

ASSIGNMENT 6.2

Hive queries:

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

Query:

```
SELECT date, temperature
FROM temperature_data
WHERE zip_code > 300000 AND zip_code < 399999;
```

Output:

```
hive> SELECT date, temperature
      > FROM temperature_data
      > WHERE zip_code > 300000 AND zip_code < 399999;
OK
1990-10-03      15
1991-10-01      22
1990-12-02       9
1991-10-03      16
1990-10-01      23
1991-12-02      10
1993-10-03      16
1994-10-01      23
1991-12-02      10
1991-10-03      16
1990-10-01      23
1991-12-02      10
Time taken: 0.79 seconds, Fetched: 12 row(s)
hive> █
```

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

Query:

```
SELECT YEAR(date), MAX(temperature)
FROM temperature_data
GROUP BY YEAR(date);
```

Output:

```
no)
hive> SELECT YEAR(date), MAX(temperature)
> FROM temperature_data
> GROUP BY YEAR(date);
Query ID = acadgild_20171215085656_eceec9eb-69a9-47ed-9be8-f3ac19957e89
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_1513259210539_0003, Tracking URL = http://localhost:8088/proxy/application_1513259210539_0003/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job_1513259210539_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-15 08:57:18,199 Stage-1 map = 0%, reduce = 0%
2017-12-15 08:57:43,363 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.2 sec
2017-12-15 08:58:07,900 Stage-1 map = 100%, reduce = 78%, Cumulative CPU 7.27 sec
2017-12-15 08:58:10,143 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.02 sec
MapReduce Total cumulative CPU time: 8 seconds 20 msec
Ended Job = job_1513259210539_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.02 sec HDFS Read: 650 HDFS Write: 48 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 20 msec
OK
1990 23
1991 22
1992 11
1993 16
1994 23
1995 12
Time taken: 85.525 seconds, Fetched: 6 row(s)
hive> █
```

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

Query:

```
SELECT YEAR(date), MAX(temperature)
FROM temperature_data
GROUP BY YEAR(date)
HAVING COUNT(YEAR(date)) >= 2;
```

Output:

```
hive> SELECT YEAR(date), MAX(temperature)
> FROM temperature_data
> GROUP BY YEAR(date)
> HAVING COUNT(YEAR(date)) >= 2;
Query ID = acadgild_20171215110404_de5feed2-7f47-4e10-a06b-c7f01a98e8cd
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_1513259210539_0004, Tracking URL = http://localhost:8088/proxy/application_1513259210539_0004/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job_1513259210539_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-15 11:05:14,910 Stage-1 map = 0%, reduce = 0%
2017-12-15 11:05:40,087 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.91 sec
2017-12-15 11:06:01,667 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.19 sec
MapReduce Total cumulative CPU time: 9 seconds 190 msec
Ended Job = job_1513259210539_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 9.19 sec HDFS Read: 650 HDFS Write: 24 SUCCESS
Total MapReduce CPU Time Spent: 9 seconds 190 msec
OK
1990      23
1991      22
1993      16
Time taken: 76.864 seconds, Fetched: 3 row(s)
hive>
```

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

Query:

```
CREATE VIEW IF NOT EXISTS temperature_data_vw
AS SELECT YEAR(date), MAX(temperature)
FROM temperature_data
GROUP BY YEAR(date)
HAVING COUNT(YEAR(date)) >= 2;
```

Output:

```
hive> CREATE VIEW IF NOT EXISTS temperature_data_vw
> AS SELECT YEAR(date), MAX(temperature)
> FROM temperature_data
> GROUP BY YEAR(date)
> HAVING COUNT(YEAR(date)) >= 2;
OK
Time taken: 1.721 seconds
hive> DESCRIBE temperature_data_vw;
OK
_c0          int
_c1          int
Time taken: 0.405 seconds, Fetched: 2 row(s)
hive> SELECT * FROM temperature_data_vw;
Query ID = acadgild_20171215112020_c6a7606f-2bac-49ab-bea3-3451e386ca65
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_1513259210539_0005, Tracking URL = http://localhost:8088/proxy/application_1513259210539_0005/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job_1513259210539_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-15 11:21:20,422 Stage-1 map = 0%, reduce = 0%
2017-12-15 11:21:43,102 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.48 sec
2017-12-15 11:22:12,267 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.98 sec
MapReduce Total cumulative CPU time: 8 seconds 980 msec
Ended Job = job_1513259210539_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.98 sec HDFS Read: 650 HDFS Write: 24 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 980 msec
OK
1990      23
1991      22
1993      16
Time taken: 91.062 seconds, Fetched: 3 row(s)
```

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

Query:

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

Note: Output folder that is being specified in file path of above query should not be present before running it.

```
hive> INSERT OVERWRITE LOCAL DIRECTORY '/home/acadgild/temperature_view_data'
> ROW FORMAT DELIMITED
> FIELDS TERMINATED BY '|'
> SELECT * FROM temperature_data_vw;
Query ID = acadgild_20171215120505_b6ec2091-edfd-4996-b991-5b506fcce431
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_1513259210539_0006, Tracking URL = http://localhost:8088/proxy/application_1513259210539_0006/
Kill Command = /home/acadgild/hadoop-2.6.0/bin/hadoop job -kill job_1513259210539_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2017-12-15 12:06:00,624 Stage-1 map = 0%, reduce = 0%
2017-12-15 12:06:30,629 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 6.3 sec
2017-12-15 12:06:56,263 Stage-1 map = 100%, reduce = 78%, Cumulative CPU 10.38 sec
2017-12-15 12:07:00,020 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 11.86 sec
MapReduce Total cumulative CPU time: 11 seconds 860 msec
Ended Job = job_1513259210539_0006
Copying data to local directory /home/acadgild/temperature_view_data
Copying data to local directory /home/acadgild/temperature_view_data
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 11.86 sec HDFS Read: 650 HDFS Write: 24 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 860 msec
OK
Time taken: 88.684 seconds
hive>
```

Output file contents:

```
[acadgild@localhost ~]$ pwd
/home/acadgild
[acadgild@localhost ~]$ ls
airline_usecase      date.txt~           employee_expenses.txt  pig_queries.txt
assignment-5-1_queries.txt  derby.log           hadoop                 Public
assignment5.2_problem1.pig Desktop             hive_queries.txt      session5_dataset
assignment5.2_problem2.pig Documents          hive_queries.txt~     temperature_dataset.txt
assignment5.2_problem3.pig Downloads         hive-site.xml         temperature_view_data
assignment5.2_problem4.pig eclipse           metastore_db          Templates
assignment-5.2_queries.txt eclipse-jee-neon-M3-linux-gtk-x86_64.tar.gz  Music
date.txt             employee_details.txt  Pictures
[acadgild@localhost ~]$ cd temperature_view_data/
[acadgild@localhost temperature_view_data]$ ls
000000 0
[acadgild@localhost temperature_view_data]$ cat 000000 0
1990|23
1991|22
1993|16
[acadgild@localhost temperature_view_data]$
```