Snigdha Anirudha Pandit
UBID: spandit3

Arya Satam
UBID: aryanami

Shreyas Waikar
UBID: swaikar

A. MILESTONE 1

I. INTRODUCTION

To make wise decisions and enhance public safety in Chicago law enforcement organizations, legislators, and academics need to analyze crime data effectively to make wise decisions and enhance public safety in Chicago. Nfail to provide approaches to assessing time series data, spatial patterns, and crime trends are frequently disjointed and fall short of providing the thorough understanding required for production-making. We aim to create a robust data model that can effectively gather, store, and analyze crime data to give stakeholders insightful information to tackle these difficulties.

II. PROBLEM STATEMENT

Our main goal is to address the ineffectiveness and fragmentation of the existing approaches to crime data analysis. Stakeholders need help accessing and understanding intricate crime trends, geographic patterns, and temporal variations. This makes it more difficult for them to deploy focused crime prevention strategies and allocate resources efficiently. Moreover, the timeliness and relevance of the insights derived from the data are restricted by the absence of real-time analysis capabilities. The dataset used contains crime data in Chicago from the year 2001 to the present. Our database system will be robust so that the dataset could be understood and analyzed as follows:

A. *Data Collection And Storage:*
We will compile comprehensive crime data and keep it in a well-organized relational database to guarantee user accessibility and data integrity.

B. *Geospatial Analysis:*
The Location Details tables to map Chicago's crime rates. We will be able to identify high-crime areas and look into spatial clusters of specific crimes by combining crime data with geographic information. This will provide valuable insights for both law enforcement and urban development.

C. *Trend Analysis:*
A comparison of changes in different types of crimes over time using the Crimes and Crime Types tables. This will identify trends, such as increases or decreases in specific crimes over time, by querying the database and aggregating data based on crime types and timestamps.

D. *Time Series Analysis:*
We will examine weekly, monthly, and daily crime trends using the Time table. We can identify recurrent patterns and criminal activity peak times by analyzing temporal patterns and long-term trends. This will enable more proactive policing techniques.

III. REASONS FOR CHOOSING DATABASE OVER EXCEL

A. *Query Flexibility and Performance:*
Complex queries can be efficiently performed owing to the robust query languages and indexing mechanisms provided by database systems. This makes it possible for users to quickly extract insightful information from the data.

B. *Security and Integrity of Data:*
With features like transactions, access control, and encryption, databases provide means of guaranteeing data security and integrity. This guarantees confidentiality and integrity.

C. *Concurrent Access:*
Databases allow multiple users to access them at once, which makes it easier for researchers, legislators, and law enforcement to share and analyze data together. This feature, which enables team members to efficiently collaborate on data analysis tasks and share insights in real-time, is especially crucial for group projects.

IV. INTENDED USERS OF THE DATABASE

The database system will serve the following users:

A. *Law Enforcement Agencies:* The analysis outcome can be used by police departments to better allocate resources in areas of high crime rate and also help plan crime prevention measures.
B. *Academic Researchers:* Researchers and Academicians can use the data to conduct academic studies, create prediction models and also study criminology.
C. *PolicyMakers:* Assist policymakers to make better decisions related to public safety and planning.

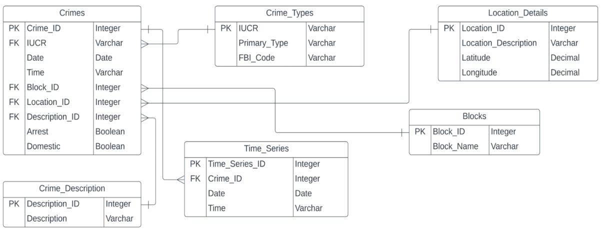
V. WHO WILL ADMINISTER THE DATABASE?

The integrity, security, and performance of the database will be upheld by a group of database administrators (DBAs) and data engineers will uphold the integrity, security, and performance. Assuring the stability and scalability of the system, they will oversee tasks like data input, schema design, optimization, backups, and security protocols to fulfill users' analytical demands. Researchers and data analysts will examine the database for various stakeholders based on their requirements.

