Analyzing Crime Data with HiveQL & Using Hadoop

Authors: Jesus Perez, Abhinav Singh, Tien Cheng

Department of Information Systems, California State University, Los Angeles

CIS5200 System Analysis and Design

jpere331@calstatela.edu, asingh122@calstatela.edu, tcheng8@calstatela.edu

Abstract: This project explores how HiveQL and Hadoop can be used to analyze large-scale public crime datasets from Los Angeles and Chicago. These datasets combine to a size of approximately 2.36 GB, offering a wealth of information for analysis. The Chicago dataset, extracted from the Chicago Police Department's CLEAR system, includes detailed records with time and block-level location data, enabling a comprehensive temporal-spatial analysis. Similarly, the LA dataset provides a robust foundation for comparing crime trends in another major U.S. urban center. The analysis focused on identifying crime trends, spatial distribution, and arrest statistics over time. This analysis is significant as it can inform policymakers and law enforcement about the evolving nature of crime, facilitate resource allocation, and enhance public safety strategies in two of the United States' largest urban centers, as well as contribute to special task forces when pursuing fugitives.

1. Introduction

We chose to analyze crime data from Los Angeles and Chicago due to their social relevance and data richness. Both cities provide extensive historical crime datasets, making them ideal for comparative analysis using Big Data tools. Our objective was to utilize HiveQL in a cloud-based Hadoop ecosystem to manage, process, and analyze this data efficiently.

Understanding crime patterns is crucial for city planners, law enforcement, and community organizations. By leveraging Big Data tools, stakeholders can make data-informed decisions for resource allocation and public safety. Our work builds upon basic SQL skills and scales them to large datasets using distributed computing.

2. Related Work

Being that the project is based on Big Data cloud computing, it was fairly simple to find other works of crime data from our LinkedIn community page.

One of those works is The Chicago Crime ML Project from Python developer Jigor Purohit. While both projects handle block-level, time-stamped data, the Chicago ML initiative is model-driven, emphasizing predictive analytics, anomaly detection, and classification using Python-based tools and machine learning frameworks like Scikit-learn. The CIS 5200 project is more focused on data warehousing and querying large datasets via HiveQL and Hadoop, aiming to uncover temporal-spatial trends and inform strategic law enforcement planning. The ML project is forward-looking in prediction, whereas CIS 5200 is exploratory and

comparative, providing a broader context for inter-city public safety efforts.

The second project is from the Pomona College Tableau Dashboard project from Tony He, where it takes a vastly different approach compared to the CIS 5200 project. While CIS 5200 dives into granular, urban-level crime data from Chicago and Los Angeles, the Pomona project explores macro-level crime trends across the entire United States, with a specific spotlight on North Carolina. Using Tableau for interactive data visualization, Pomona's project focuses on historical trends and policy implications, such as the effects of capital punishment laws, rather than real-time analysis or deep spatial-temporal granularity. In contrast, CIS 5200 employs big data tools (HiveQL, Hadoop) to perform district/block-level analytics, serving operational needs like task force planning and resource allocation. Pomona is policy-oriented and historical, while CIS 5200 is tactical, data-intensive, and localized.

The third project is The Washington Crime Power BI Project from Okechukwu Irokwe that offers a localized but visually rich analysis that differs from the CIS 5200 project's big data methodology. Focused solely on Washington State, this project uses Power BI to explore crime by time-of-day, neighborhood clusters, and offense types, offering real-time, accessible insights for public service communication. Compared to the dual-city, distributed data processing approach of CIS 5200, which leverages HiveQL and Hadoop for high-volume, comparative analytics between Chicago and Los Angeles, the Power BI project prioritizes operational transparency and public engagement over back-end data complexity. Where CIS 5200 is built for scale, depth, and cross-city comparison, the Washington project excels in dashboard storytelling for community-level monitoring.

3. Background / Existing Work

Big Data tools like Hadoop, HDFS, and Hive have become standard for processing large datasets that are difficult to handle with traditional databases. Hive provides a familiar SQL-like interface for querying structured data stored in HDFS.

Los Angeles and Chicago have publicly released their crime data spanning over a decade. These datasets are typically large, noisy, and come in varying formats. Previous projects have explored visualization and minor analytics using Excel or Python, but few have scaled the process using Big Data frameworks. Our project bridges this gap by demonstrating how to clean, query, and analyze these datasets using Hive in a distributed environment.

4. Your Work

Our project is structured as a tutorial guiding users through each phase of the Big Data pipeline:

1. Data Collection

We sourced publicly available datasets:

- *Los Angeles Crime Data (2010–2019)*
- *Los Angeles Crime Data (2020–Present)*
- *Chicago Crime Data (2001–Present)*

2. Data Upload to HDFS

The CSV files were transferred from local machines to the Hadoop cluster using SCP. Once logged into the cluster via SSH, the files were moved into HDFS under project-specific directories.

3. Hive Table Creation

We created two Hive tables, one for each city, by defining schemas that matched the datasets' fields. The tables were stored as delimited text files within HDFS and configured to skip header lines.

