Project Title: Crime Statistics in UK

Team Name: Team Maverick

Authors:

1. Kavipriya Ramasamy (kavipriyar@iisc.ac.in)
2. Srividya Lakshmanan (srividhyal@iisc.ac.in)
3. Kokane Manoj Bhausaheb (manojkokane@iisc.ac.in)
4. Aravind SS (aravindss@iisc.ac.in)

Problem Statement:

Definition:

Analyse, process, and optimize the handling of crime statistics using Apache Spark to derive actionable insights. Identify crime trends, hotspots, and patterns to enhance public safety and law enforcement strategies.

Problem Motivation:

By analysing crime data, valuable insights can be drawn to improve public safety, refine law enforcement strategies, and optimize resource allocation. Managing the large volume of static datasets requires efficient big data processing techniques to generate actionable insights.

Design Goals and Features Supported:

- Efficient data processing for large crime datasets.
- Capability to filter, aggregate, and analyse crime statistics based on factors such as location, time, and crime type.
- Identification of crime hotspots and temporal trends.
- Interactive querying of crime statistics for dynamic reporting.
- Scalable pipeline accommodating future integration of new data.
- Support for exploratory data analysis and visualization tools.
- Scalability/Performance Goals: Handle millions of records efficiently and deliver insights in real-time.

High-Level Design:

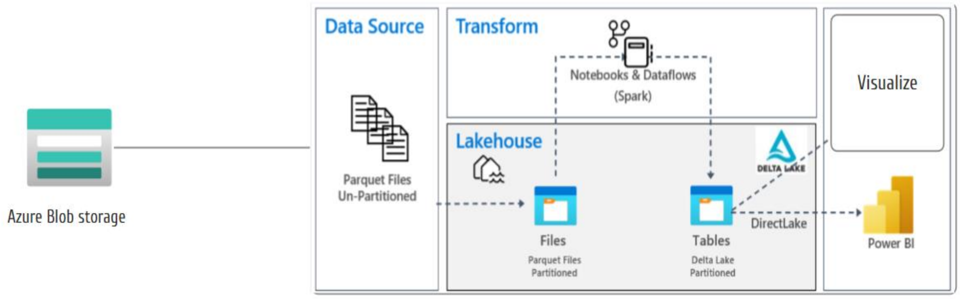


Fig.1 Architecture flow diagram

- **Initial Setup**
  The initial setup for this project involves configuring the Apache Spark environment in fabric to handle the large crime dataset efficiently. This includes setting up a Spark environement (in Microsoft Fabric) to enable distributed data processing.
  The goals for this phase are:
  - Ensure the Spark cluster is operational and optimized for handling large-scale data.
  -Configure storage options such lakehouse in Microsoft Fabric to store the data in different stages.
  -Schema such as Dimesion,Landing, Stage and Reporting are created to organize the data in different steps.
  -Configure the storage account in azure account and create the container to hold the input data. -Data is organized into different directories based on the month and year
  -Identifying connectivity and authentication to access the storage account in Microsoft Fabric environment

- **Data Ingestion**
  The crime dataset is ingested from a static source, such as a CSV file. API ingestion for dimension data directly from the official source data. (https://data.police.uk/docs/method/forces/)

  The ingestion process involves:
  - Loading the dataset into Spark DataFrames for initial processing.
  - Ensuring schema inference or manual schema definition to handle various data types (e.g., categorical, numerical).
  - Validating the dataset by checking for structural integrity and completeness, such as verifying headers and row consistency. The Spark framework facilitates efficient ingestion by leveraging distributed reads and parallelism, which is crucial for large datasets.

