* **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  
  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  
     
  **Perform taxi trip analysis by solving the questions below:**  
  What is the total Number of trips ( equal to the number of rows)?  
  What is the total revenue generated by all the trips? The fare is stored in the column total\_amount.  
  What fraction of the total is paid for tolls? The toll is stored in tolls\_amount.  
  What fraction of it is driver tips? The tip is stored in tip\_amount.  
  What is the average trip amount?  
  What is the average distance of the trips? Distance is stored in the column trip\_distance.  
  How many different payment types are used?  
  For each payment type, display the following details:  
  Average fare generated  
  Average tip  
  Average tax – tax is stored in column mta\_tax  
  On average which hour of the day generates the highest revenue?
* **Answer**
* **Total number of trips (rows):**

SELECT COUNT(\*) AS total\_trips FROM taxidata;

Total revenue generated (sum of total\_amount):

* **Total revenue generated (sum of total\_amount):**

SELECT SUM(total\_amount) AS total\_revenue FROM taxidata;

Fraction of total revenue paid for tolls:

* **Fraction of total revenue paid for tolls:**

SELECT SUM(tolls\_amount) / SUM(total\_amount) AS tolls\_fraction FROM taxidata;

Fraction of total revenue paid as driver tips:

* **Fraction of total revenue paid as driver tips:**

SELECT SUM(tip\_amount) / SUM(total\_amount) AS tips\_fraction FROM taxidata;

Average trip amount:

* **Average trip amount:**

SELECT AVG(total\_amount) AS avg\_trip\_amount FROM taxidata;

Average trip distance:

* **Average trip distance:**

SELECT AVG(trip\_distance) AS avg\_trip\_distance FROM taxidata;

Number of different payment types:

* **Number of different payment types:**

SELECT COUNT(DISTINCT payment\_type) AS num\_payment\_types FROM taxidata;

For each payment type, average fare, tip, and tax:

* **For each payment type, average fare, tip, and tax**

SELECT payment\_type,

AVG(fare\_amount) AS avg\_fare,

AVG(tip\_amount) AS avg\_tip,

AVG(mta\_tax) AS avg\_tax

FROM taxidata

GROUP BY payment\_type;

Hour of the day that generates the highest average revenue:

* **Hour of the day that generates the highest average revenue:**

SELECT HOUR(TO\_TIMESTAMP(pickup\_datetime)) AS hour\_of\_day,

AVG(total\_amount) AS avg\_revenue

FROM taxidata

GROUP BY HOUR(TO\_TIMESTAMP(pickup\_datetime))

ORDER BY avg\_revenue DESC

LIMIT 1;