A real-life scenario:

Let's consider the Chicago Police Department(CPD'S) efforts to manage a recent rise in violent crimes in particular communities as an example(South Side). The South Side of the city has seen a noticeable rise in robberies and gunshots in the recent months, according to the CDP. THey made the decision to employ data analytics to better understand the underlying patterns and trends that are causing an increase in violent crime in order to successfully address this problem. The CDP then took targeted measures to stop the rise based on the analysis results. The measures taken include (1) Increasing the number of police during peak hours (2) Implementation of policing programs to build awareness and trust (3) Collaboration with social service providers to address the root causes of crime, which include deprivation and opportunity gaps. Over time, the South Side saw a decrease in violent crime rates, according to the CDP.

VI. E/R DIAGRAM



VII. RELATIONS AND ATTRIBUTES

To understand the huge dataset in a better way, the group members decided to manually study the data. We looked for possible connections in the dataset which could aid in finding meaningful insights. For the purpose of Milestone#1, We decided to create different csv files and manually create the tables. We plan to achieve this in a better way moving forward. The created tables are as follows:

Table 1: Crimes Table

Attributes: Crime_ID, IUCR, Date, Time, Block_ID, Location_ID, Description_ID, Arrest, Domestic

Relations: It is the main table. Here each record corresponds to a distinct crime incident. Since the Crime_ID is a primary key, every record in this table can be uniquely identified by it.

Table 2: Crime_Types Table

Attributes: IUCR, Primary_Type, FBI_Code

Relations: The IUCR field connects the Crime_Types Table, and is a categorization table for crimes. The IUCR code functions as a foreign key in the Crimes table and a primary key in the Crime_Types table as it is specific to each type of crime and is used in the Crimes table. Since there is only one type of crime, there is a **one-to-one relationship**.

Table 3: Blocks Table

Attributes: Block_ID, Block_Name

Relations: Provides details about the blocks where crimes have been committed. The Block_ID connects it to the Crimes table. The Blocks table and the Crimes table denote a **one-to-many relationship**, suggesting that while each crime is associated with a single block, each block may be associated with multiple crimes.

Table 4: Location_Details

Attributes: Location_ID, Location_Description, Latitude, Longitude

*Relations:*Contains comprehensive data regarding crime scenes. The Location_ID connects it to the Crimes table. Moreover, there is a **one-to-many relationship** here, meaning that a single location detail may be related to several different criminal incidents.

Table 5: Crime_Descriptions Table

Attributes: Description_ID, Description

Relations: Contains the crime descriptions. It is linked by the Description_ID to the Crimes table. This is another instance of a **one-to-many relationship** in which several crimes can be connected to a single description.

Table 6: Time_Series Table

Attributes: Time_Series.ID, Crime_ID, Date, Time

Relations: This table looks to be intended for monitoring crime incidents over a period of time. It exhibits a **one-to-many relationship** with the Crimes table via the Crime_ID link. This means that several records for a single crime incident over various time periods may be stored in the table, possibly capturing updates or modifications to the crime status.

VIII. DESCRIPTION OF ATTRIBUTES

The following table represents the attributes of the real time dataset which is being used by the team. The dataset can be found on the following link: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present-Map/ahwe-kpsy

Attribute Name	Data Type	Description
ID	Integer	Each record in the dataset was allocated a unique numerical identification. This ensures that each incident may be uniquely referenced and accessed.
Case Number	Varchar	The Chicago Police Department assigns a unique alphanumeric code to each occurrence, often known as the Records Division Number (RD Number). It acts as an official reference for retrieving case details.
Date	Date	The date and time at which the incident occurred or was reported. In some circumstances, the exact timing may be an estimate based on available data.
Block	Varchar	To protect privacy, this field contains a partially masked address for the incident's location, specifying simply the block where the incident happened. Complete addresses are not provided, but the information is detailed enough to determine the general location of the incident.

