BIG DATA HADOOP & SPARK TRAINING

ACADGILD

ASSIGNMENT 6.2

BY :-

SAHIL KHURANA

PROBLEM STATEMENT

Task 1- Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.

Task 2- Calculate maximum temperature corresponding to every year from temperature_data table.

Task 3- Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

Task 4- Create a view on the top of last query, name it temperature_data_vw.

Task 5- Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

Associated Data Files

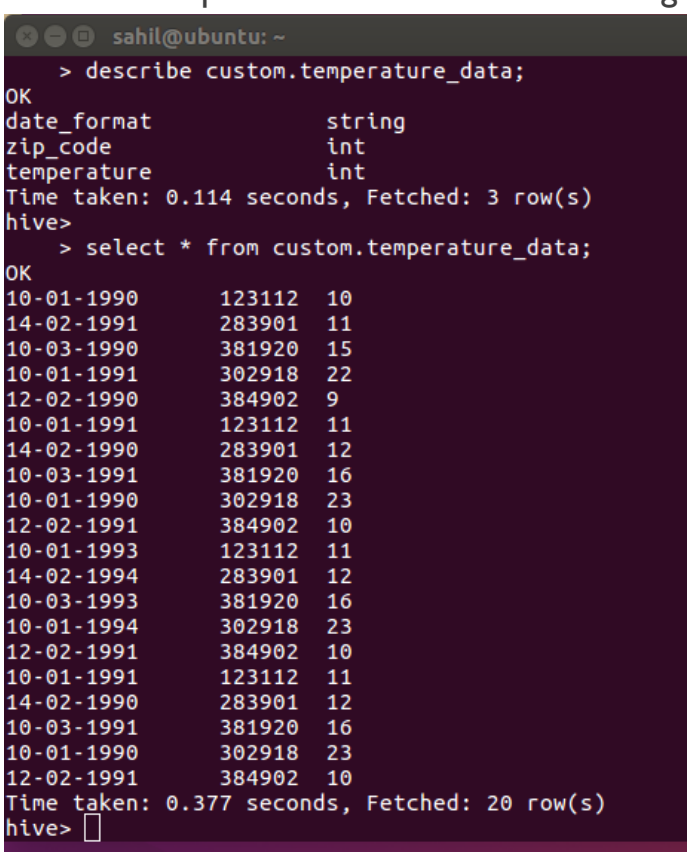
https://drive.google.com/file/d/0Bxr27gVaXO5sa0JBamZXdkpYUFk/view?usp=sharing

dataset_Session_14.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
```

Note:- To solve the Assignment 6.1, I have created a VM with Ubuntu 16.04 OS and configured Hadoop 2.8.2 and hive-2.3.2 on the same.

Dataset is imported in the hive table in Assignment 6.1



```
sahil@ubuntu: ~
    > describe custom.temperature_data;
OK
date_format             string
zip_code                int
temperature             int
Time taken: 0.114 seconds, Fetched: 3 row(s)
hive>
    > select * from custom.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: 0.377 seconds, Fetched: 20 row(s)
hive>
```

Task 1- Fetch date and temperature from temperature_data where zip code is greater than 300000 and less than 399999.

Commands for Task 1;-

I discover that there is a very interesting way to write command in hive.
So, I tried the same in Task 1.

```
hive>
    > from custom.temperature_data select temperature,zip_code,
    > case
    > when zip_code >300000 then 'Zip Code greater than 300000'
    > when zip_code <399999 then 'Zip Code less than 399999'
    > end
    > as ZIP_Code_Bracket;
```

```
sahil@ubuntu: ~
hive>
    > from custom.temperature_data select temperature,zip_code,
    > case
    > when zip_code >300000 then 'Zip Code greater than 300000'
    > when zip_code <399999 then 'Zip Code less than 399999'
    > end
    > as ZIP_Code_Bracket;
OK
10      123112  Zip Code less than 399999
11      283901  Zip Code less than 399999
15      381920  Zip Code greater than 300000
22      302918  Zip Code greater than 300000
9       384902  Zip Code greater than 300000
11      123112  Zip Code less than 399999
12      283901  Zip Code less than 399999
16      381920  Zip Code greater than 300000
23      302918  Zip Code greater than 300000
10      384902  Zip Code greater than 300000
11      123112  Zip Code less than 399999
12      283901  Zip Code less than 399999
16      381920  Zip Code greater than 300000
23      302918  Zip Code greater than 300000
10      384902  Zip Code greater than 300000
11      123112  Zip Code less than 399999
12      283901  Zip Code less than 399999
16      381920  Zip Code greater than 300000
23      302918  Zip Code greater than 300000
10      384902  Zip Code greater than 300000
Time taken: 0.753 seconds, Fetched: 20 row(s)
hive>
    >
```

Task 2- Calculate maximum temperature corresponding to every year from temperature_data table.

Commands for Task 2;-

hive>
> from custom.temperature_data select date_format, MAX(temperature) group by date_format;

```
sahil@ubuntu: ~
hive>
    > from custom.temperature_data select date_format, MAX(temperature) group by date_format;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a d
rk, tez) or using Hive 1.X releases.
Query ID = sahil_20171204105840_8ed2234e-99d3-40a3-a981-bd27146465d6
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_1512401278388_0001, Tracking URL = http://ubuntu:8088/proxy/application_1512401278388_0001/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512401278388_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-04 10:59:06,316 Stage-1 map = 0%,  reduce = 0%
2017-12-04 10:59:17,407 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.68 sec
2017-12-04 10:59:25,104 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.21 sec
MapReduce Total cumulative CPU time: 6 seconds 210 msec
Ended Job = job_1512401278388_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.21 sec   HDFS Read: 8898 HDFS Write: 398 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 210 msec
OK
10-01-1990      23
10-01-1991      22
10-01-1993      11
10-01-1994      23
10-03-1990      15
10-03-1991      16
10-03-1993      16
12-02-1990      9
12-02-1991      10
14-02-1990      12
14-02-1991      11
14-02-1994      12
Time taken: 48.254 seconds, Fetched: 12 row(s)
hive>
```

Task 3- Calculate maximum temperature from temperature_data table corresponding to those years which have at least 2 entries in the table.

Commands for Task 3;-

