Hadoop Spark Project

Problem Statement:

Imagine that you are working as an analyst for a famous Taxi App company. Your organization provides hassle-free travel to people all around the world. You have been provided with a Spark–Hadoop setup to perform certain analytical tasks.

Dataset:

https://intellipaat-course-attachments.s3.ap-south1.amazonaws.com/Hadoop/Hadoop+Datasets-20200609T120700Z-001.zip

Dataset Description:

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 To Be Performed:

Use Spark shell to perform the following tasks:

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 given as 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:
   a. Average fare generated
   b. Average tip
   c. Average tax – tax is stored in the column, mta_tax
9. On average, which hour of the day generates the highest revenue?