**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 for the months of November and December 2017.

<https://drive.google.com/drive/folders/1E3-5J-NXrV8Z7fKvXUncGPtOzL50RseJ?usp=sharing>

**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.

create table if not exists nyc\_taxifare(VendorID int,tpep\_pickup\_datetime varchar(20),tpep\_dropoff\_datetime varchar(20),passenger\_count int,trip\_distance float,RatecodeID int,store\_and\_fwd\_flag varchar(5),PULocationID int,DOLocationID int,payment\_type int,fare\_amount float,extra float,mta\_tax float,tip\_amount float,tolls\_amount int,improvement\_surcharge float,total\_amount float)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE

TBLPROPERTIES ("skip.header.line.count"="2");

**------------------------------------------------------------------------------------------**

1. Check if data table has been loaded successfully into the HIVE table.

load data local inpath '/home/hadoop/Downloads/casestudy/'

overwrite into table nyc\_taxifare;

**------------------------------------------------------------------------------------------**

1. Write a query that summarises the number of records of each provider.

A picture containing text

Description automatically generated

1. Let's check if there are any records in which the pickup\_timestamp is after the dropoff\_timestamp. (use unix\_timestamp).

Text

Description automatically generated

1. EDA of components associated with Trip Details from nyc\_taxifare.

A picture containing text, electronics

Description automatically generated

1. EDA of components associated with Fare Details from nyc\_taxifare.

A picture containing text, electronics, screenshot

Description automatically generated

1. 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.

A picture containing text

Description automatically generated

1. Checking the rate\_code parameter. Write a query to show records of each rate\_code.

Text

Description automatically generated

1. Checking the payment\_type parameter and analyse the queries.

Text

Description automatically generated

1. Checking the extra\_charges attribute and analyse the queries.

Text

Description automatically generated

Shape

Description automatically generated

1. Checking MTA tax attribute and analyse the count of each group of mta\_tax\_charge.

A screenshot of a computer

Description automatically generated with medium confidence

1. Checking store\_forward\_flag parameter and analyse the count of each group of store\_foreward\_flag.

Text

Description automatically generated

1. Checking if non-zero tip amount has been registed for cash payment trips.

A picture containing text

Description automatically generated

1. Checking improvement\_surcharge other than $0.30 has been recorded.

Text

Description automatically generated with medium confidence

1. Creating the orc\_parted\_taxifare table of ORC table format using TBLPROPERTIES ("orc.compress" = "SNAPPY").

set hive.exec.dynamic.partition=true;

set hive.exec.dynamic.partition.mode=nonstrict;

set hive.exec.max.dynamic.partitions.pernode=1000;

drop table if exists nyc\_taxifare\_orc;

create table if not exists nyc\_taxifare\_orc(VendorID int,tpep\_pickup\_datetime varchar(20),tpep\_dropoff\_datetime varchar(20),passenger\_count int,trip\_distance float,RatecodeID int,store\_and\_fwd\_flag varchar(5),PULocationID int,DOLocationID int,payment\_type int,fare\_amount float,extra float,mta\_tax float,tip\_amount float,tolls\_amount int,improvement\_surcharge float,total\_amount float)

partitioned by (mnth int,m\_day int)

row format delimited

fields terminated by ','

stored as orc

TBLPROPERTIES ("orc.compress"="SNAPPY");

1. Populating the orc\_parted\_taxifare PARTITION (mnth, m\_day) table.

insert into table nyc\_taxifare\_orc

partition(mnth,m\_day)

select VendorID,tpep\_pickup\_datetime,tpep\_dropoff\_datetime,passenger\_count,trip\_distance,RatecodeID,store\_and\_fwd\_flag,PULocationID,DOLocationID,payment\_type,fare\_amount,extra,mta\_tax,tip\_amount,tolls\_amount,improvement\_surcharge,total\_amount,month(tpep\_pickup\_datetime) as mnth,day(tpep\_pickup\_datetime) as m\_day from nyc\_taxifare;

**------------------------------------------------------------------------------------------**

1. Compare the average fare\_charge for November and December.

Text

Description automatically generated with medium confidence

1. Explore the ‘number of passengers per trip’ - how many trips are made by each level of ‘Passenger\_count’? Text

   Description automatically generated

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]]

A screenshot of a computer

Description automatically generated with medium confidence

1. Which is the most preferred mode of payment?

A picture containing text

Description automatically generated

1. 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’.

A picture containing text, electronics

Description automatically generated

1. 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.

Text

Description automatically generated with medium confidence

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.

A picture containing text

Description automatically generated

1. What is the correlation between the number of passengers and tip paid?

Text

Description automatically generated with medium confidence

1. Create five buckets of ‘tip paid’: [0-5), [5-10), [10-15) , [15-20) and >=20. Calculate the percentage share of each bucket.

A picture containing text, screenshot

Description automatically generated

1. Which month has a greater average ‘speed’ - November or December?

A picture containing graphical user interface

Description automatically generated

1. 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.

A screenshot of a computer

Description automatically generated with medium confidence