**CASE STUDY: NYC Taxi Data Analysis**

This case study is cantered on the concepts of Ingesting and Analysing Big Data on the APACHE-HIVE platform.

The dataset provided contains the detailed trip level data of trips made by taxis in New York City. Analysis is focused on the yellow taxis.

<https://drive.google.com/drive/folders/1YGQBagh9X__Y65THEzuZBfiy3-lV2JBE?usp=share_link>

**Data Description:**

* vendorid--A code indicating the TPEP provider that provided the record. 1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.
* tpep\_pickup\_timestamp--The date and time when the meter was engaged.
* tpep\_dropoff\_timestamp--The date and time when the meter was disengaged.
* passenger\_count--The number of passengers in the vehicle. This is a driver-entered value.
* trip\_distance--The elapsed trip distance in miles reported by the taximeter.
* rate\_code--The final rate code in effect at the end of the trip. 1= Standard rate 2=JFK 3=Newark 4=Nassau or Westchester 5=Negotiated fare 6=Group ride.
* store\_forward\_flag--This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka store and forward, because the vehicle did not have a connection to the server. Y= store and forward trip N= not a store and forward trip.
* pickup\_location--TLC Taxi Zone in which the taximeter was engaged.
* dropoff\_location--TLC Taxi Zone in which the taximeter was disengaged.
* payment\_type--A numeric code signifying how the passenger paid for the trip. 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip.
* fare\_charge--The time-and-distance fare calculated by the meter.
* extra\_charge--Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
* mta\_tax\_charge--$0.50 MTA tax that is automatically triggered based on the metered rate in use.
* tip\_amount--Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
* tolls\_charge--Total amount of all tolls paid in trip.
* improvement\_surcharge--$0.30 improvement surcharge assessed trips at the flag drop. The improvement surcharge began being levied in 2015.
* total\_charge--The total amount charged to passengers. It does not include cash tips.

USE Hive Queries to

1. Creating the intial data table titled nyc\_taxifare to be used for Preliminary Analysis.
2. Check if data table has been loaded successfully into the HIVE table.
3. Write a query that summarises the number of records of each provider.
4. Let's check if there are any records in which the pickup\_timestamp is after the dropoff\_timestamp. (use unix\_timestamp).
5. EDA of components associated with Trip Details from nyc\_taxifare.
6. EDA of components associated with Fare Details from nyc\_taxifare.
7. Since passenger\_count is an attribute registered by the driver it can be a source of Erroneous data. Write a query to analyse passenger\_count.
8. Checking the rate\_code parameter. Write a query to show records of each rate\_code.
9. Checking the payment\_type parameter and analyse the queries.
10. Checking the extra\_charges attribute and analyse the queries.
11. Checking MTA tax attribute and analyse the count of each group of mta\_tax\_charge.
12. Checking store\_forward\_flag parameter and analyse the count of each group of store\_foreward\_flag.
13. Checking if non-zero tip amount has been registed for cash payment trips.
14. Checking improvement\_surcharge other than $0.30 has been recorded.
15. Creating the orc\_taxifare table of ORC table format.
16. Compare the average fare\_charge for November and December.
17. Explore the ‘number of passengers per trip’ - how many trips are made by each level of ‘Passenger\_count’?

Do most people travel solo or with other people?

Let's have a look at how many trips are made by each level of passenger\_count.

1. Let's compare if the passengers prefer to travel solo [i.e, passenger\_count=1] or in groups [i.e, passenger\_count [2-6]]
2. Which is the most preferred mode of payment?
3. What is the average tip paid? Compare the average tip with the 25th, 50th and 75th percentiles and comment whether the ‘average tip’ is a representative statistic (of the central tendency) of ‘tip amount paid’.
4. Explore the ‘Extra’ (charge) variable - what is the fraction of total trips where an extra charge is levied?

Let us observe the extra\_charge attribute in a grouped table w.r.t number of records.

1. The number of trips where the extra\_charge was levied is marginally lower than the number of trips for which it was not.

Let us write a query to compare the Fraction of trips for which the extra\_charge was levied.

1. Create five buckets of ‘tip paid’: [0-5), [5-10), [10-15) , [15-20) and >=20. Calculate the percentage share of each bucket.
2. Which month has a greater average ‘speed’ - November or December?
3. Analyse the average speed of the most happening days of the year i.e. 31st December (New year’s eve) and 25th December (Christmas Eve) and compare it with the overall average.
4. Create partitioning of table using total charge as total\_charge category.