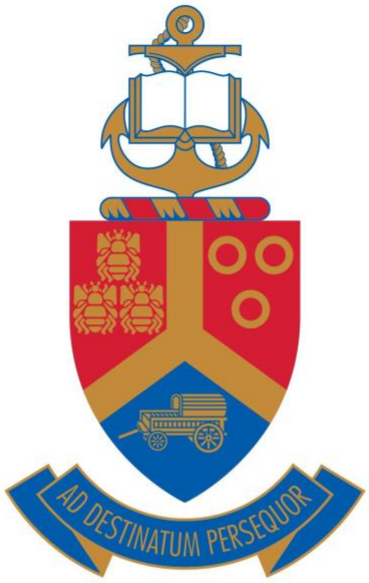
MIT 805

Big Data

Assignment1 part 2



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Student Number: u23970911

Name: Mapamela Wendy

Date: 09 October 2023

Table of Contents

1. Introduction

1.1. Overview of the data: 311 Call Center Inquiry

The dataset used for this assignment is collected by New York City (NYC) through the 311 call center, where residents seek non-emergency municipal services or report issues . The data is available on NYC Open data portal, that is used to handle NYC government information and data. The data is collected from residents through 311 call center for the purpose of delivering efficient and high quality service to the NYC residents. The dataset is 19.11 gigabytes and consists of 92.8 million records classified into 9 columns which will be used in this assignment for analysis. The data is collected from 2010 to date and gets updated daily. Based on the V's (volume, velocity, veracity, validity, variety and value) discussed in part 1 of this assignment (see appendix 1 for detailed information), this data can be considered big data.

1.2. Value of the dataset

The NYC 311 call center generates big data and from an organisational perspective, the main problem is finding a way to effectively leverage this data to improve the city's service delivery and operational efficiency. Through the use of an appropriate big data framework the City can track which agencies receive most calls and if those inquiries are resolved, this can help them optimise their systems and potentially reduce delays in resolutions.

2. Proposed Big Data Framework

To conduct an analysis of large datasets, a big data framework is required. Apache Hadoop and Apache Spark are both distributed data processing frameworks capable of handling big data. Hadoop enables parallel analysis of datasets by clustering multiple computers to enhance processing speed [1]. The framework consists of four modules namely; Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), Map Reduce and Hadoop Common. HDFS is a distributed file system that runs on standard hardware, it has a high fault tolerance and provides better throughput compared to traditional systems [1]. YARN oversees cluster nodes and schedules jobs and tasks [1]. Map Reduce enables parallel computation on data. The data is converted into key value pairs and output is reduced to provide desired outcome [1]. Hadoop common consists of Java libraries. Even though Hadoop is excellent at handling big data, it has its drawbacks; the framework runs on Java and its interface is not user friendly this may be a steeper learning curve. Due to map reduce jobs being slower on Hadoop, Apache Spark was introduced [2].

Spark can be integrated into python and its high-level API's make it more user friendly [2]. Due to its in-memory processing engine it performs better than Hadoop. The framework also allows multiple workloads to be integrated seamlessly and it is equipped with a rich set of libraries [2]. For the purposes of analysis for this assignment, PySpark library was used. This library is built into python and runs on a java virtual machine. PySpark's integration with python allowed the use of google colab which is integrated with google drive. Google drive was used as a cloud storage (Data can be found here: https://drive.google.com/file/d/1dHU1Dhu5jpG9Yvbky7txQpHcKEOMesd5/view), its seamless integration with Google colab made it easier to store, access and analyze the data. The integration also enhanced efficiency eliminating the need for local data transfers.

3. Methodology

To perform map reduce, initially data is stored in a file and the file serves as primary storage for the data [2]. The data is then picked by a Java virtual machine that is running on Spark and distributed across the nodes of Spark cluster for parallel processing. Spark then introduces Resilient Distributed Datasets(RDDs) that are responsible for managing & processing distributed data [2]. Dataset in RDD is then broken down into smaller parts and computed on different nodes of the cluster.

