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.

SOLUTION:

-----created database custom and table temperature_data----------

hive> create database custom;

OK

Time taken: 0.329 second>

OK

Time taken: 0.065 seconds

hive> create table temperature_data(

> date_temp STRING,

> zip_code INT,

> temperature INT)

> row format delimited fields terminated by ',';

OK

Time taken: 0.446 seconds

-----Loading input file to table temperature_data------

hive> LOAD DATA LOCAL INPATH '/home/acadgild/Downloads/dataset_Session 14.txt' into table temperature_data;

Loading data to table custom.temperature_data

OK

Time taken: 1.152 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

Time taken: 0.53 seconds, Fetched: 20 row(s)

TASK 2:

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

hive> select date_temp,temperature from temperature_data where zip_code 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: 0.487 seconds, Fetched: 12 row(s)

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

hive> select substring(date_temp,7,10),max(temperature) from temperature_data group by substring(date_temp,7,10);

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. spark, tez) or using Hive 1.X releases.

Query ID = acadgild_20180514053527_2cadc4f9-d88f-4dcb-8944-719040c0b997

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_1526194450906_0009, Tracking URL = http://localhost:8088/proxy/application_1526194450906_0009/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1526194450906_0009

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-05-14 05:35:49,524 Stage-1 map = 0%, reduce = 0%

2018-05-14 05:36:11,105 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.99 sec

2018-05-14 05:36:28,628 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.28 sec

MapReduce Total cumulative CPU time: 8 seconds 280 msec

Ended Job = job_1526194450906_0009

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.28 sec HDFS Read: 9277 HDFS Write: 167 SUCCESS

Total MapReduce CPU Time Spent: 8 seconds 280 msec

OK

1990 23

1991 22

1993 16

1994 23

Time taken: 63.321 seconds, Fetched: 4 row(s)

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

hive> select substring(date_temp,7,10), count(substring(date_temp,7,10)),max(temperature) from temperature_data group by substring(date_temp,7,10) having count(substring(date_temp,7,10))>=2;

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. spark, tez) or using Hive 1.X releases.

Query ID = acadgild_20180514055412_d053ee6a-9b5f-4376-9eab-1dd7f6ba93bd

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_1526194450906_0011, Tracking URL = http://localhost:8088/proxy/application_1526194450906_0011/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1526194450906_0011

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-05-14 05:54:35,544 Stage-1 map = 0%, reduce = 0%

2018-05-14 05:54:51,895 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.88 sec

2018-05-14 05:55:09,935 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.48 sec

MapReduce Total cumulative CPU time: 8 seconds 480 msec

Ended Job = job_1526194450906_0011

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1   Cumulative CPU: 8.48 sec   HDFS Read: 10347 HDFS Write: 175 SUCCESS

Total MapReduce CPU Time Spent: 8 seconds 480 msec

OK

1990 7 23

1991 9 22

1993 2 16

1994 2 23

Time taken: 60.077 seconds, Fetched: 4 row(s)

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

hive> create VIEW temperature_data_vw as select substring(date_temp,7,10), count(substring(date_temp,7,10)),max(temperature) from temperature_data group by substring(date_temp,7,10) having count(substring(date_temp,7,10))>=2;

OK

Time taken: 0.753 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 execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = acadgild_20180514063341_72dfb47a-3d92-443d-ab71-7e455525049d

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_1526194450906_0012, Tracking URL = http://localhost:8088/proxy/application_1526194450906_0012/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1526194450906_0012

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-05-14 06:34:12,463 Stage-1 map = 0%, reduce = 0%

2018-05-14 06:34:34,081 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 5.48 sec

2018-05-14 06:34:54,138 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 10.11 sec

MapReduce Total cumulative CPU time: 10 seconds 110 msec

Ended Job = job_1526194450906_0012

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 10.11 sec HDFS Read: 10608 HDFS Write: 175 SUCCESS

Total MapReduce CPU Time Spent: 10 seconds 110 msec

OK

1990 7 23

1991 9 22

1993 2 16

1994 2 23

Time taken: 74.802 seconds, Fetched: 4 row(s)

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/hiveoutput/put1' 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 execution engine (i.e. spark, tez) or using Hive 1.X releases.

Query ID = acadgild_20180514092424_b4313de6-bc24-4263-b58c-014d16daa476

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_1526194450906_0014, Tracking URL = http://localhost:8088/proxy/application_1526194450906_0014/

Kill Command = /home/acadgild/hadoop-2.7.2/bin/hadoop job -kill job_1526194450906_0014

Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1

2018-05-14 09:24:47,012 Stage-1 map = 0%, reduce = 0%

2018-05-14 09:25:06,644 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.59 sec

2018-05-14 09:25:26,514 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 9.52 sec

MapReduce Total cumulative CPU time: 9 seconds 520 msec

Ended Job = job_1526194450906_0014

Moving data to local directory /home/acadgild/hiveoutput/put1

MapReduce Jobs Launched:

Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 9.52 sec HDFS Read: 10326 HDFS Write: 40 SUCCESS

Total MapReduce CPU Time Spent: 9 seconds 520 msec

OK

Time taken: 64.416 seconds

/home/acadgild/hiveoutput

[acadgild@localhost hiveoutput]$ cd put1

[acadgild@localhost put1]$ ls -ltr

total 4

-rw-r--r--. 1 acadgild acadgild 40 May 14 09:25 000000_0

[acadgild@localhost put1]$ cat 000000_0

1990|7|23

1991|9|22

1993|2|16

1994|2|23

[acadgild@localhost put1]$