Hive – Vehicle data

SELECT COUNT(*) FROM VehicleData;

```
Logging initialized using configuration in jar:file:/home/ec2-user/apache-hive-2
.0.1-bin/lib/hive-common-2.0.1.jar!/hive-log4j2.properties
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
hive> CREATE TABLE VehicleData (
    > barrels08 FLOAT, barrelsA08 FLOAT,
    > charge120 FLOAT, charge240 FLOAT,
    > city08 FLOAT)
    > ROW FORMAT DELIMITED FIELDS
    > TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 1.707 seconds
```

```
hive> LOAD DATA LOCAL INPATH '/home/ec2-user/vehicles.csv'
    > OVERWRITE INTO TABLE VehicleData;
Loading data to table default.vehicledata
OK
Time taken: 1.354 seconds
hive> SELECT COUNT(*) FROM VehicleData;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the futu
re versions. Consider using a different execution engine (i.e. spark, tez) or us
ing Hive 1.X releases.
Query ID = ec2-user_20200930054111_156afd1c-3e32-490c-87a7-3229079e8877
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_1601443711508_0001, Tracking URL = http://ip-172-31-56-147.ec
2.internal:8088/proxy/application_1601443711508_0001/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_16014437115
08_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-09-30 05:41:24,951 Stage-1 map = 0%,  reduce = 0%
2020-09-30 05:41:32,623 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.27 sec
2020-09-30 05:41:40,388 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.58 sec
MapReduce Total cumulative CPU time: 2 seconds 580 msec
Ended Job = job_1601443711508_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.58 sec   HDFS Read: 11775010 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 580 msec
```

Output: 34175 (highlighted in below screenshot)

```
OK
34175
Time taken: 30.8 seconds, Fetched: 1 row(s)
```

SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData;

```
hive> SELECT MIN(barrels08), AVG(barrels08), MAX(barrels08) FROM VehicleData;
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 = ec2-user_20200930054553_2eeef564-ae36-45a1-8f1d-9ee335af2f36
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_1601443711508_0002, Tracking URL = http://ip-172-31-56-147.ec2.internal:8088/proxy/application_1601443711508_0002/
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1601443711508_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2020-09-30 05:46:01,819 Stage-1 map = 0%,  reduce = 0%
2020-09-30 05:46:09,380 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 1.57 sec
2020-09-30 05:46:15,994 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 2.92 sec
MapReduce Total cumulative CPU time: 2 seconds 920 msec
Ended Job = job_1601443711508_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 2.92 sec   HDFS Read: 11777415 HDFS Write: 37 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 920 msec
```

Output: (highlighted in below screenshot)

```
OK
0.059892          17.820177449476272          47.06831
Time taken: 24.366 seconds, Fetched: 1 row(s)
```

SELECT (barrels08/city08) FROM VehicleData;

```
2.99505615234375
1.0767260158763212
0.8716353310479058
0.8716353310479058
1.1440213918685913
2.4970455169677734
0.9155832926432291
0.550305407980214
0.6539881115867978
0.4898370901743571
0.6278285526093983
0.7875944438733553
0.7163524150848388
0.8716353310479058
0.8716353310479058
1.1440213918685913
Time taken: 0.204 seconds, Fetched: 34175 row(s)
```

INSERT OVERWRITE DIRECTORY 'ThreeColExtract'
SELECT barrels08, city08, charge120
FROM VehicleData;

```
hive>
    > INSERT OVERWRITE DIRECTORY 'ThreeColExtract'
    > SELECT barrels08, city08, charge120
    > FROM VehicleData;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consid
spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20200930055156_04d5e78a-df9e-455f-bfb3-e30a7a1af758
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1601443711508_0004, Tracking URL = http://ip-172-31-56-147.ec2.internal:8088/pro
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1601443711508_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-09-30 05:52:04,754 Stage-1 map = 0%,  reduce = 0%
2020-09-30 05:52:13,329 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.07 sec
MapReduce Total cumulative CPU time: 2 seconds 70 msec
Ended Job = job_1601443711508_0004
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://localhost/user/ec2-user/ThreeColExtract/.hive-staging_hive_2020-09-30_05-51-
Moving data to: ThreeColExtract
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.07 sec   HDFS Read: 11770539 HDFS Write: 627873 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 70 msec
OK
Time taken: 17.772 seconds
```