The Spark architecture consists of a master node, cluster manager and worker nodes [3]. The master node acts as a driver and works with the cluster manager to split and distribute jobs to worker nodes [4]. Worker nodes are responsible for executing the tasks assigned by worker nodes of processing the information [4]. During the mapping phase transformation is applied to the data to assign key value pair, after mapping reduce function is then used to aggregate keys and combine data from different partitions.

4. Results and Visualisation

Map Reduce algorithm was used in order to analyse and derive insights from the NYC call center dataset, the code can be found here: https://github.com/SalizwaMapamela/MIT-805-Assignment-1-Part-2 and data was visualised using PowerBI. Figure 1 below represents top agencies with the most inquiries as well as top resolutions provided for the inquiries. It can be observed that there are 17.8 million records that are not assigned to any agency, this information was not removed during the pre-processing stage. This is important to understand

why these were not allocated as they may be the main reason for resolution backlogs in the organisation. This may also be due to incorrect data handling processes which need to be investigated further to ensure efficiency and reduce backlog. The 20.22% of unresolved inquiries may be as a result of lack of accountability from the organisation since some of the inquiries were not allocated to the relevant agency. It can also be observed on figure 2 and figure 3 that the most common inquiries are Parking ticket lookup, this is a request for information and 30.9% resolution for information provided may be attributed to this.

