New York Taxi

Data engineering 2 - Project Report

Submitted By

Kartikeya Sharma (11018760)
Aditya Raj Singh (11018674)
Mohammadreza Yazdankhah (11026643)

Abstract:

The dataset contains records of taxi trips, capturing various attributes and measurements related to each trip. This dataset can be utilized for diverse analytical purposes within the transportation domain. It encompasses information such as pickup and drop-off timestamps, trip distances, fare amounts, payment types, and tipping amounts.

By leveraging this dataset, we can perform a wide range of analyses and derive valuable insights. For instance, transportation analysis can be conducted to understand transportation patterns, including the number of trips taken by different vendors, total distances traveled, and average fare amounts. This analysis can shed light on the demand for taxi services and help identify popular routes.

Furthermore, vendor performance evaluation can be carried out to assess the performance of different taxi service providers based on metrics like trip ratings, tip amounts, and customer feedback. This evaluation can assist in identifying high-performing vendors and implementing strategies to enhance overall service quality.

Pricing strategy optimization is another application of this dataset, where analyses can be performed to identify optimal pricing strategies based on factors such as trip distances, fare amounts, and payment types. This optimization can maximize revenue generation and improve customer satisfaction simultaneously.

Additionally, passenger behavior analysis can be conducted to understand passenger preferences, payment patterns, and tipping behavior. This analysis can facilitate customer segmentation, enabling personalized marketing initiatives and tailored service offerings.

Lastly, congestion analysis can be performed to identify areas or time periods with high traffic congestion based on attributes such as pickup and drop-off locations, trip distances, and congestion surcharges. This analysis can aid transportation authorities in implementing targeted congestion management strategies.

Overall, the dataset provides a rich source of information for in-depth analysis and exploration within the transportation domain, facilitating various applications such as demand forecasting, fraud detection, and route optimization.

Reason to choose this dataset

- **Rich and Diverse Data**: The dataset provides a wide range of data points that enable exploration and analysis across various domains within the transportation industry. It encompasses factors such as vendor performance, pricing strategies, passenger behavior, congestion, and more, allowing for comprehensive and multifaceted analysis.
- **Real-World Application**: Taxi services are an integral part of urban transportation systems, and analyzing taxi trip data can provide valuable insights into demand patterns, customer preferences, and operational efficiency. The dataset reflects real-world scenarios and can be used to derive practical solutions and improvements in the transportation sector.
- **Availability and Manageability**: The dataset is readily available and manageable for analysis purposes. It contains enough records to conduct meaningful analyses, yet it is not overly large, making it convenient to work with in terms of storage and processing requirements.
- **Learning and Exploration**: The dataset offers ample opportunities for learning and exploration in various analytical techniques and methodologies. It allows for the application of predictive modeling, statistical analysis, data visualization, and other data-driven approaches to gain insights and make informed decisions in the transportation domain.

Tools used:

Area	Tools
Database	MySQL workbench
Languages	Python
IDE	Visual Studio Code, Tableau prep Builder
Streaming Platform	Apache Kafka
Visualization	Tableau
Documentation	Microsoft Word Adobe PDF

Contents

Abstract: 2
Reason to choose this dataset 3
Tools used: 4
List of Figures 6
Github URL: 7
Data set: 7
Introduction 7
Data Source 7
Scope and Coverage 7
Data Collection and Availability 8
Data Privacy and Anonymization 8
Potential Applications 8
Solution: 9
Introduction 9
Streaming Data Flow 9
2.1 Data Retrieval 9
2.2 Kafka Integration 10
2.3 MySQL Database: 10
2.4 Tableau Visualization: 11
Batch Data Flow 12
3.1 Data Acquisition 12
3.2 Tableau Prep Builder: 13
3.3 MySQL Database: 13
3.4 Tableau Visualization: 14
Chapter 5: 15
Summary: 15
Outlook and Future Work: 15
Bibliography: 17
Youtube: 17
Websites: 17

List of Figures

Figure 1 - Data Source Scope and Coverage 7
Figure 2 - Lambda Architecture 9
Figure 3 - Connecting to NYC API 10
Figure 4 - Starting ZooKeeper 10
Figure 5 - Starting Kafka 10
Figure 6 - MySQL Database Overview 1 11
Figure 7 - Live Location 12
Figure 8 - Data Acquisition 12
Figure 9 - Tableau Prep Builder 13
Figure 10 - MySQL Database Overview 2 13
Figure 11 - Yellow Taxi Dashboard 14
Figure 12 - Green Taxi Dashboard 14

Github URL:

https://github.com/kartik03091991/Data_Engineering_2.git

Data set:

Introduction

The NYC Green and Yellow Taxi dataset is a comprehensive collection of transportation data that captures detailed information about taxi rides in New York City. This dataset provides a valuable resource for researchers, analysts, and policymakers to gain insights into various aspects of the city's taxi services, including travel patterns, passenger demographics, fare structures, and environmental impact.