IUCR	Varchar	An alphanumeric code used to classify incidents under Illinois Uniform Crime Reporting guidelines. This code is directly linked to the incident's major and secondary descriptors, allowing for a clearer categorization and analysis of crime data.
Primary Type	Varchar	This data field represents the incident's primary type, as determined by the IUCR code. Examples include theft, battery, and criminal damage.
Description	Varchar	A thorough explanation of the incident, including further information regarding the nature of the crime. This is a subdivision of the major kind that provides more specific information (for example, aggravated battery, car burglary).
Location Description	Varchar	Description of the exact place where the incident took place, such as an apartment, sidewalk, or parking lot. This provides context for the incident.
Arrest	Boolean	A boolean (checkbox) field that indicates if an arrest was made in connection with the occurrence. It helps to grasp the report's immediate outcome.
Domestic	Boolean	Another boolean parameter that determines if the occurrence was related to domestic abuse as defined by the Illinois Domestic abuse Act. This aids in the identification of cases with domestic concerns.
Beat	Integer	The smallest police geographical region (beat) in which the incident occurred. Beats are the primary building blocks of the Chicago Police Department's patrol approach, and each has its own patrol car.
District	Integer	Identifies the police district where the incident occurred. Chicago is organized into 22 police districts, with each functioning as a major administrative policing region.
Ward	Integer	Chicago is divided into 50 wards, which are served by an advocate on the City Council. The City Council area where the incident occurred.
Community Area	Integer	Identifies the region in which the incident occurred. Chicago has 77 defined community zones for statistical and planning purposes.
FBI Code	Integer	It is a type of crime classification used by the FBI's National Incident-Based Reporting System (NIBRS). This uniform categorization aids in comparing crime statistics from various jurisdictions.
X Coordinate	Integer	It is a numerical field that indicates the location of the occurrence in the State Plane Illinois East NAD 1983 projection system. These coordinates are slightly altered for privacy purposes, ensuring that the location is within the same block as the real incident.
Y Coordinate	Integer	It is a numerical field that indicates the location of the occurrence in the State Plane Illinois East NAD 1983 projection system. These coordinates are slightly altered for privacy purposes, ensuring that the location is within the same block as the real incident.
Year	Integer	The year that the occurrence was reported. This field enables the examination of crime statistics across time.
Updated On	Timestamp	The date and time at which the record was last updated in the dataset. This allows consumers to understand how current the information is.
Latitude	Decimal	It is a geographical location that indicates the incident's shifted location for mapping reasons. These data, like the X and Y coordinates, have been modified for privacy reasons.
Longitude	Decimal	It is a geographical location that indicates the incident's shifted location for mapping reasons. These data, like the X and Y coordinates, have been modified for privacy reasons.
Location	Varchar	It is an integrated field that integrates geographic location information (latitude and longitude) into a format suited for mapping and geographic analysis, while respecting privacy concerns by simply indicating the block of the occurrence.

IX. WHAT HAPPENS WHEN PRIMARY KEY IS DELETED

In the case of following we noticed that,

A. *No Action*: If there are any corresponding foreign key values, the deletion of the primary key will be blocked and an error will be generated.
B. *Cascade:* In this particular command, if a primary key is deleted, all the related or connected foreign key values will be deleted.
C. *Set Null:* If set to Set Null, all the related foreign key values will be set to NULL when a primary key is deleted.
D. *Set Default:* A default value will be allocated to foreign key if the primary key is deleted.

X. QUERY APPLICATION

After creating and uploading the 6 different dataset on postgreAdmin. We performed a few basic queries to ensure the basic working of our system.***[Fig. 1-4]***

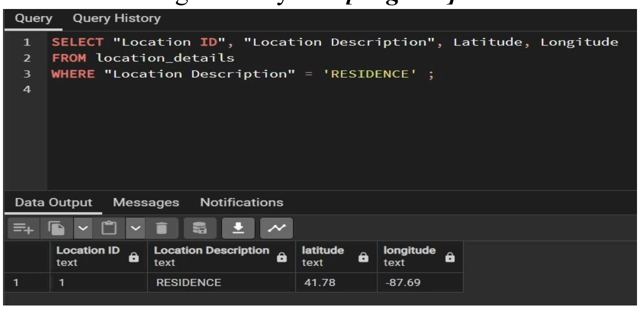


Fig. 1

Query | Query History

```sql
SELECT * FROM crime_types
WHERE "fbi_code" = '11';
```

