Problem statement:

In this case study, we are giving a real world example of how to use HIVE on top of the HADOOP for different exploratory data analysis. In here, we have a predefined dataset (yellow_tripdata_2015-01-06.csv) **having more than 15 columns.**

Queries for Hive Case study

Tasks:

1.Create a table named taxidata . Required ddl script is given below.

```
Create database challenge;

Use challenge;

CREATE TABLE IF NOT EXISTS taxidata(vendor_id string, pickup_datetime
string,dropoff_datetime string, passenger_count int,

trip_distance Float,pickup_longitude Float, pickup_latitude Float, rate_code
int,store_and_fwd_flag string, dropoff_longitude Float,

dropoff_latitude Float,payment_type string, fare_amount Float, extra Float,mta_tax Float,
tip_amount Float, tolls_amount Float,

total_amount Float, trip_time_in_secs int )ROW FORMAT DELIMITED FIELDS
TERMINATED BY ','

TBLPROPERTIES ("skip.header.line.count"="1");
```

2.Load data from the csv file - yellow_tripdata_2015-01-06.csv

```
load data local inpath"/home/cloudera/sidd/Challenge/Mini project 3/yellow_tripdata.csv" into  TABLE
taxidata;
```

Perform taxi trip analysis by solving the questions below:

1. **What is the total number of trips (equal to the number of rows)?**

```
Hive>  select count(*) as number_of_trips from taxidata;
```

```
hive> set hive.cli.print.header =true;
hive> select count(*)as number_of_trips from taxidata ;
Query ID = cloudera_20220929111717_ac3ca37e-b7e8-441f-80a1-70dfbdb8d6d4
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_1664473883634_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:18:41,018 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:19:10,120 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 6.69 sec
2022-09-29 11:19:33,238 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 12.3 sec
MapReduce Total cumulative CPU time: 12 seconds 300 msec
Ended Job = job_1664473883634_0001
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 12.3 sec   HDFS Read: 1522584 HDFS Write: 6 SUCCESS
Total MapReduce CPU Time Spent: 12 seconds 300 msec
OK
number_of_trips
10000
Time taken: 119.113 seconds, Fetched: 1 row(s)
```

2. What is the total revenue generated by all the trips? The fare is stored in the column total_amount.

Hive> select round(sum(total_amount),2) as total_revenue from taxidata;

```
hive> select round(sum(total_amount),2) as  total_revenue from taxidata;
Query ID = cloudera_20220929113434_8fadf074-a5f7-4666-bf3e-f7be2e46b1a8
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_1664473883634_0003, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0003/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0003
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:34:45,858 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:35:05,686 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.72 sec
2022-09-29 11:35:31,097 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 15.51 sec
MapReduce Total cumulative CPU time: 15 seconds 510 msec
Ended Job = job_1664473883634_0003
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 15.51 sec   HDFS Read: 1523264 HDFS Write: 10 SUCCESS
Total MapReduce CPU Time Spent: 15 seconds 510 msec
OK
total_revenue
160546.81
Time taken: 67.138 seconds, Fetched: 1 row(s)
hive>
```

3. What fraction of the total is paid for tolls? The toll is stored in tolls_amount.

Hive> select round((sum(tolls_amount)/sum(total_amount)),2) as fraction from taxidata;

```
hive> select round((sum(tolls_amount)/sum(total_amount)),2) as fraction from taxidata;
Query ID = cloudera_20220929113636_cab82108-6802-4292-b03f-56a12858a846
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_1664473883634_0004, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0004/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0004
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:37:02,459 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:37:21,600 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.7 sec
2022-09-29 11:37:40,814 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.72 sec
MapReduce Total cumulative CPU time: 14 seconds 720 msec
Ended Job = job_1664473883634_0004
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 14.72 sec   HDFS Read: 1524269 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 720 msec
OK
fraction
0.02
Time taken: 62.221 seconds, Fetched: 1 row(s)
```

4. What fraction of it is driver tips? The tip is stored in tip_amount.

Hive> select round((sum(tip_amount)/sum(total_amount)),2) as fraction from taxidata;

```
hive> select round((sum(tip_amount)/sum(total_amount)),2) as fraction from taxidata;
Query ID = cloudera_20220929114444_e24a0491-5d74-4d83-8a23-419476c410e3
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_1664473883634_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:45:05,821 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:45:26,450 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.04 sec
2022-09-29 11:45:49,960 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.93 sec
MapReduce Total cumulative CPU time: 14 seconds 930 msec
Ended Job = job_1664473883634_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 14.93 sec   HDFS Read: 1524265 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 930 msec
OK
fraction
0.11
Time taken: 66.962 seconds, Fetched: 1 row(s)
```

5. What is the average trip amount?

Hive> select round(avg(tip_amount),2) as average_trip_amount from taxidata;

```
hive> select round(avg(tip_amount),2) as average_trip_amount from taxidata;
Query ID = cloudera_20220929115151_649c8ebc-342e-431d-9c08-254c5c0001ef
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_1664473883634_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:51:42,349 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:52:08,698 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 8.6 sec
2022-09-29 11:52:33,037 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 16.64 sec
MapReduce Total cumulative CPU time: 16 seconds 640 msec
Ended Job = job_1664473883634_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 16.64 sec   HDFS Read: 1523531 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 640 msec
OK
average_trip_amount
1.73
Time taken: 79.85 seconds, Fetched: 1 row(s)
```