Data Source

The NYC Green and Yellow Taxi dataset is sourced from the New York City Taxi and Limousine Commission (TLC). The TLC is the agency responsible for regulating the city's taxi industry, ensuring passenger safety, and overseeing the operation of licensed taxis. As part of its regulatory functions, the TLC collects and maintains a rich dataset containing information on each taxi trip taken within the city.

Scope and Coverage

The dataset covers both green and yellow taxis, which are two distinct categories of taxis operating in New York City. Green taxis primarily serve the boroughs outside of Manhattan, while yellow taxis predominantly operate within Manhattan. The dataset encompasses a wide range of attributes associated with each taxi trip, including:

VendorID	lpep_pickup_datetime	lpep_dropoff_datetime	store_and_fwd_flag	RatecodeID	PULocationID	DOLocationID	passenger_count	trip_distance	fare_amount	extra	mta_tax	tip_amount	tolls_amount	ehail_fee	imp
2	12/1/2022 0:32	12/1/2022 0:37	N	1	166	24	2	0.77	5.5	0.5	0.5	1	0		
1	12/1/2022 0:26	12/1/2022 0:31	N	1	74	41	1	0.6	5	0.5	0.5	0	0		
1	12/1/2022 0:20	12/1/2022 0:49	N	1	260	17	1	0	22.2	0	0.5	0	0		
2	12/1/2022 0:20	12/1/2022 0:28	N	1	80	256	1	1.71	8	0.5	0.5	0	0		
2	12/1/2022 0:09	12/1/2022 0:13	N	1	179	179	1	0.62	5	0.5	0.5	4	0		
2	12/1/2022 0:27	12/1/2022 0:34	N	1	74	262	1	2.62	9.5	0.5	0.5	3.39	0		
2	12/1/2022 0:51	12/1/2022 1:00	N	1	74	263	1	2.66	10	0.5	0.5	0	0		
2	12/1/2022 0:29	12/1/2022 0:50	N	1	134	222	1	9.07	28.5	0.5	0.5	0	0		
2	12/1/2022 0:14	12/1/2022 0:23	N	1	92	70	1	2.44	9.5	0.5	0.5	0	0		
2	12/1/2022 0:20	12/1/2022 0:41	N	1	95	193	1	9.23	28	0.5	0.5	0	0		
2	12/1/2022 0:14	12/1/2022 0:27	N	1	244	143	1	5.48	16.5	0.5	0.5	5	0		
2	12/1/2022 0:45	12/1/2022 1:14	N	1	166	146	1	6.69	25	0.5	0.5	7.26	0		
2	12/1/2022 0:10	12/1/2022 0:17	N	1	168	168	1	1.33	7	0.5	0.5	1	0		
2	12/1/2022 0:20	12/1/2022 0:20	N	5	260	260	1	0	10	0	0	2.58	0		
2	11/30/2022 23:59	12/1/2022 0:04	N	1	181	181	5	0.81	5	0.5	0.5	1.58	0		
2	12/1/2022 0:15	12/1/2022 0:20	N	1	129	82	1	0.85	5.5	0.5	0.5	0	0		
2	12/1/2022 0:30	12/1/2022 1:13	N	1	129	28	2	8.31	34.5	0.5	0.5	0	0		
2	11/30/2022 23:08	11/30/2022 23:24	N	1	80	140	5	5.72	18	0.5	0.5	5.51	0		
2	12/1/2022 0:38	12/1/2022 0:44	N	1	260	83	5	0.98	6	0.5	0.5	0	0		
1	12/1/2022 0:10	12/1/2022 0:19	N	1	92	56	1	2.3	9	3.25	0.5	0	0		
1	12/1/2022 0:30	12/1/2022 0:31	N	5	92	92	2	0.1	0	0	0	10	0		

Figure 1 - Data Source Scope and Coverage

a) Trip Details: The dataset includes information about the pickup and drop-off locations, trip distance, trip duration, and the number of passengers.

b) Timestamps: The dataset captures the precise timing of each trip, allowing for temporal analysis and understanding of travel patterns across different times of the day, days of the week, and seasons.

c) Fare Information: The dataset contains data related to fare calculation, including base fare, distance-based fare, time-based fare, tolls, surcharges, and tips. This information can be utilized to examine fare structures and fluctuations.

d) Payment Types: The dataset includes the types of payment used for each trip, such as credit card, cash, or other forms of payment.

f) Geographical Information: The dataset provides latitude and longitude coordinates for the pickup and drop-off locations, enabling spatial analysis and visualization of taxi routes and hotspots.

