AWS Big Data Project

Problem statement:

Imagine that you are working as an analyst in a famous taxi app company. Your organization provides hassle-free travel to people all around the world. You are provided with an AWS account to perform certain analytical tasks using AWS cloud services.

Link for the dataset and resources:
https://intellipaat-course-attachments.s3.ap-south-1.amazonaws.com/Hadoop/Hadoop+Datasets-20200609T120700Z-001.zip

Dataset description:

In here, you have a predefined dataset (**yellow.csv**), having more than 15 columns.

The dataset has different attributes as follows:

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

Tasks:

- Stream the taxi data to a Kinesis Stream
- Link the Kinesis Stream to a Kinesis Firehose delivery stream
- Deliver the streamed data to Amazon S3
- Visualize the relationship between the total fare and the distance covered in trips
- Visualize the percentage of payment types using a pie chart
- Visualize the relationship between the passenger count and the tip amount

Use EMR to find out the following details:

1. What is the total number of trips?
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? (Toll fares are stored in tolls_amount)
4. What fraction of the total is paid as driver tips? (Tips are stored in tip_amount)
5. What is the average trip amount?
6. What is the average distance of the trips? (Distances are 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 average, which hour of the day generates the highest revenue?