**A Data-Driven Analysis of New York City Traffic Accidents**

Submitted by

**Sravani Kairam Konda (18) - 02130417 (Section-2)**

**Ankitaben Bhratkumar Mungalpara (40) – 02129569 (Section-1)**

**Ramana Sai Payili (32) – 02096697 (Section-2)**

Under the Guidance of

Prof. **AshokKumar R Patel**

Submitted to the Department of Computer Science.

University Of Massachusetts

North Dartmouth, Massachusetts

**Abstract**  
  
This dataset encapsulates motor vehicle collisions reported by the New York City Police Department during the period from January to August 2020. Each entry furnishes comprehensive details on individual collisions, including temporal aspects such as date and time, spatial attributes like borough, zip code, street name, and geographic coordinates (latitude/longitude), as well as information pertaining to involved vehicles, victims, and contributing factors.

The objective of this project is to conduct a thorough analysis of the New York City motor vehicle collision dataset, which serves as a valuable resource for understanding and addressing traffic safety challenges in this bustling metropolis. Leveraging the capabilities of Azure tools, the analysis workflow involves the use of Azure Data Factory for efficient data ingestion, storage of raw data in Azure Data Lake Storage Gen2, application of Azure Databricks for transformative processes, subsequent storage of processed data, and the utilization of Azure Synapse Analytics for in-depth and comprehensive analytics. This endeavor aims to extract meaningful insights from the dataset, contributing to the enhancement of traffic safety measures and accident prevention strategies in the city of New York.

**Introduction**

In this project, we aim to delve into the NYC Open Data on motor vehicle collisions reported by the New York City Police Department for the period of January to August 2020. The dataset encompasses a wide array of information about accidents, including temporal, spatial, vehicular, victim-related, and contributing factors. The primary objectives of this analysis are to identify patterns, trends, and key insights regarding the occurrences of traffic accidents within this time frame.

**1.1 Motivation**

The city of New York, renowned for its vibrant urban landscape and incessant vehicular activity, faces unique challenges in ensuring the safety of its residents and commuters. With an ever-evolving transportation network, understanding the dynamics of motor vehicle collisions becomes imperative for implementing effective safety measures and accident prevention strategies. The dataset, spanning from January to August 2020 and meticulously compiled by the New York City Police Department, serves as a valuable repository of information that holds the key to unraveling patterns, identifying risk factors, and ultimately enhancing the overall traffic safety landscape.

Analyzing this dataset is not merely an academic exercise but a proactive step towards fostering a safer and more resilient city. By delving into the nuances of each collision - from the temporal and spatial dimensions to the intricate details of vehicles, victims, and contributing factors - we gain insights that can inform evidence-based policies, targeted interventions, and community awareness programs. The motivation behind this project is rooted in the desire to harness technology, specifically Azure tools, to unlock the potential of this dataset. Through advanced data analytics and storage solutions, we aim to transform raw information into actionable intelligence, contributing to a data-driven approach for traffic safety management in one of the world's most dynamic urban environments. The ultimate goal is to reduce accidents, improve emergency response, and enhance the overall quality of life for the residents and visitors navigating the bustling streets of New York City.

**1.2 Problem Statement**

New York City grapples with persistent challenges in traffic safety, necessitating a comprehensive analysis of motor vehicle collisions from January to August 2020. The existing dataset, meticulously compiled by the New York City Police Department, offers a rich source of information for addressing these challenges. The lack of an organized approach to data analysis, coupled with the diverse factors contributing to collisions, hinders proactive safety measures.

The problem is the absence of a systematic framework for extracting meaningful insights from the dataset. Key challenges include data volume, diverse contributing factors, and the need for a seamless Azure-based workflow for ingestion, transformation, storage, and analytics. Our objective is to formulate an efficient analytical framework utilizing Azure tools, encompassing Azure Data Factory for streamlined data processing, Azure Databricks for efficient transformation, and Azure Synapse Analytics for in-depth analytics. By overcoming these challenges, we aim to uncover patterns, identify risk factors, and provide actionable insights, contributing to evidence-based policies and interventions for enhanced traffic safety management in New York City.