Data Collection and Availability

The TLC collects the taxi data through its electronic trip record (ETR) system, which is installed in all licensed taxis. The ETR system records trip data in real-time, capturing relevant information as the journey progresses. The collected data is stored in a structured format, allowing for easy extraction and analysis.

The NYC Green and Yellow Taxi dataset is publicly available and accessible through the TLC's website or the NYC Open Data portal. The dataset is updated on a regular basis, usually on a monthly or quarterly basis, ensuring that users have access to the most recent information for their analysis.

Data Privacy and Anonymization

To protect passenger privacy, the TLC follows strict protocols for data anonymization. Personally, identifiable information (PII) such as passenger names, addresses, and contact details are removed from the dataset. Furthermore, specific measures are taken to aggregate and generalize location data, ensuring that individual trips cannot be easily linked to identifiable individuals.

Potential Applications

The NYC Green and Yellow Taxi dataset offers numerous opportunities for analysis and research. Some potential applications of the dataset include:

Transportation Planning: The dataset can aid in identifying travel patterns, demand hotspots, and transportation bottlenecks, assisting in the formulation of effective transportation policies and infrastructure planning.

Fare Analysis: By analysing fare structures, tipping behaviour, and payment trends, researchers can gain insights into the economic aspects of taxi services and explore fare optimization strategies.

Environmental Impact: The dataset can be utilized to evaluate the environmental impact of taxi services, such as carbon emissions and fuel consumption, enabling policymakers

Solution:

Introduction

This report presents a comprehensive analysis of the data flow process for the NYC Green and Yellow Taxi dataset. The dataset is utilized for two distinct data processing approaches: streaming data flow and batch data flow. The streaming data flow involves real-time data retrieval from the NYC API and subsequent storage and visualization, while the batch data flow involves processing pre-existing data files obtained from the NYC website. This report discusses the implementation details and tools used for each approach, highlighting their strengths and limitations.

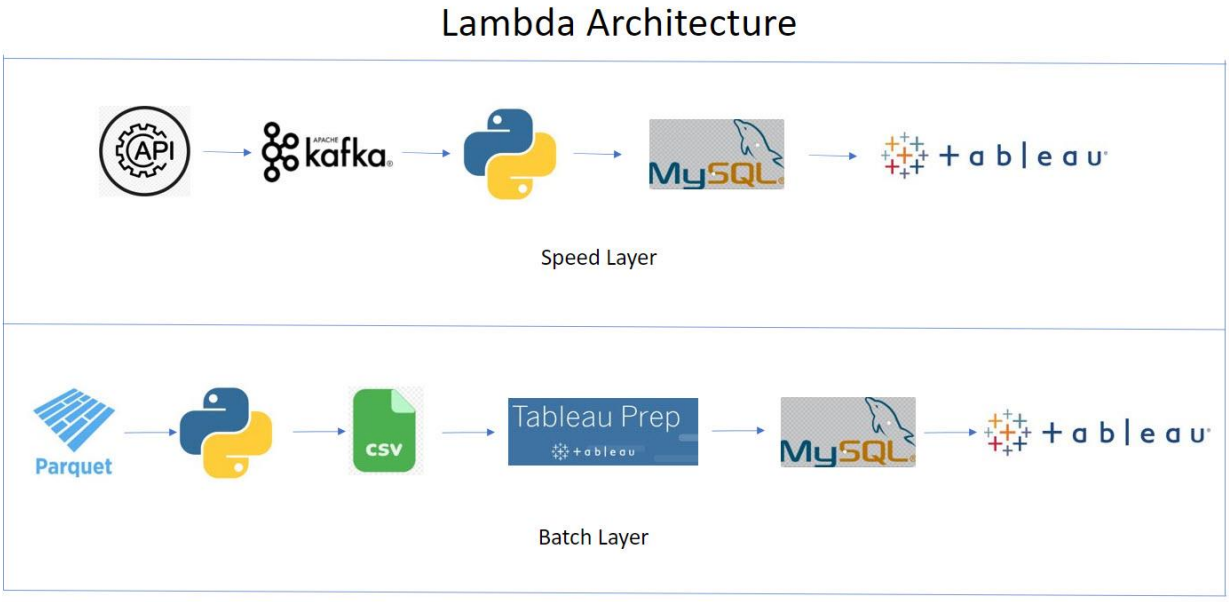

Figure 2 - Lambda Architecture

Streaming Data Flow

2.1 Data Retrieval:

To retrieve real-time data from the NYC Green and Yellow Taxi dataset, the NYC API is utilized. Python programming language is used in conjunction with libraries such as requests to make API calls and obtain the desired data. The received data is then converted into a format suitable for processing and storage.

```
from time import sleep
from json import dumps
from kafka import KafkaProducer
import pandas as pd
from kafka  import KafkaConsumer
import pandas as pd
from sodapy import Socrata
import csv

topic_name='test1'
producer = KafkaProducer(bootstrap_servers=['localhost:9092']) # ,value_serializer=lambda x: dumps(x).encode('utf-8'))
client = Socrata(domain  = "data.cityofnewyork.us",
                 app_token = "[redacted]",
                 username="[redacted]",
                 password="[redacted]")

results = client.get("djnb-wcxt", limit=500)
results_df = pd.DataFrame.from_records(results)
#results_df.fillna(value=None, inplace=True)
```

