

SESSION-8 – BASIC HIVE ASSIGNMENT

Task 1

Create a database named 'custom'.

Create a table named temperature_data inside custom having below fields:

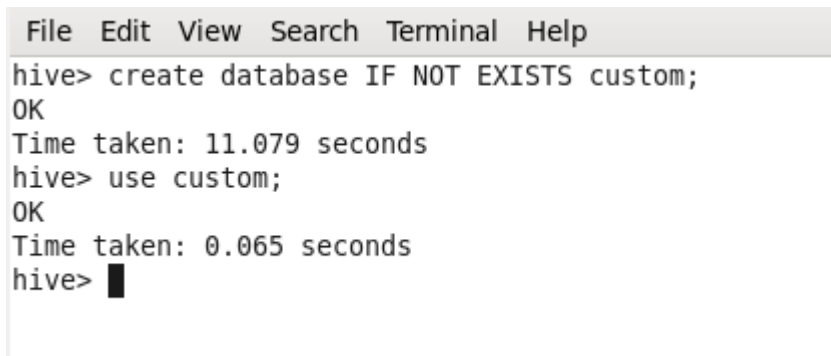
- 1. date (mm-dd-yyyy) format*
- 2. zip code*
- 3. temperature*

The table will be loaded from comma-delimited file.

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

Solution :

First Create a database custom using the command as shown in screenshot and use the custom database to create table inside it as shown below:

A screenshot of a Hive terminal window. The window has a menu bar with 'File', 'Edit', 'View', 'Search', 'Terminal', and 'Help'. The terminal text shows the following commands and outputs:

```
hive> create database IF NOT EXISTS custom;  
OK  
Time taken: 11.079 seconds  
hive> use custom;  
OK  
Time taken: 0.065 seconds  
hive> █
```

Then temperature_data table is created as shown below in screenshot and data from a local file dataset.txt is loaded which has comma separated values in it as shown below :

File Edit View Search Terminal Help

```
[acadgild@localhost ~]$ cat /home/acadgild/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 ~]$
```

File Edit View Search Terminal Help

```
hive> CREATE DATABASE IF NOT EXISTS custom;
OK
Time taken: 0.037 seconds
hive> use custom;
OK
Time taken: 0.063 seconds
hive> CREATE TABLE temperature_data (
    > tempdate STRING,
    > zipcode BIGINT,
    > temperature INT )
    > ROW FORMAT DELIMITED FIELDS TERMINATED BY ',';
OK
Time taken: 0.747 seconds
hive>
```

```

hive> LOAD DATA LOCAL INPATH '/home/acadgild/dataset.txt' INTO TABLE temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 3.862 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: 7.564 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.**

Solution :

The date in table is in 'mm-dd-yyyy' format . To display the date in normal 'dd-mm-yyyy' format, first cast string to date and use the command "from_unixtime(unix_timestamp(column name),'date format 'dd-mm-yyyy') as shown below .

The query used to get the desired output is in the below screenshot :

Select

from_unixtime(unix_timestamp(cast(to_date(from_unixtime(unix_timestamp(temperature,'MM-dd-yyyy'))as date)),'dd-MM-yyyy'), temperature from temperature_data where zipcode > 300000 and zipcode < 399999;

```
File Edit View Search Terminal Help
hive> SELECT from_unixtime(unix timestamp(cast(to_date(from_unixtime(unix timestamp(tempdate,'MM-dd-yyyy'))as date)), 'dd-MM-yyyy'), temperature 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: 0.536 seconds, Fetched: 12 row(s)
hive>
```

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

table.

Solution :

Select max(temperature),substr(tempdate,7,10) from temperature_data group by substr(tempdate,7,10);

