University of Pretoria
School of Information Technology
Private Bag x 20, Hatfield, 0028, Pretoria, South Africa

Assignment 2: MapReduce and Visualisation

MIT805: Big Data

Author: Boikanyo Radiokana (16097492)

Lecturer: Stacey Baror

09 October 2023

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

Contents

1 Brief Overview of the Dataset

The data collected for the big data study is a collection of daily taxi trip records generated from meters that are installed in taxis operating in New York City. The data is created and authorized by a technology provider, Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and provided to the NYC Taxi and Limousine Commission (TLC) [1]. TLC collects and analyses the data to assist with routine reviews of prices and the adoption of necessary enforcement actions to ensure accurate and complete trip records [1]. An extract from January 2019 to June 2020, spanning 1 year and 6 months, specifically for Yellow Taxis, in CSV format, with a size of 9.36 gigabytes (GB) and containing 101.24 million records consisting of 18 columns, will be used for analysis. The extract can be found on Kaggle as well as the official TLC website [2]. The dataset is characterised by its large volume, high velocity, diverse variety and high veracity as detailed in Appendix A, and therefore can be considered big data.

2 Value of the Dataset

TLC currently faces challenges related to its taxi trip data, which can be used to generate value through the adoption of a big data framework. One of the main challenges is to reduce the time taxi drivers spend on the road without passengers. By leveraging the data collected, TLC can use the insights effectively to improve their taxi operations, pricing structures, traffic congestion management and customer service.

3 Proposed Big Data Framework

For the purpose of analysing big data and generating actionable insights, a big data framework is required. Some of the prominent open-source frameworks developed by Apache include Hadoop and Spark. While both Hadoop and Spark employ parallel processing to distribute tasks across a network of computers (clusters), they have some distinct properties.

The Hadoop ecosystem is primarily comprised of four components. It uses an HDFS (Hadoop File Distributed System) to store data across clusters, which are then batch processed using a MapReduce algorithm [3]. The HDFS systems employs redundancy by making copies of the data across the clusters, making it fault tolerant [4]. MapReduce is used to split tasks into smaller ones. Computing resources are allocated and managed through YARN (Yet Another Resource Negotiator), which may require manual intervention for optimisation [3]. All the above components are dependent on the Hadoop Core to access the necessary libraries [4].

Hadoop is complex to configure and requires knowledge of the Java Programming language, therefore, it is a steep learning curve, especially for data scientists who are mostly familiar with Python. Since Hadoop relies on disk storage, it performs slower than Spark, however, it is more cost-effective [3].

Spark was developed to address the limitations of MapReduce [3]. It uses RAM for in memory-computations, making it faster. It offers several API languages such as PySpark (Python), Java, R and Scala, making it accessible and easy to use. In an organisational context, the dataset used in this assignment will require real-time processing to gather real-time insights such as traffic congestion to optimise routes and dynamic pricing based on demand.

For this assignment, Spark is chosen as the Big Data Framework to be adopted due to its fast performance, ease of use, real-time processing capabilities, automated provisioning of resources and flexibility to integrate into cloud services. The framework will be implemented on Google Collab, while the data will be stored on Google Drive because of seamless integration across the platforms and access to data. The link to the data can be found here.

4 MapReduce

As previously mentioned in Section 2, the framework to be employed for implementing MapReduce is Spark, utilizing the PySpark API. The TLC dataset partitions are initially loaded into a Spark DataFrame, preprocessed to incorporate the lookup tables and extract the relevant for columns for analysis. The data distributed across the clusters is converted to Resilient Distrubuted Datasets (RDD) [5], which is required before applying the MapReduce algoritm. RDDs are partitioned across the different clusters and transformed using the Map, Shuffle and Reduce operations respectively to aggregate the dataset.

To efficiently distribute the RDDs and manage execution of tasks, Spark uses a master, work and client nodes. The master node is responsible for coordinating the spark execution, whilst the worker nodes are responsible for executing computational tasks that are assigned by the master node [6]. The client nodes are responsible for loading the data into the various clusters [6].

