FACULTY OF COMPUTATIONAL AND SOFTWARE ENGINEERING

Course: Big Data Analytics

& Business Intelligence

Course Code: SEng-9251

Group B Name	**ID**
1. Nancy Tesfaye	NSR/1221/14
2. Yordanos Eshetu	NSR/2772/14
3. Dibora Shibeshi	NSR/2277/14
4. Bontu Abera	NSR/139/14
5. Ashenafi Aberham	NSR/076/14
6. Gulelat Erena	NSR/917/14

Contents

PROJECT PROPOSAL 3

1. Introduction 3

2. Problem Statement 4

1. Data Engineering Problem 5

2. Predictive Modeling Problem 5

3. Project Objectives 5

Core Objective 5

Specific Objectives 6

4. Data Description 7

4.1 Chicago Crime Dataset 7

4.2 NYC Crime Dataset 7

4.3 Big Data Qualification 8

5. Project Scope 8

5.1 In-Scope Activities 8

5.2 Out-of-Scope Activities 9

6. Methodology (Step-by-Step) 10

7. Significance 13

Academic Significance 13

Practical Significance 13

8. Big Data Requirement Justification 13

Dataset Volume 14

Velocity and Variety 14

Why Big Data Tools Are Required 14

9. Ethical Considerations 14

10. Timeline 16

11. Current Status 17

12. Conclusion 18

PROJECT PROPOSAL

Predictive Policing / Crime Analytics

Chicago Crime and NYC Crime Datasets

1. Introduction

Modern cities generate an enormous volume of crime-related information every single day. Police departments, emergency call centers, city surveillance systems, and public reporting portals continuously record incidents, arrests, complaints, and location-based activities. As these data streams accumulate over years, they form massive historical crime datasets that contain valuable patterns about how crime evolves across time, space, and socio-environmental contexts. However, traditional policing strategies—largely dependent on human judgment, manual reports, and reactive patrols—are not equipped to extract insights from such large-scale, high-dimensional data.

Predictive Policing/Crime Analytics focuses on leveraging Big Data technologies to transform these large datasets into meaningful intelligence. This project specifically examines two of the most detailed and widely used public crime datasets in the world:

- **Chicago Crime Dataset**, which contains over 8 million incident records collected since 2001.
- **NYC Crime Dataset**, extracted from the NYPD Complaint Data Historic dataset, containing several million reports dating back to 2006.

Both datasets are longitudinal, high-volume, and geographically rich, which makes them ideal for applying Big Data processing frameworks such as **Apache Hadoop** and **Apache Spark**. Because these datasets exceed the scale of typical classroom CSV files, they allow us to demonstrate true large-scale data engineering, advanced analytics, and machine learning.

This project aims to design and implement a **complete data analytics pipeline** that can process raw crime data, engineer relevant features, detect temporal trends, and ultimately generate forecasting models. By applying modern time-series techniques such as SARIMA and Prophet, the system will be able to predict crime evolution on a daily, weekly, or monthly basis.

In addition to forecasting, the project seeks to uncover:

- **Seasonality patterns** (e.g., crime increasing in summer)
- **Temporal correlations** (e.g., holiday spikes)
- **Long-term trends** (e.g., declining theft rates)
- **High-risk crime categories** such as assault, battery, theft, or burglary

- **Cross-city comparisons** between Chicago and NYC

The focus on Chicago and NYC is intentional: both cities have dense populations, diverse neighborhoods, and strong public datasets. They also differ in policing strategies, demographics, and urban design, which creates opportunities to compare patterns across cities.

With the rise of **data-driven governance**, predictive policing systems are becoming essential for resource planning, patrol scheduling, hotspot detection, and policy evaluation. However, building such systems requires more than simply loading a CSV file. It requires:

- **Big Data ingestion techniques**
- **Distributed computation using Spark**
- **Data cleaning and preprocessing at scale**
- **Aggregation of millions of records by date**
- **Statistical stationarity testing**
- **Parameter tuning for forecasting models**
- **Visualization and interpretation of results**

This project therefore simulates a professional end-to-end crime analytics workflow as would be implemented in real police departments or city analytics teams.

Ultimately, the purpose of this research is not only to forecast crime but also to demonstrate how Big Data technologies can enhance modern policing. By integrating historical datasets with scalable ETL pipelines and predictive modeling, the system offers a data-driven approach that can support decision-makers, improve public safety strategies, and contribute to the growing field of urban analytics.