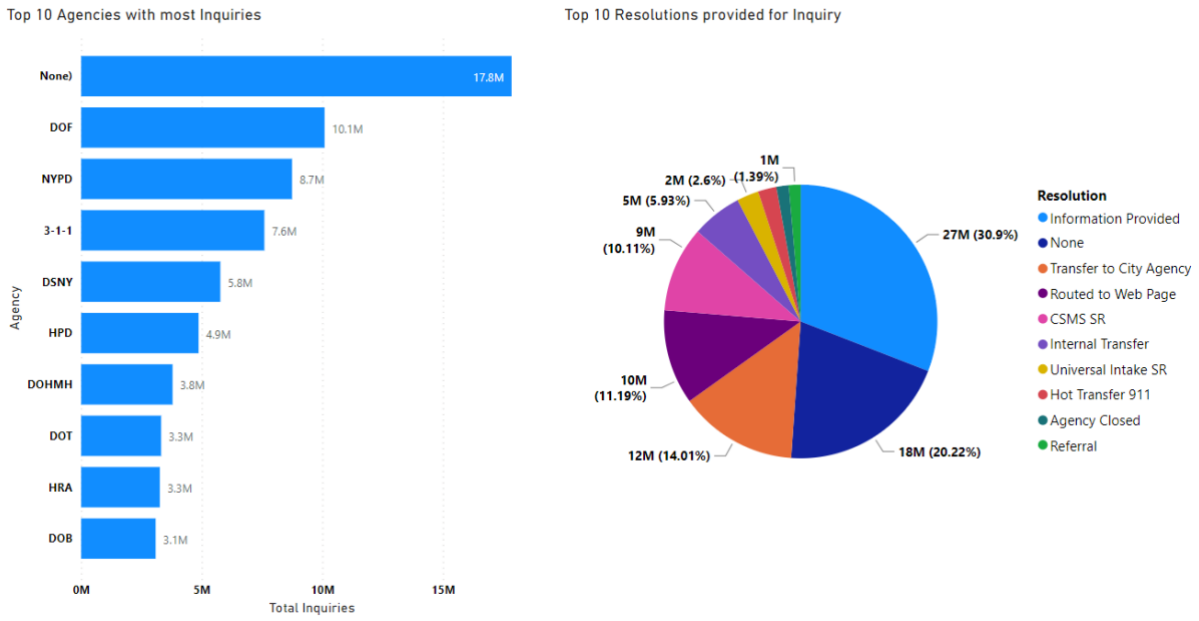


Figure 1:Agencies with the most enquiries and Top resolutions provided

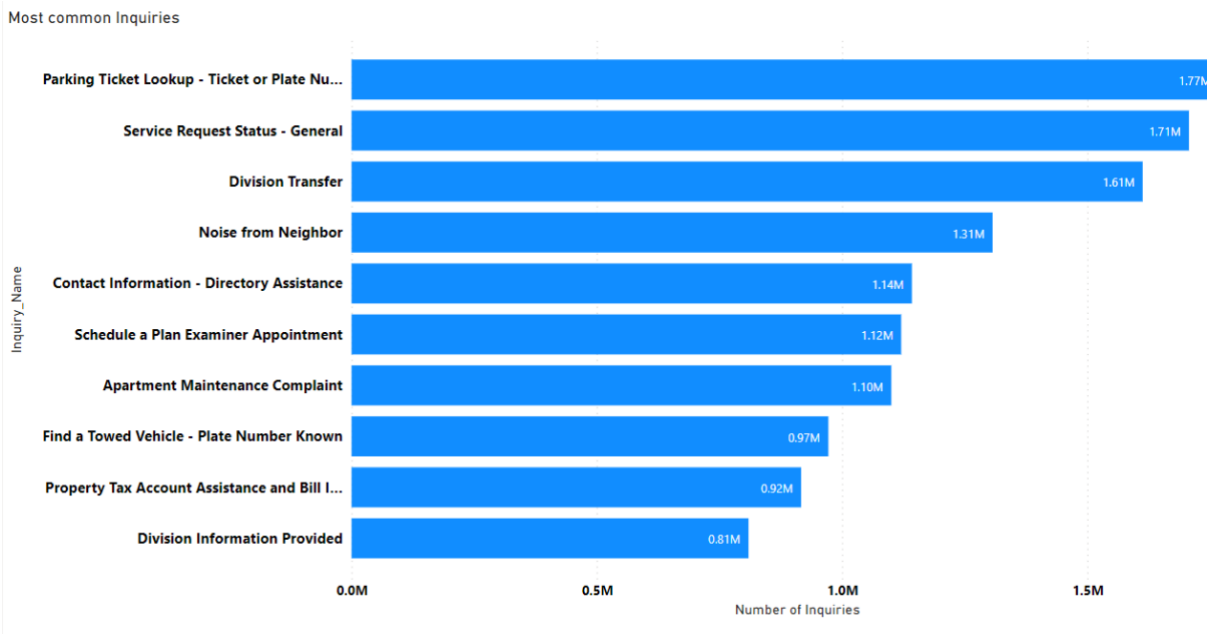


Figure 2: Most common inquiries

Figure 3: Most common inquiries

The original dataset was then explored to understand relationships and insights that may not have been visible through Map Reduce or chosen as main features. It can be observed on figure 4 that the number of call center inquiries has significantly plummeted of the years, this may due to some online systems that have been introduced by NYC to lodge complaints or request information.

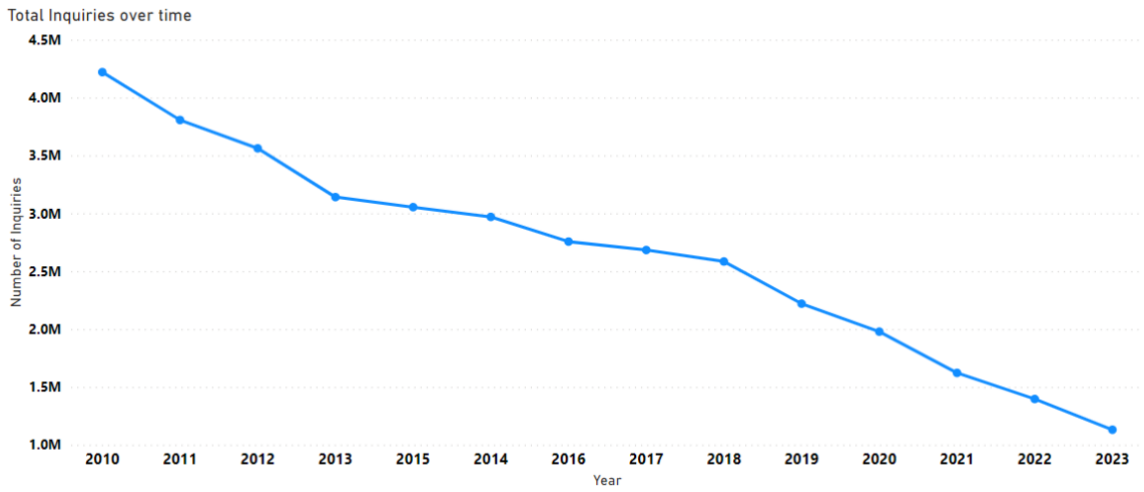


Figure 4: Total inquiries over years

Figure 5 below shows the number of inquiries received throughout the day, it can be observed that the most number of call were received between 9am and 3pm. This may be due to peak business hours when residents are actively seeking assistance and support. In order to ensure efficiency and seamless service delivery, the City should consider allocating more resources to the call center during this time-interval.

