**NYC Yellow Taxi Case Study using HiveQL**

**Tasks:**

1. *Create a table named taxidata.*

Create database hive;

Use hive;

CREATE TABLE IF NOT EXISTS taxidata

(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 )

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

STORED as TEXTFILE TBLPROPERTIES ("skip.header.line.count"="1");

1. *Load data from the csv file. (yellow\_tripdata\_2015-01-06.csv)*

LOAD DATA INPATH '/user/hive OVERWRITE INTO TABLE taxidata;

1. *Run some basic queries to check the data is loaded properly.*

Select \* from taxidata;

1. *Run the queries required to answer the following questions.*

**Problem Statement:**

* What is the total Number of trips (equal to number of rows)?

Select count (\*) from taxidata;

* What is the total revenue generated by all the trips?

Select sum(total\_amount) as total\_revenue from taxidata;

* What fraction of the total is paid for tolls?

Select sum(tolls\_amount)/sum(total\_amount) as toll\_pct from taxidata;

* What fraction of it is driver tips?

Select sum(tip\_amount)/sum(total\_amount) as tip\_pct from taxidata;

* What is the average trip amount?

Select avg(total\_amount) as avg\_tripamount from taxidata;

* For each payment type, display the following details:

1. Average fare generated
2. Average tip
3. Average tax

select payment\_type,

avg(fare\_amount) as average\_fare,

avg(tip\_amount) as average\_tip,

avg(mta\_tax) as average\_tax,

from taxidata

group by payment\_type;

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

select h24 as hour,

avg(total\_amount) as avg\_revenue

from (select hour(pickup\_datetime) as h24,

total\_amount

from taxidata) ff

group by h24

order by avg\_revenue desc;

* What is the average distance of the trips?

select

avg(trip\_distance) as avg\_distance

from trips4;

* How many different payment types are used?

select distinct payment\_type from taxidata;