Figure 3 - Connecting to NYC API

2.2 Kafka Integration:

Kafka, a distributed streaming platform, is utilized as a buffer storage solution for the streaming data flow. Python code, implemented in Visual Studio Code (VSCode), is used to create Kafka producers that receive the converted data and pass it onto Kafka topics. Kafka consumers are also implemented to retrieve data from Kafka topics for further processing.

```
PS D:\DE2\kafka> .\bin\windows\zookeeper-server-start.bat .\config\zookeeper.properties
```

Figure 4 - Starting ZooKeeper

```
PS D:\DE2\kafka> .\bin\windows\kafka-server-start.bat .\config\server.properties
```

Figure 5 - Starting Kafka

2.3 MySQL Database:

The data retrieved from the Kafka topics is stored in a MySQL database. MySQL is a popular open-source relational database management system known for its performance and scalability. Python is used to establish a connection with the MySQL database, and SQL queries are executed to create tables and store the received data.

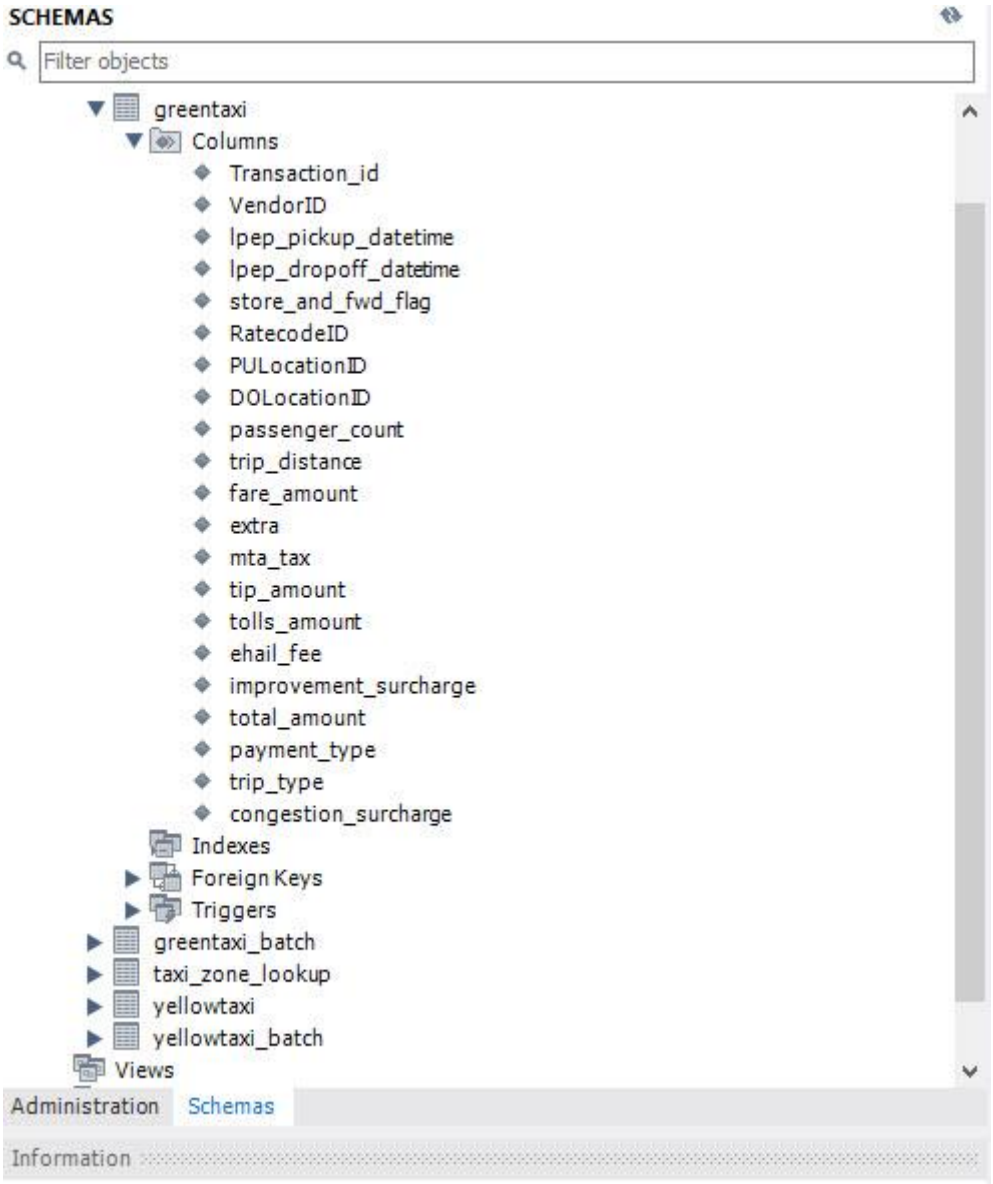

Figure 6 - MySQL Database Overview 1

2.4 Tableau Visualization:

Tableau Desktop is utilized as a visualization tool to explore and analyse the data stored in the MySQL database. Tableau connects to the database as a data source, enabling users to create interactive dashboards, visualizations, and reports.

