Assignment 8.1
Hive Basics

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.

We have used **'create database custom'** command to create custom database.
After this, **'show databases'** command shows **custom** database in database list.
We have to use **'use custom'** command to create any table or any object in custom database.

```
hive> show databases;
OK
default
Time taken: 13.545 seconds, Fetched: 1 row(s)
hive> create database custom;
OK
Time taken: 0.393 seconds
hive> show databases;
OK
custom
default
Time taken: 0.056 seconds, Fetched: 2 row(s)
hive> use custom;
OK
```

Our present local path is **'/home/acadgild'** and we could see that **'dataset.txt'** file is present in this location.

```
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ ls -l
total 144
-rw-rw-r--.  1 acadgild acadgild 28065 Jul 29 05:11 Assignment5_task1.jar
-rw-rw-r--.  1 acadgild acadgild  8374 Jul 29 00:13 Assignment5_task2.jar
-rw-rw-r--.  1 acadgild acadgild  8689 Jul 29 00:32 Assignment5_task3.jar
-rw-rw-r--.  1 acadgild acadgild   437 Aug  1 08:49 dataset.txt
drwxr-xr-x.  3 acadgild acadgild  4096 Feb  2 12:51 Desktop
drwxr-xr-x.  2 acadgild acadgild  4096 Feb  2 12:52 Documents
drwxr-xr-x.  2 acadgild acadgild  4096 Feb 13 14:24 Downloads
drwxrwxr-x.  3 acadgild acadgild  4096 Dec 29  2017 eclipse
drwxrwxr-x.  3 acadgild acadgild  4096 Jan 16  2018 eclipse-workspace
-rw-rw-r--.  1 acadgild acadgild 10824 Jul 29 12:30 employee.java
-rw-rw-r--.  1 acadgild acadgild  2978 Jul 29 15:24 Hbase Commands.txt
drwxrwxr-x. 13 acadgild acadgild  4096 Feb  9 18:06 install
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Music
-rw-rw-r--.  1 acadgild acadgild    72 Jul 29 00:02 musicdata.txt
-rw-rw-r--.  1 acadgild acadgild 16507 Jul 29 12:57 Person.java
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Pictures
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Public
-rw-rw-r--.  1 acadgild acadgild  3677 Jul 29 11:57 Sqoop_Commands_AG.txt
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Templates
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Videos
```

As we could see that dataset.txt is having date field in '**dd-mm-yyyy'** format. But we need to have date field in **'mm-dd-yyyy'** format in **temperature_data** table.

So we have created a temporary table first and load data from dataset.txt file into this temporary table. Then we have inserted data into **'temperature_data'** table from this temporary table using insert into select statement.

Assignment 8.1
Hive Basics

```
[acadgild@localhost ~]$ cat dataset.txt;
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[acadgild@localhost ~]$
```

temporary table created :

```
hive> create table temporary
    > (temp_date string,
    >  zip_code int,
    >  temperature int)
    >  row format delimited fields terminated by ',';
OK
Time taken: 2.162 seconds
hive> select * from temporary;
OK
```

We have loaded data from dataset.txt into temporary table:

```
hive> load data local inpath '/home/acadgild/dataset.txt' into table temporary;
Loading data to table custom.temporary
OK
Time taken: 3.805 seconds
hive> select * from temporary;
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: 0.625 seconds, Fetched: 20 row(s)
```

Here we have created **temperate_data** table :

```
hive> create table temperature_data
    > (temp_date string,
    > zip_code int,
    > temperature int)
    > row format delimited fields terminated by ',';
OK
Time taken: 1.595 seconds
hive> select * from temperature_data;
OK
Time taken: 5.655 seconds
```

Then we have inserted data into **'temperature_data'** table from this **temporary** table using below insert into select statement with the help of from_unixtime and unix_timestamp functions.

Assignment 8.1
Hive Basics

```
hive> insert into table temperature_data select from_unixtime(unix_timestamp(temp_date,'dd-mm-yyyy'),'mm-dd-yyyy'),zip_code,temperature
from temporary;
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_20180803143352_b47b29ea-f326-4bc6-9459-47ed35e5d737
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1533273394630_0021, Tracking URL = http://localhost:8088/proxy/application_1533273394630_0021/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533273394630_0021
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-08-03 14:34:22,102 Stage-1 map = 0%,  reduce = 0%
2018-08-03 14:34:42,145 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.78 sec
MapReduce Total cumulative CPU time: 4 seconds 780 msec
Ended Job = job_1533273394630_0021
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
Moving data to directory hdfs://localhost:8020/user/hive/warehouse/custom.db/temperature_data/.hive-staging_hive_2018-08-03_14-33-52_146_
4659454401720967057-1/-ext-10000
Loading data to table custom.temperature_data
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 4.78 sec   HDFS Read: 4872 HDFS Write: 499 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 780 msec
OK
Time taken: 56.232 seconds
```

