**BIG DATA CRIME ANALYTICS: A PYSPARK AND NOSQL STUDY OF NYPD ARRESTS**

**Team 2: Jainam Jagani, Aditya Baxi**

**Under the guidance of Dr. Zhang**

**George Mason University, AIT-614-004 Big Data Essentials**

**October 24, 2023**

**INTRODUCTION**

The New York City Police Department (NYPD) stands at the forefront of maintaining peace and order in this dynamic urban landscape. Understanding and deciphering the intricacies of NYPD's arrest data is not merely an academic endeavor; it is an essential pursuit in the realm of public safety, social justice, and community well-being. This project embarks on a journey to delve into the wealth of information contained within the NYPD arrest dataset, comprising details of law enforcement activities, arrest records, and incident characteristics. The dataset offers a window into the diverse and multifaceted dimensions of policing in one of the world's most iconic cities.

Background: Policing in New York City is a dynamic and multifaceted undertaking, influenced by myriad factors, including the city's socioeconomic landscape, geography, demographics, and evolving policies. The NYPD arrest dataset, a product of rigorous data collection and review processes, is a reservoir of insights into law enforcement practices within the city. It encapsulates the rich tapestry of arrest activities, providing information about the type of crimes, the locations where arrests occur, and the timing of these enforcement actions. Furthermore, the dataset includes essential details about the demographics of those who come into contact with the justice system.

Scope: This project's primary objective is to harness the power of data analytics to delve deeply into the NYPD's arrest dataset. By applying advanced data analysis techniques, it seeks to unveil concealed patterns, correlations, and trends within this repository of law enforcement activity. The insights thus derived will illuminate the very essence of police operations in New York City and unveil how they may vary according to distinct criteria, such as location, time, and demographics.

Motivation: The motivation behind this project is rooted in the imperative to promote transparency and accountability within the law enforcement realm, to inform evidence-based policies, and to foster an environment of continuous improvement. Understanding the intricacies of arrests and the underlying factors at play is pivotal in the pursuit of more effective and equitable policing practices. By uncovering patterns and trends, this project aims to offer valuable insights to law enforcement agencies, policymakers, and community advocates, ultimately contributing to a safer, more just, and more informed New York City. This dataset serves as a vital resource for both the public and decision-makers to explore the nature of police enforcement activity and initiate constructive dialogues.

**RELATED WORK**

In the realm of related work, the analysis of NYPD arrest data and its associated insights has been a subject of limited exploration. Existing findings primarily revolve around basic trends and patterns that this project seeks to build upon and deepen. Below, we provide an overview of the differences in this project's approach compared to the available related work:

1. Borough-wise Arrest Patterns:

Related work has identified the varying numbers of arrests across New York City's five boroughs. Notably, Brooklyn had the highest number of arrests during the first three quarters of 2022. While this information sets a foundation, our project aims to expand upon this by delving into the underlying factors that contribute to these variations. We will explore how population density, socio-economic factors, local industries, and law enforcement strategies intersect to influence arrest rates.

2. Crime Types and Severity:

Previous research highlighted the predominance of misdemeanor arrests in New York City during the same time frame, with felonies ranking second. This project will further investigate the nature of these arrests, aiming to discern patterns in the types of misdemeanors and felonies and how they relate to factors like location, demographics, and time. Understanding the specifics of these crimes can provide deeper insights into the criminal justice landscape.

3. Demographics and Offense Breakdown:

Related work has shed light on the racial and gender demographics of those arrested in New York City during the first three quarters of 2022. Our project will extend this by examining the nuanced interplay between race, age groups, and the types of offenses committed. Additionally, we will explore variations in arrest rates within these demographics, seeking to identify any disparities or trends that might not have been apparent in earlier studies.

4. Advanced Data Analytics:

One key differentiator is the application of advanced data analytics techniques, such as PySpark and NoSQL databases, to unearth intricate patterns and relationships in the dataset. While previous work provides a broad overview, our project's emphasis on in-depth analysis using these technologies enables us to uncover subtler nuances in the data.

In essence, this project aspires to offer a more comprehensive and nuanced perspective on the NYPD arrest dataset, going beyond the surface-level observations to understand the underlying factors that shape the arrest patterns in New York City. The use of advanced analytical tools and techniques will allow for a deeper and more data-driven exploration, potentially revealing new insights that may not have been apparent in prior research.

**OBJECTIVES**

The objectives of the overall project, which involves analyzing the NYPD arrest dataset using advanced data analytics techniques, are as follows:

Explore Borough-wise Arrest Patterns: Examine the variations in arrest rates among New York City's five boroughs and identify the underlying factors contributing to these differences, such as population density, socio-economic status, local industries, and law enforcement strategies.