Data Output | Messages | Notifications

	iucr text	primary_type text	fbi_code text
1	1152	DECEPTIVE PRACTICE	11
2	1153	DECEPTIVE PRACTICE	11
3	1150	DECEPTIVE PRACTICE	11
4	1154	DECEPTIVE PRACTICE	11
5	1210	DECEPTIVE PRACTICE	11
6	1156	DECEPTIVE PRACTICE	11
7	1110	DECEPTIVE PRACTICE	11
8	1155	DECEPTIVE PRACTICE	11
9	1130	DECEPTIVE PRACTICE	11
10	1206	DECEPTIVE PRACTICE	11
11	1220	DECEPTIVE PRACTICE	11
12	1242	DECEPTIVE PRACTICE	11
13	1261	DECEPTIVE PRACTICE	11
14	1195	DECEPTIVE PRACTICE	11
15	1185	DECEPTIVE PRACTICE	11
16	1151	DECEPTIVE PRACTICE	11

Fig. 2

Query | Query History

```sql
SELECT DISTINCT Primary_Type
FROM crime_types
GROUP BY Primary_Type
HAVING COUNT(Primary_Type) > 1;
```

Data Output | Messages | Notifications

	primary_type text
1	OBSCENITY
2	BURGLARY
3	SEX OFFENSE
4	INTIMIDATION
5	ROBBERY
6	PUBLIC PEACE VIOLATION
7	CRIMINAL SEXUAL ASSAULT
8	CRIMINAL DAMAGE
9	WEAPONS VIOLATION
10	NARCOTICS

Fig. 3

Query | Query History

```sql
SELECT "Date", COUNT("Crime ID") AS crime_count
FROM time_series
GROUP BY "Date"
ORDER BY "Date";
```

Data Output | Messages | Notifications

	Date text	crime_count bigint
1	02/09/2023 12:00:00 AM	1
2	02/09/2023 12:01:00 AM	1
3	02/09/2023 12:03:00 AM	1
4	02/09/2023 12:05:00 AM	1
5	02/09/2023 12:07:00 AM	1

Fig. 4

B. MILESTONE 2

I. NORMALIZATION

Normalization is an essential process in database design. It organizes the database's attributes and relations to eliminate redundant information. The process involves dividing larger tables into smaller ones to maintain a lossless decomposition. Boyce-Codd Normal Form [BCNF] is based on functional dependencies where every determinant is a candidate key. A candidate key is an attribute or a set of attributes that can uniquely identify each row in a table. Now, looking at the dataset we have chosen, We followed these steps to achieve BCNF:

Step 1: Identify Functional Dependencies:
Analyze all the attributes in the dataset and identify the types of dependencies between the attributes.

Step 2: Identify Candidate Keys:
Find a set of attributes so they are minimal and uniquely identify tuples in each table.

Step 3: Normalize into BCNF:
Based on these keys and functional dependencies, decompose the relation to smaller tables.

After following these steps, the tables obtained are as follows:

Table 1: CrimeIncident Table-
Attributes: ID (Primary Key), Case Number, Date, Block, IUCR (Foreign Key), Arrest (Foreign Key), Time ID (Foreign Key), Updated On.
Candidate Key: {ID}
Functional Dependency: ID→{Case Number, Date, Block, IUCR, Arrest, Time ID, Updated On}

Table 2: CrimeType Table-
Attributes: IUCR(Primary Key), Primary Type, Description, FBI Code.
Candidate Key: {IUCR}
Functional Dependency: IUCR→{Primary Type, Description, FBI Code}

Table 3: LocationDetails Table-
Attributes: Block(Primary Key), Location Description
Candidate Key: {Block}
Functional Dependency: Block → {Location Description}

Table 4: ArrestDetails Table-
Attributes: Arrest ID(Primary Key), Arrest
Candidate Key: {Arrest ID}
Functional Dependency: Arrest ID → {Arrest}

Table 5: Time_Series Table-
Attributes: Time ID(Primary Key), Date
Candidate Key: Time ID
Functional Dependency: Time ID → {Date}

