Queries for Hive Case study

Tasks:

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

   Create database mohita;
   Use mohita;

   CREATE TABLE IF NOT EXISTS taxidata (vendor_id string, pickup_datetime string, dropoff_datetime string, passenger_count int, trip_distance DECIMAL(9,6), pickup_longitude DECIMAL(9,6), pickup_latitude DECIMAL(9,6), rate_code int, store_and_fwd_flag string, dropoff_longitude DECIMAL(9,6), dropoff_latitude DECIMAL(9,6),

   payment_type string, fare_amount DECIMAL(9,6), extra DECIMAL(9,6), mta_tax DECIMAL(9,6), tip_amount DECIMAL(9,6), tolls_amount DECIMAL(9,6), total_amount DECIMAL(9,6), trip_time_in_secs int)

   ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' STORED as TEXTFILE TBLPROPERTIES ("skip.header.line.count"="1");

2. Load data from the csv file - yellow_tripdata_2015-01-06.csv LOAD DATA INPATH '/user/mohita' OVERWRITE INTO TABLE taxidata;

3. Run some basic queries to check the data is loaded properly.

Query: Select * from taxidata;

4. Run the queries required to answer the following questions.

Problem statement:

Use the above data to come up with answers to these questions:

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

Query: Select count(*) from taxidata;

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

```
Select sum(total_amount) as total_revenue from taxidata;
```

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

```
Select sum(tolls_amount)/sum(total_amount) as toll_pct from taxidata;
```

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

```
Select sum(tip_amount)/sum(total_amount) as tip_pct from taxidata;
```

5. What is the average trip amount?

```
Select avg(total_amount) as avg_tripamount from taxidata;
```

6. For each payment type, display the following details

   i. Average fare generated – fare amount is stored in fare_amount
   ii. Average tip
   iii. Average tax – tax is stored in column mta_tax

```
select payment_type,
avg(fare_amount) as average_fare,
avg(tip_amount) as average_tip,
avg(mta_tax) as average_tax,
from taxidata group by
payment_type;
```

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

```
select h24 as hour, avg(total_amount)
as avg_revenue

from (select hour(pickup_datetime) as h24, total_amount
from taxidata) ff
group by h24 order by
avg_revenue desc;
```

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

```
 select  avg(trip_distance) as
avg_distance from trips4;
```

9. How many different payment types are used? Column name – payment_type.

```
select distinct payment_type from taxidata;
```