We could see all 20 records in temperature_data table with date format as ‘mm-dd-yyyy’.

```
hive> select * 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: 0.495 seconds, Fetched: 20 row(s)
```

Task 2

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

Here, we set column header to TRUE so that we can have column headers along with output.

```
hive> set hive.cli.print.header=true;
```

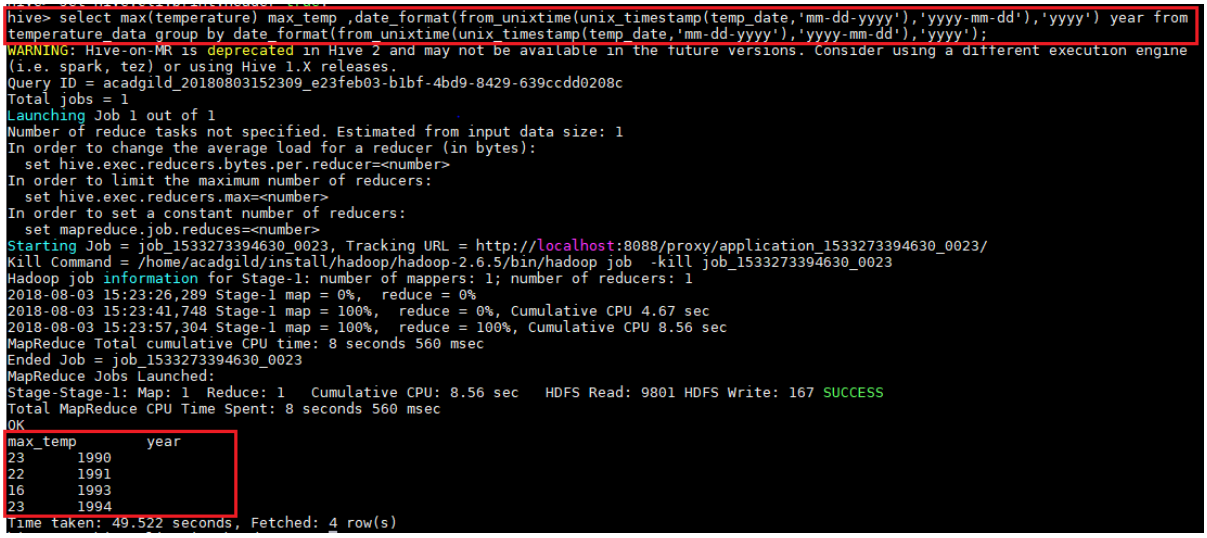
Then we have used below select query :

```
hive> select temp_date,temperature from temperature_data where zip_code >300000 and zip_code < 399999;
OK
temp_date       temperature
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
```

Assignment 8.1
Hive Basics

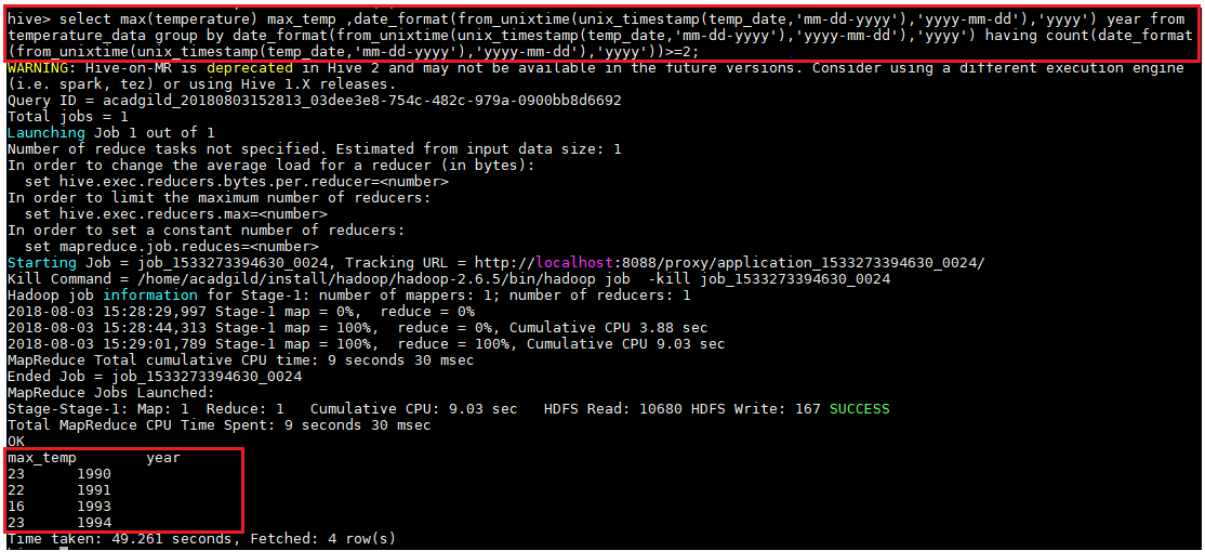
- **Calculate maximum temperature corresponding to every year from temperature_data table.**

  We have used below select query by using max_temp and year as column alias for table :
  Output shows Maximum temperature corresponding to every year.