Table 6: CrimeLocation Table-
Attributes: Block(Primary Key), Beat, District, Ward, Community Area, X Coordinate, Y Coordinate, Latitude, Longitude, Location.
Candidate Key: {Block}
Functional Dependency: Block → {Beat, District, Ward, Community Area, X Coordinate, Y Coordinate, Latitude, Longitude, Location}

II. FINALIZED ER DIAGRAM

We made a few changes in the original ER diagram based on normalization. The modified ER diagram in BCNF can be represented as the following(Fig. 5):

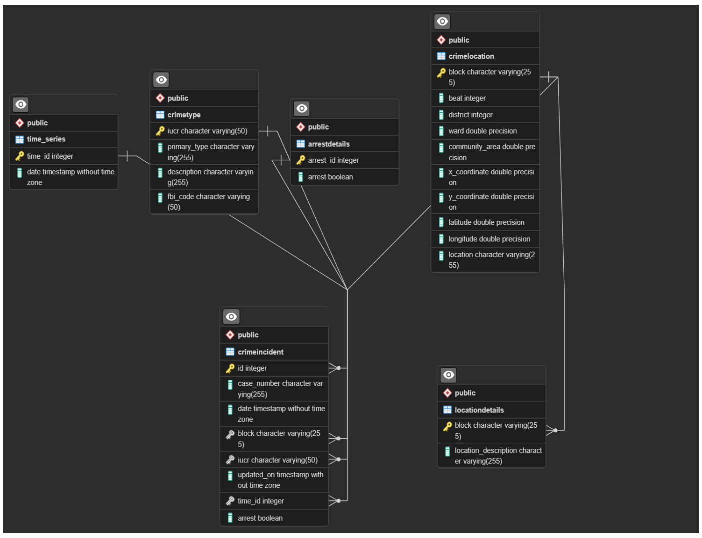


Fig. 5

III. INDEXING

The Chicago Crime dataset contains data from 2001, which makes it very large. We faced multiple challenges while working with a particular dataset. The problems faced were:

- The processing time for queries led to slow response time.
- It took us some time to analyze the dataset and maintain data integrity.
- We noticed that subqueries took a lot of time to execute.
- Because this dataset comprises Crime statistics, the dataset will grow with time, and scalability is essential; thus, performing indexing was crucial for this particular dataset.

By using Indexing, we were able to optimize query performance, especially on columns that were frequently used for joining and searching. Here are the indexes added in each table:

1. CrimeIncident Table:
   Indexes - Date, Arrest, Time ID
   Reason: Speed up join operations.
2. Time_Series Table:
   Indexes - Date
   Reason: Enable easier filtering and sorting according to date.
3. CrimeLocation Table:
   Indexes- Beat, District
   Reason: Commonly used location attributes.

INDEXING QUERIES:
-- Index on 'Date' in CrimeIncident table
CREATE INDEX idx_crimeincident_date ON CrimeIncident (Date);

-- Index on 'Arrest' in CrimeIncident table

CREATE INDEX idx_crimeincident_arrest ON CrimeIncident (Arrest);

-- Index on 'Time ID' in CrimeIncident table
CREATE INDEX idx_crimeincident_timeid ON CrimeIncident (Time_ID);

-- Index on 'Date' in Time_Series table
CREATE INDEX idx_timeseries_date ON Time_Series (Date);

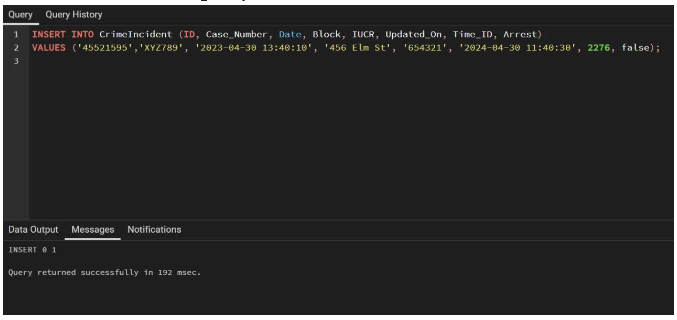
-- Index on 'Beat' in CrimeLocation table
CREATE INDEX idx_crimelocation_beat ON CrimeLocation (Beat);

