec
Ended Job = job_1533041762263_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.41 sec   HDFS Read: 10257 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 410 msec
OK
temperature_data_vw.year        temperature_data_vw.maximum temp
1990    23
1991    22
1993    16
1994    23
Time taken: 41.069 seconds, Fetched: 4 row(s)
```

Ans(e):

INSERT OVERWRITE LOCAL DIRECTORY
'/home/acadgild/Desktop/TestHadoop/hive/temperature_data_vw'
ROW FORMAT DELIMITED
FIELDS TERMINATED BY '|'
select * from temperature_data_vw**;**
Explanation:
(Exporting file into local file system at path /home/acadgild/Desktop/TestHadoop/hive/temperature_data_vw
ROW FORMAT DELIMITED – It means there is some delimiter inside every line while table creation & Every line is considered to be as a record.
FIELDS TERMINATED BY '|' – Each field is separated by a pipe '|'.
select * from temperature_data_vw**; -** Selecting the entire data from the newly created view to be exported into local file system.
)

ScreenShot:

```
hive (custom)> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/Desktop/TestHadoop/hive/temperature_data_vw'
             > ROW FORMAT DELIMITED
             > FIELDS TERMINATED BY '|'
             > select * from temperature_data_vw;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. sp
ark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180801045432_9bbae416-4c41-46ca-9274-bff335733831
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_1533041762263_0025, Tracking URL = http://localhost:8088/proxy/application_1533041762263_0025/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533041762263_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-01 04:54:48,407 Stage-1 map = 0%,  reduce = 0%
2018-08-01 04:55:00,101 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.14 sec
2018-08-01 04:55:15,491 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.97 sec
MapReduce Total cumulative CPU time: 7 seconds 970 msec
Ended Job = job_1533041762263_0025
Moving data to local directory /home/acadgild/Desktop/TestHadoop/hive/temperature_data_vw
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.97 sec   HDFS Read: 9933 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 970 msec
OK
temperature_data_vw.year        temperature_data_vw.maximum temp
Time taken: 44.028 seconds
```

********** End **********