```
File Edit View Search Terminal Help
hive> select max(temperature) , substr(tempdate,7,10) from temperature_data group by substr(tempdate,7,10);
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_20180522230656_01d24af4-11e5-4395-898d-05dac654fdb5
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_1527008225391_0001, Tracking URL = http://localhost:8088/proxy/application_1527008225391_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1527008225391_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-22 23:07:52,105 Stage-1 map = 0%, reduce = 0%
2018-05-22 23:08:30,718 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.94 sec
2018-05-22 23:08:53,752 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 7.96 sec
2018-05-22 23:08:55,159 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.84 sec
MapReduce Total cumulative CPU time: 8 seconds 840 msec
Ended Job = job_1527008225391_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.84 sec HDFS Read: 9294 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 840 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 121.661 seconds, Fetched: 4 row(s)
hive>
```

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

Solution :

Select max(temperature) , substr(tempdate,7,10) from temperature_data group by substr(tempdate,,7,10) having count(substr(tempdate,1,4)) >=2;

```
File Edit View Search Terminal Help
hive> select max(temperature) , substr(tempdate,7,10) from temperature_data group by substr(tempdate,7,10) having count(subst
r(tempdate,7,10))>=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_20180522231024_f94210ef-8177-41a8-829f-635605bc578c
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_1527008225391_0002, Tracking URL = http://localhost:8088/proxy/application_1527008225391_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1527008225391_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-22 23:10:49,061 Stage-1 map = 0%, reduce = 0%
2018-05-22 23:11:13,134 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.95 sec
2018-05-22 23:11:36,591 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 8.73 sec
2018-05-22 23:11:37,827 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.17 sec
MapReduce Total cumulative CPU time: 10 seconds 170 msec
Ended Job = job_1527008225391_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.17 sec HDFS Read: 10180 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 170 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 75.767 seconds, Fetched: 4 row(s)
hive> █
```

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

Solution :

Create view temperature_data_vw as select max(temperature) , substr(tempdate,7,10) from temperature_data group by substr(tempdate,,7,10) having count(substr(tempdate,7,10)) >=2;

```

File Edit View Search Terminal Help
hive> create view temperature_data_vw as select max(temperature) , substr(tempdate,7,10) from temperature_data group by substr(tempdate,7,10) having count(substr(tempdate,7,10))>=2;
OK
Time taken: 0.736 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_20180522231425_7d3aeef9-f464-4370-b83d-a888241fdcf7
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_1527008225391_0003, Tracking URL = http://localhost:8088/proxy/application_1527008225391_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1527008225391_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-22 23:14:43,904 Stage-1 map = 0%, reduce = 0%
2018-05-22 23:15:01,296 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.3 sec
2018-05-22 23:15:23,432 Stage-1 map = 100%, reduce = 83%, Cumulative CPU 9.28 sec
2018-05-22 23:15:24,520 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.7 sec
MapReduce Total cumulative CPU time: 9 seconds 700 msec
Ended Job = job_1527008225391_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 9.7 sec HDFS Read: 10258 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 700 msec
OK
23      1990
22      1991
16      1993
23      1994
Time taken: 59.999 seconds, Fetched: 4 row(s)
hive> █

```

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

Solution :

INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/Hiveoutput' ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' SELECT * FROM temperature_data_vw;

File Edit View Search Terminal Help

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/Hiveoutput' 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_20180522231927_688df75e-ca98-4d22-80b9-aa5e6e38e9e5
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_1527008225391_0004, Tracking URL = http://localhost:8088/proxy/application_1527008225391_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1527008225391_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-05-22 23:19:49,503 Stage-1 map = 0%, reduce = 0%
2018-05-22 23:20:11,141 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.86 sec
2018-05-22 23:20:36,866 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 9.73 sec
2018-05-22 23:20:37,914 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.76 sec
MapReduce Total cumulative CPU time: 10 seconds 760 msec
Ended Job = job_1527008225391_0004
Moving data to local directory /home/acadgild/Hiveoutput
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.76 sec HDFS Read: 9868 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 760 msec
OK
Time taken: 72.01 seconds
hive> █
```

File Edit View Search Terminal Help

```
[acadgild@localhost ~]$ ls -ls /home/acadgild/Hiveoutput
total 4
4 -rw-r--r--. 1 acadgild acadgild 32 May 22 23:20 000000_0
[acadgild@localhost ~]$ cat /home/acadgild/Hiveoutput/000000_0
23|1990
22|1991
16|1993
23|1994
[acadgild@localhost ~]$ █
```