Output:

```
[ec2-user@ip-172-31-56-147 bin]$  hdfs dfs -ls /user/ec2-user/ThreeColExtract
Found 1 items
-rwxr-xr-x   1 ec2-user supergroup      627873 2020-09-30 05:52 /user/ec2-user/ThreeColExtract/000000_0
```

Two column extract

```
CREATE TABLE VehicleDataUpdate (
barrels08 FLOAT, barrelsA08 FLOAT,
charge120 FLOAT, charge240 FLOAT,
city08 FLOAT, city08U FLOAT, cityA08 FLOAT, cityA08U FLOAT)
ROW FORMAT DELIMITED FIELDS
TERMINATED BY ',' STORED AS TEXTFILE;

LOAD DATA LOCAL INPATH '/home/ec2-user/vehicles.csv'
OVERWRITE INTO TABLE VehicleDataUpdate;
```

```
hive> CREATE TABLE VehicleDataUpdate (
    > barrels08 FLOAT, barrelsA08 FLOAT,
    > charge120 FLOAT, charge240 FLOAT,
    > city08 FLOAT, city08U FLOAT,cityA08 FLOAT,cityA08U FLOAT)
    > ROW FORMAT DELIMITED FIELDS
    > TERMINATED BY ',' STORED AS TEXTFILE;
OK
Time taken: 1.911 seconds
hive>
    > LOAD DATA LOCAL INPATH '/home/ec2-user/vehicles.csv'
    > OVERWRITE INTO TABLE VehicleDataUpdate;
Loading data to table default.vehicledataupdate
OK
Time taken: 1.386 seconds
```

INSERT OVERWRITE DIRECTORY 'TwoColExtract'
SELECT city08U, cityA08U
FROM VehicleDataUpdate;

```
hive> INSERT OVERWRITE DIRECTORY 'TwoColExtract'
    > SELECT city08U, cityA08U
    > FROM VehicleDataUpdate;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider
spark, tez) or using Hive 1.X releases.
Query ID = ec2-user_20200930060645_2ac4bb74-c1a3-4606-8c3e-88081c89e67d
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1601443711508_0005, Tracking URL = http://ip-172-31-56-147.ec2.internal:8088/proxy
Kill Command = /home/ec2-user/hadoop-2.6.4/bin/hadoop job  -kill job_1601443711508_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2020-09-30 06:06:55,064 Stage-1 map = 0%,   reduce = 0%
2020-09-30 06:07:03,121 Stage-1 map = 100%,   reduce = 0%, Cumulative CPU 2.06 sec
MapReduce Total cumulative CPU time: 2 seconds 60 msec
Ended Job = job_1601443711508_0005
Stage-3 is selected by condition resolver.
Stage-2 is filtered out by condition resolver.
Stage-4 is filtered out by condition resolver.
Moving data to: hdfs://localhost/user/ec2-user/TwoColExtract/.hive-staging_hive_2020-09-30_06-06-45_3
Moving data to: TwoColExtract
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1   Cumulative CPU: 2.06 sec   HDFS Read: 11770779 HDFS Write: 287749 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 60 msec
OK
Time taken: 18.996 seconds
```

Output:
```
[ec2-user@ip-172-31-56-147 apache-hive-2.0.1-bin]$ hdfs dfs -ls /user/ec2-user/TwoColExtract
Found 1 items
-rwxr-xr-x   1 ec2-user supergroup      287749 2020-09-30 06:07 /user/ec2-user/TwoColExtract/000000_0
```