Project Proposal

Chicago Crime Dataset

Group No. 3:

Members:

1. Farzeem Jiwani
2. Agha Ahmad
3. Faheem Mohammad

Predictive Analysis on Chicago Crime Dataset

I. Problem Definition

Crime is an important socio-economic matter that affects public safety and needs to be addressed. The primary objective is to perform predictive analysis on the dataset in order to comprehend whether crime is a function of locality, time, climate, or any external features, along with analyzing its trends over the years. This could be helpful to law enforcement to take appropriate measures about the *When-What-Where* of the crime i.e. when and what type of crime could happen in what locality.

A few questions that we would try to answer would be:

1. How have arrests evolved over the years?
2. What time of the day do most crimes occur?
3. What are the types of locations where crimes occur most frequently?
4. How do domestic crimes compare to other crimes?
5. Are some types of crimes more likely to happen in specific locations or a specific time of the day or specific day of the week than other types of crimes?
6. Weekly/ Monthly/ Yearly crime analysis

II. Suggested Solution

The dataset contains about 7.4 million records that need to be processed fast. We would be using the Apache Spark framework as its in-memory processing capabilities would make it easier to handle the voluminous data.

A brief about the pipeline is as follows:

1. **Data Storage:**
   We would be using HDFS for data storage i.e. uploading the input CSV into HDFS using the Sandbox HDP implementation setup.

2. **Data Pre-Processing:**
   a. Dropped missing/null values that account for <1% of data
   b. Filter out irrelevant features from the dataset
   c. Perform Random sampling techniques (if needed) to balance the data

3. **Data Exploration:**
   In this step, we would analyze important trends for crime detection and prevention. The analysis will also help identify useful features for building predictive models.

4. **Data Querying:**
   In this step, we would be using relational queries to perform data analysis using Spark SQL on top of the dataset to provide insights and understand complex relationships amidst the features.

5. **Predictive Modelling:**
   We intend to use a few models to perform predictive analysis on the final processed dataset and compare their classification performance.

Below is a tabular overview of the tasks described above and the corresponding technology stack:

Sr. No.	**Task**	**Technology**
I.	Data Storage	HDFS
II.	Data Pre-processing	MapReduce, PySpark
III.	Exploratory Data Analysis	PySpark, Seaborn, Folium
IV.	Data Querying	PySpark SQL
V.	Predictive Modelling	Scikit-learn

III. Dataset Descriptions

The dataset reflects reported incidents of crime (except for murders where data exists for each victim) that occurred in the City of Chicago from 2001 to the present. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

Link: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/data

Dataset Characteristics:
Size: 1.7gb
Rows: 7.4 million (~7,406,836)
Columns: 22

Below is the list of some of the key attributes of the dataset along with their description:

Attribute/ Feature	**Description**
Date	Best estimate Date when the incident occurred.
Block	Partially redacted address where the incident occurred.
IUCR	The Illinois Uniform Crime Reporting Code is used to classify criminal incidents when taking individual reports. Reference
Primary Type	Primary description of the IUCR code.
Description	Secondary description of the IUCR code.
Location Description	Description of the location of the incident.
Arrest	Indicates whether an arrest was made.
Domestic	Indicates whether the incident was domestic-related.
District	Indicates the police district where the incident occurred.
Year	Year the incident occurred.
Latitude	Latitude of the location where the incident occurred.
Longitude	Longitude of the location where the incident occurred.
Location	Combination of Latitude and Longitude.