**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 (yellow\_tripdata\_2015-01-06.csv)** having more than 15 columns. 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 Float,
6. pickup\_longitude Float,
7. pickup\_latitude Float,
8. rate\_code int,
9. store\_and\_fwd\_flag string,
10. dropoff\_longitude Float,
11. dropoff\_latitude Float,
12. payment\_type string,
13. fare\_amount Float,
14. extra Float,
15. mta\_tax Float,
16. tip\_amount Float,
17. tolls\_amount Float,
18. total\_amount Float,
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

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