-- Index on 'District' in CrimeLocation table
CREATE INDEX idx_crimelocation_district ON CrimeLocation (District);

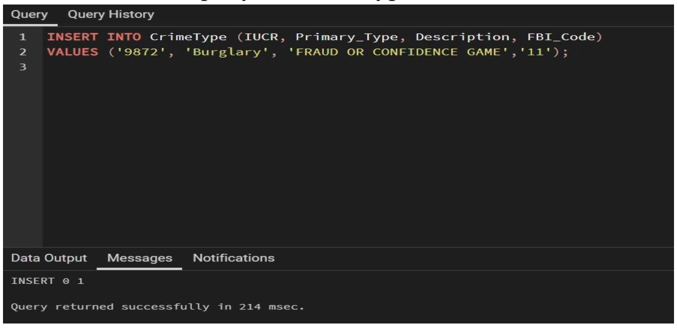
IV. SQL QUERIES

Following are the SQL queries we implemented on the BCNF normalized dataset:

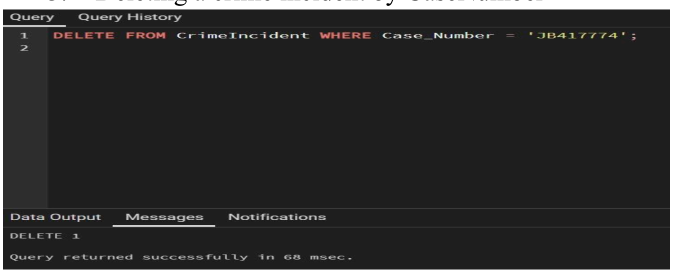
1. INSERT query on CrimIncident Table:



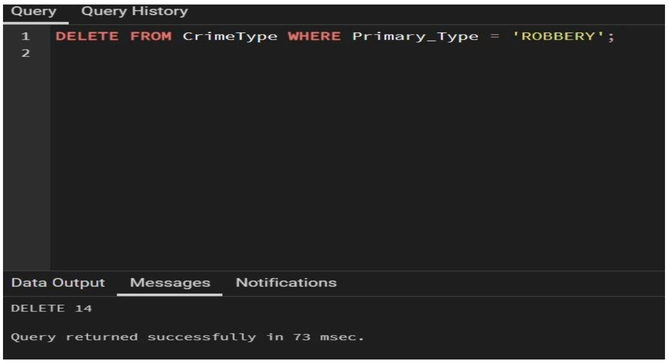
2. INSERT query on CrimeType Table:



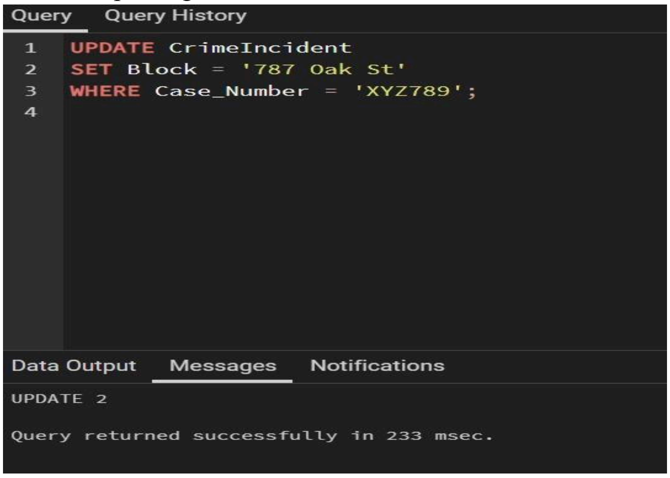
3. Deleting a crime incident by CaseNumber