6. What is the average distance of the trips? Distance is stored in the column trip_distance.

Hive> select round(avg(trip_distance),2) average_distance from taxidata;

```
hive> select round(avg(trip_distance),2)  average_distance from taxidata;
Query ID = cloudera_20220929115353_e97e468a-e1dc-4ce4-a5d3-9b32a85049b9
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_1664473883634_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 11:53:30,578 Stage-1 map = 0%,  reduce = 0%
2022-09-29 11:53:50,142 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.16 sec
2022-09-29 11:54:15,421 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.93 sec
MapReduce Total cumulative CPU time: 14 seconds 930 msec
Ended Job = job_1664473883634_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 14.93 sec   HDFS Read: 1523534 HDFS Write: 5 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 930 msec
OK
average_distance
3.25
Time taken: 69.47 seconds, Fetched: 1 row(s)
```

7. How many different payment types are used?

Hive> select distinct(payment_type) as payment_types from taxidata;

```
hive> select distinct(payment_type) as payment_types from taxidata;
Query ID = cloudera_20220929120606_7d8596b8-0093-4a2e-870f-06b6e1538dfc
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_1664473883634_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 12:07:14,734 Stage-1 map = 0%,  reduce = 0%
2022-09-29 12:07:47,032 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.93 sec
2022-09-29 12:08:10,002 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 14.42 sec
MapReduce Total cumulative CPU time: 14 seconds 420 msec
Ended Job = job_1664473883634_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 14.42 sec   HDFS Read: 1522409 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 14 seconds 420 msec
OK
payment_types
1
2
3
4
Time taken: 79.972 seconds, Fetched: 4 row(s)
```

8. For each payment type, display the following details:

- **Average fare generated**
- **Average tip**
- **Average tax – tax is stored in column mta_tax**

Hive> select payment_type, round(avg(fare_amount),2) as average_fare, round(avg(tip_amount),2) as average_tip, round(avg(mta_tax),2) as average_tax from taxidata group by payment_type;

```
Time taken: 0.245 seconds, Fetched: 5 row(s)
hive>  select Hour, round(avg(total_amount),2) as avg_revenue from (select hour(pickup_datetime) as Hour, total_amount from taxidata) a  group by Hour order by avg_revenue desc;
Query ID = cloudera_20220929121010_1c63e431-1004-4468-bd26-04dba212c5c1
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_1664473883634_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 12:10:25,040 Stage-1 map = 0%,  reduce = 0%
2022-09-29 12:10:55,288 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 11.94 sec
2022-09-29 12:11:30,081 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 19.91 sec
MapReduce Total cumulative CPU time: 20 seconds 890 msec
Ended Job = job_1664473883634_0012
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_1664473883634_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1664473883634_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-09-29 12:12:09,761 Stage-2 map = 0%,  reduce = 0%
2022-09-29 12:12:26,681 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 3.71 sec
2022-09-29 12:12:42,258 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 8.17 sec
MapReduce Total cumulative CPU time: 8 seconds 170 msec
Ended Job = job_1664473883634_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 20.89 sec   HDFS Read: 1522800 HDFS Write: 174 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 8.17 sec   HDFS Read: 4935 HDFS Write: 26 SUCCESS
Total MapReduce CPU Time Spent: 29 seconds 60 msec
OK
hour    avg_revenue
22      16.24
23      16.11
0       15.32
Time taken: 160.965 seconds, Fetched: 3 row(s)
```

9. On average which hour of the day generates the highest revenue?

Hive> select Hour, round(avg(total_amount),2) as avg_revenue from (select hour(pickup_datetime) as Hour, total_amount from taxidata) a group by Hour order by avg_revenue desc;

```
hive> select payment_type, round(avg(fare_amount),2) as average_fare, round(avg(
tip_amount),2) as average_tip, round(avg(mta_tax),2) as average_tax from taxidat
a group by payment_type;
Query ID = cloudera_20220929121010_aad279af-9506-4e33-ad05-c7f9abb074bf
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_1664473883634_0013, Tracking URL = http://quickstart.cloudera
:8088/proxy/application_1664473883634_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1664473883634_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-09-29 12:11:07,716 Stage-1 map = 0%,  reduce = 0%
2022-09-29 12:11:47,610 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 9.04 se
c
2022-09-29 12:12:10,428 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.05
 sec
MapReduce Total cumulative CPU time: 17 seconds 50 msec
Ended Job = job_1664473883634_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.05 sec   HDFS Read: 152509
2 HDFS Write: 65 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 50 msec
OK
1       13.56   2.7     0.5
2       11.39   0.0     0.5
3       13.21   0.0     0.42
4       12.22   0.0     0.5
Time taken: 108.284 seconds, Fetched: 4 row(s)
hive>
```