The Mapper function transforms the RDD by splitting the input and processes it to generate a key-value pair as an output. Each key is assigned a value of 1. Sequentially, the key-value pair undergoes a shuffling process where the data is redistributed across the partitions to group or sort. The final step is the Reducer, which aggregates the data with the same key-value pair and produces a reduced output dataset.

Various features were used to apply the MapReduce algorithm to generate data-insights that will be useful to address the challenges experienced by TLC. The link to the github repository can be found here.

5 Visualisation & Analysis of Results

The following visuals were created using the output from the MapReduce algorithm. The output from the algorithm was exported to CSV files and imported to PowerBI to create the visuals. Figure 1 (left) depicts a geographical and graphical representation to illustrate the routes between towns. Figure 1 (right) shows the top 10 town routes with the highest number of taxi trips. From the figure on the right , it can be observed that majority (85 Million - 84%) of the trips have a Pickup Location as Manhattan and the destination as Manhattan (Manhattan-Manhattan). Another observation that can be made from the graphs is that the top trips either have Manhattan as a destination or as a pickup location. This is reflective of the fact that the town is densely populated and is the economic hub for New York City. This is also indicated on the figure on the left, with Manhattan having the largest node size.

The figure on the right also indicates that 1M trips have both locations captured as "Unknown-Unknown", which can cause serious challenges from an audit and passenger security perspective. The location is an important aspect to evaluate traffic congestion and can potentially introduce challenges when providing passengers with trip information for disputes or inquiries. It is important that the meters installed within the taxis are regularly maintained to ensure that accurate information is collected at all times enabling generation of reliable insights.

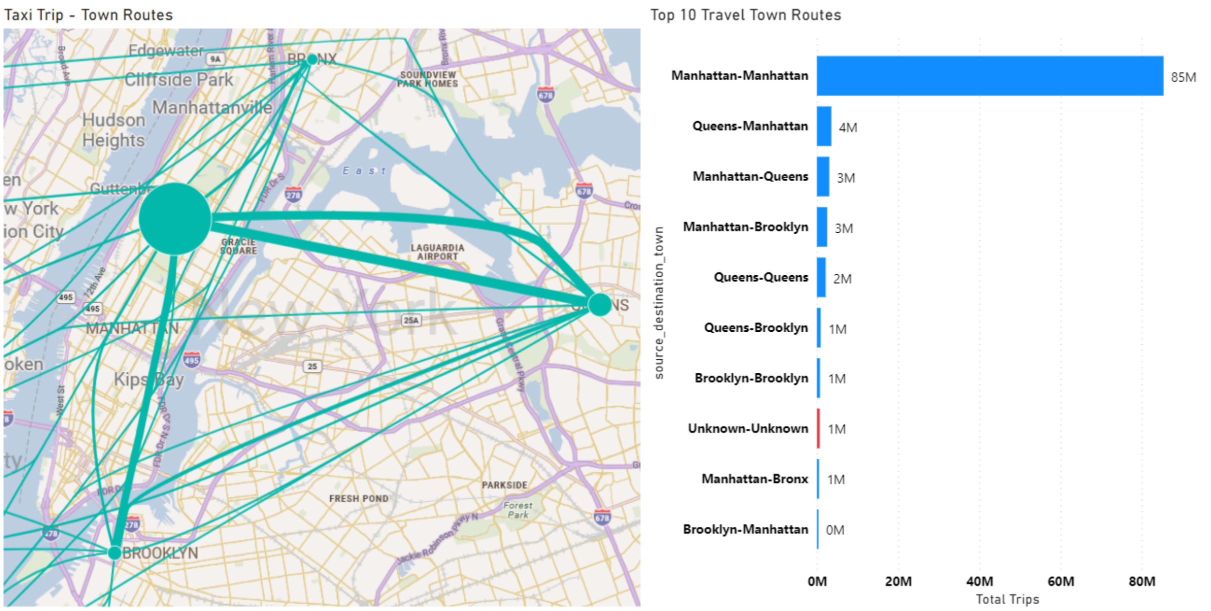

Figure 1: Town Routes

Figure 2 illustrates the number of trips per day of the week, with the trend for the years. From the graph, it can be observed that the trip volumes increase from Monday to Thursday and decrease from Friday to Sunday. This may be due to the to the fact that during the week people are using private transportation like Yellow taxis to get to work on time. The decline observed from Friday through to the weekend could be attributed to individuals opting for alternative modes of transportation, such as utilizing public transit systems.

