**BY: -  
Himanshu Kumar (10525006)**

**Technical Report for Chicago City Crime (2001-2018) Datamart**

Table of Contents

[1. Introduction 2](#_Toc27243360)

[1.2. Reasons for selecting the subject area AND DATA 2](#_Toc27243361)

[1.3. Vision and Goals 6](#_Toc27243362)

[1.4. Key StakeHolders 6](#_Toc27243363)

[1.5. Business requirements 7](#_Toc27243364)

[2. SCHEMA 8](#_Toc27243365)

[3. ETL 10](#_Toc27243366)

[4. VISUALIZATIONS AND REPORTS 15](#_Toc27243367)

[4.1. Visualizations 16](#_Toc27243368)

[4.2. Reports 16](#_Toc27243369)

[5. Include XML and Schema 18](#_Toc27243370)

[6. Graph Databases 18](#_Toc27243371)

[6.1. COMAPRISON to realtional databases 20](#_Toc27243372)

[7. Bibliography 25](#_Toc27243373)

[Appendix A – VISUALIZATIONS Code 25](#_Toc27243374)

[Appendix B – Neo 4J code 25](#_Toc27243375)

# 1. Introduction

For a city to prosper, it’s inevitable that its residents feel safe and put their 100% in whatever they are doing. There should not be fear in their minds while going from home to work or vice versa and there should not be any feeling of insecurities in their minds with respect to their families. A city’s crime data gives an insight into the state of its affairs and is a very important factor for law enforcement agencies, lawmakers, potential investors & residents. While it gives law enforcement agencies a chance to make policing and crime control more effective, it provides a big picture of the crime scene to its city residents.

Keeping the above points in mind, here we are trying to build a data mart for the reported crimes in the city of Chicago for the duration 2001-2018.Please note that “murder” as a crime is not covered here because it comes under “heinous crimes” category and there is a separate database for that.

**Source of Data**: For preparing the data mart, data from Chicago city police website has been referred. The link for the same is:

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

## 1.2. Reasons for selecting the subject area AND DATA

Since we are analyzing the reported crime records for a city, below subject areas become necessary to shape the data mart: -

1. **CASE\_DIM:** Case dimension table has been named as CASE\_DIM in

our data mart. It primarily stores all data that identifies the case. Below are the records stored under this dimension: -

1. Case\_ID: Unique Identifier given to the case. It is the primary key for the dimension table.
2. Case\_No: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
3. Arrest: It’s a flag that signifies whether arrest has been made for the reported crime or not. 0 for FALSE and 1 for TRUE.
4. Domestic: It’s a flag that tells whether the case comes under domestic crime. If flag is 1 then TRUE, if it’s 0 then FALSE.
5. **ADDRESS\_DIM:** The address and location details of the place of crime and the jurisdictional police district are stored in ADDRESS\_DIM.It has the following records:
6. ADDRESSID: Contains the combination of geographical Latitude and Longitude of the place of crime. This is the primary key for the dimension table.
7. BLOCK: Block associated with the place of crime.
8. LOC\_DESC: Categorizes the type of location where crime was committed. Few examples of location descriptions are Residence, Apartment, Bank.
9. BEAT: Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has dedicated police beat car.
10. DISTRICT: Indicates the police district where the incident occurred. Three to five beats make a police sector and three sectors make a police district.
11. WARD: The ward (City Council district) where the incident occurred.
12. COMMUNITY\_AREA: Indicates the community area where the incident occurred. Chicago has 77 community areas.
13. X\_COORDINATE: The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
14. Y\_COORDINATE: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
15. LATITUDE: The latitude of the location where the incident occurred.
16. LONGITUDE: The longitude of the location where the incident occurred.

1. **DATE\_DIM:**  Date and time records associated with a crime are stored in Date Dimension table. It consists of: -
2. DATEKEY: Date and time of crime registration. It’s in the format YYYY-MM-DD hh:mm: ss.ms. This field is directly taken from source data file. This is the primary key for the dimension table.
3. CASE\_DATE: Date of crime registration. This is extracted from DATEKEY field. Its in the format YYYY-MM-DD.
4. CASE\_YEAR: Year of crime registration. This is being directly taken from source data.
5. CASE\_MONTH: Month of crime registration. This is extracted from DATEKEY field. This is in digital form where 1 represents January,2 represents February and so on.
6. CASE\_WEEK: This shows whether the day of crime registration is a weekday or weekend. This field is also extracted from DATEKEY.
7. CASE\_DAY: Day of month of crime registration. This is also extracted from DATEKEY.It is being represented by numbers varying from 1 to 31 depending on the month.
8. CASE\_QUARTER: Quarter of the year when the crime was registered. This is also extracted from DATEKEY field and is assigned as ‘Q1’ for the period Jan-March,’Q2’ for April-June and so on.
9. **CRIMETYPE\_DIM:** This dimension holds information about the category of crimes. It consists of: -
10. IUCR\_CODE\_ID: - This is the primary key for the dimension table. The Illinois Uniform Crime Reporting code is used by law enforcement agencies to categorize the crimes as per defined codes. List of IUCR codes can be found at <https://data.cityofchicago.org/d/c7ck-438e>.
11. PRIMARY\_TYPE: - Primary description of the IUCR code. Few examples are: Homicide, Crime Sexual Assault, Robbery, Stalking.
12. CASE\_DESCRIPTION: - This is the secondary description of the IUCR code, a subcategory of the primary description. Examples can be First Degree Murder, Reckless Homicide, Cyber Stalking, Forcible entry.

Apart from the dimension tables mentioned above, a Fact table CRIME\_FACT consists of the following records: -

1. CASE\_ID\_FACT: - It is a foreign key mapped to dimension CASE\_DIM’s column CASE\_ID.
2. ADDRESSID\_FACT: - It is a foreign key mapped to dimension ADDRESS\_DIM’s column ADDRESSID.
3. DATEKEY\_FACT: - It is a foreign key mapped to dimension DATE\_DIM’s column DATEKEY.
4. CRIMETYPE\_ID\_FACT: - It is a foreign key mapped to dimension CRIMETYPE\_DIM’s column IUCR\_CODE\_ID.

The FACT table has a composite primary key comprised of the above four foreign keys.

## 1.3. Vision and Goals

**VISION: -**

The long-term vision of designing this data mart is to present the crime graph of the city of Chicago to different stake holders. The city police commissioner office would like to analyze the data in order to strengthen policing and community education in areas of high crime rates. Similarly, a ward councilor would want to highlight high crime rate in his/her ward in the council meeting so as to push the law enforcement agencies to act effectively and fastly. This data mart would provide a single access point to all stake holders who are directly or indirectly associated with crime rates in the city of Chicago.