**Overview of Technologies  
  
2.1 Azure Data Factory**

In Azure Data Factory, the New York City motor vehicle collisions dataset is ingested through a streamlined process. Linked services establish connections to external data sources, and a pipeline orchestrates the workflow. The "Copy Data" activity within the pipeline efficiently moves the raw dataset to Azure Data Lake Storage Gen2, with configuration options for source-destination mapping. Monitoring tools ensure the successful execution of the data ingestion, and logging mechanisms capture any issues for prompt resolution. Scheduling features allow for periodic updates, creating a systematic and automated workflow for ingesting the dataset into the Azure environment.

**2.2 Data Lake Gen 2**

The significance of Data Lake Gen 2 lies in its indispensable role within our architecture, where scalability and security features are paramount. This adaptive solution serves as a crucial component by offering secure storage for both structured and unstructured data. This becomes particularly crucial when dealing with the substantial volumes of diverse information generated in the context of traffic accidents. The adaptability of Data Lake Gen 2 ensures that it can efficiently handle the vast and varied data sets associated with such incidents, making it an essential element in our system's architecture.

**2.3 Azure Databricks**

Through the utilization of Azure Databricks, which is built on Apache Spark, we've successfully managed to handle big data processing and implement analytics seamlessly. This platform fosters collaborative efforts between data scientists and engineers, creating an environment conducive to interdisciplinary teamwork. The dynamic processing power of Azure Databricks adjusts in real-time based on demand, a feature particularly beneficial when dealing with datasets that exhibit fluctuations in processing requirements, as is often the case with our chosen dataset.

**2.4 Synapse Analytics**

Synapse Analytics stands out as a key facilitator, enabling instant analytics while also functioning as a versatile data warehouse. Its seamless integration with various Azure services elevates our analytics environment, transforming it into a comprehensive information repository and a real-time processor capable of handling substantial data loads. This interoperability enhances the overall efficiency and performance of our analytics infrastructure.

**Pipeline**

A diagram of data processing

Description automatically generated

**Data Source:**

The dataset, containing motor vehicle collisions reported by the New York City Police Department, is the initial data source. This dataset could be in a structured format like CSV or Parquet and might be stored in an on-premises system, cloud storage, or another data repository.

**Data Ingestion with Azure Data Factory:**

Set up an Azure Data Factory pipeline with a Copy Data activity.

Define the source dataset (the NYC Traffic Accidents dataset) and the destination, specifying Azure Data Lake Storage Gen2 as the storage location for raw data.Configure the data movement settings, such as file format, compression, and any relevant parameters.

**Raw Data Storage in Azure Data Lake Storage Gen2:**

The raw dataset is copied to Azure Data Lake Storage Gen2, preserving its original format.

Azure Data Lake Storage Gen2 provides a scalable and secure storage solution for large volumes of raw data.

**Data Transformation with Azure Databricks:**

Set up a second Azure Data Factory pipeline for data transformation.

Integrate Azure Databricks into the pipeline to perform data cleaning, preprocessing, and transformation tasks. Utilize Databricks notebooks or jobs to execute transformations on the raw data stored in Azure Data Lake Storage Gen2. Save the transformed data back into Azure Data Lake Storage Gen2, specifying a separate folder or container for the processed data.

**Storage of Transformed Data in Azure Data Lake Storage Gen2:**

The cleaned and transformed dataset is stored in Azure Data Lake Storage Gen2, separate from the raw data. This organized storage allows for clear differentiation between raw and processed datasets.

**Analytics with Azure Synapse Analytics:**

Set up a third Azure Data Factory pipeline for analytics using Azure Synapse Analytics.

Leverage Azure Synapse SQL Pools or Spark Pools to perform comprehensive analytics on the transformed dataset. Define analytics queries to extract meaningful insights, identify patterns, and perform statistical analyses. Store the analytical results, aggregates, or derived datasets in Azure Synapse Analytics for further reporting and visualization.