2. Problem Statement

Major cities such as **Chicago** and **New York City** generate extremely large volumes of crime records every year. These datasets are not small or neatly structured; instead, they consist of millions of semi-structured entries that must be cleaned, standardized, and transformed before any meaningful analysis can be performed. Because of their size, complexity, and irregular formatting, traditional tools (Excel, small Python scripts, basic databases) cannot process them efficiently.

To analyze crime patterns over long time periods and support *Predictive Policing/Crime Analytics*, a scalable Big Data approach is required.

The core problem addressed in this project is therefore twofold:

1. Data Engineering Problem

The raw Chicago and NYC crime datasets contain:

- inconsistent date formats
- missing or malformed entries
- noise and redundant fields
- millions of rows requiring distributed processing

Before the data can be used for prediction, it must be extracted, cleaned, and aggregated using Big Data technologies such as **Hadoop** and **Apache Spark**. Without this preprocessing, statistical insights and forecasting models cannot be generated.

2. Predictive Modeling Problem

Once cleaned and structured, the data must be modeled to identify:

- temporal crime patterns
- seasonal fluctuations
- long-term trends
- daily/weekly/monthly crime cycles

This requires advanced time-series forecasting methods capable of handling large datasets, irregular patterns, and seasonality. The project aims to build models that can forecast future crime levels, supporting data-driven decision-making in policing.

The project seeks to solve the challenge of transforming massive, real-world crime datasets from Chicago and NYC into structured analytical formats and developing predictive models that reveal how crime evolves over time. This requires both **Big Data processing** and **statistical forecasting**, forming a complete end-to-end crime analytics pipeline.

3. Project Objectives

Core Objective

To develop a scalable **Big Data–powered predictive crime analytics system** that processes, analyzes, and forecasts crime patterns using the **Chicago Crime** and **NYC Crime** datasets from the *Predictive Policing/Crime Analytics (Group B)* project scope.

Specific Objectives

1. Data Acquisition

- Gather multi-year crime records from official open-data portals for Chicago and New York City, focusing on datasets large enough to demonstrate true Big Data processing requirements.

2. Distributed Data Storage

- Store the raw datasets within a distributed file system (e.g., HDFS storage) to support large-scale processing and parallel computation.

3. Big Data ETL Pipeline

- Develop an Apache Spark-based ETL workflow that:
  - parses and standardizes inconsistent date/time fields
  - handles missing or malformed entries
  - extracts relevant attributes (location, offense type, timestamps)
  - aggregates the data into structured analytical formats

4. Construction of Time-Series Datasets

- Transform the cleaned data into daily, weekly, and monthly crime counts for both cities, enabling temporal analysis and forecasting.

5. Exploratory Crime Analysis

- Perform descriptive analytics to reveal:
  - long-term crime trends
  - seasonal variations
  - peak activity periods
  - anomalies and irregular patterns

6. Predictive Modeling

- Apply advanced time-series forecasting models, including SARIMA or similar statistical methods, to predict future crime levels for each city.

7. Cross-City Comparative Analysis

- Compare crime patterns between Chicago and NYC to identify:
  - shared seasonal behaviors

  - differences in long-term trends
  - city-specific crime dynamics

8. Reporting and Visualization

- Present findings through:
  - interactive dashboards
  - charts and time-series visualizations
  - written analytical reports
  - model performance summaries

4. Data Description

4.1 Chicago Crime Dataset

The **Chicago Crime dataset** is one of the largest publicly available municipal crime datasets in the United States. It contains:

- **More than 7 million crime incident records** spanning from **2001 to the present**
- Semi-structured entries with inconsistencies typical of long-term public datasets
- Key fields such as:
  - **Date and time** of the offense
  - **Primary crime category** (e.g., theft, assault, narcotics)
  - **Arrest indicator** showing whether an arrest was made
  - **Geographical coordinates** (latitude, longitude)
  - **Police district and community area identifiers**

The dataset's size, continuous updates, and diverse field formats make it a strong candidate for Big Data techniques, particularly distributed processing and large-scale time-series construction.

4.2 NYC Crime Dataset

The **NYC (New York City) Complaint-Level Crime dataset** contains:

- **Several million crime complaint reports** from **2006 to the present**
- Records that vary by year in format and completeness, which requires careful cleaning and standardization
- Key fields including:
  - **Incident date and reporting time**
  - **Offense classification** (felony, misdemeanor, violation)
  - **Borough information** (e.g., Manhattan, Bronx, Queens)
  - **Precinct and jurisdiction details**

Like the Chicago dataset, its size, longitudinal structure, and variability require scalable Big Data pipelines for processing and analysis.

