

Vijay's Assignment – Hive 1

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.

```
hive> create database custom;
OK
Time taken: 9.879 seconds
hive> use custom;
OK
Time taken: 0.045 seconds

hive> create table temperature_data(date1 string, zipcode string, temperature int) row format delimited fields terminated by
',' stored as textfile;
OK
Time taken: 1.409 seconds
hive> load data local inpath '/home/acadgild/hive1.txt' into table temperature_data;
Loading data to table custom.temperature_data
OK
Time taken: 2.968 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
```

Task 2

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

```
hive> select date1, 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: 1.353 seconds, Fetched: 12 row(s)
```

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

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadmild@localhost:~
File Edit View Search Terminal Help
at org.apache.hadoop.hive.cli.CliDriver.main(CliDriver.java:686)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:498)
at org.apache.hadoop.util.RunJar.main(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:28 cannot recognize input near 'as' 'date' '.' in selection target
hive> select substr(date1,7,4), max(temperature) from temperature_data group by substr(date1,7,4);
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 = acadmild_20190114184054_f6dc622-d664-4727-96c3-e02alc0f8c2d
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_1547486987092_0003, Tracking URL = http://localhost:8088/proxy/application_1547486987092_0003/
Kill Command = /home/acadmild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1547486987092_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-14 18:41:12,086 Stage-1 map = 0%, reduce = 0%
2019-01-14 18:41:25,075 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.34 sec
2019-01-14 18:41:37,670 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 5.84 sec
MapReduce Total cumulative CPU time: 5 seconds 840 msec
Ended Job = job_1547486987092_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 5.84 sec HDFS Read: 9100 HDFS Write: 167 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 840 msec
OK
1990      23
1991      22
1993      16
1994       23
Time taken: 45.748 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.

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadmild@localhost:~
File Edit View Search Terminal Help
Total MapReduce CPU Time Spent: 7 seconds 990 msec
OK
1990      7      23
1991      9      22
1993      2      16
1994      2      23
Time taken: 47.974 seconds, Fetched: 4 row(s)
hive> select substr(date1,7,4), count(*), max(temperature) from temperature_data group by substr(date1,7,4) having count(*) >
= 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 = acadmild_20190114184852_bb88b049-e422-44ca-933e-d714bcfef2cd
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_1547486987092_0005, Tracking URL = http://localhost:8088/proxy/application_1547486987092_0005/
Kill Command = /home/acadmild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1547486987092_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-14 18:49:09,115 Stage-1 map = 0%, reduce = 0%
2019-01-14 18:49:22,986 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.14 sec
2019-01-14 18:49:37,671 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.45 sec
MapReduce Total cumulative CPU time: 7 seconds 450 msec
Ended Job = job_1547486987092_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.45 sec HDFS Read: 9955 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 450 msec
OK
1990      7      23
1991      9      22
1993      2      16
1994      2      23
Time taken: 46.03 seconds, Fetched: 4 row(s)
hive>

```

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

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
1991 9 22
1993 2 16
1994 2 23
Time taken: 46.03 seconds, Fetched: 4 row(s)
hive> create view temperature_data_vw as select substr(date1,7,4), count(*), max(temperature) from temperature_data group by
substr(date1,7,4) having count(*) >=2;
OK
Time taken: 0.567 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 execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20190114185159_dee6dabd-8a79-4c40-a59b-1092515efedb
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_1547486987092_0006, Tracking URL = http://localhost:8088/proxy/application_1547486987092_0006/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1547486987092_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-14 18:52:13,561 Stage-1 map = 0%, reduce = 0%
2019-01-14 18:52:24,964 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.3 sec
2019-01-14 18:52:40,484 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.65 sec
MapReduce Total cumulative CPU time: 7 seconds 650 msec
Ended Job = job_1547486987092_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.65 sec HDFS Read: 9986 HDFS Write: 175 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 650 msec
OK
1990 7 23
1991 9 22
1993 2 16
1994 2 23
Time taken: 42.088 seconds, Fetched: 4 row(s)
hive>