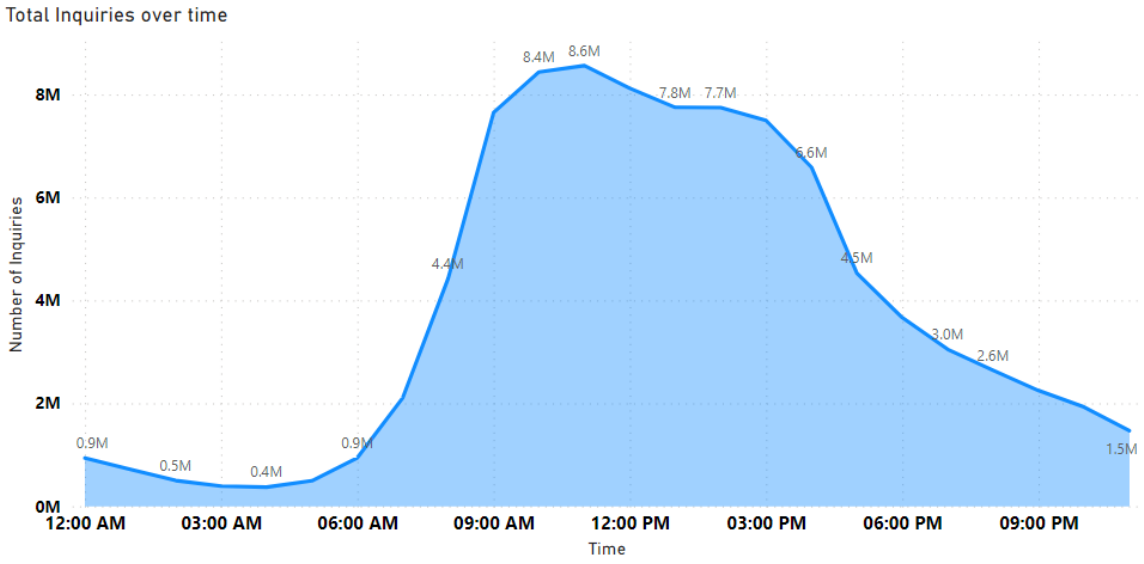


Figure 5: Total inquires over time

5. Conclusion

Using a big data framework like Apache Spark to analyse the NYC 311 call center dataset has facilitated the extraction of valuable insights. These insights can help the city optimise their service delivery processes, address resolution backlogs and better allocate resources to meet the needs of the residents effectively. It is crucial for the city to continue to leverage big data analytics to enhance its operations. This includes refining data handling processes, addressing unassigned inquiries and staying adaptable to evolving communication trends. By leveraging big data the city can better serve its residents and maintain a high standard of operational efficiency.

References

[1] AWS, "AWS," [Online]. Available: https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/. [Accessed 05 October 2023].

[2] AWS, "AWS," [Online]. Available: https://aws.amazon.com/big-data/what-is-spark/. [Accessed 4 October 2023].

[3] A. Garg, "IntelliPaat," 22 September 2023. [Online]. Available: https://intellipaat.com/blog/tutorial/spark-tutorial/spark-architecture/#:~:text=The%20Apache%20Spark%20framework%20uses,real%2Dtime%20processing%20as%20well.. [Accessed 4 October 2023].

