**Big Data Systems and Analytics**

**Big Data Analytics for Urban Crime and Public**

**Safety in Smart Cities: A MySQL- Based Approach**

Table of Contents

[**Chapter 1: Introduction and Background** 4](#_Toc214967966)

[**1.1** **Introduction** 4](#_Toc214967967)

[**1.2** **Background to the Study** 4](#_Toc214967968)

[**1.3** **Problem Statement** 4](#_Toc214967969)

[**1.4** **Aim and Objectives** 4](#_Toc214967970)

[**Chapter 2: Big Data Context** 5](#_Toc214967971)

[**2.1** **Introduction to Big Data** 5](#_Toc214967972)

[**2.2** **The 5Vs of Big Data in this Study** 5](#_Toc214967973)

[**2.2.1 Volume** 5](#_Toc214967974)

[**2.2.3 Variety** 5](#_Toc214967975)

[**2.2.4 Veracity** 6](#_Toc214967976)

[**2.2.5 Value** 6](#_Toc214967977)

[**2.3 Big Data in Smart Cities** 6](#_Toc214967978)

[**Chapter 3: System Designs and Methodology** 7](#_Toc214967979)

[**3.1 Data Preparation and Feature Selection** 7](#_Toc214967980)

[**3.2 System Architecture** 7](#_Toc214967981)

[**3.3 Data Preprocessing Workflow** 8](#_Toc214967982)

[**3.3.1 Data loading in python** 8](#_Toc214967983)

[**3.3.2 Cleaning steps** 8](#_Toc214967984)

[**3.4 Creating MySQL Database Connection** 9](#_Toc214967985)

[***Fig 7: MySQL Successful Database connection.*** 9](#_Toc214967986)

[**3.5 Database Design** 9](#_Toc214967987)

[3.6 **Data Import using MySQL terminal** 10](#_Toc214967988)

[**3.7 Analytical Queries and Techniques** 10](#_Toc214967989)

[**Chapter 4: Analysis and Results** 12](#_Toc214967990)

[**4.1 Overview of Analysis** 12](#_Toc214967991)

[**4.2 Crime Frequency by Category** 12](#_Toc214967992)

[**4.3 Yearly Crime Trends (2019-2024)** 12](#_Toc214967993)

[**4.4 Arrest Trends Across the Years** 13](#_Toc214967994)

[**4.5 Analysis of Crimes with the Highest Arrest Counts** 13](#_Toc214967995)

[**4.6 Monthly Trends** 14](#_Toc214967996)

[**4.7 Arrest Trends Across Districts** 14](#_Toc214967997)

[**4.8 Summary of Findings** 15](#_Toc214967998)

[**Chapter 5: Conclusion and Future Work** 16](#_Toc214967999)

[**5.1 Conclusion** 16](#_Toc214968000)

[**5.2 Future Work** 16](#_Toc214968001)

[**References** 17](#_Toc214968002)

# **Chapter 1: Introduction and Background**

## **Introduction**

The fast growth and development in urban areas have influenced challenges in managing city operations, public safety and security. Large volumes of both structured and unstructured dataset are now generated by modern cities through technological advancement and development of digital infrastructures, particularly from sensors, surveillance devices, open-data websites, and digital reporting systems (Kitchin, 2014). Such data are utilised in smart cities to optimise resource allocation, improve decision-making and increase service-delivery. Big data analytics can be of high value in crime analysis and prevention (Batty et al., 2012).

Using a combination of Python for cleaning and preprocessing and MySQL for structured analysis, this project demonstrates how big data analytics techniques can extract insights that support law enforcement efforts in smart city environments (Hashem et al., 2016).

## **Background to the Study**

To encourage transparency and support research-based public policy, cities are rapidly relying on open data platforms (Janssen et al., 2012). Because of its significance for operational policing and community awareness, crime data has been one of the most extensively studied datasets. However, the large volume, high velocity and variety of the daily generated dataset, makes it complex to analyse crime trends (Chen, Mao & Liu, 2014).

This study main objective is to utilise big data analytics techniques to uncover patterns in crime distribution, arrest behaviour and temporal trends across several years.

## **Problem Statement**

The depth and efficiency required to solve modern urban safety challenges are not found in traditional crime analysis methodologies, which often rely on static or summarised reports (Weisburd & Braga, 2019). For law enforcement agencies to respond efficiently, they require, fast, precise and comprehensive information. This project addresses the need for a structured approach to handling and analysing over 1.4 million crime records using big data analytics techniques.

## **Aim and Objectives**

This report aims to analyse Chicago crime data (2019–2024) using big data technique to extract meaningful insights that support crime pattern understanding in smart cities.