**GOALS: -**

* To bring down crime rate in the city of Chicago
* To act as a single source of data for various stakeholders like law enforcement agencies, investors, media houses, general public etc, as and when needed by them.
* To act as a guide for effective community education in areas of high crime rates.
* To act as a handy guide to government agencies to decide on the manpower needed in various law enforcement agencies.
* To act as a single guide for Chicago police to increase patrolling and vigilance across areas of high crime rates across the city of Chicago.

## 1.4. Key StakeHolders

Since this is an insight into the crime data of a megacity, the obvious stakeholders would be: -

* + **Chicago Police Commissioner’s Office:**

The police commissioner’s office will use the data to ascertain areas of high crime rate and adjust the officer’s rosters accordingly.

* + **Investors who are planning to invest in the city:**

Potential investors who are planning to set up their businesses or start their offices in the city would like to see the crime graph of the city before making investments because for them, safety of their businesses and their employees would be of utmost importance.

* + **Media houses who need to report crime incidents:**

Various news media like print, electronic and web cover crime incidents on a day to day basis in their articles. They would like to use this crime data to highlight the overall state of crime in the city and also educate it’s readers/audiences about the performance of law enforcement agencies in curbing the same.

* + **Residents of the city:** They need to be aware of the crime rate in their city in general and in their localities, in particular. Also, this will help them pressurize authorities and ward councilors to take corrective actions.
  + **Political representatives and provincial and federal government officials:** to formalize policies and plan funding to control crime in the city.
  + **Government agencies involved in community education:** who can formalize a plan for extensive community education in areas of high crime rate in the city.

## 1.5. Business requirements

• **Community Education to tackle crime rate**: - A city’s crime data reflects the maturity of the society living within it. If the crime rate is high, the police and administration have to be on their toes and tackle the situation not only by means of increased and strict policing but also educating the people through community education and other means. People need to be educated about laws dealing with certain crimes, about why their near and dear ones are tempted to commit crime, how can they help their near and dear ones from committing crime, how can they timely inform the authorities if they suspect some crime is about to be committed or has been already committed.

• **Data for law enforcement agencies to improve policing: -** By analyzing the crime data, the city police commissioner’s office, FBI and related government agencies can devise a plan as to which area need more police officers on field. By analyzing the crime rate during the time of day(morning/noon/evening/night), the agencies can prepare rosters to depute manpower accordingly. By analyzing the crime rate in certain police districts and beats, a decision can be taken to shift some effective police officers from areas having low crime rates to areas having high crime rates.

• **Data for Investors who are planning to invest in the city: -** Investors prefer a city where their businesses as well as their workers are safe. By analyzing the overall crime rate in the city and also crime rates across various city localities, businesses can take a decision whether to invest in the city and if yes, which can be the safest areas to do so.

• **Data for media houses to highlight the state of affairs of the city and also to pressurize law enforcement agencies to act urgently in case of high crime rates: -** Various kind of media like newspapers, television, radio, online news sites can analyze the data and publish it to pressurize city, provincial and federal administrations to act if the crime rate is going up rapidly.

• **Data for general public to see if the present administration is doing good job on the law and order front: -** General public can use the data to see if the crime rates are going up or down in their city and localities year-on-year. Accordingly, they can gauge whether the present administration responsible for maintaining law and order is working effectively and take a call whether to re-elect it or not.

# 2. SCHEMA

The Data mart has been designed in a Star Schema model.The data model is as shown below :-

**Reasons for design: -**

* + Star schema is best for performance and since this is a data mart i.e. only single Fact table is there , a Star schema best fits for the design.
  + It is best broken into four dimension tables as a further break up was not neccessaary.
  + A fact table with a composite primary key acts as a unique identifier for a given record in the real world.
  + Date Dimension table has few derived/extracted columns for better visulization and reporting for a more holistic MIS(Management Information System).
  + Locations related to crime place and concerned police areas are listed in a single Address dimension so as to store all geographical data in a single table for better performance when someone queries the geographical data.

# 3. ETL

**ETL Process:**

For Integrating any kind of data we need to transform it using some Extraction tools and process. Which is necessary for proper alignment and logical arrangement of data. It includes various steps in the process. ETL basically is used to integrate the data, it Extract the data from various other sources of information then it helps to transform the data into vital information which can be useful for providing users/ application for understanding the graphical and analytical representation of the data. Proper design is necessary for any project and Dataware for this the following techniques are explained. Extraction, Cleaning, Transform and Loading.

**Extraction:**

Data has been extracted from our source dataset of crime, In which the data is in a structured format. The dataset contains 23 different columns in which all the detailed. We have taken 50,000 rows for our process. Which contains the previous 10 years of crime records.

**Cleaning:**

Before transforming the data in the staging table, whole data need to be cleaned and pre-processed. Here we had taken the data in CSV format. Cleaning of the data was done partially in R to remove the unwanted null values and also few of the primary data was null which was handled in SSIS while populating the tables.

**Transformation and Loading:**

After the data cleaned, the file was in CSV which is in a readable format. For transformation, we have used SSIS Tools. Tables for Dimensions and facts were created in SSMS with proper datatype and relations. The cleaned data source was then uploaded into SSMS using then import functionality in SSMS in which the proper datatypes were defined for each column to avoid datatype mismatch in while loading the data. Fig.1.1 describes the import action for uploading the data type into SSSM. Once this dataset was imported into SSMS reconciliation of data was done to check all the columns that have been inserted into the table. Further, we have used SSIS (Visual Studio 2017) for integrating SSMS. New packages were created in SSIS to transform data into Dimension tables.