In-Depth Analysis of Crime Types: Investigate the nature and distribution of different types of crimes, specifically focusing on misdemeanors and felonies, to understand their prevalence, patterns, and how they relate to factors like location, demographics, and time.

Demographic and Offense Correlations: Analyze the relationships between race, age groups, and the types of offenses committed, seeking to identify disparities and trends within these demographics and exploring the potential causes.

Advanced Data Analytics: Utilize advanced data analytics tools and techniques, including PySpark and NoSQL databases, to conduct a more in-depth analysis of the NYPD arrest dataset, enabling the discovery of subtle and previously unnoticed nuances in the data.

Provide Valuable Insights: Ultimately, the project aims to provide valuable insights to various stakeholders, including law enforcement agencies, policymakers, and community advocates, to foster transparency, accountability, and evidence-based decision-making in the realm of policing and criminal justice in New York City.

**SELECTION OF DATABASE**

Brief Description:

The selected dataset is derived from a larger dataset containing information on arrests made by the NYPD during the year 2023. It includes details such as the date of arrest, type of crime, legal code, law category, borough of arrest, precinct code, jurisdiction code, age group of the perpetrator, sex and race of the perpetrator, geographical coordinates, and georeferenced information. It contains 112571 rows and 19 features/columns

Features/Columns for Selection:

For data analysis, we would like to select a subset of features/columns from the provided dataset. The features to be selected should be relevant and suitable for the specific analysis goals. Based on the nature of potential analysis tasks, here's a proposed list of features to be selected:

ARREST\_DATE: This feature provides the date of the arrest, which is essential for temporal analysis.

PD\_DESC: This column contains a description of the offense, which is crucial for understanding the nature of the crimes.

OFNS\_DESC: The description of the offense, which can provide more information on the nature of the crimes.

LAW\_CAT\_CD: This feature categorizes the arrest into different law categories (e.g., felony, misdemeanor), which is important for understanding the severity of the offenses.

ARREST\_BORO: This feature indicates the borough where the arrest took place, which is vital for geographical analysis.

AGE\_GROUP: Age group of the perpetrator, which is relevant for demographic analysis.

PERP\_SEX: The sex of the perpetrator, which can be useful for demographic analysis.

PERP\_RACE: The race/ethnicity of the perpetrator, which is essential for demographic and social analysis.

Latitude: Geographical coordinates (latitude), which are important for spatial analysis.

Longitude: Geographical coordinates (longitude), necessary for spatial analysis.

The features provides a balanced mix of temporal, categorical, and geographical attributes, enabling a comprehensive data analysis while ensuring that the selected dataset retains a substantial number of records for robust analysis.

**ARCHITECTURE AND METHODS**

A diagram of a process

Description automatically generated

Conceptual System Architecture for NYPD Arrest Data Analysis Project:

Data Source: The project starts with the NYPD arrest dataset, which is manually extracted and reviewed quarterly by the Office of Management Analysis and Planning.

Data Ingestion: Data is ingested from the source into a data processing environment. This could be a data lake, database, or distributed file system.

Data Preparation: Data preparation involves cleaning, transformation, and structuring of the dataset. It includes removing duplicates, handling missing values, and reformatting data for analysis.

Data Analysis Engine: This component consists of tools and libraries for data analysis. It can include PySpark for big data analysis, NoSQL databases for storing intermediate results, and data analysis libraries like pandas, NumPy, and scikit-learn.

Analysis Tasks: The analysis tasks involve exploratory data analysis, statistical analysis, machine learning, and data visualization.

Results and Insights: The outcomes of the analysis are insights and patterns related to arrest data. These insights may pertain to arrest patterns, demographic correlations, geographic trends, and more.

Visualization and Reporting: Data visualization tools like Matplotlib, Seaborn, or Tableau are used to create visual representations of the insights. These visualizations are often used in reports and dashboards.

Presentation and Reporting: The project's findings are presented through reports, dashboards, or presentations. Stakeholders, including law enforcement agencies, policymakers, and community advocates, can use these reports to inform decisions.

Feedback and Iteration: Feedback from stakeholders is collected and used to iterate and improve the analysis, data preparation, and visualization processes.

Deployment: If the project's insights lead to actionable recommendations, they may be deployed in the field, potentially influencing law enforcement practices.

Documentation and Knowledge Sharing: Throughout the project, documentation is maintained to ensure that methods, findings, and insights are well-documented and shared with relevant parties.

Security and Compliance: Security measures and compliance with data protection laws are considered throughout the project to protect sensitive arrest data.

This conceptual system architecture or framework outlines the high-level components and flow of your NYPD arrest data analysis project. It represents the stages from data ingestion to reporting and emphasizes the iterative nature of data analysis projects, where feedback and continuous improvement are essential.

