**GROUP 5**

**Project Title**

**“Big Data Analytics for Crime Pattern Detection and Prevention in Urban Areas”**

**1. Introduction**

In recent years, data-driven decision-making has become critical for law enforcement and public safety. This project focuses on building a scalable big data pipeline to analyze publicly available crime datasets. Using distributed processing and visualization tools, the project enables stakeholders to derive actionable insights and patterns in criminal activity across different regions.

**2. Objective**

* To build a data analytics pipeline for crime data.
* To process large volumes of data efficiently using Apache Spark on Amazon EMR.
* To design a Hive-based data warehouse for structured query access.
* To develop interactive visualizations for identifying crime patterns and trends.

**3. Problem Statement**

Despite the availability of extensive crime-related data, cities often struggle to identify high-risk areas, patterns in criminal behaviour, and resource allocation needs. This project aims to analyze historical crime records using big data technologies to extract actionable insights that can help law enforcement agencies:

* Understand crime distribution by type and location,
* Identify trends and anomalies across years and neighbourhoods,
* Support data-driven policing through dashboards.

**4. Attribute description (43 columns)**

| **Attribute** | **Description** |
| --- | --- |
| id | Unique ID for each case |
| Case Number | Police-assigned case number |
| Block | Location or street block where the incident occurred |
| NIBRS Code | Crime classification code based on National Incident-Based Reporting  System (NIBRS) standard |
| Primary Type | Main category of crime (e.g., Theft, Robbery) |
| Description | Detailed crime type |
| Location Description | Type of location (e.g., Street, Hotel, Beach) |
| Arrest | Whether arrest was made (TRUE/FALSE) |
| Domestic | Whether it's a domestic incident |
| Beat | Police beat identifier |
| Area | Local area or district |
| Ward | Political ward |
| Community Area | Named community zone |
| FBI Code | FBI classification code |
| X Coordinate | X map coordinate |
| Y Coordinate | Y map coordinate |
| Year | Year of the crime |
| Latitude | Geographic latitude |
| Longitude | Geographic longitude |
| Location | Combined location coordinates |
| weapon description | Weapon involved (e.g., Club, Knife, Vehicle) |
| Vict Age | Age of the victim |
| Vict Sex | Sex of the victim |
| Victim Race | Race of the victim |
| Date OCC | Date of occurrence |
| Time OCC | Time of occurrence |
| Date Rptd | Date the crime was reported |
| Time Rptd | Time the crime was reported |
| Date Arrested | Date of arrest, if any |
| Time Arrested | Time of arrest |
| Premises desc | Description of premises (e.g., School, Park) |
| district | Police district code |
| Suspect Age | Age of the suspect |
| Suspect Sex | Sex of the suspect |
| Suspect Race | Race of the suspect |
| Case Status | Current status (e.g., Closed, Under Investigation) |
| Crime Category | Broad crime classification (e.g., Violent, Property) |
| Secondary Description | Additional classification |
| Census Tract | U.S. Census tract number |
| Zip Code | Zip code of the incident |
| Incident Narrative | Short summary of the event |
| Priority Level | Threat level (e.g., High, Medium, Low) |
| Repeat Offense Flag | Indicates repeat offender involvement (TRUE/FALSE) |

**5. Tools & Technologies Used**

* **Python & SQL** – for scripting and querying
* **Amazon S3** – for cloud data storage
* **Amazon EMR** – to run Spark and Hive in a distributed environment
* **Apache Spark (PySpark)** – for distributed data processing
* **Apache Hive** – for structured data storage and SQL queries
* **ODBC Connector** – to link Hive to BI tools
* **Power BI / Tableau** – for data visualization and dashboarding

**6. Workflow of the Project (Pipeline)**

1. **Data Collection** – Download crime data from Data.gov (CSV).
2. **Data Storage (S3)** – Store raw and processed data in Amazon S3.
3. **Data Processing (EMR)** – Clean and preprocess data using EMR.
4. **Data Processing (Spark)** – Perform aggregations, KPI calculations, and data transformations.
5. **Data Warehousing (Hive)** – Load final datasets into Hive for querying.
6. **Visualization (ODBC + BI)** – Connect Hive to Power BI/Tableau via ODBC for dashboards.

**7. Key Performance Indicators (KPIs)/ Data Visualization**

* Top 10 Crime Categories
* Arrest VS Non-Arrest Percent
* Crime Category-wise arrest and non-arrest
* Yearly Crime Trend
* Monthly crime Trend
* Domestic Crime Trend
* % of cases in which weapons are involved
* Crime category-wise no. of cases
* Weapon usage by crime type
* Top 10 NIBRS Group A of offenses
* Theft trend over the years

**8. Methodology**

1. **Data Ingestion:** Load and store data using Spark from CSV or HDFS/S3.
2. **Data Cleaning & Preprocessing:** Handle missing values, standardize fields, extract time-based features.
3. **Exploratory Data Analysis:** Use Spark SQL and Python to compute trends and correlations.
4. **KPI Computation:** Generate aggregate metrics grouped by category, geography, and time.
5. **Visualization & Dashboarding:** Publish findings on BI tools for dynamic exploration.

**9. Expected Outcomes**

* Interactive dashboards for decision-makers to monitor crime rates and trends
* Identification of high-risk areas and demographics
* Insightful reports to optimize police patrol deployment and preventive measures

**10. Architecture Diagram:**