This pipeline ensures a seamless flow from the initial data source through data ingestion, transformation, and analytics, utilizing the strengths of each Azure service to efficiently process and analyze the NYC Traffic Accidents dataset. The structured nature of Azure Data Factory enables orchestration, scheduling, and monitoring of these activities. The end result is actionable insights derived from the comprehensive analysis of the dataset.

**Technology**

**1.Data Ingestion with Azure Data Factory:**

We started a project to handle data, and the first big task was getting information into our system using a tool called Azure Data Factory (ADF). Think of ADF as a traffic cop directing the flow of data. We connected ADF to our data sources, like where we store information. In our case, it's Azure Blob Storage and Azure Data Lake Storage—these are like big virtual containers for our data.

We understood the importance of Data Lake Storage, which is a bit like a super-organized digital filing cabinet. With ADF, we smoothly put our raw data into this filing cabinet, keeping it in its original form and organized the way it was.

Then, we defined something called "datasets" in ADF. Imagine a dataset as a blueprint that tells us how our data looks and where it's stored. We made sure ADF could understand the different shapes and sizes of our data, especially since it's stored in this organized filing cabinet.

Now, the exciting part—creating pipelines. Think of a pipeline as a set of instructions for ADF. We created these instructions to move data from one place to another, like copying it from where it's stored to a temporary area for processing. The cool thing is, we could see and design these instructions easily in ADF without needing to write a lot of complicated code.

We also made sure our data journey had safety nets. Just like when you play a video game and get extra lives, we built in features to retry tasks if they didn't work the first time. We set up a system to keep an eye on everything too, so if something went wrong, we'd know about it and fix it quickly.

All of this sets the stage for our big goal: efficiently bringing in the data about car accidents in New York City from January to August 2020. We've securely stored this information in Azure Data Lake Storage Gen2—a special spot that's ready for the next steps in our project, where we'll analyze and make sense of the data.  
  
**2. Data Transformation with Databricks**

After successfully bringing in the data using Azure Data Factory, our next important move was to make that raw dataset structured and ready for analysis. We chose Azure Databricks, an analytics platform based on Apache Spark, for this task. Databricks is known for its powerful capabilities in handling big data efficiently.

In the Databricks environment, we created a dedicated notebook—a digital space for our data transformation work. We used PySpark, which is like the language Databricks speaks, to write code that cleaned up the data, filled in missing values, and reshaped it to suit our analysis needs. The interactive nature of Databricks made it easy for us to try different things, test them, and improve our code as needed.

One cool thing about this phase was how Spark handles things. It's like having a team of workers, and each one does a specific job at the same time. This parallel processing feature of Spark made our data transformation tasks super-fast, especially when dealing with large amounts of data. This not only sped up the process but also made sure our solution could handle more data if needed.

In addition, Databricks provided a collaborative space for our team. Several team members could work on the notebook at the same time, making it a hub for teamwork and sharing knowledge. We kept everything organized with version control, which helped us track changes, go back to previous versions if necessary, and keep a neat record of our code.

Considering the dataset about New York City accidents, our goal was to use Databricks to make smart and efficient changes to the data. We wanted to fix any issues, get rid of inconsistencies, and make the dataset optimized for our next steps—analyzing the data. We did various analyses, like figuring out which borough had the most accidents, understanding the factors contributing to fatal accidents, finding out when accidents peaked, and checking how accidents were distributed across different days of the week.

After all these transformations and analyses, we made sure to export the refined data back to Azure Data Lake Storage. This sets the stage for the next steps in our project—further analysis and creating visualizations to understand and communicate the insights we've uncovered.

**3. Advanced Analytics with Spark**

In Databricks, we utilized PySpark code within a notebook to perform advanced analytics on the NYC Traffic Accidents dataset. Our focus was on deriving insightful metrics, such as the average entries by gender per discipline and the total medals won by each country. Leveraging Spark's distributed computing capabilities, the platform efficiently handled complex computations across the dataset, demonstrating both its effectiveness and scalability. The analyses provided valuable insights into gender distribution across disciplines and a comprehensive view of country-wise medal counts. This approach showcases the power of Spark in managing large-scale data-intensive tasks, particularly relevant for detailed analyses of event datasets like the NYC Traffic Accidents.