```
hive>
    > SELECT date_format,MAX(T.temperature) as temperature
    > FROM (select date_format,temperature from custom.temperature_data) T
    > GROUP BY date_format
    > HAVING count(T.date_format) >= 2;
```

```
sahil@ubuntu: ~/Desktop
hive>
    > SELECT date_format,MAX(T.temperature) as temperature
    > FROM (select date_format,temperature from custom.temperature_data) T
    > GROUP BY date_format
    > HAVING count(T.date_format) >= 2;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a dif
. spark, tez) or using Hive 1.X releases.
Query ID = sahil_20171204125928_d069849d-bd7a-4dda-8509-2991d4ee9263
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_1512401278388_0013, Tracking URL = http://ubuntu:8088/proxy/application_1512401278388_0013/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512401278388_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-04 12:59:42,024 Stage-1 map = 0%,  reduce = 0%
2017-12-04 12:59:54,385 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.99 sec
2017-12-04 13:00:03,086 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.85 sec
MapReduce Total cumulative CPU time: 7 seconds 850 msec
Ended Job = job_1512401278388_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.85 sec   HDFS Read: 10068 HDFS Write: 217 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 850 msec
OK
10-01-1990      23
10-01-1991      22
10-03-1991      16
12-02-1991      10
14-02-1990      12
Time taken: 35.33 seconds, Fetched: 5 row(s)
hive>
```

Output:-

10-01-1990 23
10-01-1991 22
10-03-1991 16
12-02-1991 10
14-02-1990 12

Task 4- Create a view on the top of last query, name it temperature_data_vw.

Commands for Task 4;-

```
hive> CREATE VIEW temperature_data_vw AS
    > SELECT date_format,MAX(t1.temperature) as temperature
    > FROM (select date_format,temperature from custom.temperature_data) t1
    > GROUP BY date_format
    > HAVING count(t1.date_format) > 2;

hive> select * from temperature_data_vw;
```

```
buntu: ~/Desktop
hive>
    >
    >
    > CREATE VIEW temperature_data_vw AS
    > SELECT date_format,MAX(T.temperature) as temperature
    > FROM (select date_format,temperature from custom.temperature_data) T
    > GROUP BY date_format
    > HAVING count(T.date_format) >= 2;
OK
Time taken: 0.583 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 dif
rk, tez) or using Hive 1.X releases.
Query ID = sahil_20171204130257_2dce6812-55b2-4f59-be4a-4cd91bba23fd
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_1512401278388_0014, Tracking URL = http://ubuntu:8088/proxy/application_1512401278388_0014/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512401278388_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-04 13:03:07,699 Stage-1 map = 0%,  reduce = 0%
2017-12-04 13:03:14,129 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.37 sec
2017-12-04 13:03:21,737 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.33 sec
MapReduce Total cumulative CPU time: 4 seconds 330 msec
Ended Job = job_1512401278388_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.33 sec   HDFS Read: 10188 HDFS Write: 217 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 330 msec
OK
10-01-1990      23
10-01-1991      22
10-03-1991      16
12-02-1991      10
14-02-1990      12
Time taken: 25.062 seconds, Fetched: 5 row(s)
hive>
```

Task 5- Export contents from temperature_data_vw to a file in local file system, such that each file is '|' delimited.

Commands for Task 5;-

hive -e 'select * from temperature_data_vw;' | sed 's/[[:space:]]\+/\|/g' > /home/sahil/Desktop/temperature_data_vw.txt;

```
sahil@ubuntu:~/Desktop$ hive -e 'select * from temperature_data_vw;' | sed 's/[[:space:]]\+/\|/g'  > /home/sahil/Desktop/temperature_data_vw.txt;
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.2-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.8.2/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.clas
s]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]

Logging initialized using configuration in jar:file:/usr/local/apache-hive-2.3.2-bin/lib/hive-common-2.3.2.jar!/hive-log4j2.properties Async: true
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. spa
rk, tez) or using Hive 1.X releases.
Query ID = sahil_20171204130528_750ea91d-2d91-4b50-8983-07fce77888e2
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_1512401278388_0015, Tracking URL = http://ubuntu:8088/proxy/application_1512401278388_0015/
Kill Command = /usr/local/hadoop-2.8.2//bin/hadoop job  -kill job_1512401278388_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-04 13:05:50,494 Stage-1 map = 0%,  reduce = 0%
2017-12-04 13:05:57,482 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.39 sec
2017-12-04 13:06:06,254 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.87 sec
MapReduce Total cumulative CPU time: 5 seconds 870 msec
Ended Job = job_1512401278388_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.87 sec   HDFS Read: 10188 HDFS Write: 217 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 870 msec
OK
Time taken: 39.52 seconds, Fetched: 5 row(s)
sahil@ubuntu:~/Desktop$ cat /home/sahil/Desktop/temperature_data_vw.txt
10-01-1990|23
10-01-1991|22
10-03-1991|16
12-02-1991|10
14-02-1990|12
sahil@ubuntu:~/Desktop$
```

sahil@ubuntu:~/Desktop$ cat /home/sahil/Desktop/temperature_data_vw.txt
10-01-1990|23
10-01-1991|22
10-03-1991|16
12-02-1991|10
14-02-1990|12