Objectives

1. To collect, clean, and preprocess a large dataset of crime records using Python.
2. To design and implement an SQL-based analytical system for querying crime data.
3. To apply big data techniques in examining temporal, spatial, and categorical crime patterns.
4. To discuss the importance of the 5Vs of Big Data (Volume, Velocity, Variety, Veracity, and Value) in context to the dataset and analytical processes.
5. To generate visual and tabular outputs that support crime trend interpretation.

# **Chapter 2: Big Data Context**

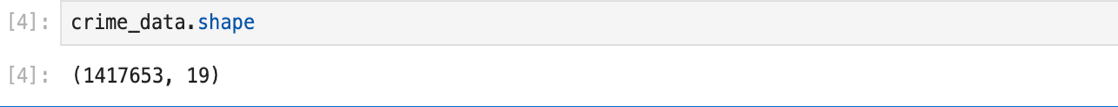
## **Introduction to Big Data**

Big data means extremely large and complex dataset that cannot be processed using traditional data processing technique. Big data utilises advanced computational methods to find hidden patterns, correlations, and trends. In smart cities, big data plays a vital role to support decision making in areas that includes public safety, healthcare, transportation and the environment. Big Data is described by five fundamental characteristics, known as the Big 5 Vs of Big Data.

## **The 5Vs of Big Data in this Study**

### **2.2.1 Volume**

The quantity of data generated is referred to as volume. The Chicago dataset used in this research consist of 1,417,653 records making it large enough to require and advanced processing tool. Pandas library in Python was utilised for cleaning and MYSQL was used for storage and analysis.



***Fig 1: Python code showing the volume of the dataset.***

**2.2.2 Velocity**

The speed at which data set is generated and processed is referred to as velocity. The Chicago crime dataset, which represents fast-streaming data in smart city systems, is updated every day in real-world applications. The research is based on a dataset that is usually updated in near real-time, although past data was used.

Millions of rows can be imported in seconds by using MySQL's high-velocity ingestion feature, "LOAD DATA INFILE".

### **2.2.3 Variety**

Variety represents the different types and forms of data. There are twenty distinct features in the dataset, including:

* Numerical features (Districts, ward, latitude, longitude),
* Categorical features (location description, crime type),
* Boolean features (domestic, arrest)
* Mixed datetime formats that need to be cleaned.

High data variety is highlighted by the necessity to handle missing values, convert booleans, and convert dates.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 2: MYSQL terminal showing distinct features in the dataset.***

### **2.2.4 Veracity**

Veracity simply means the accuracy, reliability and validity of data. Few challenges were faced while looking at the veracity of the datasets, which includes:

* Inconsistent datetime formats (e.g., 24 hrs and 12hrs timestamp),
* Missing latitude/longitude values,
* Boolean format (True, False),
* Fields incorrectly typed during SQL import.

Data cleaning in Python (using “pd.to\_datetime”, replacing NaN values, converting Booleans) improved accuracy and reduced noise, increasing dataset veracity.

### **2.2.5 Value**

Data value refers to the value and the useful insights the data provides. The ultimate goal of processing crime data is to determine its value, that lies in the ability to identify crime hotspots, peak crime hours, understands trends and patterns in crime types and locations and, ultimately, enhance long-term strategic planning in smart cities to ensure public safety.

## **2.3 Big Data in Smart Cities**

Big Data refers to large and complex sets of data that cannot be stored, processed, analysed or manipulated using traditional tools and methods. Big Data allows to detect patters, predict behaviours and make data driven decisions at a considerable speed and scale. In the context of urban crime and public safety, Big Data becomes essential for identifying crime hotspots, understanding temporal crime patterns and detecting anomalies.

# **Chapter 3: System Designs and Methodology**

## **3.1 Data Preparation and Feature Selection**

When imported into python, the raw Chicago crime datasets had 22 features (indexed 0-21). The dataset has 21 features when viewed outside of Python; this difference arises from Python's zero-based indexing, which only represents position rather than count.

Three features were dropped during the cleaning and preprocessing stage due to irrelevance, high missing values and lack of analytical value. The dataset had 19 columns in Python (index positions 0–18) after feature selection, which equals 18 columns when imported into MySQL.

By dropping irrelevant or empty features, this intentional reduction increases analytical efficiency. Thus, the final schema saved in MySQL consist of 18 cleansed and standardised features, which serves as the foundation for all subsequent SQL queries and analysis.