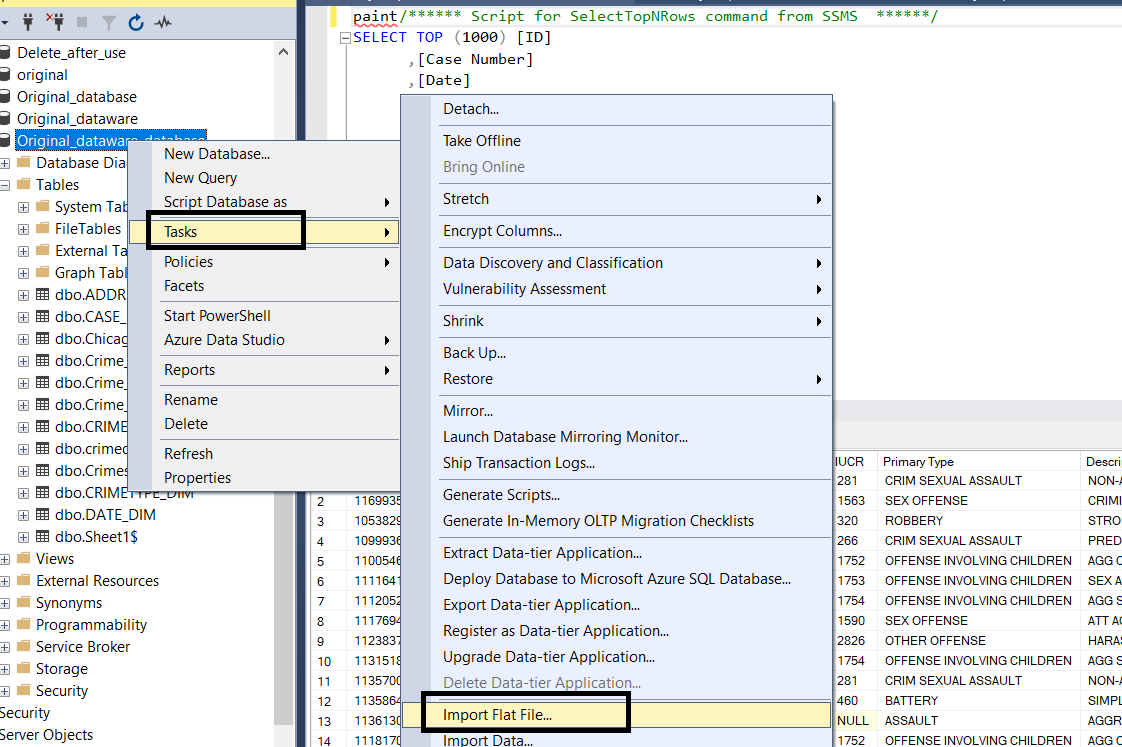
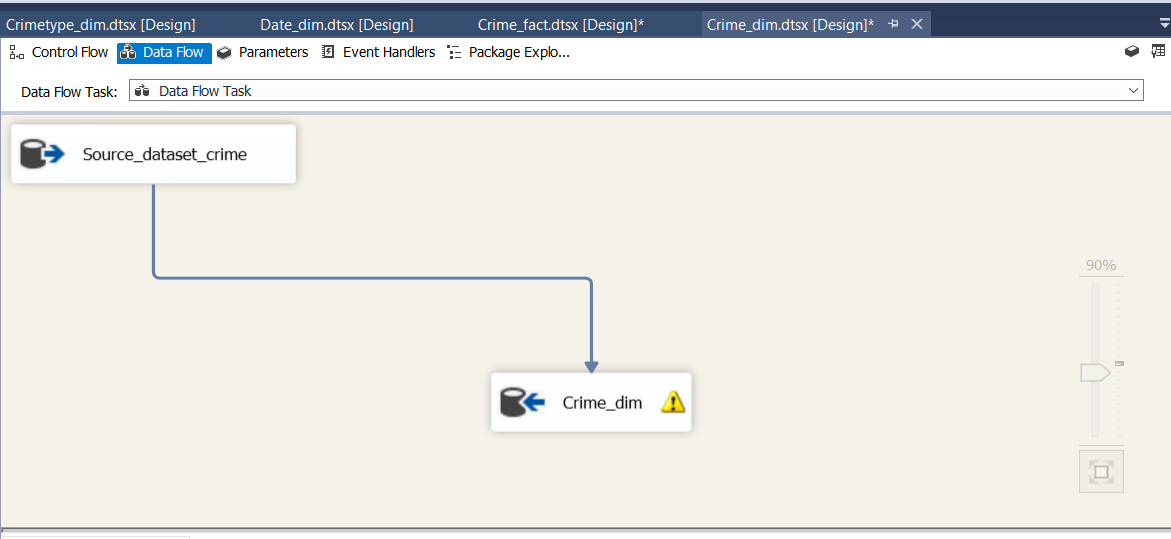


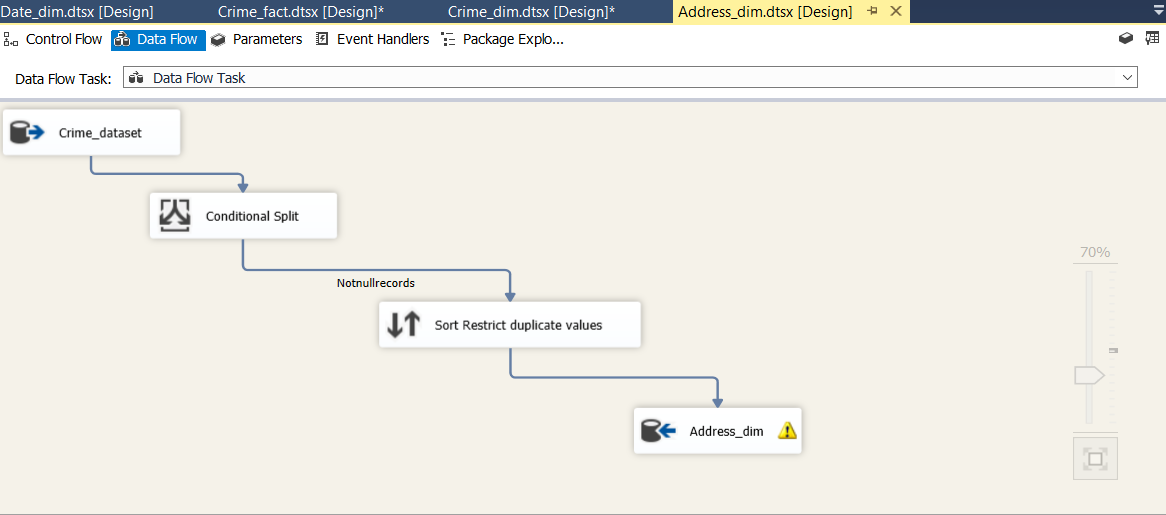
Fig. 1.1

Tables for dimension and facts were populated through the data source uploaded in the SSMS. Once the dataset was available then using it as the Source in SSIS Tables where populated. Populating the crime\_dim table it was done by creating a new package, Source selected as the dataset from SSMS and populated using the unique case id into crime\_dim table (fig 1.2), crime fact (refer fig 1.5).



