Exercise 6a: Analysis the Weather Data set using Hive Query Language (View, Group By, and Order By)

This exercise try to Analysis the Weather data using Hive Query Language (View, Group By, Join, and Order By).

Step 01: Display Available Database
Step 02: Create Database as "weather"
Step 03: Use "weather" Database

```
[cloudera@quickstart ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j.p
roperties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> show databases;
OK
default
one
supermarket
Time taken: 1.751 seconds, Fetched: 3 row(s)
hive> create database weather;
OK
Time taken: 0.307 seconds
hive> show databases;
OK
default
one
supermarket
weather
Time taken: 0.09 seconds, Fetched: 4 row(s)
hive> use weather;
OK
Time taken: 0.157 seconds
hive>
```

Step 04: Display available Tables

```
hive> show tables;
OK
Time taken: 0.235 seconds
hive>
```

Step 05: Create Table "d_weather2020" with following scheme

Country as string

Date as Date
Summary as string
icon as string
temperatureHigh as float
temperatureLow as float
humidity as float
pressure as float
windSpeed as float
visibility as float
ozone as float
Lat as int
Long as int

```
hive> create table d_weather2020(country STRING,date DATE,summary STRING,icon STRING,temperatureHigh FLOAT,humidity FLOAT,pressure FLOAT,windspeed FLOAT,visibility FLOA
T,ozone FLOAT,lat INT,long INT)row format delimited fields terminated by ',' tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.523 seconds
```

Step 06: Copy 'daily_weather_2020.csv' into Hadoop local (/home/cloudera/)
/home/cloudera/daily_weather_2020.csv

Step 07: Load 'daily_weather_2020.csv' data into table 'd_weather2020'

```
hive> LOAD DATA LOCAL INPATH'/home/cloudera/daily_weather_2020.csv' INTO TABLE d_weathers2020;
Loading data to table weather.d_weathers2020
Table weather.d_weathers2020 stats: [numFiles=1, totalSize=3725862]
OK
Time taken: 1.946 seconds
hive>
```

Step 08: Display the content in table 'd_weather2020'

```
    > describe d_weathers2020;
OK
country                 string
date                    date
summary                 string
icon                    string
temperaturehigh         float
temperaturelow          float
humidity                float
pressure                float
windspeed               float
visibility              float
ozone                   float
lat                     int
long                    int
Time taken: 0.263 seconds, Fetched: 13 row(s)
hive>
```

Q01: Create View ‘temp25to45’ by select which record has‘temperatureHigh’ between 25 degree Celsius and 45 degree Celsius.

```
hive>
    >
    > CREATE VIEW  temp25to45 AS SELECT temperatureHigh FROM d_weathers2020 WHERE temperatureHigh BETWEEN 25 AND 45;
OK
Time taken: 0.879 seconds
hive> SELECT * FROM temp25to45;
OK
Time taken: 1.129 seconds
hive>
```

Q02: Show how many temperature readings in India.

```
hive> SELECT COUNT(temperatureHigh),COUNT(temperatureLow) FROM d_weathers2020 WHERE Country=='India';
Query ID = cloudera_20220902213737_5c837332-441f-4b26-bb4e-c22a2754965c
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-02 21:38:03,698 Stage-1 map = 0%,  reduce = 0%
2022-09-02 21:38:17,578 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.05 sec
2022-09-02 21:38:31,146 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.57 sec
MapReduce Total cumulative CPU time: 5 seconds 570 msec
Ended Job = job_1661880724832_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.57 sec   HDFS Read: 3736615 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 570 msec
OK
0       0
Time taken: 42.499 seconds, Fetched: 1 row(s)
```

Q03: Group the d_weather2020 table records by country name. In addition this group should have less than 25 degree Celsiusin temperatureLow column.

```
hive> SELECT AVG(temperatureLow) FROM d_weathers2020 WHERE temperatureLow<25 GROUP BY country;
Query ID = cloudera_20220902215858_acc056da-a152-4fbf-b7b3-5208b73324e9
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_1661880724832_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0006
```

```
14.100000381469727
18.8799991607666
23.309999465942383
15.600000381469727
14.729999542236328
18.59000015258789
19.010000228881836
13.140000343322754
7.550000190734863
15.420000076293945
18.309999465942383
18.959999084472656
20.139999389648438
22.760000228881836
18.969999313354492
16.1200008392334
19.229999542236328
10.380000114440918
14.069999694824219
15.300000190734863
13.079999923706055
16.610000610351562
17.93000030517578
12.899999618530273
10.220000267028809
18.510000228881836
23.639999389648438
24.889999389648438
23.049999237060547
20.15999984741211
23.959999084472656
Time taken: 53.096 seconds, Fetched: 1370 row(s)
```