4.3 Big Data Qualification

Both datasets clearly meet Big Data criteria due to:

- **Volume:** Millions of rows spanning decades
- **Variety:** Mixed data types, inconsistent formatting, and evolving schemas
- **Velocity:** Frequently updated and continuously growing datasets
- **Complexity:** Requires distributed systems to process, transform, and model efficiently

This project treats these datasets not as simple CSV tables but as **massive, evolving data streams** requiring Hadoop/Spark-level computation to extract meaningful trends and predictive insights.

5. Project Scope

5.1 In-Scope Activities

This project focuses on building a complete Big Data analytical pipeline for crime forecasting using Chicago and NYC crime datasets. The scope includes the following components:

1. Big Data Engineering with Apache Spark

Implementation of distributed data processing workflows capable of handling multi-million-record datasets. Spark will be used to clean, transform, aggregate, and prepare data for analytics at scale.

2. Data Storage and Management

Raw and processed datasets will be stored in a Hadoop-based environment. This ensures scalability, accessibility, and the ability to reprocess or extend the dataset as needed.

3. ETL (Extract–Transform–Load) Pipeline Development

A full ETL sequence will be designed to:

- Ingest raw CSV or parquet files
- Standardize formats across both cities
- Handle missing, duplicated, or inconsistent records
- Generate unified, analysis-ready time-series datasets

This pipeline will be automated so new data can be incorporated without redesign.

4. Time-Series Modeling and Forecasting

The project will construct analytical models for daily, weekly, and monthly crime counts. Techniques may include:

- Seasonal ARIMA (SARIMA)
- Exponential Smoothing
- Trend and seasonality decomposition

These models will be used to identify long-term tendencies, cyclical patterns, and future projections.

5. Visualization and Reporting

Interactive and static visualizations will be developed to present:

- Historical crime patterns
- Seasonal effects
- Cross-city comparisons
- Forecasted future crime trends

A final report and presentation will summarize findings, limitations, and real-world implications.

5.2 Out-of-Scope Activities

To maintain ethical and academic boundaries, the following activities are **explicitly excluded** from the project:

1. Identification of Individual Offenders

No attempt will be made to track, classify, or identify specific persons involved in criminal incidents.

2. Predictive Profiling

The project will not develop models that infer characteristics about individuals, groups, or communities beyond aggregated crime patterns.

3. Geospatial Policing Recommendations

No operational policing decisions—such as officer deployment, hotspot targeting, or patrol routing—will be generated unless considered as an optional extension for general academic analysis only.

6. Methodology (Step-by-Step)

This project follows a structured Big Data analytics workflow designed to manage, process, and model large-scale crime datasets from Chicago and NYC. Each step ensures that the system can handle millions of records while producing accurate, reproducible analytical results.

STEP 1 — Data Acquisition

Raw datasets covering multiple decades of crime activity will be obtained directly from:

- The Chicago Data Portal
- The NYC Open Data Portal

The files will be downloaded in CSV or parquet format and placed into a clearly organized directory structure (e.g., /raw_data/chicago, /raw_data/nyc). This ensures traceability, version control, and ease of ingestion into the Big Data pipeline.

STEP 2 — Big Data Storage Setup

To support distributed processing, the datasets will be placed in one of the following environments:

Option A — Hadoop Distributed File System (HDFS)
Used when deploying the system on a Big Data cluster.

Option B — Local Storage Structured for Parallel Reads
Folders will be optimized for Spark by:

- Splitting data into manageable chunks
- Ensuring consistent naming
- Avoiding nested unstructured folders

This step ensures Spark can scan the datasets in parallel for high-performance processing. And we used option B.

STEP 3 — ETL Pipeline Using Apache Spark

A full Extract–Transform–Load (ETL) pipeline will be developed to convert raw crime tables into unified and analysis-ready datasets.

EXTRACT

- Load multi-million-row crime logs into Spark using distributed readers.
- Infer schema or manually define datatypes where needed.

TRANSFORM

Comprehensive data cleaning and restructuring will be performed, including:

- Handling missing values and invalid entries
- Converting and standardizing date/time fields
- Normalizing crime type names across Chicago and NYC
- Dropping irrelevant or redundant fields
- Filtering for major and comparable crime categories (e.g., theft, assault, burglary)

This stage ensures consistency between both datasets, enabling accurate multi-city comparison.

LOAD

- Save cleaned datasets into /processed/ directories
- Generate and store daily, weekly, and monthly aggregated crime counts
- Output files stored as CSV or parquet for efficient future reads

STEP 4 — Time-Series Dataset Creation

After ETL, crime incidents will be transformed into structured time-series tables. For each crime category and for total crime counts, Spark/Pandas will produce:

- **Daily crime time-series**
- **Weekly crime time-series**
- **Monthly crime time-series**

These files will later serve as input to forecasting models.