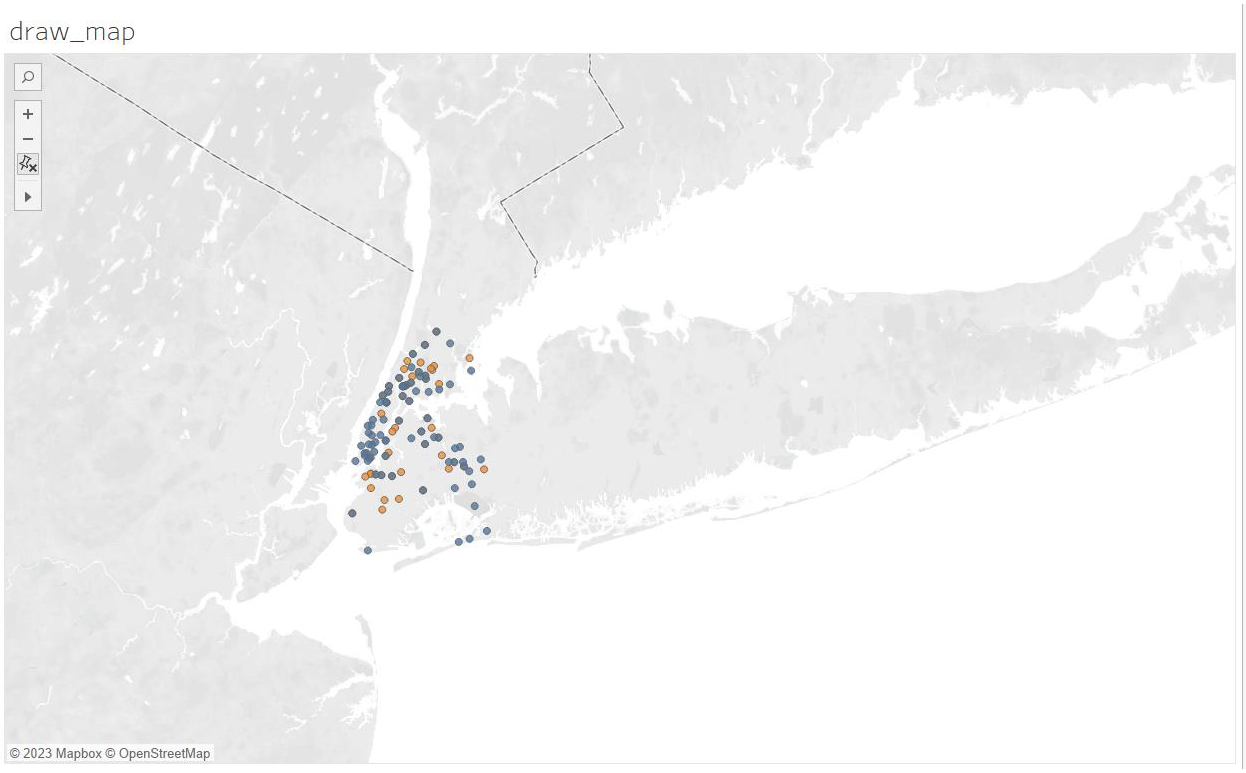

Figure 7 - Live Location

Batch Data Flow

3.1 Data Acquisition:

For the batch data flow, pre-existing data files in the Parquet format are obtained from the NYC website. Python's pandas library is used to read the Parquet files and convert them into CSV files, which are easier to work with in subsequent processing steps.

```
from kafka import KafkaProducer, KafkaConsumer
import mysql.connector

consumer = KafkaConsumer( 'test1', bootstrap_servers=['localhost:9092'])

# Consume data from Kafka
for message in consumer:
    # Convert byte object to string using UTF-8 encoding
    string_object = message.value.decode('utf-8')
    Data = string_object.split(',')
    print(Data)
    #print(Data[0])
    VendorID = None if Data[0] == 'nan' else Data[0]
    lpep_pickup_datetime = None if Data[1] == 'nan' else Data[1]
    lpep_dropoff_datetime = None if Data[2] == 'nan' else Data[2]
    store_and_fwd_flag = None if Data[3] == 'nan' else Data[3]
    RatecodeID = None if Data[4] == 'nan' else Data[4]
    PULocationID = None if Data[5] == 'nan' else Data[5]
    DOLocationID = None if Data[6] == 'nan' else Data[6]
    passenger_count = None if Data[7] == 'nan' else Data[7]
    trip_distance = None if Data[8] == 'nan' else Data[8]
    fare_amount = None if Data[9] == 'nan' else Data[9]
    extra = None if Data[10] == 'nan' else Data[10]