- **Data Preprocessing**

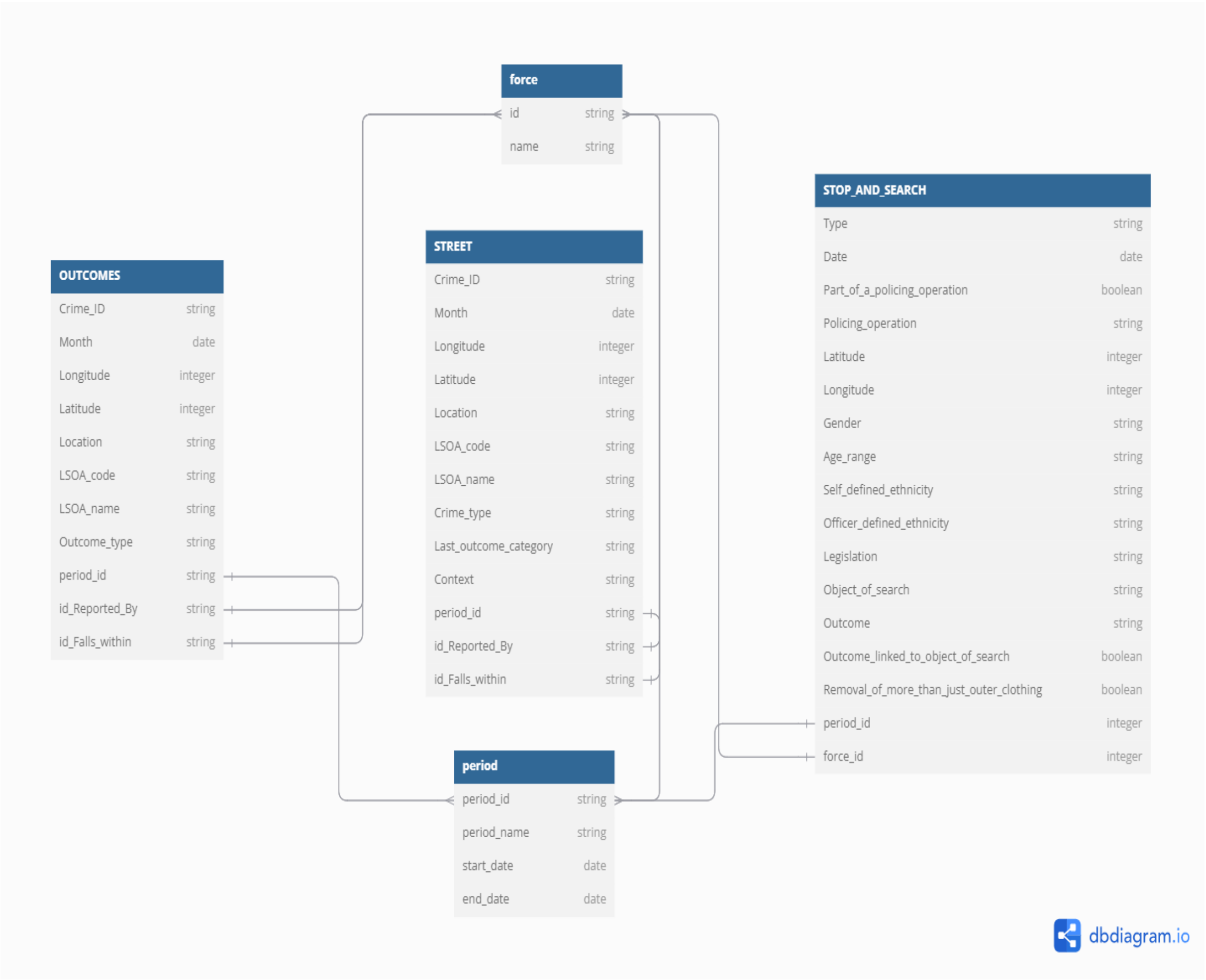
  Data preprocessing ensures the dataset is clean, uniform, and ready for analysis. The key steps in this phase include:

  - **Cleaning Missing or Corrupt Data**: Dropping or imputing missing values depending on the significance of the data point.

- **Column Transformation**: Transforming columns such as timestamps into a standardized datetime format or geolocation into latitude/longitude.
- **Normalization**: Ensuring consistency in categorical data, such as crime type names, by removing duplicates or standardizing labels.
- **Deduplication**: Eliminating duplicate rows to ensure accurate analysis. This step is critical for maintaining data quality and enabling reliable analysis downstream.

- **Transformation**
  Transformation involves modifying the dataset into a structure suitable for analysis and visualization. This includes:
  - **Data Partitioning**: Partitioning the dataset by time (e.g., year, quarter) and location to optimize query performance.
  - **Aggregation**: Summarizing data, such as calculating the total number of crimes per location or period, to identify trends. These transformations ensure the dataset is ready for advanced analysis using Spark SQL and machine learning pipelines.