## **3.2 System Architecture**

The tools used in this study are:

|  |  |
| --- | --- |
| **Tools** | **Purpose** |
| Python (pandas) | For data cleaning and preprocessing |
| MYSQL Workbench | Database creation, data import and analysis |
| Local terminal (MacOS) | large velocity data import with LOAD DATA LOCAL INFILE |
| Dataset Source | [Chicago 2019-2024 crime records](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/explore/query/SELECT%0A%20%20%60id%60%2C%0A%20%20%60case_number%60%2C%0A%20%20%60date%60%2C%0A%20%20%60block%60%2C%0A%20%20%60iucr%60%2C%0A%20%20%60primary_type%60%2C%0A%20%20%60description%60%2C%0A%20%20%60location_description%60%2C%0A%20%20%60arrest%60%2C%0A%20%20%60domestic%60%2C%0A%20%20%60beat%60%2C%0A%20%20%60district%60%2C%0A%20%20%60ward%60%2C%0A%20%20%60community_area%60%2C%0A%20%20%60fbi_code%60%2C%0A%20%20%60x_coordinate%60%2C%0A%20%20%60y_coordinate%60%2C%0A%20%20%60year%60%2C%0A%20%20%60updated_on%60%2C%0A%20%20%60latitude%60%2C%0A%20%20%60longitude%60%2C%0A%20%20%60location%60%0AWHERE%0A%20%20%60date%60%0A%20%20%20%20BETWEEN%20%222019-01-01T00%3A00%3A00%22%20%3A%3A%20floating_timestamp%0A%20%20%20%20AND%20%222024-12-31T23%3A45%3A00%22%20%3A%3A%20floating_timestamp%0AORDER%20BY%20%60date%60%20DESC%20NULL%20FIRST/page/filter) |

***Table 1: Tools used.***

A screen shot of a computer

AI-generated content may be incorrect.

***Fig 3: A flow diagram of the tools used.***

## **3.3 Data Preprocessing Workflow**

### **3.3.1 Data loading in python**

The raw CSV file was read into Python for inspection and cleaning.

A screenshot of a data

AI-generated content may be incorrect.

***Fig 4: Python code showing data loading***

### **3.3.2 Cleaning steps**

Key preprocessing steps included:

1. Converting date feature to datetime format
2. Handling missing values
3. Converting Boolean features

A screen shot of a computer code

AI-generated content may be incorrect.

***Fig 5: python code showing data cleaning***

***A screenshot of a computer program

AI-generated content may be incorrect.***

***Fig 6: Python output showing the cleaned dataset***

## **3.4 Creating MySQL Database Connection**

Setting up a secured connection in MySQL server was essential before creating the database schema and importing the cleaned dataset. MySQL workbench provided a graphic user interface for managing server connections which was used to achieve this. A new connection named ‘SmartCityCrimeDB’ was created using the standard TCP approach.

As seen in Fig 7 below, the system confirmed that the connection was secured and valid after testing the connection in workbench.

A screenshot of a computer

AI-generated content may be incorrect.

## ***Fig 7: MySQL Successful Database connection.***

## **3.5 Database Design**

**3.5.1 Table Schema**

A MySQL table named ‘crime\_data’ was created. As shown in Fig 8, this table had 18 distinct features with the necessary datatype assigned to each of them. This table consist of a primary key which was assigned to the identifier feature ‘id’.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 8: Table creation in MySQL***

## 3.6 **Data Import using MySQL terminal**

Due to the large size of the dataset and MySQL workbench restrictions, MySQL terminal was used on MacOS to import the data.

A screenshot of a computer program

AI-generated content may be incorrect.

***Fig 9: MYSQL Terminal output showing the successful “LOAD DATA” execution.***

## **3.7 Analytical Queries and Techniques**

A list of structured SQL queries was created to extract categorical and temporal insights from the cleaned dataset to support the study's goals. These queries provided the foundation for the analytical framework that was utilised to find trends in arrests, crime patterns, and changes in distribution across districts and years. The fundamental methods used in the analysis are represented by the SQL statements mentioned below; the specific findings are shown later in Chapter Four.

Queries used:

* Query to identify the top 10 crime types.
* Query to calculate yearly crime totals.
* Query to show monthly crime patterns.
* Query to analyze total arrests per year.
* Query to analyse crime distribution by district.
* Query to calculate arrest rates by crime category.
* Query to extract trends across the 2019–2024 period.

# **Chapter 4: Analysis and Results**

## **4.1 Overview of Analysis**

Following the successful import of **1,417,653 crime records** into MySQL, a series of analytical queries were run to explore:

* Dominant crime categories
* Yearly crime trends
* Arrest patterns
* Monthly Variations
* District distributions

## **4.2 Crime Frequency by Category**

Theft was identified as the highest crime type in Chicago with over 312,000 reported offences between 2019 and 2024. Also, Battery and Criminal Damage were identified as the second and third highest crime type. Assault, motor vehicle theft and deceptive practices also had high crime count indicating combination of violence, car theft and financial fraud crime. This crime distribution shows an even balance of personal and property crime alongside new trends, such as an increase in auto theft and persistent offences involving firearms in Chicago.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 9: Analysis showing the top 10 Crime types.***