```

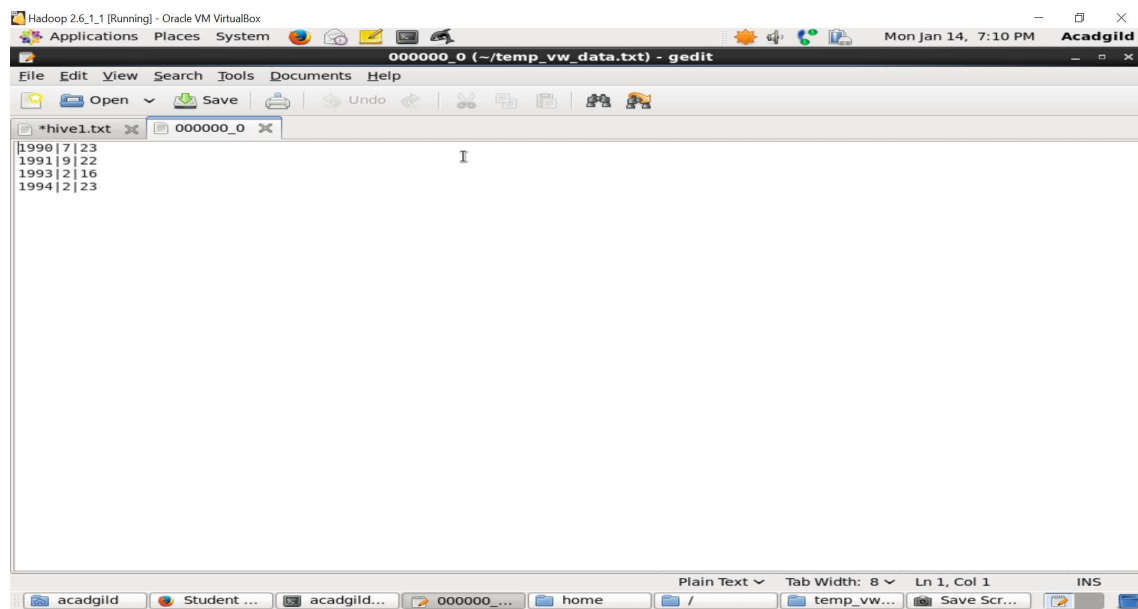
- Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

```

Hadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
Applications Places System
acadgild@localhost:~
File Edit View Search Terminal Help
Moving data to local directory /temp_vw_data.txt
Failed with exception Unable to move source hdfs://localhost:8020/tmp/hive/acadgild/46d5261e-7c03-47ca-b800-5368969d47df/hive
2019-01-14 19:06:17.086.7811121011295101086-1/-mr-10000 to destination /temp_vw_data.txt
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/46d5261e-7c03-47ca-b800-5368969d47df/hive_2019-01-14_19-06-17_086.7811121011295101086-1/-mr-10000 to de
stination /temp_vw_data.txt
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.45 sec HDFS Read: 9580 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 450 msec
hive> insert overwrite local directory '/home/acadgild/temp_vw_data.txt' 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 execu
tion engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = acadgild_20190114190827_ffa0bfca-e5ec-4fa0-8521-a083b2168f15
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_1547486987092_0008, Tracking URL = http://localhost:8088/proxy/application_1547486987092_0008/
Kill Command = /home/acadgild/install/hadoop/hadoop-2.6.5/bin/hadoop job -kill job_1547486987092_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2019-01-14 19:08:41,178 Stage-1 map = 0%, reduce = 0%
2019-01-14 19:08:56,428 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.43 sec
2019-01-14 19:09:13,962 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 9.58 sec
2019-01-14 19:09:16,248 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.11 sec
MapReduce Total cumulative CPU time: 11 seconds 110 msec
Ended Job = job_1547486987092_0008
Moving data to local directory /home/acadgild/temp_vw_data.txt
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.11 sec HDFS Read: 9608 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 110 msec
OK
Time taken: 50.066 seconds
hive>

```

Output



The screenshot shows a gedit text editor window titled "000000_0 (~/temp_vw_data.txt) - gedit". The window has a menu bar with "File", "Edit", "View", "Search", "Tools", "Documents", and "Help". Below the menu bar is a toolbar with icons for "Open", "Save", "Undo", "Redo", "Copy", "Paste", "Print", and "Run". The main text area contains the following text:

```
1990|7|23
1991|9|22
1993|2|16
1994|2|23
```

The status bar at the bottom of the window displays "Plain Text", "Tab Width: 8", "Ln 1, Col 1", and "INS". The taskbar at the very bottom shows several open applications, including "acadgild", "Student ...", "acadgild...", "000000_...", "home", "temp_vw...", and "Save Scr...".