Data Model:



force	
id	string
name	string

OUTCOMES	
Crime_ID	string
Month	date
Longitude	integer
Latitude	integer
Location	string
LSOA_code	string
LSOA_name	string
Outcome_type	string
period_id	string
id_Reported_By	string
id_Falls_within	string

STREET	
Crime_ID	string
Month	date
Longitude	integer
Latitude	integer
Location	string
LSOA_code	string
LSOA_name	string
Crime_type	string
Last_outcome_category	string
Context	string
period_id	string
id_Reported_By	string
id_Falls_within	string

STOP_AND_SEARCH	
Type	string
Date	date
Part_of_a_policing_operation	boolean
Policing_operation	string
Latitude	integer
Longitude	integer
Gender	string
Age_range	string
Self_defined_ethnicity	string
Officer_defined_ethnicity	string
Legislation	string
Object_of_search	string
Outcome	string
Outcome_linked_to_object_of_search	boolean
Removal_of_more_than_just_outer_clothing	boolean
period_id	integer
force_id	integer

period	
period_id	string
period_name	string
start_date	date
end_date	date

Big Data Platforms Used:

- Microsoft Fabric: Microsoft Fabric is an end-to-end analytics and data platform designed for enterprises that require a unified solution. It encompasses data movement, processing, ingestion, transformation, real-time event routing, and report building.
- Apache Spark: Distributed data processing and analysis.
- Azure Blob Storage: Scalable data storage.
- PowerBI: Visualization and exploratory data analysis.

Scalability/Performance Metrics:

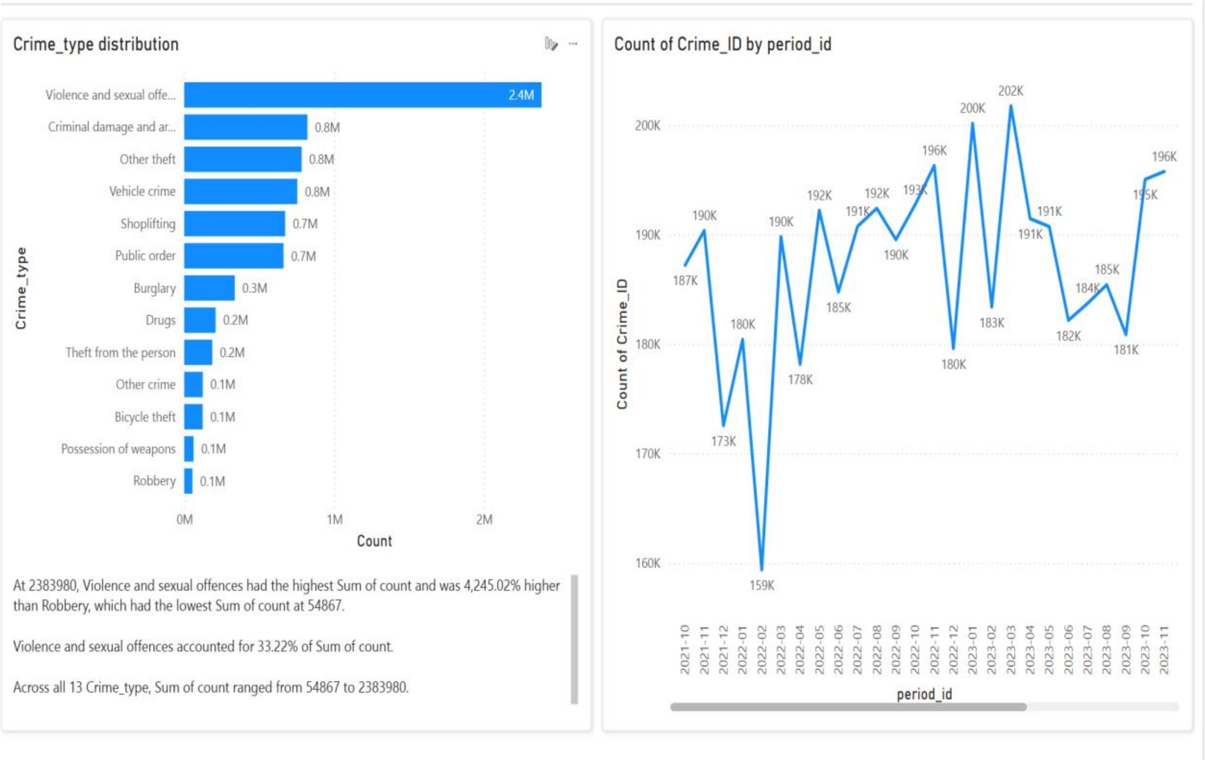
1.**Big Data Processing:**

- PySpark enables distributed computation for scalable processing of large datasets.
- Partitioning by time and leveraging lazy execution optimize query performance.
- Incremental pipeline with staging layer to hold the historical data and ingesting only the incremental data

