An Analytical Platform for Urban Crime Patterns Using Integrated Datasets

Team Name: Data Catalyst

Authors and Email Address:

- Anfaal Obaid Waafy (anfaalwaafy@iisc.ac.in)
- Manikanda Sakthi Subramanium (manikandasa1@iisc.ac.in)
- Harshit Agarwal (harshita@iisc.ac.in)
- Vimalraj K (vimalrajk@iisc.ac.in)

PROBLEM

Definition

This project will build a scalable data engineering pipeline to integrate historical crime data with weather and demographic datasets. The goal is an analytical platform for large-scale exploratory data analysis (EDA) to uncover crime patterns and inform public safety.

Motivation

Effective urban planning and law enforcement require insights from massive, disparate datasets. This project tackles the core data engineering challenge: building a foundational platform to ingest, join, and analyze large-scale urban data, enabling data-driven decision-making.

Design Goals, Features Supported

The primary goal is to build an end-to-end batch processing pipeline that transforms raw urban data for analysis.

- **Ingestion**: Ingest and clean crime, weather, and census data.
- **Integration**: Join and aggregate data by time and location using Spark.
- **Analysis**: Use Spark SQL to perform large-scale EDA on the integrated dataset.
- **Visualization**: Develop an interactive dashboard to display trends and correlations.

SOLUTION APPROACH

High-level Design

A batch pipeline using Apache Spark will be implemented. Raw data (CSVs, APIs) will be landed in the Hadoop Distributed File System (HDFS). A Spark job will then perform the complete ETL process—cleaning, joining, and aggregating data. The final, integrated table will be queried via Spark SQL to power a visualization dashboard.

Big Data Platforms Used

- **Apache Spark (Spark SQL & DataFrames)**: The central platform for all large-scale ETL, data transformation, and final analytical querying.
- **Hadoop Distributed File System (HDFS)**: Will serve as the data lake for storing raw data and the final processed analytical table.

Data Sources and Data Models

Data Sources:

- **Primary (Crime)**: Chicago Crime Data from 2001 to Present. This dataset contains over 8 million records of reported crimes, including type, date, and location.
  - *Link*: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
- Secondary (Weather): Historical daily weather summaries for Chicago O'Hare International Airport (Station ID: USW00094846) from NOAA's Climate Data Online. This provides daily temperature, precipitation, and other meteorological data.
  - *Link*: https://www.ncei.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail
- Secondary (Demographics): U.S. Census American Community Survey (ACS). We will extract socioeconomic data (e.g., median income, population) at the census tract level.
  - Example API Call: https://api.census.gov/data/2022/acs/acs5?get=NAME,B01003_001E,B19013_001E&for=tract:*&in=state:17&in=county:031

Data Models: The final output of our data engineering pipeline will be a single, wide, and denormalized analytical table. Each row in this table will represent a unique crime incident, enriched with the corresponding daily weather data and the location-specific demographic information from the census tract where the crime occurred. Table will be stored in HDFS using using optimized storage and query performance techniques.

EVALUATION APPROACH

Experiment Plan

The project will be validated across four phases:

1. **Data Ingestion**: Ingest and clean raw data from all sources into HDFS.
2. **ETL Execution**: Run the Spark ETL job and validate the final Parquet table.
3. **Analysis**: Run complex Spark SQL queries to uncover meaningful trends.
4. **Visualization**: Build an interactive dashboard to display the findings.

Success Metrics

- **Analytical Depth**: Generate and visualize at least five significant, non-obvious insights using Spark SQL.
- **Scalability**: Demonstrate the pipeline's efficiency by measuring performance scaling as compute resources are varied.
- **Deliverables**: Produce a final dataset that correctly integrates the three data sources (crime, weather, census) and a dashboard with map and time-series views.