Assignment 8 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.

create database custom;

```
hive> create database custom;
OK
Time taken: 20.219 seconds
hive> show databases;
OK
custom
default
simplidb
Time taken: 1.223 seconds, Fetched: 3 row(s)
hive> use custom;
OK
Time taken: 0.217 seconds
hive>
```

CREATE TABLE temperature_data(

DATE_ STRING,
zip_code BIGINT,
temperature INT
)
row format delimited fields terminated by ',';

```
hive> CREATE TABLE temperature_data(
    > DATE_ STRING,
    > zip_code BIGINT,
    > temperature INT
    > )
    > row format delimited fields terminated by ',';
OK
Time taken: 5.111 seconds
hive> SHOW TABLES;
OK
temperature_data
Time taken: 0.181 seconds, Fetched: 1 row(s)
hive> DESC temperature_data
    > ;
OK
date_                   string
zip_code                bigint
temperature             int
Time taken: 1.968 seconds, Fetched: 3 row(s)
hive>
```

LOAD DATA LOCAL INPATH '/home/acadgild/dataset_Session14.txt' into table temperature_data;

```
hive> LOAD DATA LOCAL INPATH '/home/acadgild/dataset_Session14.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 7.56 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: 7.204 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.

Select date_,temperature from temperature_data where zip_code >= 300000 and zip_code < 399999;

```
hive> Select date_,temperature from temperature_data where zip_code >= 300000 and zip_code < 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: 3.903 seconds, Fetched: 12 row(s)
```

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

select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP BY YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY'))));

```
hive> select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP BY YEAR(to_date(from_unixtime(UNIX_
TIMESTAMP(date_, 'DD-MM-YYYY'))));
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 u
sing Hive 1.X releases.
Query ID = acadgild_20181208030250_da94c9c2-a355-469f-ad4c-2fa2150b3982
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_1544146716043_0001, Tracking URL = http://localhost:8088/proxy/application_1544146716043_0001/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544146716043_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-08 03:03:47,341 Stage-1 map = 0%,  reduce = 0%
2018-12-08 03:04:13,564 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.48 sec
2018-12-08 03:04:40,092 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 9.21 sec
2018-12-08 03:04:42,434 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 10.29 sec
MapReduce Total cumulative CPU time: 10 seconds 290 msec
Ended Job = job_1544146716043_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 10.29 sec   HDFS Read: 9589 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 10 seconds 290 msec
OK
1989    23
1990    22
1992    16
1993    23
Time taken: 113.284 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.

select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP BY YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY'))));

```
hive> select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP BY YEAR(to_date(from_unixtime(UNIX_
TIMESTAMP(date_, 'DD-MM-YYYY')))) HAVING count(*) > 1;
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 u
sing Hive 1.X releases.
Query ID = acadgild_20181208030856_433dfcc7-19f2-472a-9424-8e04777e0894
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_1544146716043_0002, Tracking URL = http://localhost:8088/proxy/application_1544146716043_0002/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544146716043_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-08 03:09:22,360 Stage-1 map = 0%,  reduce = 0%
2018-12-08 03:09:42,173 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.48 sec
2018-12-08 03:10:03,045 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 10.14 sec
2018-12-08 03:10:05,666 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.18 sec
MapReduce Total cumulative CPU time: 12 seconds 180 msec
Ended Job = job_1544146716043_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.18 sec   HDFS Read: 10523 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 180 msec
OK
1989    23
1990    22
1992    16
1993    23
Time taken: 71.349 seconds, Fetched: 4 row(s)
hive>
```

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

create view temperature_data_vw as select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP BY YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))) HAVING count(*) > 1;

```
hive> create view temperature_data_vw as select YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))),MAX(temperature) from temperature_data GROUP
BY YEAR(to_date(from_unixtime(UNIX_TIMESTAMP(date_, 'DD-MM-YYYY')))) HAVING count(*) > 1;
OK
Time taken: 1.077 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 u
sing Hive 1.X releases.
Query ID = acadgild_20181208034053_b3ef6707-c50b-4c16-b76e-c86a8bf15469
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_1544146716043_0003, Tracking URL = http://localhost:8088/proxy/application_1544146716043_0003/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544146716043_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-08 03:41:17,940 Stage-1 map = 0%,  reduce = 0%
2018-12-08 03:41:34,905 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.09 sec
2018-12-08 03:41:58,185 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 10.24 sec
2018-12-08 03:41:59,474 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.27 sec
MapReduce Total cumulative CPU time: 11 seconds 270 msec
Ended Job = job_1544146716043_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.27 sec   HDFS Read: 10603 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 270 msec
OK
1989    23
1990    22
1992    16
1993    23
Time taken: 68.576 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.

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

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/hiveassignmet' ROW FORMAT DELIMITED FIELDS TERMINATED BY '|' SELECT * FROM custom.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 u
sing Hive 1.X releases.
Query ID = acadgild_20181208035856_b0a9846b-cdf0-446b-b933-b1c9d0cb8b9f
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_1544146716043_0004, Tracking URL = http://localhost:8088/proxy/application_1544146716043_0004/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1544146716043_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-12-08 03:59:21,618 Stage-1 map = 0%,  reduce = 0%
2018-12-08 03:59:44,757 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 5.79 sec
2018-12-08 04:00:07,496 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 9.74 sec
2018-12-08 04:00:11,178 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.73 sec
MapReduce Total cumulative CPU time: 12 seconds 730 msec
Ended Job = job_1544146716043_0004
Moving data to local directory /home/acadgild/hiveassignmet
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.73 sec   HDFS Read: 10218 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 730 msec
OK
Time taken: 76.435 seconds
hive>
```

```
[acadgild@localhost ~]$ cd hiveassignmet
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost hiveassignmet]$ pwd
/home/acadgild/hiveassignmet
[acadgild@localhost hiveassignmet]$ ll
total 4
-rw-r--r--. 1 acadgild acadgild 32 Dec  8 04:00 000000_0
[acadgild@localhost hiveassignmet]$ cat 000000_0
1989|23
1990|22
1992|16
1993|23
[acadgild@localhost hiveassignmet]$
```