**4. Synapse Analytics for Data Warehousing**

In the move to Azure Synapse Analytics, we harnessed its capabilities for seamless data integration and warehousing. By establishing a Synapse workspace and crafting a dedicated database, we organized and stored our transformed data. The user-friendly interface of Synapse Analytics enabled interactive loading of data into tables, visually representing the ingested and transformed information. This step underscored the significance of a centralized data repository, ensuring efficient querying and analysis. It serves as a crucial foundation for the next phases of our project, allowing for streamlined data access and exploration within Azure Synapse Analytics.

**5. Data Visualization on Azure Synapse Analytics**

Having successfully processed the NYC Traffic Accidents dataset through ingestion, transformation, and warehousing using Azure Synapse Analytics, our focus shifted to extracting meaningful insights through data visualization and analytics. Azure Synapse Studio, functioning as a robust tool, provided a unified workspace for processing both big data and data warehousing.

Within Synapse Studio, we formulated SQL queries tailored to our analytical objectives, such as counting accidents by borough and analyzing daily trends. The adaptability of Synapse Studio's SQL query editor facilitated intricate analyses, demonstrating its effectiveness across various analytical requirements.

An exceptional feature was the seamless integration of Apache Spark in Synapse Studio, enabling the execution of Spark SQL queries for advanced analytics. This capability proved beneficial for handling complex computations in our extensive Traffic Accidents dataset through distributed computing.

Synapse Studio's visualization tools played a crucial role in transforming raw data into actionable insights. We designed interactive charts, graphs, and dashboards to effectively communicate complex findings.

By leveraging Azure Synapse Analytics, we translated collision data into meaningful insights for informed decision-making. This encompassed exploring accident counts by borough, daily trends, the distribution of contributing factors for fatal accidents, and hourly patterns. Synapse Studio's integration of data exploration and visualization streamlined our workflow, making it an indispensable tool for end-to-end analytics. The collaborative environment within Synapse Studio facilitated teamwork, ensuring transparency and reproducibility in our analysis workflows.

**Results and Analysis:**

***Step 1: Data Ingestion to Data Lake Storage using Azure Data Factory***

1. Define Source:

A screenshot of a computer

Description automatically generated

1. Define Sink:

A screenshot of a computer

Description automatically generated

1. Validate the pipeline:

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

***Step 2: Data Transformation on Azure Databricks***

1. Azure Storage Account Configuration for Spark on Azure Databricks:

A screen shot of a computer

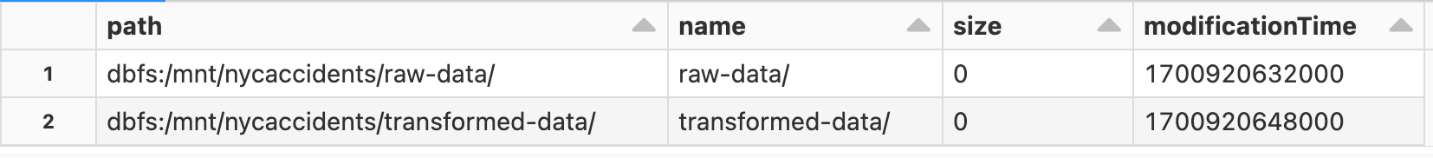
Description automatically generated

1. Mount External Storage in Databricks:

A close-up of a white background

Description automatically generated

1. Listing the mounted directories:



1. Load Dataset:

A close-up of a computer code

Description automatically generated

1. Printing Data:

A close-up of a number of persons injured

Description automatically generated

1. Data Transformation:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

***Data Analysis:***

1. Top Accident Counts by Borough:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

1. Analysis of Top Contributing Factors for Fatal Accidents:

A screen shot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

1. Analysis of Peak Accident Times:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. Accident Distribution by Day of the Week:

