Hive Basics Assignment

Task

Create a database named 'custom'.

```
hive> show databases;
OK
acadgild_db
default
Time taken: 2.743 seconds, Fetched: 2 row(s)
hive>
    >
    >
    >
    >
    >
    >
    >
    >
    > CREATE DATABASE IF NOT EXISTS custom;
OK
Time taken: 1.728 seconds
hive> show databases;
OK
acadgild_db
custom
default
Time taken: 0.193 seconds, Fetched: 3 row(s)
hive>
```

Assign-2

Create a table named temperature_data inside custom having below fields:
1. date (mm-dd-yyyy) format
2. zip code
3. temperature

```
hive> show databases;
OK
acadgild_db
custom
default
Time taken: 0.114 seconds, Fetched: 3 row(s)
hive> use custom;
OK
Time taken: 0.166 seconds
hive> create table if not exists temperature_data(
    > TDate string,
    > Zipcode bigint,
    > temperature int
    > )
    > row format delimited
    > fields terminated by ',';
OK
Time taken: 3.893 seconds
hive> show tables;
OK
temperature_data
Time taken: 0.695 seconds, Fetched: 1 row(s)
hive>
```

Assign -3

The table will be loaded from comma-delimited file.
Load the dataset.txt (which is ',' delimited) in the table.

```
hive> load data local inpath '/home/acadgild/dataset_Session.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 6.858 seconds
hive> select * from temerature_data;
FAILED: SemanticException [Error 10001]: Line 1:14 Table not found 'temerature_data'
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: 7.08 seconds, Fetched: 20 row(s)
hive>
```

Task-2

Assignment-1

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

```
hive> select Tdate,temperature from temperature_data where Zipcode between 300000 and 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: 2.344 seconds, Fetched: 12 row(s)
hive>
```

Assign-2

Calculate maximum temperature corresponding to every year from temperature_data table.

```
hive> select substring(Tdate,7,10) as Year, max(temperature) from temperature_data group by substring(Tdate,7,10);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a differ
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20180601042009_3f40354d-1a13-4757-b4fb-7aca9d3d4ca7
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_1527696602505_0023, Tracking URL = http://localhost:8088/proxy/application_1527696602505_0023/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527696602505_0023
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-01 04:21:16,831 Stage-1 map = 0%,  reduce = 0%
2018-06-01 04:21:37,856 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.17 sec
2018-06-01 04:21:59,309 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 16.79 sec
2018-06-01 04:22:01,712 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.98 sec
MapReduce Total cumulative CPU time: 19 seconds 980 msec
Ended Job = job_1527696602505_0023
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 19.98 sec   HDFS Read: 9135 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 19 seconds 980 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 116.604 seconds, Fetched: 4 row(s)
hive>
```


Assign-3

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

```
hive> select substring(Tdate,7,10) as Year, max(temperature) from temperature_data group by substring(Tdate,7,10) having coun
t(substring(Tdate,7,10)) >1;
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_20180601042852_b6be0131-e6e9-41a8-b1ee-7f8a0a3086fa
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_1527696602505_0024, Tracking URL = http://localhost:8088/proxy/application_1527696602505_0024/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527696602505_0024
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-01 04:29:18,299 Stage-1 map = 0%,  reduce = 0%
2018-06-01 04:29:37,923 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.84 sec
2018-06-01 04:29:58,626 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 14.52 sec
2018-06-01 04:29:59,922 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.93 sec
MapReduce Total cumulative CPU time: 17 seconds 930 msec
Ended Job = job_1527696602505_0024
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.93 sec   HDFS Read: 10166 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 930 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 71.754 seconds, Fetched: 4 row(s)
hive>
```


Assign-4

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

```
hive> create view temperature_vw as select substring(Tdate,7,10) as Year, max(temperature) from temperature_data group by sub
string(Tdate,7,10) having count(substring(Tdate,7,10)) >1;
OK
Time taken: 1.333 seconds
hive> select * from temperature_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_20180601043305_fe2ac764-7945-406e-8135-4b6274f6a381
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_1527696602505_0025, Tracking URL = http://localhost:8088/proxy/application_1527696602505_0025/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527696602505_0025
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-01 04:33:31,070 Stage-1 map = 0%,  reduce = 0%
2018-06-01 04:33:47,729 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.16 sec
2018-06-01 04:34:10,083 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 17.77 sec
2018-06-01 04:34:12,584 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 22.12 sec
MapReduce Total cumulative CPU time: 22 seconds 120 msec
Ended Job = job_1527696602505_0025
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 22.12 sec   HDFS Read: 10220 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 22 seconds 120 msec
OK
1990    23
1991    22
1993    16
1994    23
Time taken: 69.355 seconds, Fetched: 4 row(s)
hive>
```


Assign -5

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

```
hive>
    > insert overwrite local directory '/home/acadgild/'
    > row format delimited
    > fields terminated by '|'
    > select * from temperature_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_20180601043925_6f72349e-d0c5-402a-ad95-c24429128bb7
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_1527696602505_0026, Tracking URL = http://localhost:8088/proxy/application_1527696602505_0026/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job  -kill job_1527696602505_0026
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2018-06-01 04:39:49,169 Stage-1 map = 0%,  reduce = 0%
2018-06-01 04:40:08,635 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.11 sec
2018-06-01 04:40:26,458 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 16.96 sec
2018-06-01 04:40:31,400 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 23.63 sec
MapReduce Total cumulative CPU time: 23 seconds 630 msec
Ended Job = job_1527696602505_0026
Moving data to local directory /home/acadgild
Failed with exception Unable to move source hdfs://localhost:8020/tmp/hive/acadgild/f5934ec5-2e93-47ad-9179-773bc5c283ee/hive
_2018-06-01_04-39-25_694_1483284674026987577-1/-mr-10000 to destination /home/acadgild
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.MoveTask. Unable to move source hdfs://localhost:8
020/tmp/hive/acadgild/f5934ec5-2e93-47ad-9179-773bc5c283ee/hive_2018-06-01_04-39-25_694_1483284674026987577-1/-mr-10000 to de
stination /home/acadgild
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 23.63 sec   HDFS Read: 9776 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 23 seconds 630 msec
hive>
```