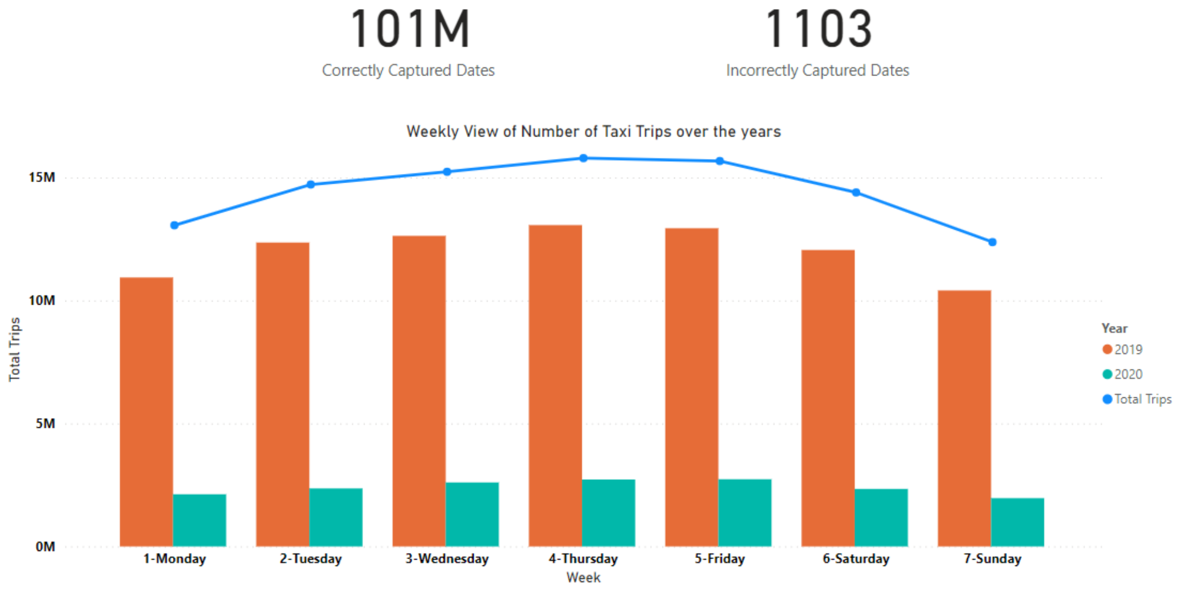

Figure 2: Town Routes

Furthermore, it can be observed in Figure 3 that the peak hours for trip requests are between 8am and 6pm, which are considered as the prime working hours. The trend is similar for the both years, which suggests that there is a consistent high demand for taxis during working hours.

