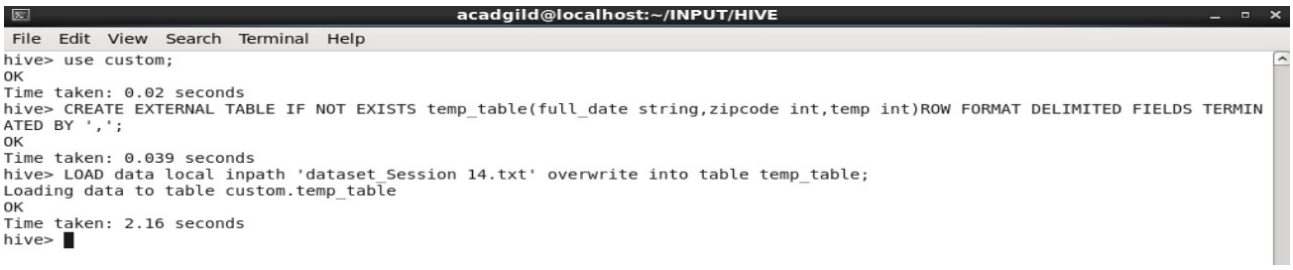
Bigdata Assignment 2.8

- entered into hive shell by using command -

start-all.sh
hive

- Then in the 'custom' database , we created a external temporary table named temp_table and loaded the data into the tablefrom the text file loacted in the local file system.

Command used -
use custom;
CREATE EXTERNAL TABLE IF NOT EXISTS temp_table(full_date string,zipcode , temp int) ROW FORMAT DELIMITED FIELDS TERMIBATED BY ',' ;
LOAD data local inpath 'dataset_Session 14.txt' overwrite into table temp_table;



- Then we checked the contents of the table as it can be ssen int he below screenshot the date format is dd-MM-yyyy.

select * from temp_table;

```
hive> select * from temp_table;
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: 3.518 seconds, Fetched: 20 row(s)
hive>
```

- We created actual table to store the correct date format data .

CREATE EXTERNAL TABLE IF NOT EXISTS temperature_data(full_date timestamp, zipcode int, temp int) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' ;

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS temperature_data(full_date timestamp,zipcode int,temp int)ROW FORMAT DELIMITED FIEL
DS TERMINATED BY ',';
OK
Time taken: 0.067 seconds
hive>
```

- Inserted the data into new table 'temperature_data' from the previous table temp_table using unix_timestamp function.

INSERT OVERWRITE TABLE temperature_data SELECT from_unixtime(unix_timestamp(full_date,'dd-MM-yyyy')), zipcode , temp from temp_table;

```
hive> INSERT OVERWRITE TABLE temperature_data SELECT from_unixtime(unix_timestamp(full_date,'dd-MM-yyyy')),zipcode,temp from
temp_table;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180605094655_6e5c9c3e-cf52-4255-8131-8a7ec6dffece
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_1528171677641_0002, Tracking URL = http://localhost:8088/proxy/application_1528171677641_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1528171677641_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2018-06-05 09:47:04,217 Stage-1 map = 0%,  reduce = 0%
2018-06-05 09:47:12,866 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.26 sec
MapReduce Total cumulative CPU time: 3 seconds 260 msec
Ended Job = job_1528171677641_0002
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_2018-06-05_0
9-46-55_771_4445495260754980132-1/-ext-10000
Loading data to table custom.temperature_data
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 3.26 sec   HDFS Read: 5086 HDFS Write: 679 SUCCESS
Total MapReduce CPU Time Spent: 3 seconds 260 msec
OK
Time taken: 18.845 seconds
```

- After inserting the data , we checked the content of the table

select * from temperature_data;

```
acadgild@loca
File  Edit  View  Search  Terminal  Help
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: 0.2 seconds, Fetched: 20 row(s)
hive>
```

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

Ans - **select full_date , temp from tempearture_data where zipcode>300000 and zipcode<399999;**

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
hive> select full_date,temp from temperature_data where zipcode>300000 and zipcode<399999;
OK
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
Time taken: 3.421 seconds, Fetched: 12 row(s)
hive>
```