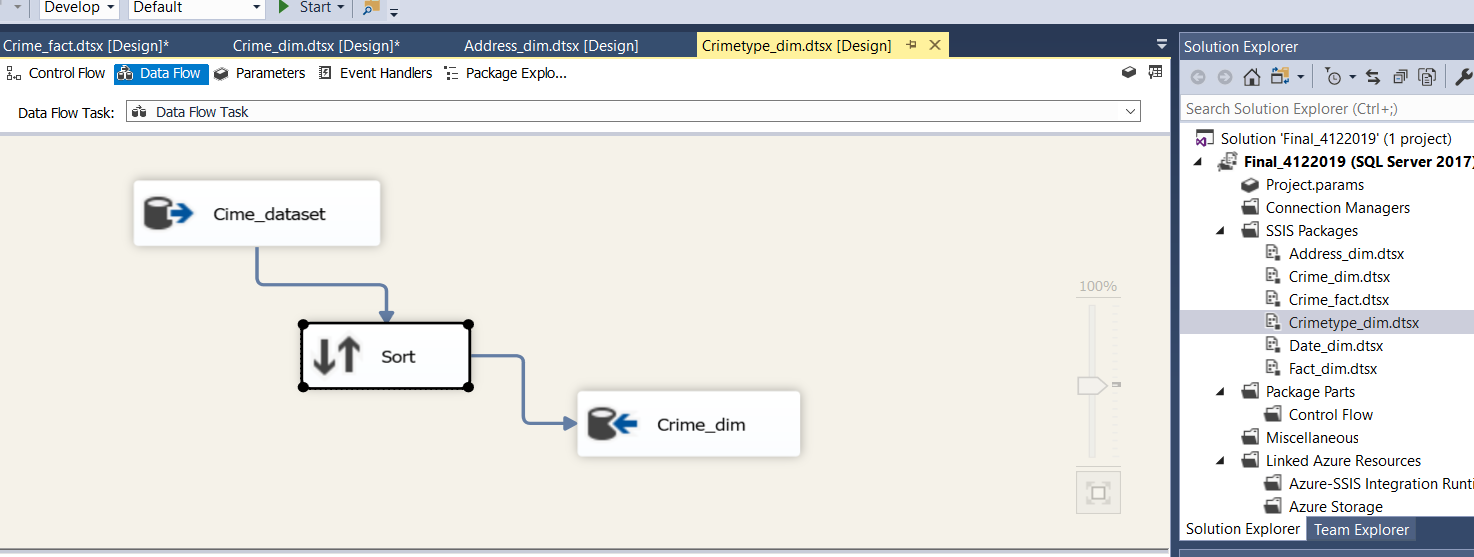
**Fig 1.2 Crime dimension population**

Similarly, for address\_dim (fig 1.3) we have used SSIS expression for restricting the null and duplicate values to the dimension table.



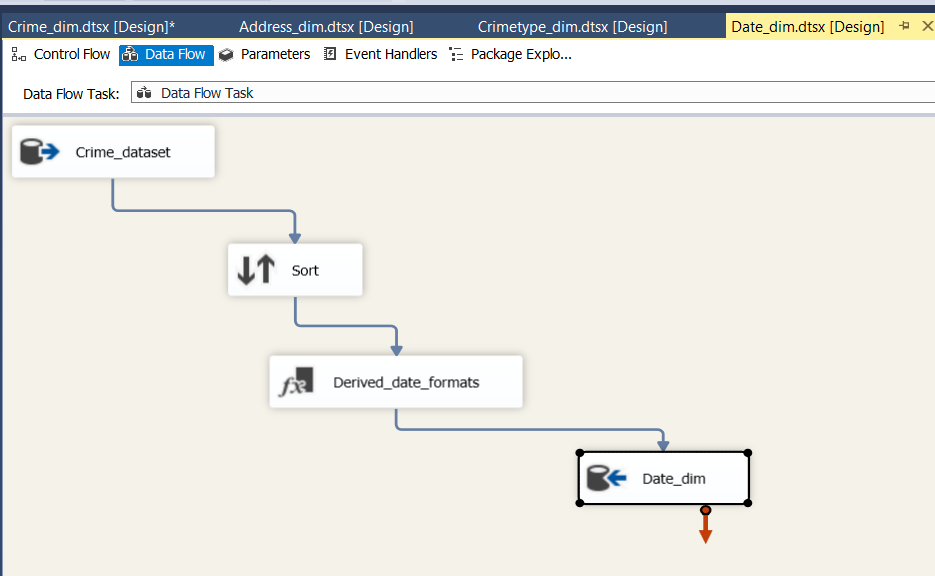
**Fig 1.3 Address dimension**

Crimetype\_dim (refer fig 1.4), It contains sort expression in which duplicate records are suppressed and only unique records with a primary key are passed.