    mta_tax = None if Data[11] == 'nan' else Data[11]
    tip_amount = None if Data[12] == 'nan' else Data[12]
    tolls_amount = None if Data[13] == 'nan' else Data[13]
    improvement_surcharge = None if Data[14] == 'nan' else Data[14]
    total_amount = None if Data[15] == 'nan' else Data[15]
    payment_type = None if Data[16] == 'nan' else Data[16]
    trip_type = None if Data[17] == 'nan' else Data[17]
    congestion_surcharge = None if Data[18] == 'nan' else Data[18]



    print(VendorID,lpep_pickup_datetime,lpep_dropoff_datetime,store_and_fwd_flag,RatecodeID,PULocationID,DOLocationID,passenger_count,trip_distance,fare_amount,extra,mta_tax,ti
#consumer.close()
```

Figure 8 - Data Acquisition

3.2 Tableau Prep Builder:

Tableau Prep Builder is employed to perform data preparation tasks for the batch data. The CSV files obtained in the previous step are imported into Tableau Prep Builder, where various data transformations, cleaning, and shaping operations can be applied. This ensures that the data is in an optimal format for visualization.

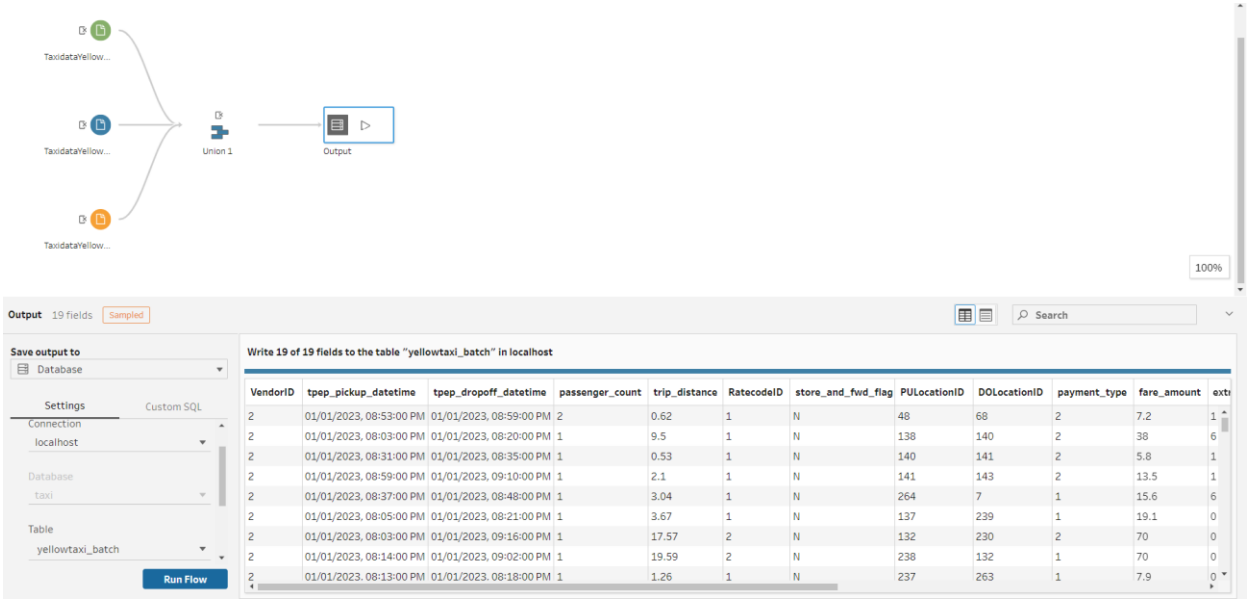

Figure 9 - Tableau Prep Builder

3.3 MySQL Database:

Similar to the streaming data flow, a MySQL database is utilized to store the processed data from Tableau Prep Builder. The data is structured into tables, and SQL commands are executed to create the necessary schema and insert the data into the corresponding tables.