2. Calculate maximum temperature corresponding to every year from temperature_data table.

Ans – **select year(full_date) , MAX(temp) from temperature_data GROUP BY year(full_date);**

```
acadgild@localhost:~/INPUT/HIVE
File  Edit  View  Search  Terminal  Help
hive> select year(full_date),MAX(temp) from temperature_data GROUP BY year(full_date);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180605094938_b6df2306-c384-4c50-8ab8-1fa151ef8876
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_1528171677641_0003, Tracking URL = http://localhost:8088/proxy/application_1528171677641_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1528171677641_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-05 09:49:46,872 Stage-1 map = 0%,  reduce = 0%
2018-06-05 09:49:54,601 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.19 sec
2018-06-05 09:50:03,346 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.63 sec
MapReduce Total cumulative CPU time: 6 seconds 630 msec
Ended Job = job_1528171677641_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.63 sec   HDFS Read: 9420 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 630 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 27.118 seconds, Fetched: 4 row(s)
hive>
```

3. Calculate maximum temperature from temperature_data table corresponding

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

Ans - **select year(full_date) , MAX(temperature)_data GROUP BY year(full_date) HAVING COUNT(year(full_date))>=2;**

```
acadgild@localhost:~/INPUT/HIVE
File  Edit  View  Search  Terminal  Help
hive> select year(full_date) , MAX(temp) from temperature_data GROUP BY year(full_date) HAVING COUNT(year(full_date))>=2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180605095949_736300f1-1d0d-4069-b84d-bd490493c1ac
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_1528171677641_0004, Tracking URL = http://localhost:8088/proxy/application_1528171677641_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1528171677641_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-05 09:59:58,221 Stage-1 map = 0%,  reduce = 0%
2018-06-05 10:00:05,682 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.0 sec
2018-06-05 10:00:15,459 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.23 sec
MapReduce Total cumulative CPU time: 8 seconds 230 msec
Ended Job = job_1528171677641_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.23 sec   HDFS Read: 10539 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 230 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 27.817 seconds, Fetched: 4 row(s)
hive>
```

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

Ans – **CREATE VIEW tempearture_data_vw as select year(full_date) , MAX(temp) from tempearture_data GROUP BY year(full_date) HAVING COUNT(year(full_date))>=2;**

```
hive> CREATE VIEW temperature_data_vw AS select year(full_date) , MAX(temp) from temperature_data GROUP BY year(full_date) HA
VING COUNT(year(full_date))>=2;
```

```
acadgild@localhost:~/INPUT/HIVE
File  Edit  View  Search  Terminal  Help
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 execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180605100627_c381d463-b8d2-4f98-8100-81ee728102aa
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_1528171677641_0005, Tracking URL = http://localhost:8088/proxy/application_1528171677641_0005/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1528171677641_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-05 10:06:36,040 Stage-1 map = 0%,  reduce = 0%
2018-06-05 10:06:42,478 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.93 sec
2018-06-05 10:06:52,280 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.94 sec
MapReduce Total cumulative CPU time: 7 seconds 940 msec
Ended Job = job_1528171677641_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.94 sec   HDFS Read: 10605 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 940 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 26.661 seconds, Fetched: 4 row(s)
```

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

Ans – **INSERT OVERWRITE LOCAL DIRECTORY 'home/acadgild/INPUT/HIVE/tempearture_data_vw.txt' ROW FORMAT DELIMITED FIELDS TERMINATED BT '|' select * from temperature_data_vw;**

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/INPUT/HIVE/temperature_data_vw.txt' ROW FORMAT DELIMITED FIELDS TERMIN
ATED BY '|' select * from temperature_data_vw;
```

The content of the file is shown in the below screenshot.

```
acadgild@localhost:~/INPUT/HIVE/temperature_data_vw.txt
File  Edit  View  Search  Terminal  Help
[acadgild@localhost HIVE]$ ls
dataset_Session 14.txt  temperature_data_vw.txt
[acadgild@localhost HIVE]$ cd temperature_data_vw.txt/
[acadgild@localhost temperature_data_vw.txt]$ ls
000000_0
[acadgild@localhost temperature_data_vw.txt]$ cat 000000_0
1990|23
1991|22
1993|16
1994|23
[acadgild@localhost temperature_data_vw.txt]$
```