Assignment 7 - HIVE

TASK 1:
a) To create a database named 'custom'
-> With the fields:
a) date
b) zip code
c) temperature
-> Table should be loaded from comma-delimited file.
-> Load dataset.txt(which is ',' delimited) in the table.

COMMAND: CREATE DATABASE CUSTOM;
EXPLANATION: Open the console use command 'hive', this will open your hive platform. Then type the above command which will create a database inside your hive.

SOLUTION:
```
hive> create database custom;
OK
Time taken: 0.119 seconds
```

COMMAND: USE CUSTOM;
EXPLANATION: This command will make u use the created data base "custom".

SOLUTION:
```
hive> use custom;
OK
Time taken: 0.139 seconds
```

COMMAND: CREATE TABLE TEMPERATURE_DATA(DT STRING, ZIPCODE INT, TEMP INT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ',';

EXPLANATION: The above command will create a table called temperature_data with the schema for the dataset with date, zipcode, temperature with fields seperated by 'comma' .

SOLUTION:
```
hive> create table temperature_data(dt string, zipcode int, temp int)
   > row format delimited
   > fields terminated by ',';
OK
Time taken: 1.781 seconds
```

COMMAND: load data local inpath 'file:///home/acadgild/dataset.txt' into table temperature_data

EXPLANATION: This command will load the 'dataset.txt' file into the created table 'temperature_data' from the local directory.

SOLUTION:
```
hive> load data local inpath 'file:///home/acadgild/dataset.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 1.621 seconds
```

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.

A) COMMAND: SELECT DT, TEMP FROM TEMPERATURE_DATA WHERE ZIPCODE BETWEEN 300000 AND 399999;
EXPLANATION: This command while select the date and temperature for the zipcode between 300000 and 399999 from the table temperature_data with the loaded file 'dataset.txt' and display the output in the console.

SOLUTION:

```
Time taken: 1.621 seconds
hive> select dt, temp from temperature_data where zipcode between 300000 and 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
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: 7.837 seconds, Fetched: 24 row(s)
```

B) COMMAND: SELECT DT, MAX(TEMP) FROM TEMPERATURE_DATA GROUP BY DT;
EXPLANATION: This command will select the maximum temperature corresponding to every year from the table temperature_data and display the output.

SOLUTION:

```
hive> select dt, max(temp) from temperature_data group by dt;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181213113744_2bd7da2f-f6a3-4095-b672-a2087936fac1
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_1544641458887_0024, Tracking URL =
http://localhost:8088/proxy/application_1544641458887_0024/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544641458887_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-13 11:38:19,567 Stage-1 map = 0%,  reduce = 0%
2018-12-13 11:38:35,723 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.79 sec
2018-12-13 11:38:53,743 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.66 sec
MapReduce Total cumulative CPU time: 6 seconds 660 msec
Ended Job = job_1544641458887_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.66 sec   HDFS Read: 9247 HDFS Write: 414 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 660 msec
OK
 NULL
10-01-1990 23
10-01-1991 22
10-01-1993 11
10-01-1994 23
10-03-1990 15
10-03-1991 16
10-03-1993 16
12-02-1990 9
12-02-1991 10
14-02-1990 12
14-02-1991 11
14-02-1994 12
Time taken: 70.196 seconds, Fetched: 13 row(s)
```

C) COMMAND: SELECT DT, MAX(TEMP) FROM TEMPERATURE_DATA GROUP BY DT HAVING COUNT(DT)>2;

EXPLANATION: This command will select the maximum temperature for the corresponding years which have atleast 2 entries in the dataset and display the output.

SOLUTION:

```
hive> select dt, max(temp) from temperature_data group by dt having count(dt)>2;
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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181213114147_58857c71-ca79-4fff-89fb-a62e6605dca9
```

```
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_1544641458887_0025, Tracking URL =
http://localhost:8088/proxy/application_1544641458887_0025/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544641458887_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-13 11:42:06,622 Stage-1 map = 0%,  reduce = 0%
2018-12-13 11:42:21,094 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.11 sec
2018-12-13 11:42:40,612 Stage-1 map = 100%,  reduce = 68%, Cumulative CPU 8.3 sec
2018-12-13 11:42:41,974 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.05 sec
MapReduce Total cumulative CPU time: 9 seconds 50 msec
Ended Job = job_1544641458887_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.05 sec   HDFS Read: 10405 HDFS Write: 217 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 50 msec
OK
10-01-1990 23
10-01-1991 22
10-03-1991 16
12-02-1991 10
14-02-1990 12
Time taken: 56.686 seconds, Fetched: 5 row(s)
```

D) COMMAND: CREATE VIEW TEMPERATURE_DATA_VW AS
SELECT DT, MAX(TEMP) FROM TEMPERATURE_DATA GROUP BY DT HAVING COUNT(DT)>2;

EXPLANATION: This command will create a view naming temperature_data_vw on top of the table temperature_data with entries of maximum temperature corresponding to years having atleast two entries.

SOLUTION:

```
hive> create view temperature_data_vw as
    > select dt, max(temp) from temperature_data group by dt having count(dt)>2;
OK
Time taken: 0.628 seconds
```

E) EXPLANATION: To export the above view created into file in local directory. First we need to create a directory in the local directory using the below command in the console by opening another terminal.

COMMAND: MKDIR /HOME/ACADGILD/EXPORT_FILE_HIVE

SOLUTION:

```
[acadgild@localhost ~]$ mkdir /home/acadgild/export_file_hive;
You have new mail in /var/spool/mail/acadgild
```

COMMAND: LS

EXPLANATION: Using the command 'ls' you can see the directory 'export_file_hive' is created.

SOLUTION:

```
[acadgild@localhost ~]$ ls
dataset.txt  eclipse           export_hive_file  Music          Templates
Desktop      eclipse-workspace  file.txt.save     Pictures       test.txt
Documents    Exported_Jar       install           Public         Videos
Downloads    export_file_hive   jar file          television.txt
```

EXPLANATION: After creating the directory in local by using the below command the table inside the view temperature_data_vw with fields delimited by '|' will be exported from hive to local directory.

COMMAND:

```
INSERT OVERWRITE LOCAL DIRECTORY '/HOME/ACADGILD/EXPORT_FILE_HIVE'
    ROW FORMAT DELIMITED
    FIELDS TERMINATED BY '|'
        SELECT * FROM TEMPERATURE_DATA_VW;
```

SOLUTION:

```
hive> insert overwrite local directory '/home/acadgild/export_file_hive'
   > 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. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20181213120649_2462a88e-1be4-4fae-a094-6e1f75415719
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_1544641458887_0027, Tracking URL =
http://localhost:8088/proxy/application_1544641458887_0027/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544641458887_0027
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-13 12:07:09,475 Stage-1 map = 0%,  reduce = 0%
2018-12-13 12:07:24,526 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.95 sec
2018-12-13 12:07:44,955 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 7.67 sec
2018-12-13 12:07:46,084 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.45 sec
MapReduce Total cumulative CPU time: 8 seconds 450 msec
Ended Job = job_1544641458887_0027
Moving data to local directory /home/acadgild/export_file_hive
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.45 sec   HDFS Read: 10103 HDFS Write: 70 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 450 msec
OK
Time taken: 57.642 seconds
```

EXPLANATION: WE CAN SEE THIS EXPORTED FILE BY GOING INTO THE DIRECTORY

"/home/acadgild/export_file_hive'