Hive Mini Project-2

25 October 2022 11:45

This dataset include only 500 rows from original dataset so the output may be differ from original datasets

PART-1:Examine the Data

1.Find the total number of tickets for the year.

```
hive>
    > select count(*) as total_record from parking_violation_orc;
Query ID = cloudera_20221024230808_10d3f86f-0ecb-4f10-864e-18f77724aa85
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_1666676796839_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-24 23:09:29,027 Stage-1 map = 0%,  reduce = 0%
2022-10-24 23:10:02,913 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 13.64 sec
2022-10-24 23:10:46,451 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 15.23 sec
2022-10-24 23:10:47,509 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.44 sec
MapReduce Total cumulative CPU time: 16 seconds 440 msec
Ended Job = job_1666676796839_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.44 sec   HDFS Read: 31741 HDFS Write: 4 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 440 msec
OK
total_record
499
Time taken: 151.761 seconds, Fetched: 1 row(s)
```

2.Find out how many unique states the cars which got parking tickets came from.

```
hive> --Find out how many unique states the cars which got parking tickets came from;
hive> select count(distinct registration_state) as unique_state from parking_violation_orc;
Query ID = cloudera_20221024234545_0505492a-e3ac-42eb-b252-f89c6a65550a
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_1666676796839_0002, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0002/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0002
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-24 23:46:29,227 Stage-1 map = 0%,  reduce = 0%
2022-10-24 23:47:17,992 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.23 sec
2022-10-24 23:47:32,209 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.02 sec
MapReduce Total cumulative CPU time: 6 seconds 20 msec
Ended Job = job_1666676796839_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 6.02 sec   HDFS Read: 32509 HDFS Write: 3 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 20 msec
OK
unique_state
25
Time taken: 128.416 seconds, Fetched: 1 row(s)
```

3.Some parking tickets don't have addresses on them, which is cause for concern. Find out how many such tickets there are(i.e. tickets where either "Street Code 1" or "Street Code 2" or "Street Code 3" is empty)

```
hive> select count(*) as empty_address from parking_violation_orc where street_code1 is null or street_code2 is null or street_code3 is null;
Query ID = cloudera_20221025001212_477d9f11-de1b-4dd3-9582-8701444721e1
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_1666676796839_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 00:12:58,092 Stage-1 map = 0%,  reduce = 0%
2022-10-25 00:13:52,251 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.84 sec
2022-10-25 00:14:06,260 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 7.07 sec
MapReduce Total cumulative CPU time: 7 seconds 70 msec
Ended Job = job_1666676796839_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 7.07 sec   HDFS Read: 35770 HDFS Write: 2 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 70 msec
OK
empty_address
0
```

PART -2 Aggregation tasks

1.How often does each violation code occur? (frequency of violation codes - find the top 5)

```
hive> select violation_code ,count(*) as frequency from parking_violation_orc group by violation code order by frequency desc limit 5;
FAILED: ParseException line 1:91 missing EOF at 'code' near 'violation'
hive> select violation_code ,count(*) as frequency from parking_violation_orc group by violation_code order by frequency desc limit 5;
Query ID = cloudera_20221025003333_6d0e7ad7-c19a-447c-9c92-b41dcf7b2463
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_1666676796839_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 00:33:30,260 Stage-1 map = 0%,  reduce = 0%
2022-10-25 00:33:50,850 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 3.39 sec
2022-10-25 00:34:07,804 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 5.54 sec
MapReduce Total cumulative CPU time: 5 seconds 540 msec
Ended Job = job_1666676796839_0005
Launching Job 2 out of 2
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_1666676796839_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0006
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-25 00:34:24,437 Stage-2 map = 0%,  reduce = 0%
2022-10-25 00:34:37,007 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.5 sec
2022-10-25 00:35:23,915 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.69 sec
MapReduce Total cumulative CPU time: 5 seconds 690 msec
Ended Job = job_1666676796839_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 5.54 sec   HDFS Read: 31607 HDFS Write: 837 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.69 sec   HDFS Read: 5866 HDFS Write: 30 SUCCESS
Total MapReduce CPU Time Spent: 11 seconds 230 msec
OK
violation_code  frequency
36      63
21      58
38      42
40      39
20      39
Time taken: 135.916 seconds, Fetched: 5 row(s)
```

2.1 How often does each vehicle body type get a parking ticket? How about the vehicle make? (find the top 5 for both)

```
hive> select vechile_body_type ,count(days_in_parking_effect) as total_count from parking_violation_orc group by vechile_body_type order by total_count desc limit 5;
Query ID = cloudera_20221025013333_4443d069-748a-45b9-b589-19daf13654f3
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_1666676796839_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 01:33:52,599 Stage-1 map = 0%,  reduce = 0%
2022-10-25 01:34:36,078 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.7 sec
2022-10-25 01:34:55,603 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 11.12 sec
MapReduce Total cumulative CPU time: 11 seconds 120 msec
Ended Job = job_1666676796839_0013
Launching Job 2 out of 2
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_1666676796839_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-25 01:35:15,448 Stage-2 map = 0%,  reduce = 0%
2022-10-25 01:35:27,557 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.62 sec
2022-10-25 01:35:42,789 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.06 sec
MapReduce Total cumulative CPU time: 5 seconds 60 msec
Ended Job = job_1666676796839_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 11.12 sec   HDFS Read: 32079 HDFS Write: 743 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.06 sec   HDFS Read: 5882 HDFS Write: 40 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 180 msec
OK
vechile_body_type       total_count
SUBN    174
4DSD    147
VAN     66
DELV    34
SDN     19
Time taken: 125.187 seconds, Fetched: 5 row(s)
```