Q04: Calculate average humidity of each country and order the result by 'country' name.

```
hive> SELECT AVG(humidity) FROM d_weathers2020 GROUP BY Country ORDER BY Country;
FAILED: SemanticException [Error 10004]: Line 1:67 Invalid table alias or column reference 'Country': (possible column names are: _c0)
hive> SELECT Country,AVG(humidity) FROM d_weathers2020 GROUP BY Country ORDER BY Country;
Query ID = cloudera_20220902220404_3b7b1a86-c799-433a-96c9-b48ea7a500d4
Total jobs = 2
Launching Job 1 out of 2
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_1661880724832_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0007
```

```
9971    70.3499984741211
9972    70.61000061035156
9973    70.05999755859375
9974    69.5
9975    69.23999786376953
9976    67.7699966430664
9977    66.56999969482422
9978    66.80000305175781
9979    69.0199966430664
998     40.279998779296875
9980    70.56999969482422
9981    71.0
9982    70.7699966430664
9983    70.2300033569336
9984    67.36000061035156
9985    67.19999694824219
9986    68.81999969482422
9987    69.56999969482422
9988    68.87000274658203
9989    68.66000366210938
999     41.72999954223633
9990    69.30999755859375
9991    66.55000305175781
9992    67.94999694824219
9993    70.8499984741211
9994    70.29000091552734
9995    71.12999725341797
9996    70.06999969482422
9997    72.45999908447266
9998    69.81999969482422
9999    70.44000244140625
Time taken: 98.753 seconds, Fetched: 30688 row(s)
hive>
```

Q05: Find each country minimum temperature and order by ‘country’ name.

```
hive> SELECT Country,MIN(temperatureLow) FROM d_weathers2020 ORDER BY Country;
FAILED: SemanticException [Error 10025]: Line 1:7 Expression not in GROUP BY key 'Country'
hive> SELECT Country,MIN(temperatureLow) FROM d_weathers2020  GROUP BY Country ORDER BY Country;
Query ID = cloudera_20220902221212_e9035083-ac62-42e8-a128-9d868c6910f4
Total jobs = 2
Launching Job 1 out of 2
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_1661880724832_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0009
```

Q06: Find each country maximum temperature and order by ‘icon’.

```
Time taken: 88.722 seconds, Fetched: 30688 row(s)
hive> SELECT Country,MAX(temperatureLow),icon  FROM d_weathers2020  GROUP BY Country  ORDER BY icon;
FAILED: SemanticException [Error 10025]: Line 1:35 Expression not in GROUP BY key 'icon'
hive> SELECT Country,MAX(temperatureHigh),icon  FROM d_weathers2020  GROUP BY Country  ORDER BY icon;
FAILED: SemanticException [Error 10025]: Line 1:36 Expression not in GROUP BY key 'icon'
hive> SELECT Country,MAX(temperatureHigh),icon  FROM d_weathers2020  GROUP BY Country,icon ORDER BY icon;
Query ID = cloudera_20220902221717_d7b15180-f74c-49c8-bfe5-d92bfd6bc1c6
Total jobs = 2
Launching Job 1 out of 2
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_1661880724832_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0011
```

```
24646   NULL    Windy throughout the day.
16163   NULL    Windy throughout the day.
458     NULL    Windy throughout the day.
24690   NULL    Windy throughout the day.
11323   NULL    Windy throughout the day.
11210   NULL    Windy throughout the day.
13398   NULL    Windy throughout the day.
28683   NULL    Windy throughout the day.
12219   NULL    Windy throughout the day.
24699   NULL    Windy throughout the day.
29760   NULL    Windy until evening.
21850   NULL    Windy until evening.
24449   NULL    Windy until evening.
17925   NULL    Windy until evening.
21843   NULL    Windy until evening.
24701   NULL    Windy until evening.
28347   NULL    Windy until evening.
29303   NULL    Windy until evening.
16213   NULL    Windy until evening.
18332   NULL    Windy until evening.
3953    NULL    Windy until evening.
243     NULL    Windy until evening.
3728    NULL    Windy until evening.
13328   NULL    Windy until evening.
27896   NULL    Windy until evening.
26966   NULL    Windy until evening.
2617    NULL    Windy until evening.
21091   NULL    Windy until evening.
795     NULL    Windy until evening.
29770   NULL    Windy until evening.
21454   NULL    Windy until evening.
Time taken: 91.173 seconds, Fetched: 30688 row(s)
hive>
```

Q07:Create view for clear-day entry in ‘icon’.

```
hive> CREATE VIEW clearday AS
    > SELECT * FROM d_weathers2020
    > WHERE icon=='clear-day';
OK
Time taken: 0.219 seconds
hive>
hive> select * FROM clearday;
OK
Time taken: 0.757 seconds
hive>
```

Q08: Count how many clear day in 'Aruba'.

```
hive> SELECT COUNT(*) FROM clearday  WHERE Country=='Aruba';
Query ID = cloudera_20220902223232_e0c2d8e5-d556-4d75-bd4a-975779871b5d
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0013
```