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

  We have used below select query by using max_temp and year as column alias and count function for each year for table :
  Output shows Maximum temperature corresponding to every year having count of rows for each year as at least 2.



Assignment 8.1
Hive Basics

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

We have used below create statement to create view and you could see data in this view :

```
hive> create view temperature_data_vw as select max(temperature) max_temp ,date_format(from_unixtime(unix_timestamp(temp_date,'mm-dd-yyyy
'),'yyyy-mm-dd'),'yyyy') year from temperature_data group by date_format(from_unixtime(unix_timestamp(temp_date,'mm-dd-yyyy'),'yyyy-mm-dd
'),'yyyy') having count(date_format(from_unixtime(unix_timestamp(temp_date,'mm-dd-yyyy'),'yyyy-mm-dd'),'yyyy'))>=2;
OK
max_temp        year
Time taken: 0.613 seconds
hive> 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_20180803153207_721b6cef-0825-4e1d-b1eb-046a803004f3
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_1533273394630_0025, Tracking URL = http://localhost:8088/proxy/application_1533273394630_0025/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533273394630_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-03 15:32:23,625 Stage-1 map = 0%,  reduce = 0%
2018-08-03 15:32:37,935 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.46 sec
2018-08-03 15:32:54,422 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.32 sec
MapReduce Total cumulative CPU time: 9 seconds 320 msec
Ended Job = job_1533273394630_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.32 sec   HDFS Read: 10750 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 320 msec
OK
temperature_data_vw.max_temp    temperature_data_vw.year
23      1990
22      1991
16      1993
23      1994
```

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

  We have used below insert statement to insert data into **export** directory with fields separated by '|'.

```
hive> insert overwrite local directory '/home/acadgild/export' row format delimited fields terminated by '|'  select * from temperature_d
ata_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_20180803154555_17d46827-e0d7-4e87-820e-2a3dc2b4cab4
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_1533273394630_0026, Tracking URL = http://localhost:8088/proxy/application_1533273394630_0026/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1533273394630_0026
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-08-03 15:46:12,125 Stage-1 map = 0%,  reduce = 0%
2018-08-03 15:46:27,677 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.17 sec
2018-08-03 15:46:44,071 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 9.39 sec
MapReduce Total cumulative CPU time: 9 seconds 390 msec
Ended Job = job_1533273394630_0026
Moving data to local directory /home/acadgild/export
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 9.39 sec   HDFS Read: 10352 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 390 msec
OK
```

Assignment 8.1
Hive Basics

Below you can see that file ‘000000_0’ has been generated into export directory .
Content of file ‘000000_0’ shows the output with field separated by ‘|’

```
[acadgild@localhost ~]$ ls -l
total 148
-rw-rw-r--.  1 acadgild acadgild 28065 Jul 29 05:11 Assignment5_task1.jar
-rw-rw-r--.  1 acadgild acadgild  8374 Jul 29 00:13 Assignment5_task2.jar
-rw-rw-r--.  1 acadgild acadgild  8689 Jul 29 00:32 Assignment5_task3.jar
-rw-rw-r--.  1 acadgild acadgild   437 Aug  1 08:49 dataset.txt
drwxr-xr-x.  3 acadgild acadgild  4096 Feb  2 12:51 Desktop
drwxr-xr-x.  2 acadgild acadgild  4096 Feb  2 12:52 Documents
drwxr-xr-x.  2 acadgild acadgild  4096 Feb 13 14:24 Downloads
drwxrwxr-x.  3 acadgild acadgild  4096 Dec 29  2017 eclipse
drwxrwxr-x.  3 acadgild acadgild  4096 Jan 16  2018 eclipse-workspace
-rw-rw-r--.  1 acadgild acadgild 10824 Jul 29 12:30 employee.java
drwxrwxr-x.  2 acadgild acadgild  4096 Aug  3 15:46 export
-rw-rw-r--.  1 acadgild acadgild  2978 Jul 29 15:24 Hbase Commands.txt
drwxrwxr-x. 13 acadgild acadgild  4096 Feb  9 18:06 install
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Music
-rw-rw-r--.  1 acadgild acadgild    72 Jul 29 00:02 musicdata.txt
-rw-rw-r--.  1 acadgild acadgild 16507 Jul 29 12:57 Person.java
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Pictures
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Public
-rw-rw-r--.  1 acadgild acadgild  3677 Jul 29 11:57 Sqoop_Commands_AG.txt
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Templates
drwxr-xr-x.  2 acadgild acadgild  4096 Dec 27  2017 Videos
```

```
[acadgild@localhost ~]$ ls -l  export
total 4
-rw-r--r--. 1 acadgild acadgild 32 Aug  3 15:46 000000_0
[acadgild@localhost ~]$ cat  export/000000_0
23|1990
22|1991
16|1993
23|1994
```