2.2 How about the vehicle make?

```
hive> select vechile_make ,count(days_in_parking_effect) as total_count from parking_violation_orc group by vechile_make order by total_count desc limit 5;
Query ID = cloudera_20221025013838_a2dffdad-fced-4478-af2d-2e0daab89486
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_1666676796839_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 01:39:05,530 Stage-1 map = 0%,  reduce = 0%
2022-10-25 01:39:31,948 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 25.96 sec
2022-10-25 01:39:43,767 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 28.06 sec
MapReduce Total cumulative CPU time: 28 seconds 60 msec
Ended Job = job_1666676796839_0015
Launching Job 2 out of 2
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_1666676796839_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0016
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-25 01:40:02,509 Stage-2 map = 0%,  reduce = 0%
2022-10-25 01:40:12,068 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 2.6 sec
2022-10-25 01:40:24,466 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 4.52 sec
MapReduce Total cumulative CPU time: 4 seconds 520 msec
Ended Job = job_1666676796839_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 28.06 sec   HDFS Read: 32211 HDFS Write: 1176 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 4.52 sec   HDFS Read: 6295 HDFS Write: 44 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 580 msec
OK
vechile_make    total_count
FORD    63
TOYOT   59
HONDA   46
CHEVR   34
NISSA   30
Time taken: 96.548 seconds, Fetched: 5 row(s)
```

3.A precinct is a police station that has a certain zone of the city under its command. Find the (5 highest) frequencies of:
a.) Violating Precincts (this is the precinct of the zone where the violation occurred)

```
hive> select violation_precinct ,count(*) as total_record from parking_violation_orc group by violation_precinct order by total_record desc limit 5;
Query ID = cloudera_20221025010202_a24f9da4-5cd2-498f-b3ca-06824dc4e5f0
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>
```

```
hive> select violation_precinct ,count(*) as total_record from parking_violation_orc group by violation_precinct order by total_record desc limit 5;
Query ID = cloudera_20221025010202_a24f9da4-5cd2-498f-b3ca-06824dc4e5f0
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_1666676796839_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 01:03:08,908 Stage-1 map = 0%,  reduce = 0%
2022-10-25 01:03:57,904 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 4.94 sec
2022-10-25 01:04:26,691 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 8.25 sec
MapReduce Total cumulative CPU time: 8 seconds 250 msec
Ended Job = job_1666676796839_0007
Launching Job 2 out of 2
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_1666676796839_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0008
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-25 01:04:43,643 Stage-2 map = 0%,  reduce = 0%
2022-10-25 01:05:04,739 Stage-2 map = 100%,  reduce = 0%
2022-10-25 01:05:18,010 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 6.98 sec
MapReduce Total cumulative CPU time: 6 seconds 980 msec
Ended Job = job_1666676796839_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 8.25 sec   HDFS Read: 31692 HDFS Write: 1502 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 6.98 sec   HDFS Read: 6553 HDFS Write: 30 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 230 msec
OK
violation_precinct      total_record
0       93
19      25
14      24
114     16
18      16
Time taken: 151.16 seconds, Fetched: 5 row(s)
```

b.) Issuer Precincts (this is the precinct that issued the ticket)

```
hive> select issuer_precinct ,count(*) as total_record from parking_violation_orc group by issuer_precinct order by total_record desc limit 5;
Query ID = cloudera_20221025011717_463e17d2-b536-428c-8e8f-3766f0fd5d53
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_1666676796839_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-10-25 01:17:21,812 Stage-1 map = 0%,  reduce = 0%
2022-10-25 01:17:37,970 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 18.81 sec
2022-10-25 01:17:53,163 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 20.97 sec
MapReduce Total cumulative CPU time: 20 seconds 970 msec
Ended Job = job_1666676796839_0011
Launching Job 2 out of 2
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_1666676796839_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1666676796839_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1666676796839_0012
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-10-25 01:18:05,159 Stage-2 map = 0%,  reduce = 0%
2022-10-25 01:18:12,558 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 1.6 sec
2022-10-25 01:18:24,176 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 5.4 sec
MapReduce Total cumulative CPU time: 5 seconds 400 msec
Ended Job = job_1666676796839_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.97 sec   HDFS Read: 31719 HDFS Write: 1543 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 5.4 sec   HDFS Read: 6573 HDFS Write: 31 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 370 msec
OK
issuer_precinct total_record
0       105
19      25
14      24
114     16
13      15
Time taken: 95.801 seconds, Fetched: 5 row(s)
```