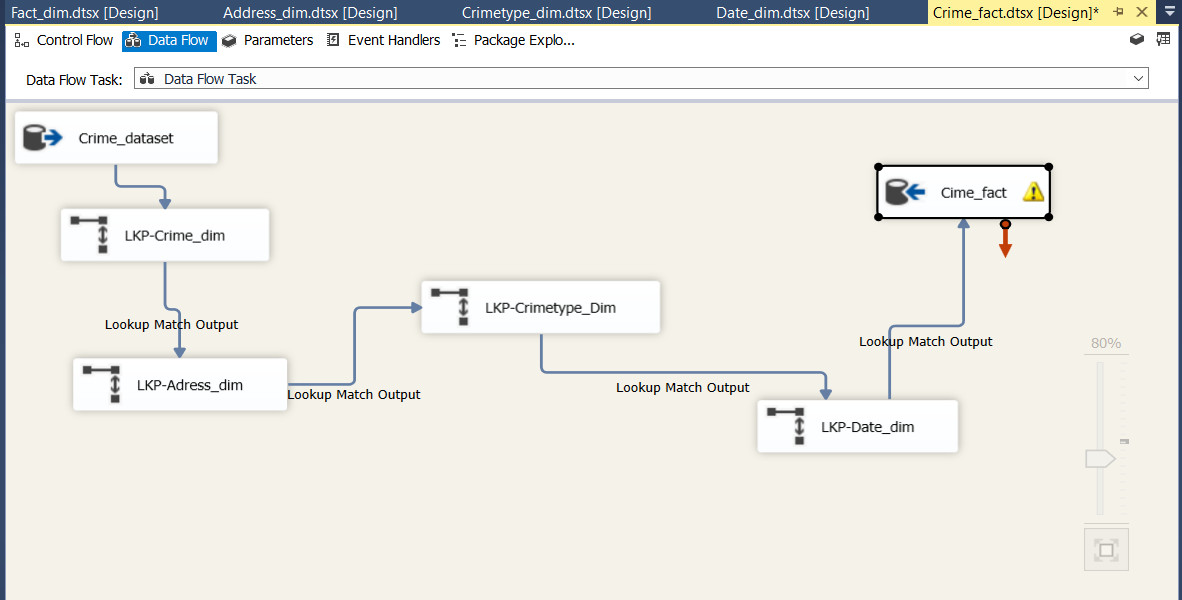


**Fig 1.4 Crime type Dimension**

Date dimension population using the crime data source and using derived columns while transforming the data into Date\_dim (Refer fig 1.5). In the derived columns cast functions are used to convert the Datekey to normal date, extract month, week and quarter to populate the dimension model.

  
 **Fig 1.5 Date\_dim**

Final population for data into crime fact table in which multiple Look-up expressions are for all the dimension tables (Fig 1.6)



**Fig 1.6 Crime\_Fact**

# 4. VISUALIZATIONS AND REPORTS

Visualizations and reports are very important aspects of a Business Intelligence system. More and more data users depend on visualizations and reports these days to plan their business strategy rather than looking around the data in the data warehouse. Visualizations and Reports are not just for communicating insights to decision-makers, they also help you spot trends in data that may not have been noticeable from database table alone.

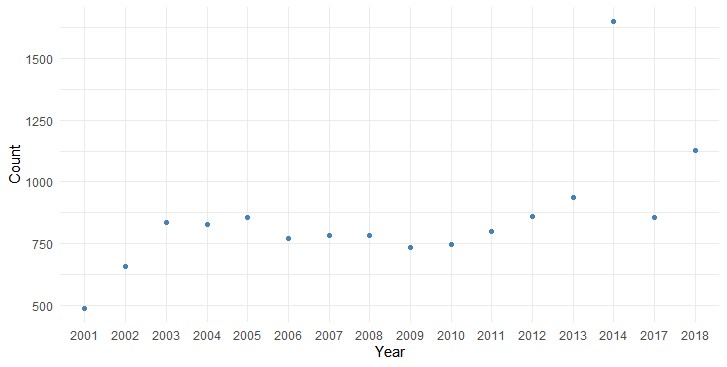
With the above advantages, it becomes imminent to represent our data mart using visualization with R and using reports with Microsoft SSRS (SQL server reporting system).

## 4.1. Visualizations

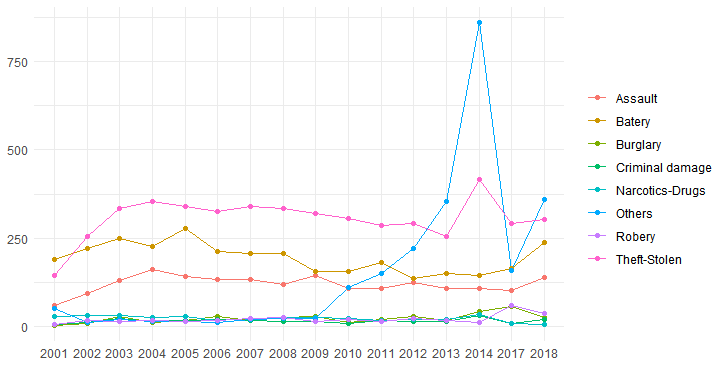
Mention the business requirement and include Visualization.

See Appendix A for code

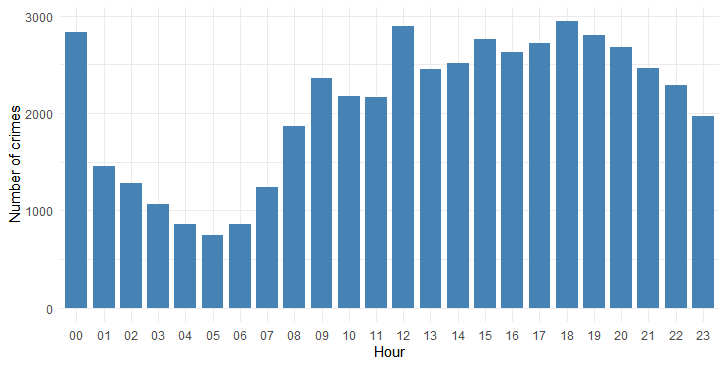
1)In the first Graph, we tried to portray the crime that evolved in the city of Chicago over time. In this Graph, X-axis represents the Year and Y-axis represents the total number of crimes.



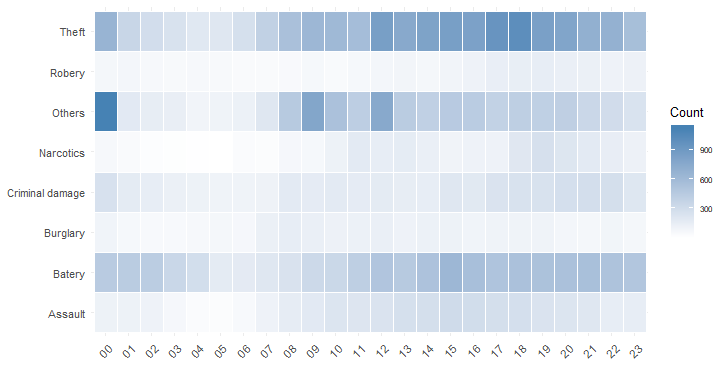
2) In the first graph, we tried to portray the crime that evolved in the city of Chicago over time from 2001 to 2018. The respective Color line represents the particular type of crime its shown in the legend.Y-axis represents the count of individual crimes.



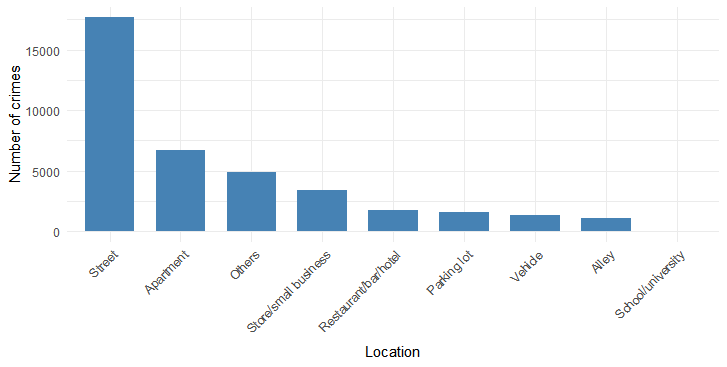
3) In this graph, we have shown the pattern that the time(hourly) of day do most crimes occur. The X-axis represents the Time in the 24hour hour format and the Y-axis represents the Total count of crime.