**DEVELOPMENT PLATFORMS:**

Software Development Platform

Software: For software development, platforms like Jupyter Notebook, PyCharm, or Visual Studio Code can be used. These platforms provide a development environment for coding, data analysis, and visualization tasks.

Programming Languages: Python is the primary programming language for data analysis in this project. Libraries like Pandas, NumPy, Matplotlib, and Seaborn are commonly used for data manipulation and visualization.

IDEs: Integrated Development Environments (IDEs) like Jupyter Notebook are popular for interactive data analysis and code sharing.

Version Control: Git and platforms like GitHub can be used for version control and collaboration.

Hardware Development Platform:

Hardware: The hardware platform could be a local development machine or cloud-based virtual machines (e.g., AWS EC2 instances, Google Cloud VMs, or Azure virtual machines) with sufficient computational resources to handle the data analysis tasks efficiently.

NoSQL Database:

One suitable NoSQL database for this project could be MongoDB. MongoDB is a versatile and widely used NoSQL database that can store and manage large datasets efficiently. It's suitable for storing the preprocessed data and intermediate results.

Big Data Engines: PySpark

PySpark is a big data processing framework that can be used for distributed data analysis. It allows you to work with large datasets and apply transformations and computations in parallel across a cluster of machines.

Spark MLlib:

Spark MLlib is a machine learning library within the Spark ecosystem. It provides a wide range of machine learning algorithms and tools for building and deploying machine learning models at scale.

Big Data Platform: Hadoop Ecosystem

Apache Hadoop, a popular big data platform, can be utilized. Hadoop consists of various components, such as Hadoop Distributed File System (HDFS) for distributed storage and Hadoop MapReduce for data processing. While PySpark and Spark MLlib are integrated with the Hadoop ecosystem, they can also work independently.

This development platform and technology stack offer a robust and scalable environment for handling and analyzing the NYPD arrest dataset. It enables you to process, analyze, and derive valuable insights from large volumes of data efficiently.

**PROJECT TASKS AND TIMELINE:**

**Phase 1: Data Preparation and Analysis (Months 1-3)**

Data Ingestion and Exploration (Week 1): Assigned to Aditya Baxi  
Ingest the NYPD arrest dataset. Conduct exploratory data analysis (EDA) to understand the data's structure and quality.

Data Cleaning and Preprocessing (Week 1): Assigned to Jainam Jagani Clean the dataset, addressing issues like duplicates and missing values. Preprocess the data, including feature engineering and transformation.

Demographic and Geographic Analysis (Week 2): Assigned to Aditya Baxi Analyze demographic factors (age, sex, race) in relation to arrests. Investigate geographical patterns by borough and precinct.

Statistical Analysis (Week 2): Assigned to Jainam Jagani Perform statistical tests to uncover insights, e.g., correlations between variables.

**Phase 2: Advanced Analysis and Reporting (Week 3-4)**

Temporal Analysis and Trends (Week 3): Assigned to Aditya Baxi Analyze temporal patterns, e.g., daily, weekly, or monthly trends in arrests. Detect any seasonal patterns.

Crime Type and Severity Analysis (Week 3): Assigned to Jainam Jagani Examine the distribution of crime types and their severity. Use machine learning models to predict crime types.

Machine Learning Models (Week 4): Assigned to Aditya Baxi Develop machine learning models for predictive analysis. Fine-tune models for accuracy.

Data Visualization and Reporting (Week 4): Assigned to Jainam Jagani Create data visualizations (e.g., charts, graphs, maps) to communicate insights. Generate comprehensive reports summarizing findings and actionable recommendations.

Project Completion and Delivery: 27th November 2024

**Project Management Timeline:**

Month 1: Data Ingestion and EDA

Month 2: Data Cleaning and Preprocessing

Month 3: Demographic and Geographic Analysis

Month 4: Temporal Analysis and Trends

Month 5: Crime Type and Severity Analysis, Machine Learning Models

Month 6: Data Visualization, Reporting, and Project Completion

**Project Duration:**

The project is expected to take approximately 4 weeks to complete, with the first two weeks dedicated to data preparation and initial analysis, and the following two weeks focused on more advanced analysis, modeling, visualization, and reporting.

**References:**

Neil Angelo Martinez, NYPD Arrest Data Analysis using Power BI, <https://medium.com/@neilangelomartinez/nypd-arrest-data-d4fc3a0501cb> , Accessed on 10/22/2023

NYPD Arrest Data (Year to Date) , [data.cityofnewyork.us](https://catalog.data.gov/dataset?publisher=data.cityofnewyork.us) , Accessed on 10/23/2023

**APPENDIX**

**A screenshot of a computer

Description automatically generated**

A computer code with many black and white text

Description automatically generated with medium confidence