STEP 5 — Exploratory Data Analysis (EDA)

A rigorous analysis will be conducted to understand the behavior of crime trends. This will include:

- Seasonal trend identification
- Year-over-year crime comparisons
- Visualization of long-term patterns
- Detection of anomalies or outliers
- Decomposition into trend, seasonality, and residual components

This stage reveals the underlying structure of the data and guides model selection.

STEP 6 — Forecasting Models

A suite of statistical forecasting models will be implemented and evaluated. These include:

- **ARIMA** for baseline trend analysis
- **SARIMA** for datasets with strong seasonality (common in crime patterns)
- **Holt-Winters Exponential Smoothing** for smoother long-term forecasts

Models will be trained on historical data and validated using hold-out test sets. Forecast graphs will show predicted crime levels for future periods.

STEP 7 — Visualization

Comprehensive visual analytics will be generated using Python libraries (Matplotlib, Plotly, Seaborn). Visual outputs include:

- Long-term crime trend lines
- Month-to-month seasonality plots
- Crime peak detection charts
- Forecasted future crime behavior
- Side-by-side Chicago vs. NYC comparisons

These visualizations will support clear communication of analytical findings.

STEP 8 — Final Documentation

A complete written report will be produced, containing:

- Overview of the Big Data pipeline
- ETL methodology and architecture diagrams
- Statistical analysis explanation
- Time-series modeling results
- Interpretation of seasonal patterns
- Limitations and future work recommendations

This ensures the project meets academic, methodological, and technical requirements for Big Data analytics.

7. Significance

Academic Significance

This project provides a comprehensive demonstration of Big Data engineering and analytical methodologies within a real-world context. By working with multi-million-record datasets from Chicago and NYC, the project:

- **Shows proficiency in Big Data ecosystems**, including Hadoop-style storage principles and distributed processing using Apache Spark.
- **Combines multiple technical disciplines**—data ingestion, cleaning, transformation, time-series modeling, and visualization—into a unified workflow.
- **Reinforces applied machine learning skills**, specifically in statistical forecasting, model evaluation, and interpretability.
- **Provides hands-on experience with end-to-end data pipelines**, mirroring processes used in professional data engineering and data science environments.

This aligns directly with academic outcomes expected in Big Data, Analytics, and Machine Learning coursework.

Practical Significance

Beyond academic benefit, the project generates insights with real societal relevance. Crime data reflects dynamic patterns influenced by location, time, and behavior. By applying predictive analytics to these datasets, the project:

- **Reveals underlying temporal patterns**, such as seasonal crime fluctuations, weekday vs. weekend differences, and long-term trends.
- **Identifies periods of heightened risk**, which can be valuable for resource allocation or strategic planning.
- **Demonstrates how forecasting tools could support public safety stakeholders**, including police departments and urban planners.
- **Highlights the potential of data-driven decision-making**, showing how cities can leverage historical information to anticipate future needs.

In essence, the project bridges technical capability with meaningful real-world application

8. Big Data Requirement Justification

The datasets used in this project fully qualify as **Big Data**, not only because of their size but also because of their structure, variability, and processing demands.

Dataset Volume

- The **Chicago Crime** dataset contains **7+ million records** spanning more than two decades.
- The **NYC Crime** dataset includes **millions of complaint reports**, updated annually and covering nearly 20 years.

These file sizes exceed what traditional tools such as Excel or basic Python can handle efficiently, especially when performing multi-level aggregation, filtering, and time-series transformations.

Velocity and Variety

- The datasets contain **semi-structured records** with inconsistent timestamps, varied crime type labels, missing fields, and multiple formats.
- Cleaning and normalizing these inconsistencies requires scalable tools capable of parallel processing.

Why Big Data Tools Are Required

Processing millions of rows using a single machine is slow, memory-intensive, and often unreliable.
Using **Apache Spark** is essential for:

- **Distributed computation:** Spark divides operations into parallel tasks across multiple cores or cluster nodes.
- **Efficient memory management:** Avoids RAM overload when reading large CSV files.
- **High-speed transformations:** Performs ETL operations (filtering, grouping, aggregating) on millions of rows in seconds.
- **Fault tolerance:** Automatically handles failures through resilient distributed datasets (RDDs).

9. Ethical Considerations

Ensuring ethical integrity is essential when working with crime-related datasets, especially those collected from real populations. This project incorporates the following safeguards:

No Personal Identifiers

The datasets from Chicago and NYC do **not contain names, addresses, phone numbers, or any form of personally identifiable information (PII)**. All analysis is conducted strictly on anonymized crime logs, ensuring full compliance with privacy requirements.