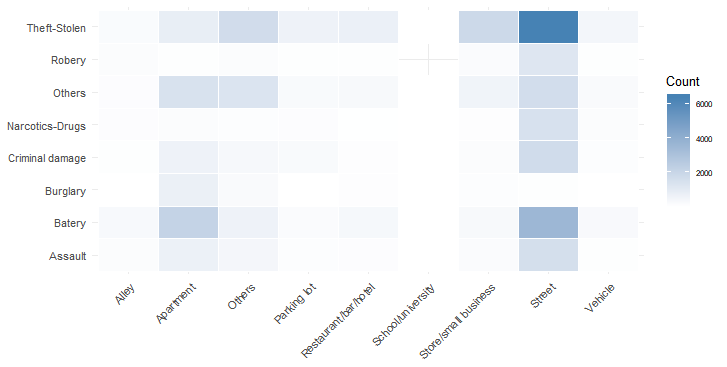
4) In this Heat Map, we have shown the pattern that the time(hourly) of day do the type of crimes occur. The X-axis represents the Hour in 24 Hour Format and Y-axis represents the Crime type. Darker the color higher the crime rate.



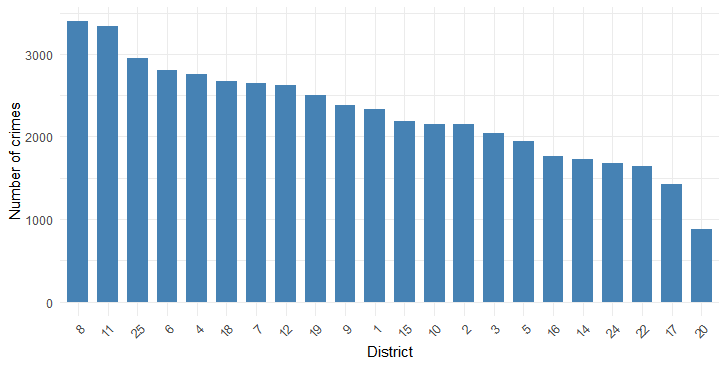
5) In which venue or locations of the city the crime is more likely to happen. Here X-axis represents the locations and Y-axis represents the number of crimes.



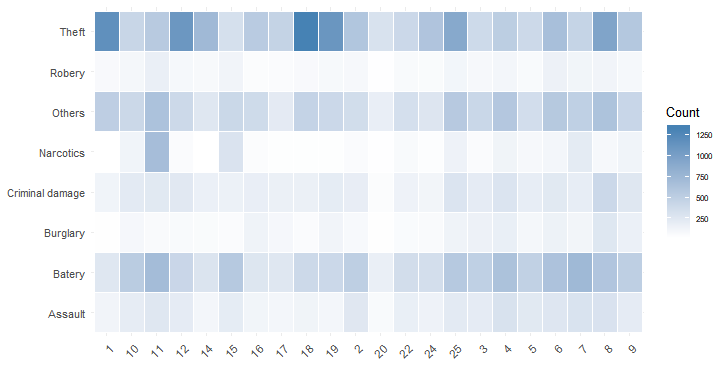
6) In Below Heat Map which venue or locations of the city the crime is more likely to happen. The darker the color, it's more likely to happen the crime.The X-axis represents common Location and Y-axis represents the type of crime.



7)Plotting the graph as per the crimes that happened in the district. The X-axis represents the District number and Y-axis represents the number of crimes.



8)In below Heat map, it represents the graph as per the type of crimes that happened in the district. Darker the Blue color it’s more likely to happen. The X-axis represents the District Number and Y-axis represents the category of crime.



## 4.2. Reports

Following 4 reports are generated in SSRS: -

1. **Year-wise and district wise breakup of crimes and % change in crime count for a given district as compared to the previous year:**

**Business Requirement: -** The Chicago Police Commissioner’s office has launched a new reward scheme to reward then district police chiefs where the % crime count for the respective district saw a decline in the last 18 years. They need a report showing the % change in overall crime count for each district for the last 18 years.

**Report Description-**This report basically shows the year of crime in 1st column followed by the police district where the crime was reported in 2nd column. 3rd column shows the total count of reported crimes for that particular year and police district.4th column shows a % change of crime counts for the police district when compared to previous year count.

****

1. **Year wise break up of different types of crimes:**

**Business Requirement:** Some print and electronic media in the city of Chicago want to write and present articles about the average number of different types of crimes for the past 18 years and hence make its readers aware of city’s crime scene both in history and present.

**Report Description:** This report shows Year of case registration in column 1, Type of crime in column 2 and count of such type of crimes for that particular year in the 3rd column. From the report it can be seen clearly that very high number of “unlawful possession of handguns” were reported in the year 2016 which came down significantly immediately the next year 2017 after stricter norms were introduced in Gun Laws in the state.



1. **Crime count by time of day:**

**Business Requirement:** The ward councilors of city of Chicago are planning to meet the city police commissioner and appeal him to adjust the availability of Beat cops during different times of day. For this, they want a comprehensive report that covers the count of crimes reported during different times of the day.

**Report Description:** This report divides a day into four time zones i.e. Morning (5 AM -11 AM), Noon (11 AM-5 PM), Evening (5 PM to midnight) and Night (midnight to 5 AM) and displays the cases registered during those zones on a yearly basis.

****

1. **Highest type of crimes for a given location type: -**

**Business Requirement:** The department of community education under the Chicago police wants to start an extensive community education program to educate people about the types of crime that usually take place in different location types like abandoned buildings, apartment, banks, streets etc. The purpose is to make them vigilant while inside their homes or while stepping out. Due to resource constraints, they just need to study the highest types of crimes that took place in different types of locations in the past 18 years to devise their course of action.

**Report Description:** There can be various location types like abandoned buildings, apartment, bank, airports etc. where various types of crimes like homicide, theft, burglary etc. are committed. This report gives yearly insight of maximum number of “type of crime” that the location type witnessed over a period of 18 years starting from 2001.