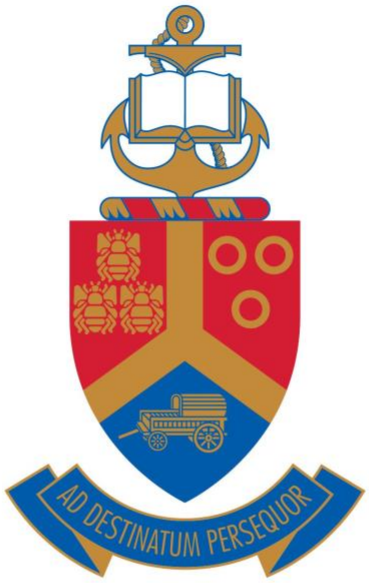
[4] N. Vaidya, "Edureka," 1 June 2023. [Online]. Available: https://www.edureka.co/blog/spark-architecture/. [Accessed 9 October 2023].

APPENDIX 1

MIT 805

Big Data

Assignment part 1



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Student Number: u23970911

Name: Mapamela Salizwa Wendy

Date: 18 August 2023

Table of Contents

1. Introduction

1.1. Overview of the data: 311 Call Center Inquiry

The dataset used for this assignment was extracted from GitHub and it is originally collected by New York City (NYC) through the 311 call center. The data is available on NYC Open data portal, that is used to handle NYC government information and data. The data is available in various formats including but not limited to CSV, Excel and XML. The data consists of 92.8 million records and the records are classified into 9 columns. It is collected from NYC residents who call 311 daily to access non-emergency municipal services, report problems to government agencies or request information (Nuance, n.d.).

The city receives multiple and unique enquiries daily and in order to speed up the data collection process they integrated interactive voice response system that uses Natural Language understanding (Nuance, n.d.). The system classifies and routes calls to the correct departments or agency in order to find solutions quicker for residents (Nuance, n.d.).

The procedure followed by the city includes collecting data from residents through 311 phone calls, provide a solution where possible or transfer calls to the correct department in order to deal with the inquiry . This ensures that the NYC government provides quick and quality service to its residents.

2. Technical aspects of the dataset

2.1.Volume

Volume in big data refers to the size or amount of information gathered from various sources, the information is usually large, disorganised and cannot be processed using traditional ways and may require robust applications and technology to handle (Khan, et al., 2014).

The data used for this assignment is 19.11 gigabytes in size and consists of approximately 92.8 million unique inquiries that are classified according to 9 features. The data is collected from 2010 to date and gets updated daily. The City receives enormous amounts of data daily from the NYC residents and thus require robust architecture and technology to record, store and process the data to ensure a quicker turn around time. They have a robust storage facility that enables them to retain data of over 10 years.

2.2.Velocity

Velocity refers to the rate at which data is generated. A rapid velocity can result in large volumes of data thus causing challenges in processing data efficiently (Khan, et al., 2014). The velocity of this dataset is very high, the city receives large volumes of calls, each with its unique inquiry . The records are updated daily with dates and time stamps thus indicating the speed at which data is generated. The data is also processed at a faster rate to ensure that solutions are provided quicker to the caller or that they are directed to the relevant department or agency that can assist quicker and enhance the city's service delivery.

2.3.Veracity and Validity

Veracity refers to the quality, integrity and meaningfulness of the data (Khan, et al., 2014). A dataset that has more records with meaningful and valuable information that can be used to analyse data has a high veracity (Framework, 2019). This dataset has approximately 17 million rows that have no value or information (see figure 1), this may be due to dummy calls or caller dropping the call before the data has been collected or classified. As part of the data processing stage these columns were removed from the dataset as they do not provide any valuable information. These columns are approximately 18% of the dataset and this indicates that the current dataset has a high veracity as it has more records that have valuable information and can be analysed.