Top Crime Location

CITY	Chicago	Los Angeles
#1	STREET	STREET
#2	RESIDENCE	SINGLE-FAMILY DWELLING
#3	APARTMENT	MULTI-UNIT DWELLING
#4	SIDEWALK	PARKING LOT
#5	RETAIL STORE	SIDEWALK
#6	ALLEY	OTHER BUSINESS
#7	PARKING LOT	GARAGE
#8	RESTAURANT	DRIVEWAY
#9	GAS STATION	VEHICLE
#10	RESIDENT-YARK	PARKING UNDERGROUND

4. Data Querying and Analysis

Using HiveQL, we ran a series of queries to extract insights:

- Crime location and time patterns
- Top crime types by frequency
- Arrest percentages per district and crime type
- Annual crime counts

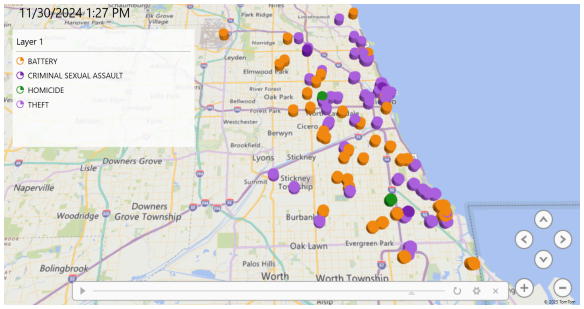
5. Exporting Results

Results were exported by inserting the output of HiveQL queries into a new Hive table designed for CSV export. The files were then retrieved from HDFS and downloaded locally using SCP.

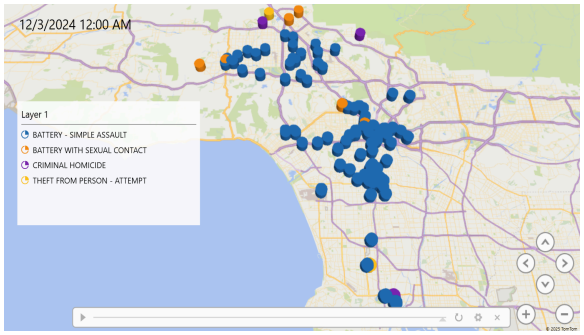
6. Visualization

We visualized the exported data in Tableau and Excel 3D Maps to identify geographic crime clusters and temporal spikes.

3D Map for Chicago: crime location and time patterns

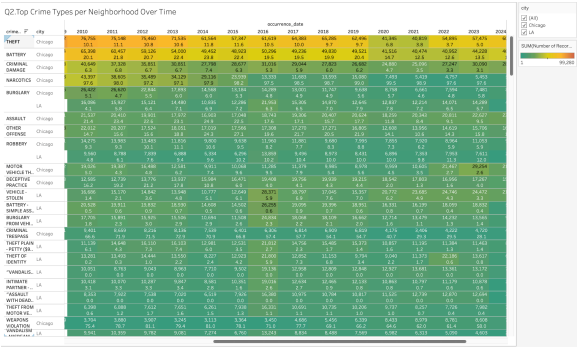


3D Map for Los Angeles: crime location and time patterns

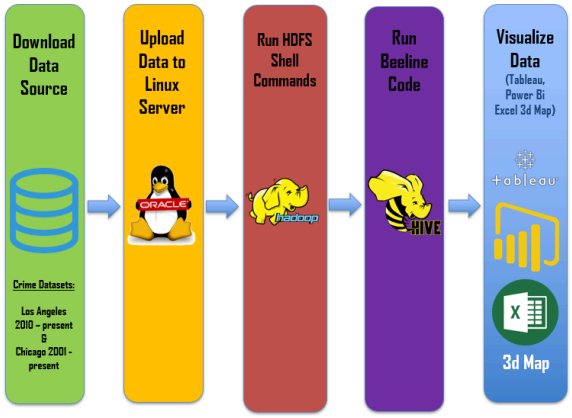


Top Types of Crimes:

Arrest Rate per Crime Type:



Workflow Diagram



5. Conclusion

This project showcases the potential of HiveQL and Hadoop for large-scale crime data analysis. We built a complete data pipeline—from ingestion and processing to querying and visualization—that can handle gigabytes of structured public data. Our comparative analysis between Los Angeles and Chicago provided insights into crime patterns, arrest efficiency, and district-level surveillance.

This work highlights the importance of Big Data tools in enabling communities and law enforcement to make informed decisions. We gained valuable experience in working with distributed systems and realized how structured query processing can be scaled using cloud resources.

6. References

- https://github.com/shjepz/BigDataCrimeAnalysis
- **LA Crime Data (2020–Present):** [https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/about_data]
- **LA Crime Data (2010–2019):** [https://data.lacity.org/Public-Safety/Crime-Data-from-2010-to-2019/63jg-8b9z/about_data]
- **Chicago Crime Data (2001–Present):** [https://data.cityofchicago.org/Public-Safety/Crimes-2025/t7ek-mgzi/about_data]
    - **Related Work**
- **Chicago Crime ML Project**
  Jigar Purohit – Data Science & ML Analysis
- **Pomona Tableau Dashboard**
  Tony He – Tableau Crime Analysis
- **Washington Crime Power BI Analysis**
  Okey Irokwe – Power BI Report