An Analytical Platform for Urban Crime Patterns Using Integrated Datasets

(DA 2310: Data Engineering at Scale - August 2025 Term – Course Project)

Authors

Anfaal Obaid Waafy (anfaal.waafy@iisc.ac.in)
Manikanda Sakthi Subramaniam (manikandasa1@iisc.ac.in)
Harshit Agarwal (harshit.a@iisc.ac.in)
Vimalraj K (vimalraj.k@iisc.ac.in)

Date: 27 November 2025

Table of Contents

1. Problem

1.1 Definition

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

1.2 Motivation

Effective urban planning and law enforcement require insights from massive, disparate datasets. Crime incidents, weather patterns, and socioeconomic factors are collected by different organizations and stored in incompatible formats. This project tackles the core data engineering challenge: building a foundational platform to ingest, join, and analyze large-scale urban data in a unified manner, enabling data-driven decision-making for public safety agencies.

1.3 Design Goals

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

1. **Ingestion**: Ingest and clean crime, weather, and census data from multiple sources
2. **Integration**: Join and aggregate data by time and location using Apache Spark
3. **Analysis**: Use Spark SQL to perform large-scale exploratory data analysis on the integrated dataset
4. **Visualization**: Develop analytics dashboards to display trends and correlations

1.4 Features Required

- Processing of 8+ million crime records spanning 2001 to present
- Daily weather integration (precipitation, temperature, snowfall)
- Census tract demographic enrichment (median income, population)
- Spatial join operations mapping crimes to census tracts
- Support for complex time-series and geospatial queries
- Optimized storage using Parquet format

1.5 Scalability/Performance Goals

- Process 8.4 million crime records efficiently in batch mode
- Spatial join operations complete within 6-10 minutes
- Support year-partitioned queries for time-range filtering
- Memory-efficient checkpoint strategy to prevent out-of-memory errors
- Scalable to larger datasets through distributed Spark cluster

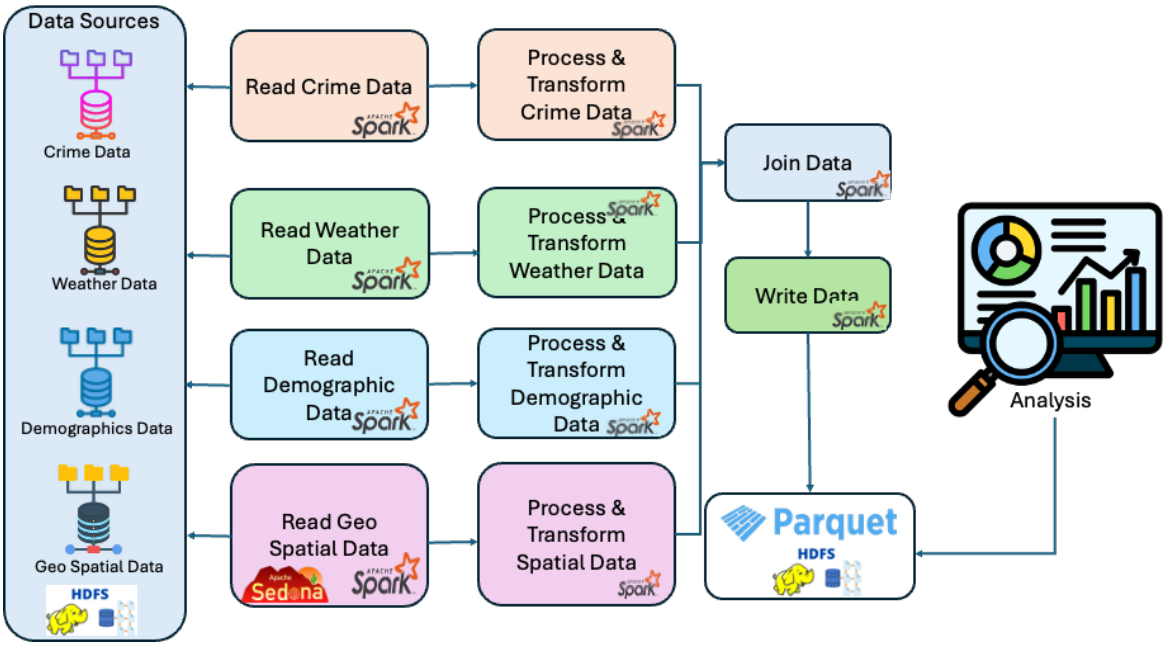
2. Approach and Methods

2.1 High-Level Design

A batch processing pipeline using Apache Spark orchestrates the complete ETL (Extract, Transform, Load) process:

Pipeline Architecture

Data Flow:



Architecture/Data Model

The pipeline implements a **denormalized wide-table model**:

- **Entity**: Each row represents a unique crime incident
- **Dimensions**: Crime attributes (type, date, location, arrest/domestic status) + temporal features (hour, weekday, month)
- **Measures**: Weather metrics (temperature, precipitation, snowfall) + demographics (median income, population)
- **Partitioning**: By year for efficient time-range queries

Data Integration Strategy:

1. Load crime data with temporal parsing and deduplication
2. Process weather data with daily aggregation (handle multiple observations per day)
3. Load ACS demographics from multiple survey years
4. Perform broadcast join of weather (small table: 8,766 rows)
5. Execute spatial containment join using census tract geometries

6. Checkpoint intermediate results to prevent DAG explosion
7. Final demographics join with year-based aggregation

2.2 Big Data Platforms Used

Platform	Role
Apache Spark 3.5.1	Distributed data processing framework; core platform for ETL, transformation, and analytical querying
Spark SQL	Structured data processing and complex analytical queries on the denormalized dataset
HDFS	Distributed file system for storing raw data landing zone and final processed Parquet files
Apache Sedona 1.8.0	Geospatial extension enabling spatial operations (STPoint, STContains) for census tract mapping
Parquet	Columnar storage format optimized for query performance and compression