****

# 5. Include XML and Schema

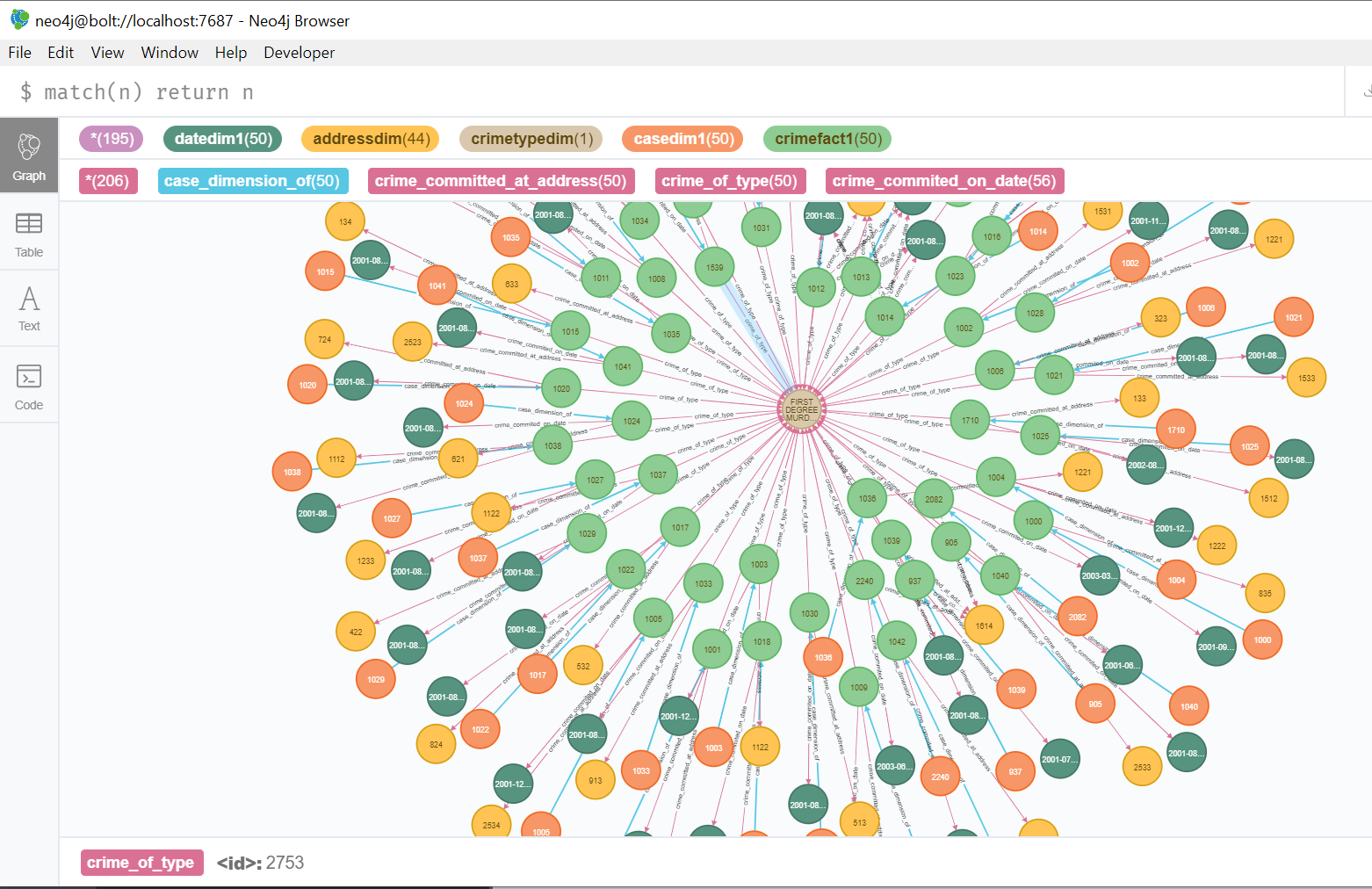
Details should be given here.

# 6. Graph Databases

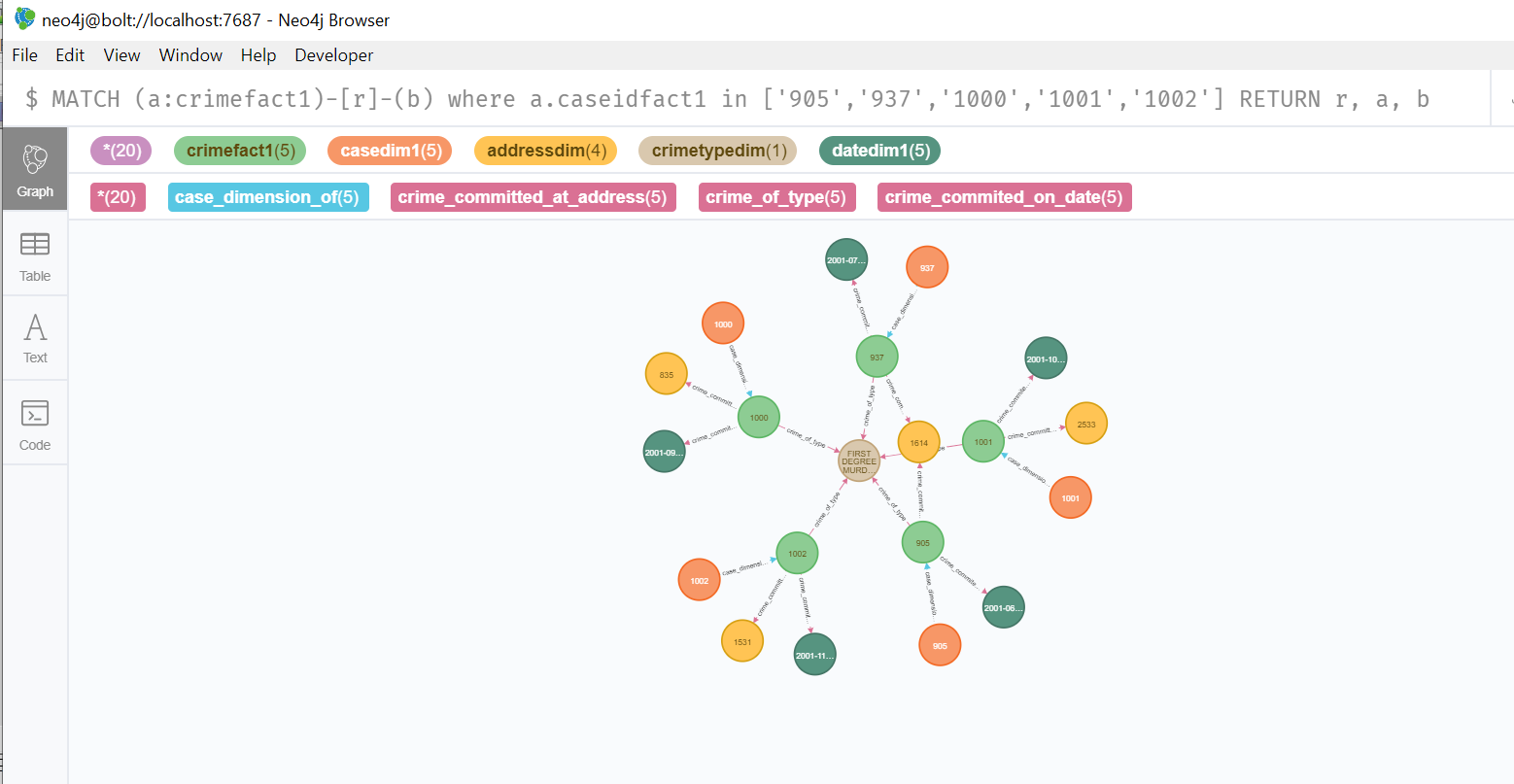
A graph is a pictorial representation of a set of objects where some pairs of objects are connected by links. It is composed of two elements - nodes (vertices) and relationships (edges).

