


ASSIGNMENT 8.1

Task 1 : Create a database named 'custom'

 acadgild@localhost:~

```
hive> show databases;  
Sun Feb 11 07:45:37 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:37 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:38 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:38 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:41 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:41 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:41 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
Sun Feb 11 07:45:41 IST 2018 WARN: Establishing SSL connection without server  
requirements SSL connection must be established by default if explicit opti  
icate property is set to 'false'. You need either to explicitly disable SSL  
rification.  
OK  
default  
retail  
Time taken: 7.171 seconds, Fetched: 2 row(s)  
hive> █
```

Command: CREATE database custom;

```
acadgild@localhost:~  
hive> create database custom;  
OK  
Time taken: 0.297 seconds  
hive> show databases;  
OK  
custom  
default  
retail  
Time taken: 0.043 seconds, Fetched: 3 row(s)  
hive> █
```

Create a table named temperature_data inside custom databases.

Command: CREATE TABLE temperature_data
 (date_temp string, zipcode int, temperature int)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
 STORED AS TEXTFILE;

```
acadgild@localhost:~  
hive> use custom;  
OK  
Time taken: 0.024 seconds  
hive> show tables;  
OK  
Time taken: 0.037 seconds  
hive> CREATE TABLE temperature_data  
      > ( date_temp string,  
      > zipcode int,  
      > temperature int)  
      > ROW FORMAT DELIMITED FIELDS TERMINATED BY ','  
      > STORED AS TEXTFILE;  
OK  
Time taken: 0.123 seconds  
hive> show tables;  
OK  
temperature_data  
Time taken: 0.04 seconds, Fetched: 1 row(s)  
hive> █
```

Load the dataset.txt (which is ',' delimited) in the table.

Command: LOAD DATA LOCAL INPATH "/home/acadgild/dataset.txt" INTO TABLE temperature_data;

```
acadgild@localhost:~  
hive> LOAD DATA LOCAL INPATH "/home/acadgild/dataset.txt" INTO TABLE temperature_data;  
Loading data to table custom.temperature_data  
OK  
Time taken: 0.531 seconds  
hive> select * from temperature_data;  
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: 0.191 seconds, Fetched: 20 row(s)  
hive>
```

Task 2 : Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.

Command: select date_temp, temperature from temperature_data
Where zipcode >30000 AND zipcode < 399999;

acadgild@localhost:~

```
hive> select date_temp, temperature from temperature_data  
> where zipcode > 300000 AND zipcode < 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

Time taken: 0.655 seconds, Fetched: 12 row(s)

hive> █

Calculate maximum temperature corresponding to every year from temperature_data table.

Command: SELECT year,MAX(t1.temperature) as temperature
FROM (select SUBSTRING(date_temp,7,4) year,
temperature from temperature_data) t1
GROUP BY year;

```
acadgild@localhost:~
Time taken: 27.138 seconds, Fetched: 4 row(s)
hive>
hive> SELECT year,MAX(t1.temperature) as temperature
> FROM
> (select SUBSTRING(date_temp,7,4) year, temperature from temperature_data) t1
> GROUP BY year;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine in 3.X releases.
Query ID = acadgild_20180212134123_f3c7e98b-c696-409b-8178-8f93bc99113b
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_1518419499395_0006, Tracking URL = http://localhost:8088/proxy/application_1518419499395_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job -kill job_1518419499395_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-02-12 13:41:32,345 Stage-1 map = 0%, reduce = 0%
2018-02-12 13:41:39,620 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.85 sec
2018-02-12 13:41:46,926 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 4.82 sec
MapReduce Total cumulative CPU time: 4 seconds 820 msec
Ended Job = job_1518419499395_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 4.82 sec HDFS Read: 9178 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 820 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 25.572 seconds, Fetched: 4 row(s)
```

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

Command: SELECT year, MAX(t1.temperature) as temperature
FROM (select SUBSTRING(date_temp,7,4) year, temperature from temperature_data) t1
GROUP BY year
HAVING count(t1.year) >= 2;

```
acadgild@localhost:~
hive> select year, MAX(t1.temperature) as temperature
> from
> (select SUBSTRING(date_temp,7,4) year, temperature from temperature_data) t1
> GROUP BY year
> HAVING count(t1.year) >= 2;

WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execut
1.X releases.
Query ID = acadgild_20180212131822_498a7de0-5a70-4eb1-97ed-f1711530d1fc
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_1518419499395_0004, Tracking URL = http://localhost:8088/proxy/application_1518419499395_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.9.0/bin/hadoop job -kill job_1518419499395_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-02-12 13:18:32,339 Stage-1 map = 0%, reduce = 0%
2018-02-12 13:18:39,752 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.17 sec
2018-02-12 13:18:48,060 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.87 sec
MapReduce Total cumulative CPU time: 6 seconds 870 msec
Ended Job = job_1518419499395_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.87 sec HDFS Read: 10197 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 870 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 27.027 seconds, Fetched: 4 row(s)
hive>
```

Output:---

```
MapReduce Total cumulative CPU time: 6 seconds 870 msec
Ended Job = job_1518419499395_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.87 sec HDFS Read: 10197 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 870 msec
OK
1990      23
1991      22
1993      16
1994      23
Time taken: 27.027 seconds, Fetched: 4 row(s)
hive>
```

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

Command: CREATE VIEW temperature_data_vw AS

SELECT year, MAX(t1.temperature) as temperature

FROM (select SUBSTRING(date_temp,7,4) year, temperature from temperature_data) t1

GROUP BY year

HAVING count(t1.year) >= 2;

```
acadgild@localhost:~  
hive> CREATE VIEW temperature_data_vw AS  
> SELECT year, MAX(t1.temperature) as temperature  
> FROM (select SUBSTRING(date_temp,7,4) year, temperature from temperature_data) t1  
> GROUP BY year  
> HAVING count(t1.year) >= 2;  
OK  
Time taken: 0.423 seconds  
hive> show views;  
OK  
temperature_data_vw  
Time taken: 0.03 seconds, Fetched: 1 row(s)  
hive> █
```

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

Command: INSERT OVERWRITE LOCAL DIRECTORY
'/home/acadgild/hive_output'

row format delimited fields terminated by '|'

SELECT * FROM temperature_data_vw;

```
acadgild@localhost:~
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hive_output'
> 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 ex
1.X releases.
Query ID = acadgild_20180212212707_9cf1d1cd-e394-470b-92cf-59dd3ae0ae82
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_1518445375024_0006, Tracking URL = http://localhost:8088/proxy/application_1518445375024_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1518445375024_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-02-12 21:27:16,440 Stage-1 map = 0%, reduce = 0%
2018-02-12 21:27:24,103 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.71 sec
2018-02-12 21:27:32,995 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.53 sec
MapReduce Total cumulative CPU time: 6 seconds 530 msec
Ended Job = job_1518445375024_0006
Moving data to local directory /home/acadgild/hive_output
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.53 sec HDFS Read: 9900 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 530 msec
OK
Time taken: 27.551 seconds
hive>
```

Output:-

```
acadgild@localhost:~
[acadgild@localhost ~]$ ls /home/acadgild/hive_output
000000_0
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$ cat /home/acadgild/hive_output/000000_0
1990|23
1991|22
1993|16
1994|23
[acadgild@localhost ~]$
```