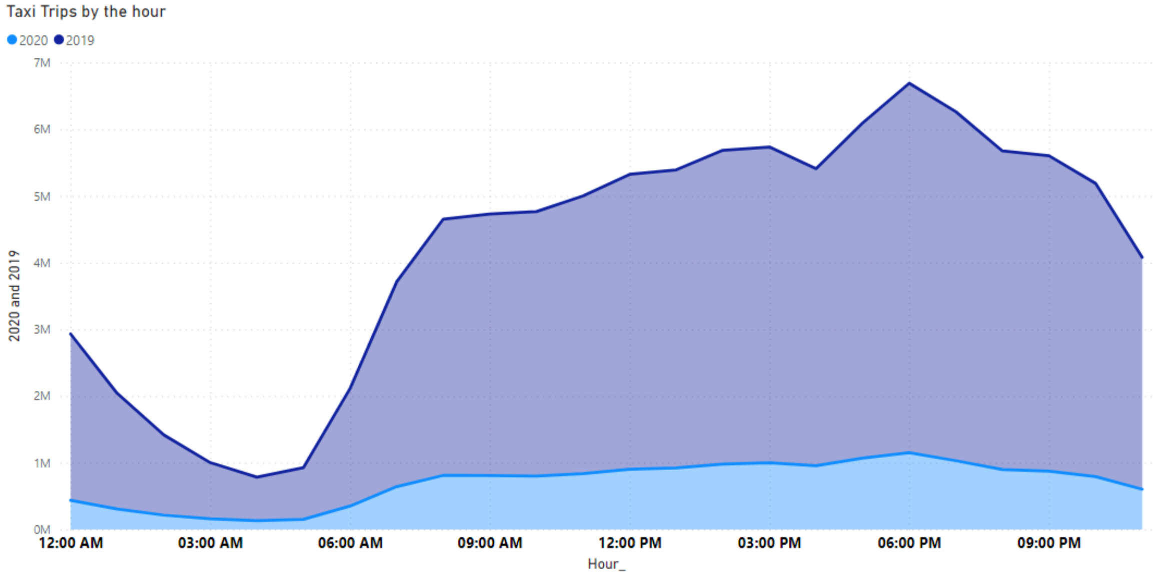

Figure 3: Hourly Analysis

From the pie chart in figure 4, it can be observed that 98.63% of the trip records were not held on the meter's memory before sending it to the respective vendor for processing. This indicates that most vehicles are able to connect to the server when taking trips, which enables the data to be processed in real-time.