Aggregated, Not Individual-Level Analysis

All computations—whether daily counts, trends, or forecasts—are performed at **aggregate levels** (e.g., per day, per month, per crime category). The project **does not analyze or profile individual offenders or victims**, avoiding any ethical risks related to surveillance or personal targeting.

Non-Prescriptive Predictions

Predictive outputs are **statistical forecasts** based on historical trends. They are not intended to:

- Predict specific individuals' actions
- Direct policing toward particular groups
- Influence policy decisions without broader contextual analysis

This ensures the modeling remains academic and informational rather than operational or punitive.

Responsible Use of Data

- All datasets are **public-release governmental data**, legally accessible for academic research.
- Processing methods follow **ethical data handling standards**, ensuring no misuse, misrepresentation, or unauthorized redistribution.
- Results are presented with appropriate caution, acknowledging limitations and avoiding overstated interpretations.

Bias Awareness

Crime datasets often contain **systemic biases** (e.g., over-policing in certain communities). To address this:

- Analyses are focused on **temporal patterns**, not demographic or spatial profiling.
- Interpretations avoid reinforcing harmful stereotypes or assumptions.

10. Timeline

The project will be executed in structured phases to ensure organized progress from raw data acquisition to final reporting. Each phase builds on the previous one, following a realistic and academically appropriate workflow.

Phase 1 — Dataset Acquisition (Week 1)

- Download multi-year crime datasets for Chicago and NYC.
- Verify dataset formats, file sizes, and schema consistency.
- Organize files into structured directory systems prepared for Big Data processing.

Phase 2 — ETL Pipeline Implementation (Weeks 2–3)

- Set up the Big Data environment (Spark/Hadoop local setup).
- Design and implement ETL scripts to clean, filter, and normalize raw crime records.
- Generate structured outputs: daily, weekly, and monthly aggregation tables.
- Validate output correctness and ensure no missing or corrupted intervals.

Phase 3 — Time-Series Construction (Week 4)

- Convert processed datasets into complete time-series structures.
- Handle gaps, duplicates, or inconsistent timestamps.
- Prepare modeling-ready files for both Chicago and NYC crime categories.

Phase 4 — Modeling and Predictions (Weeks 5–6)

- Perform statistical diagnostics (stationarity tests, decomposition, autocorrelation).
- Train forecasting models such as ARIMA, SARIMA, and Holt-Winters.
- Compare performance metrics and generate future crime predictions.
- Document analytical findings for both cities.

Phase 5 — Visualization and Final Report Writing (Weeks 7–8)

- Produce clear visualizations: trend lines, seasonal patterns, heatmaps, forecasts.
- Compare city-to-city patterns and highlight key insights.
- Compile a comprehensive academic project report covering:
    - Background
    - Methodology
    - Results
    - Discussion
    - Limitations

    - Ethical considerations
- Finalize the deliverables for submission and presentation.

11. Current Status

The project is presently in its **initial planning and design stage**. At this point, the following preliminary steps have been completed:

- **Data sources identified:**
  The Chicago Crime dataset (2001–present) and the NYC Crime Complaint dataset (2006–present) have been confirmed as the primary data sources.
- **Tool selection completed:**
  The Big Data processing framework (Apache Spark), storage approach, and analytical tools (Python, time-series libraries, visualization packages) have been chosen based on project requirements.
- **System architecture drafted:**
  A high-level ETL and modeling pipeline has been designed, outlining how raw crime logs will be ingested, cleaned, transformed, and prepared for forecasting.

No data processing, cleaning, statistical analysis, or forecasting has been performed yet. The project will now proceed into the implementation phase following the methodology outlined in this proposal.

12. Conclusion

This project aims to develop a comprehensive **Predictive Policing/Crime Analytics** framework using large-scale crime datasets from **Chicago** and **New York City**. By leveraging Big Data technologies particularly Apache Spark the system will efficiently process millions of crime records, transforming raw semi-structured logs into clean, analyzable time-series datasets.

Through the integration of ETL engineering, statistical modeling, and advanced forecasting techniques, the project will uncover meaningful crime patterns, seasonal behaviors, and long-term trends. The resulting predictive insights can support academic research, urban planning, and data-driven public safety strategies.

The intended workflow fully satisfies the instructor's Big Data requirements, demonstrating the ability to handle large datasets, apply distributed processing, and perform machine-learning-based forecasting. As the project progresses through implementation, it will culminate in a robust analytical pipeline and a well-documented final report.

Overall, this project represents a rigorous, technically grounded approach to crime analytics and predictive modeling, and it is well-positioned to deliver valuable outcomes for real-world relevance.