4. Deleting a crime type by PrimaryType

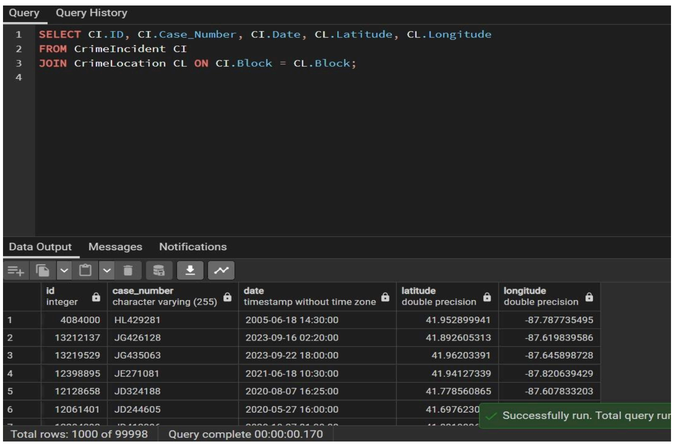


5. Updating the block of a crime incident

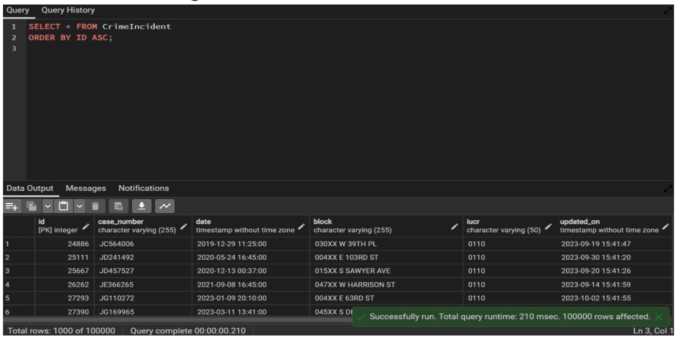


6. Updating the description of a crime type

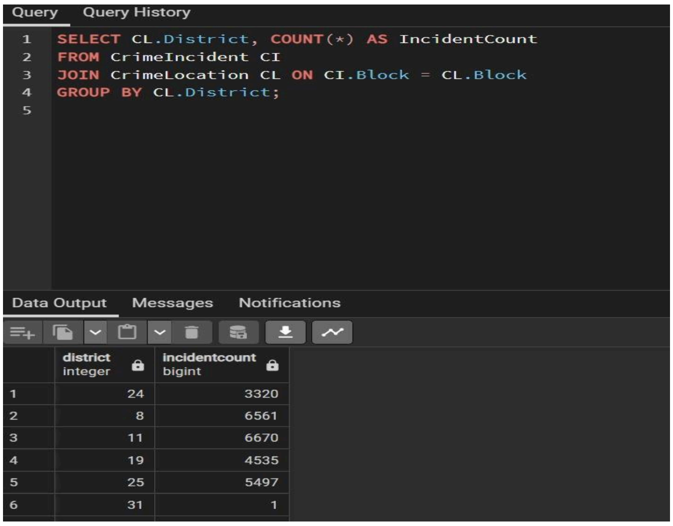
7. Query to join CrimeIncident and CrimeLocation tables



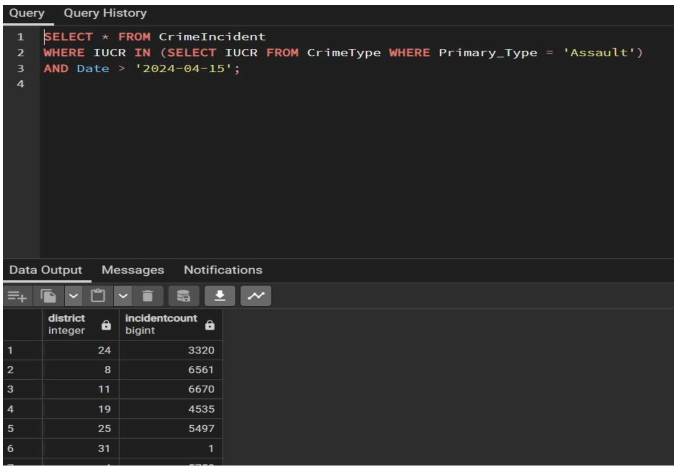
8. Query to select crime incidents ordered by ID in ascending order