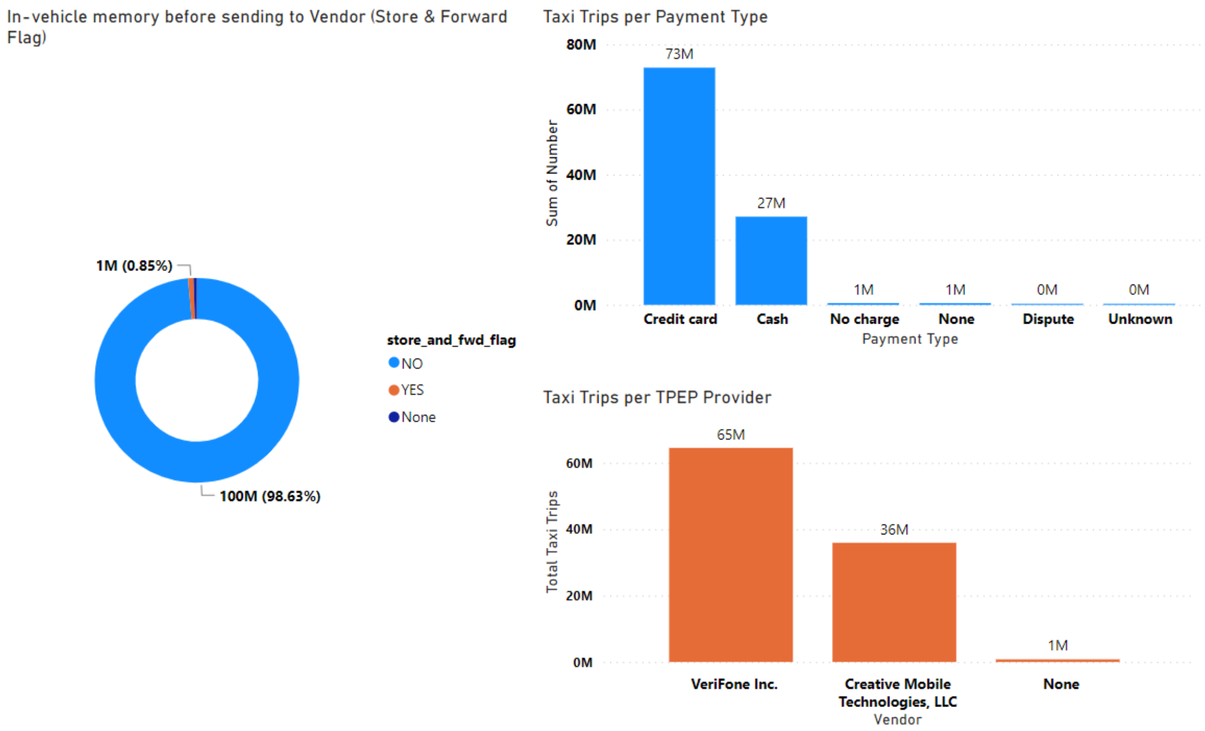

Figure 4: Town Routes

The top-right chart in Figure 4 indicates that most passengers prefer using credit cards instead of cash to pay for trips. This insight highlights the importance for the organisation to maintain secure and efficient payment systems to achieve a high rate of successfully payments.

The bottom-right graph in Figure 4 indicates that TPEP prefers payments to be processed by VeriFone Inc. instead of by Creative Mobile Technologies. This can indicate the high level of trust established with this Vendor to process electronic payments.

6 Conclusion

In this study the New York City taxi trip records spanning 1 year, 6 months, totaling 101.24 million entries were analysed. Leveraging Apache Spark and the MapReduce algorithm for processing, insights were drawn, revealing Manhattan's central role in taxi routes, credit card payment preferences, and trust in VeriFone Inc. for electronic payments. Trip volumes peaked during working hours and decreased over weekends. This data-driven approach offers solutions for reducing idle time for taxi drivers, optimizing pricing, managing traffic congestion, and enhancing customer service, contributing to more efficient taxi operations in the city. For TCP as an organization, embracing big data frameworks like Spark provides real-time processing and scalability, fostering cross-functional collaboration and cost efficiency. This strategic adoption is essential for TCP to remain competitive and maximize the value of its data assets.

References

[1] N. T. . L. Commission. "TLC Trip Record Data.", 2023. URL https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] S. Mohanasundaram. "Newyork Yellow Taxi Trip Data.", 2022. URL https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019?select=yellow_tripdata_2020-06.csv.

[3] IBM-Cloud-Education. "Hadoop vs. Spark: What's the Difference?", 2021. URL https://www.ibm.com/blog/hadoop-vs-spark/.

[4] AWS. "What's the Difference Between Hadoop and Spark?", 2023. URL https://aws.amazon.com/compare/the-difference-between-hadoop-vs-spark/.

[5] R. Zadeh. "Distributed Computing with Spark and MapReduce.", 2023. URL https://stanford.edu/~rezab/dao/notes/Intro_Spark.pdf.

[6] Databricks. "Hadoop Cluster.", 2023. URL https://www.databricks.com/glossary/hadoop-cluster.

University of Pretoria
School of Information Technology
Private Bag x 20, Hatfield, 0028, Pretoria, South Africa

Assignment 1: Data Decision, Collection & Process

MIT805: Big Data

Author: Boikanyo Radiokana (16097492)

Lecturer: Stacey Baror

18 August 2023

Contents

1 Overview of Dataset

The data collected for the big data study was created and authorized by a technology provider, Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) and provided to the NYC Taxi and Limousine Commission (TLC) [1]. The dataset is a collection of daily taxi trip records generated from meters that are installed in the taxis. Only one field is manually captured by the driver. TLC collects, updates and uploads monthly extracts of daily trip records based on the type of vehicles (Yellow, Green, For-Hire Vehicle, and High Volume For-Hire Vehicle) and published to the public. TLC collects and analyses the data to assist with routine reviews of prices and the adoption of necessary enforcement actions to ensure accurate and complete trip records [1]. The data is also used by the organisation to reduce traffic congestion and time spent by taxi drivers on the road while they are not earning any money [2]. The data is extracted from Kaggle, however, it can also be found on the official TLC website [3].

For the purpose of this study, an extract from January 2019 to June 2020 specifically ”NYC Yellow Taxis” in CSV format and size of 9,36 GigaBytes (GB) will be utilised. The data is partitioned by vehicletype-year-month folders, which offers an advantage from a data management and processing perspective. Partitioned data significantly improves system performance and allows fast and efficient retrieval, as well as parallel processing [4].