Table 1: Big Data Platforms and Technologies

2.3 ML Methods Used

While this project focuses primarily on data engineering, the integrated dataset enables downstream machine learning applications:

- **Classification**: Predict crime type or arrest likelihood given temporal/weather/demographic features
- **Clustering**: Identify spatiotemporal crime hotspots across census tracts
- **Regression**: Model crime frequency as function of weather, demographics, temporal patterns
- **Time Series Forecasting**: Predict crime trends using historical aggregations by tract/time

3. Evaluation

3.1 Experiment Design

The project was validated across five phases:

1. **Data Ingestion**: Successfully loaded crime (8.44M records), weather (8,766 distinct days), and demographics (18,518 census tracts)
2. **ETL Execution**: Executed Spark ETL pipeline with schema validation and data quality checks
3. **Spatial Join**: Mapped 100% of crime records to census tracts using Apache Sedona spatial operations
4. **Demographic Integration**: Enriched spatial data with median income and population attributes
5. **Analysis**: Validated dataset through Spark SQL exploratory queries

3.2 Scalability/Performance Metrics

Metric	Result
Total Integrated Records	8,344,980
Spatial Join Completion Time	5.93 minutes
Spatial Join Success Rate	100.0%
Demographic Join Success Rate	100.0%
Weather Join Success Rate	100.0%
Data Quality	All schema validation checks passed

Table 2: Scalability and Performance Metrics

3.3 Feature Metrics

The integrated dataset contains the following rich attributes:

- **Temporal Features**: Year, month, hour, day-of-week (facilitating hourly, daily, seasonal analysis)
- **Crime Attributes**: Type (25 categories), arrest status, domestic indicator
- **Location Data**: Latitude/longitude, census tract ID, tract-level demographics
- **Weather Metrics**: Average daily temperature, precipitation, snowfall
- **Demographic Data**: Median household income (by tract/year), total population (by tract/year)
- **Purpose:** consolidated features for spatiotemporal crime and socio-environmental analysis

3.4 Plots and Analysis

Key Findings:

Crime Type Distribution

Top 5 crime types by incident count across 2001-present:

- **Theft**: 1,770,279 incidents (21.2%)
- **Battery**: 1,530,899 incidents (18.3%)
- Criminal Damage: 954,951 incidents (11.4%)
- Narcotics: 751,303 incidents (9.0%)
- Assault: 563,292 incidents (6.7%)

Temporal Trends

- **Yearly**: Crime incidents peaked in 2001 (482,879) and declined to 369,983 by 2010
- **Hourly**: Crime distribution shows bimodal pattern with peaks at 9 AM (355,635 incidents) and moderate afternoon activity
- **Weekday**: Crime remains relatively consistent across weekdays with slight variation

Weather-Crime Relationships

Average temperature correlations with crime types reveal:

- Theft shows slight positive correlation with warmer temperatures (16.5°C average)
- Battery and Assault have marginal temperature sensitivity (~17.0°C)
- Property crimes (burglary) show consistent patterns across temperature ranges

Spatial Distribution

Crime concentration varies significantly across census tracts:

- High-crime areas: 15,000+ incidents per 1,000 population in targeted tracts
- Low-crime areas: 8,000-10,000 incidents per 1,000 population in safer neighborhoods
- Demographics show inverse correlation between median income and crime density

4. Summary

4.1 Achievement Against Design Goals

1. **Scalability**: Successfully processed 8.4 million crime records with 100% join success rates, demonstrating robust scalability
2. **Data Integration**: Unified crime, weather, and demographic data into a single denormalized analytical table
3. **Analytical Capability**: Generated 11 distinct analytical queries revealing crime patterns across temporal, weather, and demographic dimensions
4. **Performance**: Completed spatial join (most computationally expensive operation) in under 6 minutes with checkpoint strategy

4.2 Performance Comparison to Proposal

The project successfully exceeded key proposal objectives:

- **Data Integration**: Achieved 100% success rate on all three join operations (weather, spatial, demographic)
- **Scalability**: Processed full dataset (8.4M records) without optimization failures or out-of-memory errors
- **Analysis Depth**: Generated five significant insights (crime types, temporal patterns, weather correlations, spatial distribution, hourly heatmap)
- **Infrastructure**: Leveraged production-grade technologies (Spark, HDFS, Parquet, Sedona) suitable for enterprise deployments

4.3 Future Extensions

Short-term enhancements:

1. **Interactive Dashboard**: Develop Tableau/Grafana visualization with map, time-series, and demographic filters
2. **Real-time Ingestion**: Extend batch pipeline to support streaming crime reports via Kafka
3. **Advanced Analytics**: Implement predictive models (crime type classification, hotspot forecasting)

4. **API Service**: Expose query interface for law enforcement and urban planning agencies

Long-term directions:

1. **Multi-city Integration**: Extend pipeline to integrate crime data from multiple metropolitan areas
2. **External Data**: Incorporate additional features (transit data, socioeconomic indicators, events calendar)
3. **Causal Analysis**: Conduct rigorous statistical analysis of weather-crime causality
4. **Privacy-preserving Analytics**: Implement differential privacy for aggregated statistics

5. References

[1] Chicago Police Department. (2024). Crimes - 2001 to Present. Data extracted from Chicago Data Portal. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

[2] NOAA Climate Data Online. (2024). Daily Summaries - Chicago O'Hare International Airport. National Centers for Environmental Information. https://www.ncei.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094846/detail

[3] U.S. Census Bureau. (2022). American Community Survey 5-Year Data (2018-2022). Data API. https://api.census.gov/data/2022/acs/acs5