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

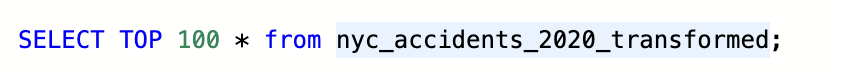
Description automatically generated

***A screenshot of a computer

Description automatically generated***

***Step 3: Data Visualization on Azure Synapse Analytics***

1. Top 100 Records from NYC Traffic Accidents:



A table with numbers and a few black text

Description automatically generated with medium confidence

1. Accident Count by Borough:

A close up of text

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

1. Daily Accident Trends:

A screenshot of a computer error

Description automatically generated

A graph of a number and a number

Description automatically generated with medium confidence

1. Contributing Factors Analysis for Injured Persons in NYC Traffic Accidents:

A screenshot of a computer code

Description automatically generated

A graph with blue lines and white text

Description automatically generated

1. Monthly Distribution Analysis of NYC Traffic Accidents:

A screen shot of a computer

Description automatically generated

A graph of blue rectangular bars

Description automatically generated with medium confidence

**Conclusion:**

In our analysis of motor vehicle collisions in NYC, we've unearthed significant insights into the patterns, trends, and key factors influencing traffic accidents. By thoroughly exploring various dimensions such as time, location, vehicles involved, victims, and contributing factors, we've gained a nuanced understanding of how and where collisions occur.

The use of Azure tools played a pivotal role in this analysis, streamlining the entire data processing journey from ingestion to visualization. Azure's capabilities ensured an efficient and seamless analytical pipeline. We utilized Azure Data Factory for collecting and moving the data, Azure Databricks for transforming and preparing it, and Azure Synapse Analytics for in-depth analysis.

These insights are more than just numbers; they represent a valuable resource for informed decision-making. By understanding the intricacies of collision occurrences, authorities and policymakers in New York City can implement targeted traffic safety measures and accident prevention strategies. Whether it's improving road infrastructure, enhancing law enforcement strategies, or conducting public awareness campaigns, these insights empower stakeholders to address specific challenges and improve overall traffic safety in the dynamic urban environment of New York City. Ultimately, the goal is to make data-driven decisions that contribute to a safer and more resilient transportation system for everyone in the city.

**Future Enhancements:**

**Real-time Monitoring and Alerts:**

* Implement a system for continuous real-time analysis of traffic data.
* Integrate alerts to notify authorities of unusual patterns or emergency situations.

**Predictive Analytics:**

* Develop models to predict collision hotspots and periods of increased risk.
* Use historical data and external factors for proactive resource allocation.

**Machine Learning for Contributing Factors:**

* Apply machine learning to identify and predict contributing factors to collisions.
* Tailor interventions based on common factors identified by the models.

**Public Awareness Campaigns:**

* Launch data-driven public awareness campaigns using insights from the analysis.
* Communicate safety tips and high-risk areas through visualizations.

**Integration with Traffic Management Systems:**

* Integrate findings with traffic management systems for optimized signal timings.
* Improve traffic flow by rerouting in real-time based on analysis.

**Collaboration with Law Enforcement:**

* Strengthen collaboration with law enforcement using actionable insights.
* Enhance strategic resource deployment and targeted operations.

**Data Privacy and Ethical Considerations:**

* Prioritize data privacy and ethical considerations in data collection and analysis.
* Ensure compliance with regulations and transparent communication about data usage.

**Continuous Data Quality Improvement:**

* Establish mechanisms for continuous improvement of data quality.
* Regularly update and enrich the dataset to address discrepancies or inaccuracies.

**Mobile Applications for Reporting:**

* Develop or enhance mobile apps for real-time reporting of road conditions.
* Crowdsource information to complement existing datasets for analysis.

**References:**

* <https://spark.apache.org/>
* <https://www.kaggle.com/datasets/melodyyiphoiching/nyc-traffic-accidents>
* <https://learn.microsoft.com/en-us/azure/data-factory/>
* <https://learn.microsoft.com/en-us/azure/synapse-analytics/overview-what-is>