2 Technical Aspects of the Dataset

2.1 Volume

In big data, volume refers to large amounts of data that cannot be processed or queried using traditional databases such as relational databases utilising SQL [5].

The size of the NYC Yellow Taxi extract is 9.36 GB, which is made up of 101.4 Million rows and 18 columns over a period of 1 year and 6 months. The graph below shows the records generated per month over the specified period. It is important to note the significant decline in records from February 2020, due to the lockdown as a result of the Covid Pandemic shown in Figure 1 below.

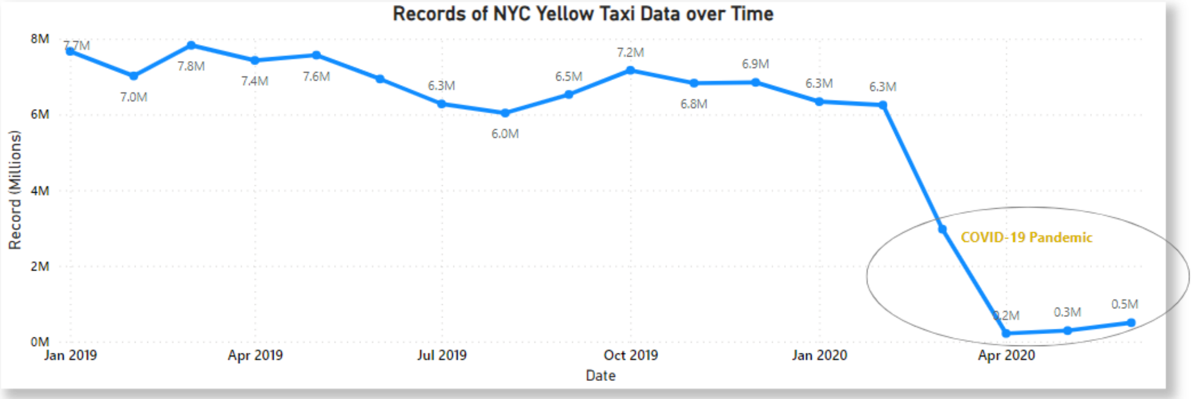

Figure 1: NYC Yellow Taxi Records

TLC has over 200 000 licensed vehicles (Yellow, Green, For-Hire Vehicle, and High Volume For-Hire Vehicle) that offer e-hailing services in New York City, and completing approximately 1 Million trips per day [1]. This can also be translated to approximately 1 Million records per day, which requires secure and scalable storage, capable of handling large volumes of data.

In addition to the high-volume raw data trips that are published on a monthly basis, TLC continues to improve their data hub to ensure that they are still able to make over 10 years' worth of data available to the public [6].

Over the years, TLC updates the data structure by adding columns which ultimately increases the volume of the existing ”high-volume raw data” over time [6].

2.2 Veracity

Veracity refers to the integrity, quality, accuracy and meaningfulness of the data [5]. Table 1 shows the columns that have null values, which can be considered as ”noise” and consequently result in records that are not meaningful. It can be observed from the table that the number of null values is 246 601 for all the columns excluding *congestion_surcharge*.

The null records only contribute 0.3% of the total number of records. 0.3% is immaterial and therefore, these records can be dropped to enhance the quality of the results derived from the dataset.

column_name	**nulls**	**% of records**	**method of correction**
rate_id	246 601	0.3%	drop records
vendor_id	246 601	0.3%	drop records
congestion_surcharge	119 536	0.1%	populate with mean
passenger count	246 601	0.3%	drop records
payment_type	246 601	0.3%	drop records
store_and_fwd_flag	246601	0.3%	drop records
tip_amount	246 601	0.3%	drop records

Table 1: Data Quality Checks

Out of a total of 18 columns, 17 columns are generated by the meter installed in the taxi, whilst the *passenger_count* is populated by the driver. Human-populated fields are prone to errors which can negatively affect the accuracy of the data. However, the quality of *passenger_count* clearly conforms to the integer datatype because all records, excluding the nulls, are integers.

All the records consistently conform to the datatypes of the specific columns. Overall, the data has high veracity and can be used to provide meaningful insights.