## **4.3 Yearly Crime Trends (2019-2024)**

The analysed data showed that crime rates dropped in 2020 and 2021 due to COVID-19 restriction. However, after pandemic, crime increased in 2022 to pre-pandemic levels and the year 2023 had the highest crime count. The data shows a U-shaped pattern, with decreased crime count during pandemic years, followed by a significant increase after the pandemic

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 10: Analysis showing crimes per year***

## **4.4 Arrest Trends Across the Years**

The arrest counts from 2019 - 2024 shows how temporal tends are affected by police word and events in real world. With a total of 56,139 arrests, 2019 had the largest number of arrests prior to the COVID-19 pandemic. in 2020 and 2021, arrest counts decreased with 25,165 arrests in 2021 being the lowest. Arrests did, however, progressively increase after 2022, reaching 35,576 in 2024. This slow recovery indicates that the city is seeing a return to regular police activity and movement patterns. These patterns demonstrate how big data analytics can determine how changes in society and policy can directly affect the results of law enforcement. These trends show how big data analytics can identify how changes in society and policy can directly affect the results of law enforcement.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 11: Analysis showing arrest per year***

## **4.5 Analysis of Crimes with the Highest Arrest Counts**

By aggregating the count of arrests for each crime type between 2019 and 2024, the query identifies which crime categories had the most arrests. The analysis indicates that Battery, Narcotics, Weapons Violation, Theft, Other Offense and Assult are the top crime type with the most arrest.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 12: Top 6 crime category with the highest number of arrest (2019-2024)***

## **4.6 Monthly Trends**

The months with the highest crime occurrence were identified in July (132,206), and August (130,419) which aligns with criminological research that warmer weather correlates with increased and social activity leading to more crime reports.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig13: Analysis showing monthly trends***

## **4.7 Arrest Trends Across Districts**

Districts 11, 7, and 6 made the most arrests between 2019 and 2024. These districts correspond to areas with historically high crime intensity, reflecting both elevated criminal activity and enhanced law enforcement presence.

A screenshot of a computer

AI-generated content may be incorrect.

***Fig 14: Analysis showing Districts with the highest number of arrests***

## **4.8 Summary of Findings**

The analysis reveals that:

* The top three offences that dominate urban crime categories are theft, battery, and criminal damage.
* Major social events, such as the consequences of a pandemic, influence changes in crime trends.
* Arrest trends change by category of crime.
* Crime rates show significant seasonal trends.

These findings demonstrate how big datasets can enhance strategic decision-making in smart city systems.

# **Chapter 5: Conclusion and Future Work**

## **5.1 Conclusion**

This utilising of big data analytics to a vast urban crime dataset of vast urban crime dataset of over 1.4 million records from Chicago (2019–2024) is explored in this research. Important insights such as crime patterns, arrest behaviours, and spatial-temporal trends were obtained through SQL-driven analysis and Python-based preprocessing.

The 5V’s of Big Data were implemented throughout this research workflow to efficiently work with such a vast in volume dataset, while also increasing the validity through preprocessing. Ultimately, statistical aggregation and trend analysis produced significant value.

The results demonstrate that crime in Chicago is marked by seasonal changes, spatial distribution, and high crime types, especially theft. The rates of arrests for different crime types are varied significantly, reflecting both operational policing strategies and the nature of offences. These findings highlight how important data driven methods are for modern urban security. They help make better long-term plans, policies, and allocation of resources.

## **5.2 Future Work**

This study recommends the following future works:

1. To find possible crime hotspots and predict high crime rates, utilising machine learning and predictive modelling.
2. To automate a real-time dashboard using tools such as: Power BI, Tableau or Python to help police departments monitor crime patterns continuously.

# **References**

Batty, M. et al., 2012. Smart cities of the future. The European Physical Journal Special Topics, 214(1), pp.481–518.

Chen, M., Mao, S. & Liu, Y., 2014. Big data: A survey. Mobile Networks and Applications, 19(2), pp.171–209.

Gandomi, A. & Haider, M., 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), pp.137–144.

Hashem, I.A.T. et al., 2016. The rise of “big data” on cloud computing: Review and open research issues. Information Systems, 47, pp.98–115.

Janssen, M., Charalabidis, Y. & Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), pp.258–268.

Kitchin, R., 2014. The data revolution: Big data, open data, data infrastructures and their consequences. London: SAGE.

Laney, D., 2001. 3D data management: Controlling data volume, velocity, and variety. META Group Research Note, pp.1–4.

McAfee, A. & Brynjolfsson, E., 2012. Big data: The management revolution. Harvard Business Review, 90(10), pp.60–68.

Provost, F. & Fawcett, T., 2013. Data Science and its relationship to big data and data-driven decision making. Big Data, 1(1), pp.51–59.

Weisburd, D. & Braga, A.A., 2019. Police innovation: Contrasting perspectives. Cambridge: Cambridge University Press.