Practice Project 1 - Yellow taxi trip analysis using Hive

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 (2018_Yellow_Taxi_Trip_Data.csv) having more than 15 columns and more than 100000 records in it. The dataset has different attributes like

1. vendor_id string,
2. pickup_datetime string,
3. dropoff_datetime string,
4. passenger_count int,
5. trip_distance DECIMAL(9,6),
6. pickup_longitude DECIMAL(9,6),
7. pickup_latitude DECIMAL(9,6),
8. rate_code int,
9. store_and_fwd_flag string,
10. dropoff_longitude DECIMAL(9,6),
11. dropoff_latitude DECIMAL(9,6),
12. payment_type string,
13. fare_amount DECIMAL(9,6),
14. extra DECIMAL(9,6),
15. mta_tax DECIMAL(9,6),
16. tip_amount DECIMAL(9,6),
17. tolls_amount DECIMAL(9,6),
18. total_amount DECIMAL(9,6),
19. trip_time_in_secs int

Perform taxi trip analysis by solving the questions below:

1. What is the total Number of trips (equal to the number of rows)?
2. What is the total revenue generated by all the trips? The fare is stored in the column total_amount.
3. What fraction of the total is paid for tolls? The toll is stored in tolls_amount.

4. What fraction of it is driver tips? The tip is stored in tip_amount.
5. What is the average trip amount?
6. What is the average distance of the trips? Distance is stored in the column trip_distance.
7. How many different payment types are used?
8. For each payment type, display the following details:

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

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