2.3 Velocity

Velocity in big data not only refers to the speed of the incoming data but also considers how fast the incoming data is processed and analysed to enable decision-making [5].

The data set has two timestamp columns which can be used to determine the velocity at which the data is generated and processed. From the data, a time difference of a few minutes to seconds, and milliseconds is observed between the records. Other records have the same pickup timestamps, which indicates that different drivers can collect passengers at the same time.

The rate at which the data has to be processed also suggests that it has to be near real-time because the pricing depends on the drop-off location and the number of passengers. TLC also has a smartphone functionality that enables the booking of trips and offers upfront pricing, suggesting the need for near real-time processing. From these observations, we can conclude that the data has a high velocity.

2.4 Variety

In big data, the various types of data collected from the source are referred to as variety. The data collected can be structured, unstructured or semi-structured in various formats [5].

The data is generated by three different vendors, namely: Creative Mobile Technologies, LLC, and VeriFone Inc. The data is made available in both CSV and JSON format which can be accessed using an API. The metadata is made available in a PDF file (unstructured), whilst the Taxi Zone lookup table for the location ID is stored in a CSV file (structured).

An expansion of the Taxi Zone Maps which show the Location Ids on a map are stored in a JPG file. Overall, the data has a large variety of data types including both structured and unstructured data.

2.5 Value

The value of the data is determined by the insights derived from the data to make decisions and thus bring value to an organisation [5]. The collected data is currently being used by TLC to monitor traffic congestion, and evaluate and review taxi rates, driver pay and hours [2]. TLC uses the data to analyse traffic congestion to reduce passenger travel time which can improve customer service.

To ensure compliance with city rules, TLC uses the data to monitor drivers who drive taxis around with no passengers and adopts the necessary enforcement actions [6].

For future work, the data can also be used to predict the estimated time of arrival considering the traffic congestion on the different routes and thus improving the customer journey.

3 Predicted Correlations and Relationships

The following relationships and correlations are expected within the data:

- Correlation between Pick-up/Drop-off Location and Fare: Certain locations have higher fares due to traffic and distance.
- Trip Duration and Distance: The longer the trip, the larger the distance.
- Correlation between the Volume of Trips and Time:
  - Larger trip volumes are expected during the day as compared to the night.
  - Larger trip volumes are expected during recess or festive seasons.
- Payments types are correlated to the location.
- Most trips are paid using cash and not credit cards.
- Correlation between the tip and the time of the day, most customers pay tips at night as compared to during the day.
- Rate: Negotiated fare is mostly done on Cash payments.
- Trips with the "store and forward" flag have a longer trip duration. Trips with no network connectivity take longer.

4 Conclusion

The NYC Yellow Taxi dataset was analysed based on 5 V's (Volume, Velocity, Variety, Veracity Value). From the analysis, using an evidence-based approach, it was demonstrated that data has high volume, high velocity, and large variety, it is reliable and can be used to generate value for the City of New York taxi industry. Therefore, the data collected can be classified as big data.

References

[1] N. T. . L. Commission. "TLC Trip Record Data.", 2023. URL https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[2] N. T. . L. Commission. "New York City Taxi and Limousine Commission 2019 Annual Report.", 2020. URL https://www.nyc.gov/assets/tlc/downloads/pdf/annual_report_2019.pdf.

[3] S. Mohanasundaram. "Newyork Yellow Taxi Trip Data.", 2022. URL https://www.kaggle.com/datasets/microize/newyork-yellow-taxi-trip-data-2020-2019?select=yellow_tripdata_2020-06.csv.

[4] P. Search. "Data Patitioning.", 2023. URL https://www.polymersearch.com/glossary/data-partitioning.

[5] M. F. Uddin, N. Gupta, et al. "Seven V's of Big Data understanding Big Data to extract value." In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education*, pp. 1–5. IEEE, 2014.

[6] N. T. . L. Commission. "New York City Taxi and Limousine Commission 2022 Annual Report.", 2023. URL https://www.nyc.gov/assets/tlc/downloads/pdf/annual_report_2022.pdf.