2. **Runtime**: a. Less than few minutes to refresh whole data for incremental load

b. 1.3 hours for full load(3 years of data) to refresh whole data

3. **Plots and Analysis:**

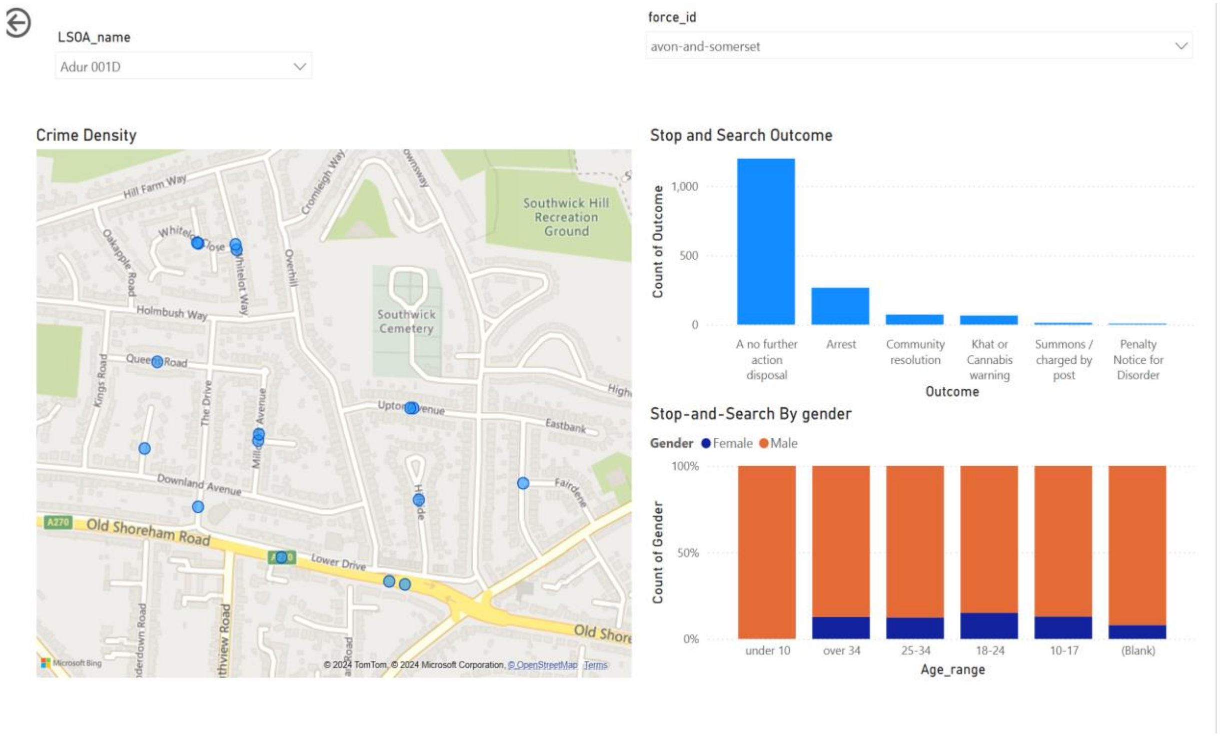


At 2383980, Violence and sexual offences had the highest Sum of count and was 4,245.02% higher than Robbery, which had the lowest Sum of count at 54867.

Violence and sexual offences accounted for 33.22% of Sum of count.

Across all 13 Crime_type, Sum of count ranged from 54867 to 2383980.