Q09: Show which day and which country has highest and lowest 'ozone' level

```
hive> SELECT MAX(ozone),MIN(ozone) FROM d_weathers2020;
Query ID = cloudera_20220903090202_28c80c2a-67ed-44be-af31-7bd5d722d9f2
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0014
```

```
hive> SELECT Date,Country FROM d_weathers2020 WHERE ozone==0.607 OR ozone== 510.7;
OK
Time taken: 0.18 seconds
hive>
```

Q10: Find the maximum and minimum temperature when we have visibility between 5 and 10.

```
hive> SELECT MAX(temperatureHigh), MIN(temperatureLow) FROM d_weathers2020 WHERE visibility BETWEEN 5 AND 10;
Query ID = cloudera_20220903090909_35bf6cd2-13ff-42bd-817b-e0ad38e60d99
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0015
```

Q11: Calculate average humidity when the weather is in cloudy order by country.

```
hive> SELECT AVG(humidity) FROM d_weathers2020 WHERE icon=='cloudy' GROUP BY country;
Query ID = cloudera_20220903091414_ebead50f-69fc-4b4d-a58c-8037259b6fd4
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_1661880724832_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 09:14:13,485 Stage-1 map = 0%,  reduce = 0%
2022-09-03 09:14:24,528 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.68 sec
2022-09-03 09:14:36,557 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.42 sec
MapReduce Total cumulative CPU time: 4 seconds 420 msec
Ended Job = job_1661880724832_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.42 sec   HDFS Read: 3736942 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 420 msec
OK
Time taken: 35.065 seconds
```

Q12: Find which country has highest temperature and lowest temperature.

```
hive> SELECT MAX(temperatureHigh),MIN(temperatureLow) FROM d_weathers2020;
Query ID = cloudera_20220903091919_35e88005-d931-46a5-b579-ff3dc309aaf7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 09:19:26,256 Stage-1 map = 0%,  reduce = 0%
```

```
hive> SELECT Country FROM d_weathers2020 WHERE temperatureHigh == 111.61 OR temperatureLow =='-66.3';
OK
Time taken: 0.111 seconds
hive>
```

Q13: Display Longitude and latitude of the lowest visibly country.

```
hive> SELECT MIN(visibility) FROM d_weathers2020;
Query ID = cloudera_20220903092424_8f2d8768-40fe-4e85-846c-6e66c6f186bc
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 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_1661880724832_0018, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0018/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0018
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 09:24:23,089 Stage-1 map = 0%,  reduce = 0%
2022-09-03 09:24:34,113 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2022-09-03 09:24:45,047 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.52 sec
MapReduce Total cumulative CPU time: 4 seconds 520 msec
Ended Job = job_1661880724832_0018
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.52 sec   HDFS Read: 3735032 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 520 msec
OK
1.12
Time taken: 34.657 seconds, Fetched: 1 row(s)
hive> SELECT Lat,Long,Country FROM d_weathers2020 WHERE visibility==0.05 ;
OK
Time taken: 0.072 seconds
hive>
```

Q14: Calculate how many rain days in each country.

```
hive> SELECT COUNT(Date) FROM d_weathers2020 WHERE icon=='rain' GROUP BY Country;
Query ID = cloudera_20220903093030_7bbcdd94-9302-40ee-b790-f0967023f143
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_1661880724832_0019, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0019
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 09:31:02,654 Stage-1 map = 0%,  reduce = 0%
2022-09-03 09:31:13,694 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.77 sec
2022-09-03 09:31:25,680 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 4.54 sec
MapReduce Total cumulative CPU time: 4 seconds 540 msec
Ended Job = job_1661880724832_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 4.54 sec   HDFS Read: 3736517 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 540 msec
OK
Time taken: 35.689 seconds
hive>
```

Q15: Group the country when it has 10 as ' visibility ' and find the average wind speed.

```
hive> SELECT AVG(windspeed) FROM d_weathers2020 WHERE visibility == 10 GROUP BY Country;
Query ID = cloudera_20220903093333_8bb542aa-0565-4d44-b7ce-7b32400c7635
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_1661880724832_0020, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1661880724832_0020/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1661880724832_0020
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-03 09:33:20,899 Stage-1 map = 0%,  reduce = 0%
2022-09-03 09:33:31,944 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.1 sec
2022-09-03 09:33:44,057 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.3 sec
MapReduce Total cumulative CPU time: 5 seconds 300 msec
Ended Job = job_1661880724832_0020
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.3 sec   HDFS Read: 3736948 HDFS Write: 215 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 300 msec
OK
1015.9000244140625
1016.0
1010.4000244140625
1019.5999755859375
1018.4000244140625
1010.5
1000.5999755859375
1009.5999755859375
1015.5
1030.199951171875
```