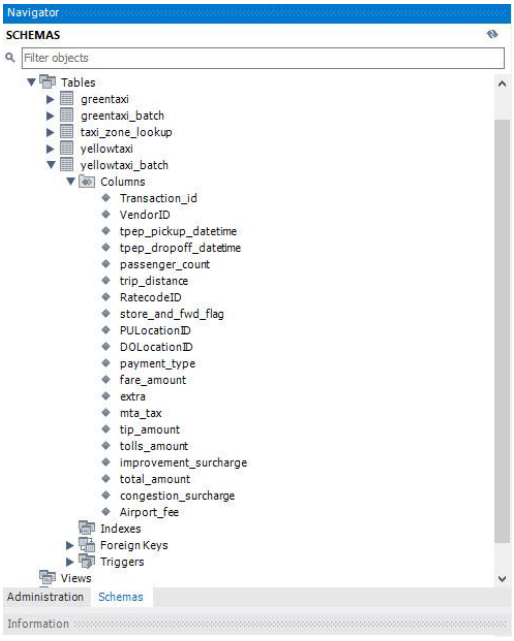

Figure 10 - MySQL Database Overview 2

3.4 Tableau Visualization:

Tableau Desktop is employed once again to connect to the MySQL database and create visualizations based on the processed batch data. The rich set of visualization options in Tableau allows for in-depth analysis and reporting of the insights derived from the dataset.