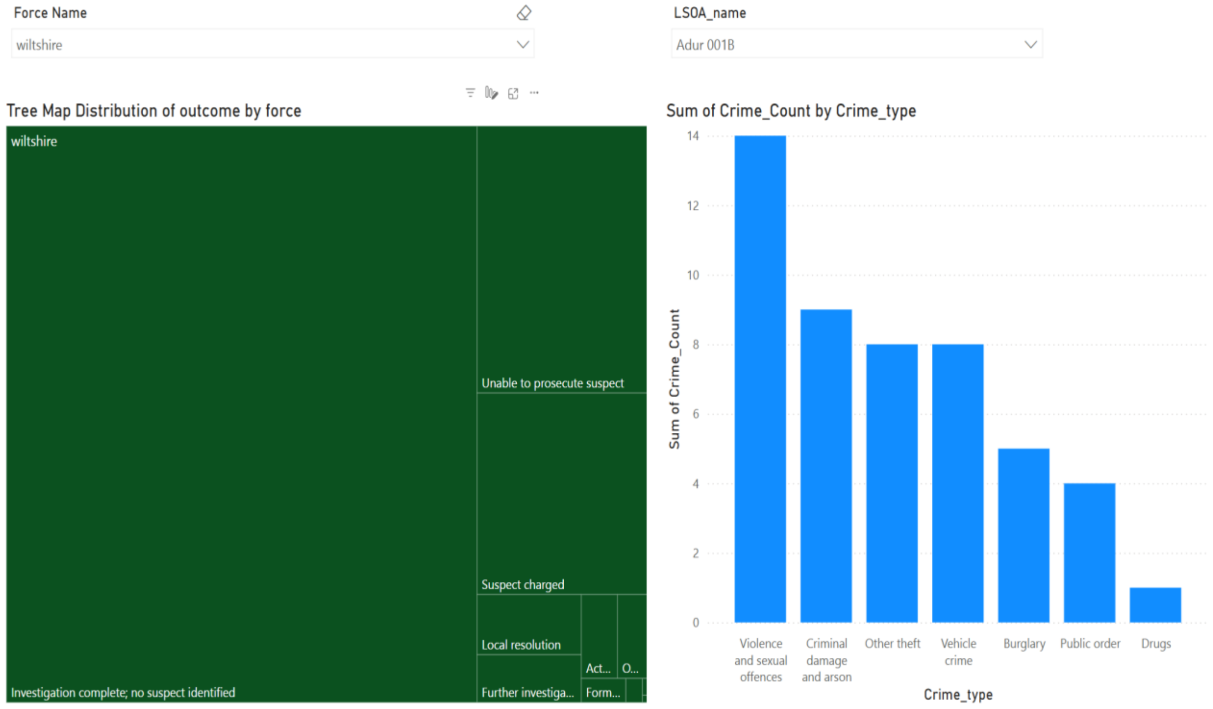
Plot a : Summary of overall crime types and total crimes over time



LSOA_name: Adur 001D

force_id: avon-and-somerset

Plot b: Crime Density and Stop and Search data



Force Name: wiltshire

LSOA_name: Adur 001B

Tree Map Distribution of outcome by force
wiltshire
Unable to prosecute suspect
Suspect charged
Local resolution
Act...
O...
Investigation complete; no suspect identified
Further investiga...
Form...
Sum of Crime_Count by Crime_type
14
12
10
8
6
4
2
0
Sum of Crime_Count
Violence and sexual offences
Criminal damage and arson
Other theft
Vehicle crime
Burglary
Public order
Drugs
Crime_type

Plot c: Force wise outcome and Area wise crimetype distribution

Summary of Approach:

The insights derived from this analysis can inform policy recommendations, enabling targeted interventions and efficient resource allocation. New data can be ingested and loaded incrementally so that latest trends are available on dashboard.

Through advanced data preprocessing, incremental data ingestion, and dynamic visualizations, the project enables stakeholders to derive actionable insights, identify crime trends, and design targeted interventions. The use of distributed computing and modular architecture ensures both efficiency and adaptability.

Looking forward, the pipeline can be enhanced to incorporate additional data sources, such as socio-economic factors, to provide deeper context and predictive capabilities. Integrating machine learning models could further refine hotspot predictions and forecast crime trends, enabling proactive measures. Additionally, expanding the reporting capabilities to include geospatial analyses and real-time alerts would make the solution even more impactful for law enforcement and policy planning. By maintaining modularity and scalability, the framework can adapt to evolving needs, ensuring its long-term relevance and effectiveness.