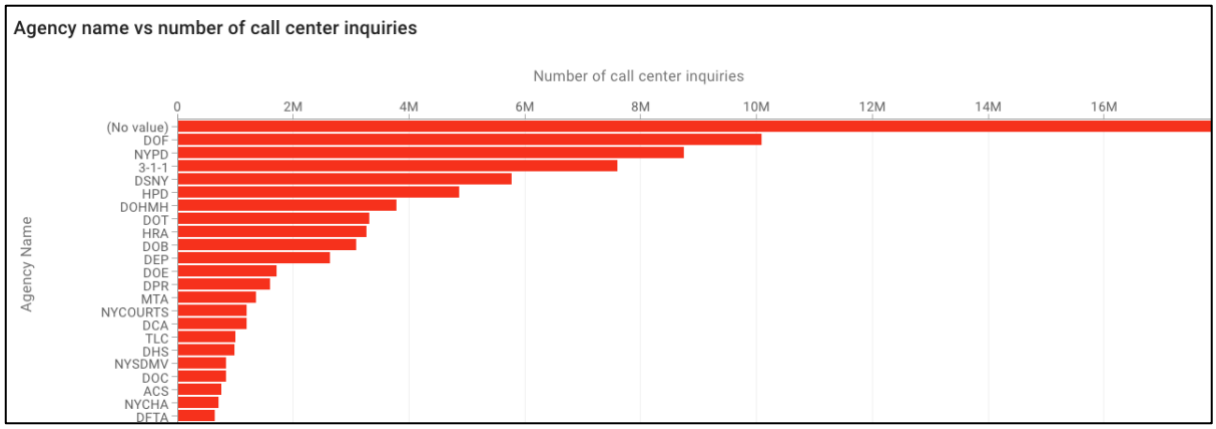


Figure 1: Graph showing Agency name vs number of call center Inquiries

Validity refers to the accuracy of the data for its intended purpose (Khan, et al., 2014). The data collected by NYC open data is accurate as there are procedures set by the organisation for accurate collection and processing of the data. However, since the dataset used for this assignment is collected telephonically it may be prone to a small level of human error as data

is not directly recorded by the source and has to be translated to the third party. The small human error does not entirely compromise the validity of the entire dataset.

2.4.Variety

The dataset used for this assignment is textual data. Although the data is not generated in different forms (audio, images) it has a high level of variety. The data is collected from different data subjects (callers) that have unique inquiries. The data set has approximately 92 million unique entries indicating the variety of the data collected. Even though some inquiries may be classified to the same agency , they may differ because some may be requests for information, some may be complaints further indicating that this dataset has a high level of variety.

2.5.Value

The value of the data refers ways in which the organisation collecting the data can gain from it. The value of the data must always outweigh the cost of collecting, storing and analysing data (Khan, et al., 2014).

The city currently has call log data dating back to 2010 that can be used to identify similar inquiries and track ways in which these were resolved over the years. This data can be used to build Machine Learning models that can assist in finding quicker and efficient ways to resolve the issues or provide information to the callers. The information on this dataset can also be used to identify which departments receive the most calls, how long they take to resolve the issues and find strategies to mitigate these issues. This would contribute greatly to the efficiency of the city and would decrease the turnaround time to resolve issues or provide information.

3. Expected Correlations & relationships

The following relationships and correlations are expected from the data:

- **Date and time:** Correlation between specific dates, times and the number of calls received. It is expected that the City will receive large volumes of calls during the day compared to evenings. Call volumes are also expected to increase during summer breaks or festive season.

- **Agency and call resolution:** Different agencies will have varying rates in resolving their inquiries. It is expected that information requests inquiries will be resolved quicker than other inquiries.
- **Agency and inquiry name:** Certain agencies might receive more calls related to certain inquiries.
- **Date-time and Agency:** Certain agencies might receive more calls during certain time periods.

4. Conclusion

The data acquired for this assignment is data obtained from NYC open data detailing calls that are received on 311 daily. It is worth noting that based on the V's (volume, velocity, veracity, validity, variety and value) discussed in this report, this data can be considered big data. The data has been processed and it will be used for the second part of this assignment which entails Map-reduction and visualisation using Hadoop.

References

1.Nuance, n.d. [Online]
Available at: https://www.nuance.com/asset/en_us/collateral/enterprise/case-study/cs-nyc311-en-us.pdf
[Accessed 15 August 2023].
2.Khan, M. A.-u.-d., Uddin, M. F. & Gupta, N., 2014. Seven V's of Big Data Understanding Big Data to extract Value. *IEEE.*
3.Framework, E. B. D., 2019. *Enterprise Big Data Framework.* [Online]
Available at: https://www.bigdataframework.org/the-four-vs-of-big-data/
[Accessed 16 August 2023].
4.Data, N. O., 2016. *NYC Open Data.* [Online]
Available at: https://data.cityofnewyork.us/City-Government/311-Call-Center-Inquiry/wewp-mm3p
[Accessed 12 August 2023].