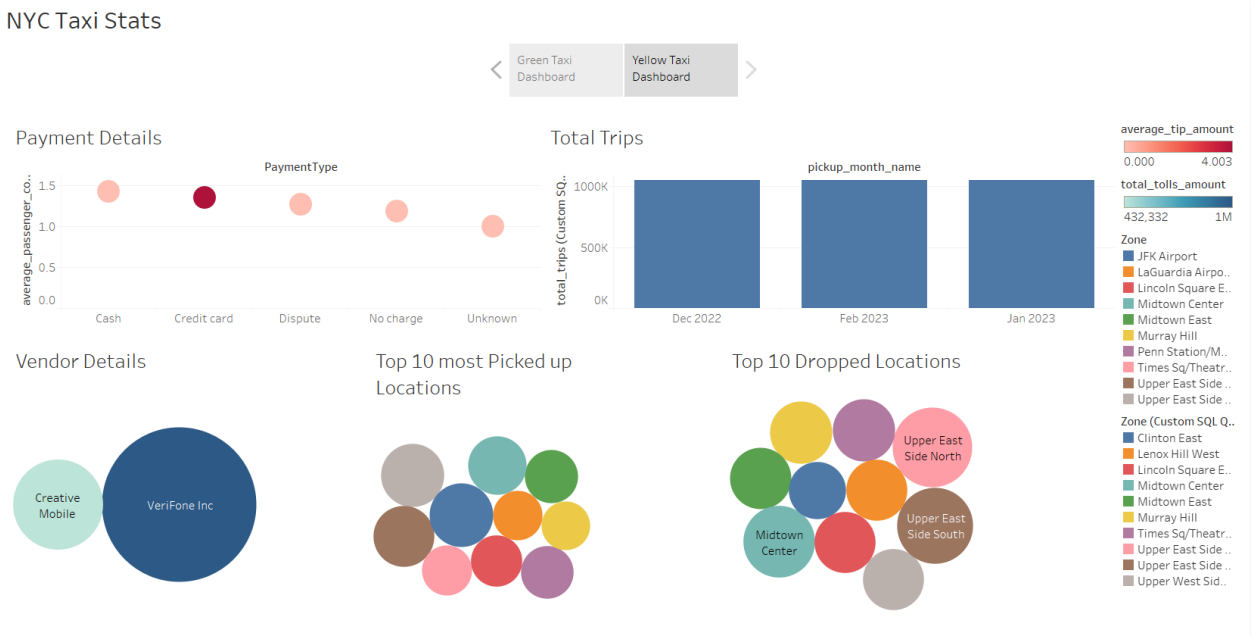

Figure 11 - Yellow Taxi Dashboard

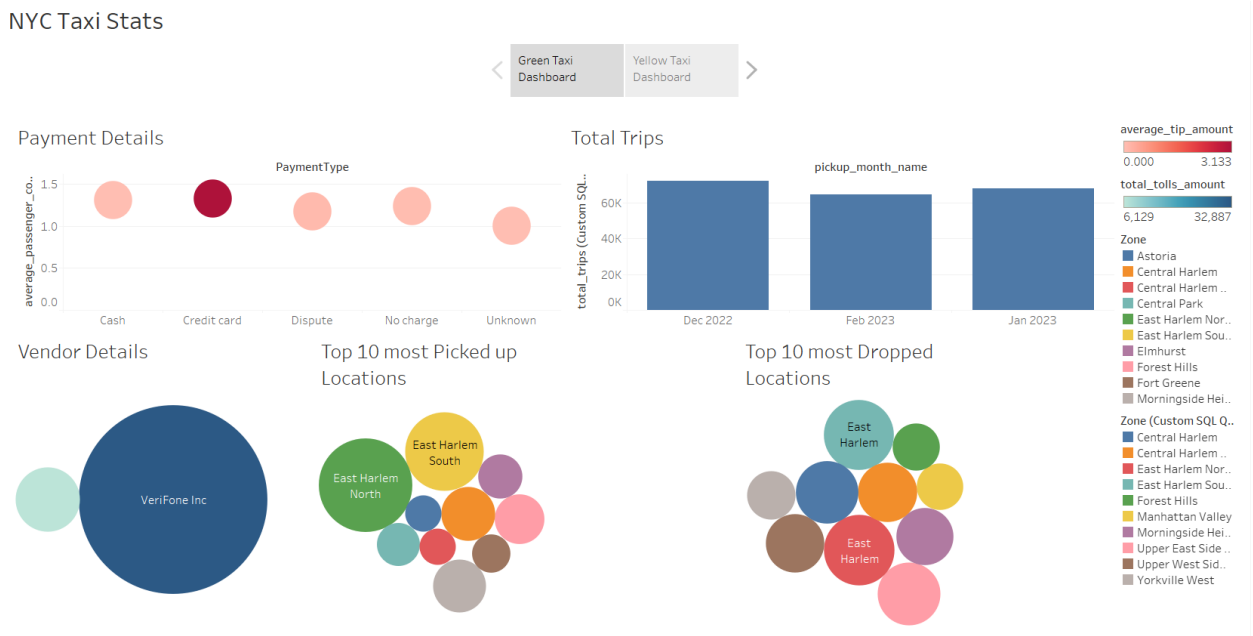

Figure 12 - Green Taxi Dashboard

Chapter 5:

Summary:

In this project, we explored the data flow process for the NYC Green and Yellow Taxi dataset, utilizing both streaming and batch processing approaches. The streaming data flow involved real-time data retrieval from the NYC API, storage in Kafka, and visualization in Tableau Desktop using a MySQL database. The batch data flow included acquiring Parquet files, conversion to CSV, data preparation in Tableau Prep Builder, storage in MySQL, and visualization in Tableau Desktop.

By implementing these data flow processes, we were able to gain valuable insights into the taxi services in NYC. We analysed travel patterns, fare structures, and passenger demographics, providing a comprehensive understanding of the taxi industry's dynamics. The visualizations created in Tableau Desktop facilitated clear and intuitive representations of the data, enabling stakeholders to make informed decisions and take necessary actions.

Outlook and Future Work:

While this project has provided valuable insights into the NYC Green and Yellow Taxi dataset, there are several areas for future exploration and improvement:

- Real-time Analytics: Enhancing the streaming data flow by incorporating real-time analytics techniques would enable immediate analysis and visualization of incoming data. This could provide more timely insights into taxi usage patterns, demand fluctuations, and route optimizations.

- Machine Learning Models: Integrating machine learning models into the data flow process can enable predictive analytics and forecasting. By training models on historical taxi data, we can make predictions about future travel patterns, fare estimations, and even optimize taxi availability based on demand.

- Advanced Data Cleaning and Transformation: Expanding the data preparation steps in Tableau Prep Builder to include more advanced cleaning and transformation techniques can help handle noisy and inconsistent data. This would ensure that the final visualization and analysis are based on high-quality and reliable information.

- Data Governance and Security: Addressing data governance and security concerns is crucial when dealing with sensitive transportation data. Implementing appropriate data access

controls, anonymization techniques, and ensuring compliance with data privacy regulations should be a priority for future work.

- Integration with Other Data Sources: Incorporating additional datasets, such as weather data, public events, or public transportation schedules, can provide a more holistic understanding of the factors influencing taxi usage. This integration can lead to more comprehensive analysis and predictive models.

In conclusion, this project has successfully demonstrated the data flow process for the NYC Green and Yellow Taxi dataset, highlighting its potential for analysing travel patterns, fare structures, and passenger demographics. The future work outlined above presents exciting opportunities for further exploration and improvement, expanding the scope and depth of insights derived from the dataset. By continually refining the data flow process and leveraging advanced analytical techniques, we can contribute to more efficient and informed decision-making in the field of transportation and urban planning.

Bibliography:

Youtube:

[1] “Install Apache Kafka on Windows PC | Kafka Installation Step-By-Step Guide” uploaded by AmpCode

URL: https://youtu.be/BwYFuhVhshI

[2] “How to install MySQL 8.0.30 Server and Workbench latest version on Windows 10” uploaded by Amit Thinks

URL: https://www.youtube.com/watch?v=2c2fUOgZMmY

[3] “What is Zookeeper and how is it working with Apache Kafka?” uploaded by Conduktor

URL: https://www.youtube.com/watch?v=t0FDmj4kaIg

Websites:

[4] Apache Kafka

URL: https://kafka.apache.org/