9. Query to count the number of crime incidents by district

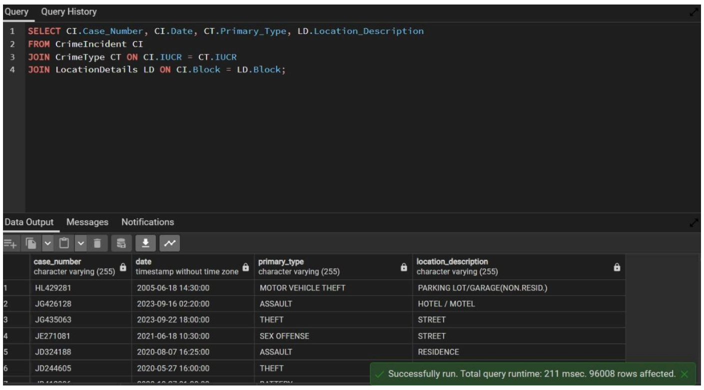


10. Query to select crime incidents with primary type 'Assault' and date after '2024-04-15'

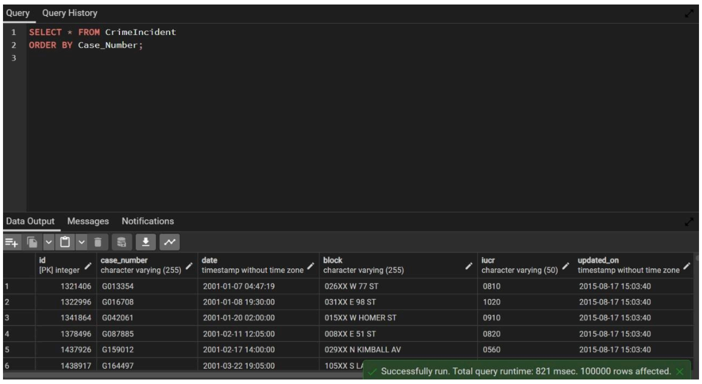


11. Query to select crime incidents with primary type 'Battery' and committed in 'Beat 123'

12. Query to join CrimeIncident, CrimeType, and LocationDetails tables



13. Query to select crime incidents ordered by case number

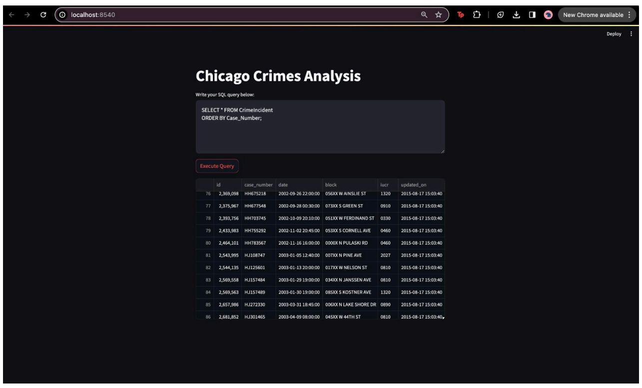


V. PROBLEMATIC QUERIES

Looking at an example from one of the tables [CrimeIncidents]. If the date column is not indexed, Postgre will need to perform an entire table scan to find matching rows, which will ultimately result in slow query performance. This issue can be resolved by adding an index on the date attribute.

Indexing is important in enhancing the speed and efficiency of data retrieval in SQL databases. This makes the application more responsive and scalable.

VI. WEBSITE



VII. CONCLUSION

This DBMS project has offered a thorough and effective way to handle and examine crime incident data. By means of rigorous database design, normalization, indexing, and query optimization, I have established a resilient database system that can manage substantial data volumes and enable sophisticated data analysis. This project exemplifies how crucial appropriate database management strategies are to maintaining data integrity, functionality, and usability in practical applications.

In the future, law enforcement organizations, researchers, and legislators can use the database system created for this project as a valuable tool for examining crime trends, pinpointing hotspots, and creating successful crime prevention plans.

By demonstrating the use of database concepts and techniques, this DBMS project is a noteworthy accomplishment in data management and analysis.

VIII. CITATIONS:

1. Dataset: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present-Map/ahwe-kpsy
2. https://www.geeksforgeeks.org/boyce-codd-normal-form-bcnf/
3. https://www.solarwinds.com/resources/it-glossary/database-index\