Graph database is a database used to model the data in the form of graph. In here, the nodes of a graph depict the entities while the relationships depict the association of these nodes. Here, we are replicating a portion of our data mart into a popular graph database called Neo4j.

50 records from each of the dimension tables and 50 records from Fact table were first exported from SQL server and imported to Neo4j to implement a part of the data mart into Neo4j.Indices and relationships for the nodes were created and the respective CQL’s can be found attached in Appendix B. Below are the graphical representations of the portion of data mart:



This contains 196 nodes and 206 relationships. When we limit it to display for only 5 CASE\_ID’s, then it looks like:



A JPEG file for the same screenshot is attached here after downloading from Neo4j browser.



The various types of relationships are created as follows: -



## 6.1. COMAPRISON to realtional databases

**Why Graph Database?**

Nowadays, most of the data exists in the form of the relationship between

different objects and more often, the relationship between the data is more

valuable than the data itself.

Relational databases store highly structured data which have several records

storing the same type of data so they can be used to store structured data

and, they do not store the relationships between the data.

Unlike other databases, graph databases store relationships and connections

as first-class entities.

The data model for graph databases is simpler compared to other databases

and, they can be used with OLTP systems. They provide features like

transactional integrity and operational availability.

**RDBMS Vs Graph Database**

Following is the table which compares various aspects of Relational

databases and Graph databases.

|  |  |  |
| --- | --- | --- |
| **Serial No.** | **RDBMS** | **Graph Database** |
| 1. | Tables | Graphs |
| 2. | Rows | Nodes |
| 3. | Columns & Data | Properties and its values |
| 4. | Constraints | Relationships |
| 5. | Joins | Traversal |

**Advantages of Neo4j over RDBMS:**

1. Flexible data model: The data model for Neo4 can be easily changed as per the j requirements of the industry.
2. Real time insights: Neo4j provides results on real time basis.
3. Connected and semi-structured data: Connected and semi structured data can be easily represented in Neo4j which is not the case with traditional RDBMS’s.
4. Unnecessary data exclusion: Neo4j only shows useful data. It would not show a node with nulls.
5. Easy retrieval: Retrieval of connected data is faster in Neo4j.
6. No joins: Neo4j does not need complex joins to retrieve connected data. Rather, it easily retrieves data from adjacent nodes or relationship details which makes it a faster database to fetch connected data when compared to relational databases.

To support our understanding of Neo4j and as a proof of concept that Neo4j is faster and easier for fetching connected data, we are hereby showcasing few examples from our Chicago Crime Data mart. We can see clearly that cypher queries don’t need to go for inner joins and are rather using existing relationships to fetch connected data:

1. **Example 1**- Find all case id’s on every date from the year ‘2002’

SQL server query:



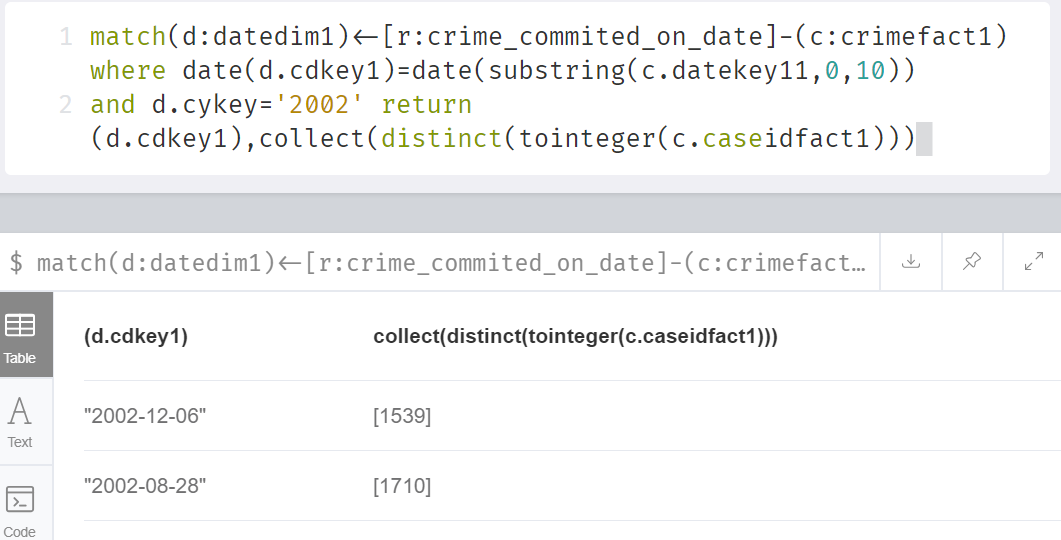
Result:

|  |  |
| --- | --- |
| case\_date | case\_id\_fact |
| 8/28/2002 | 1710 |
| 12/6/2002 | 1539 |

Cypher Query:



Result:



1. **Example 2**-Find latitude and longitude of the crime place related to case\_id=’1001’

SQL server query:



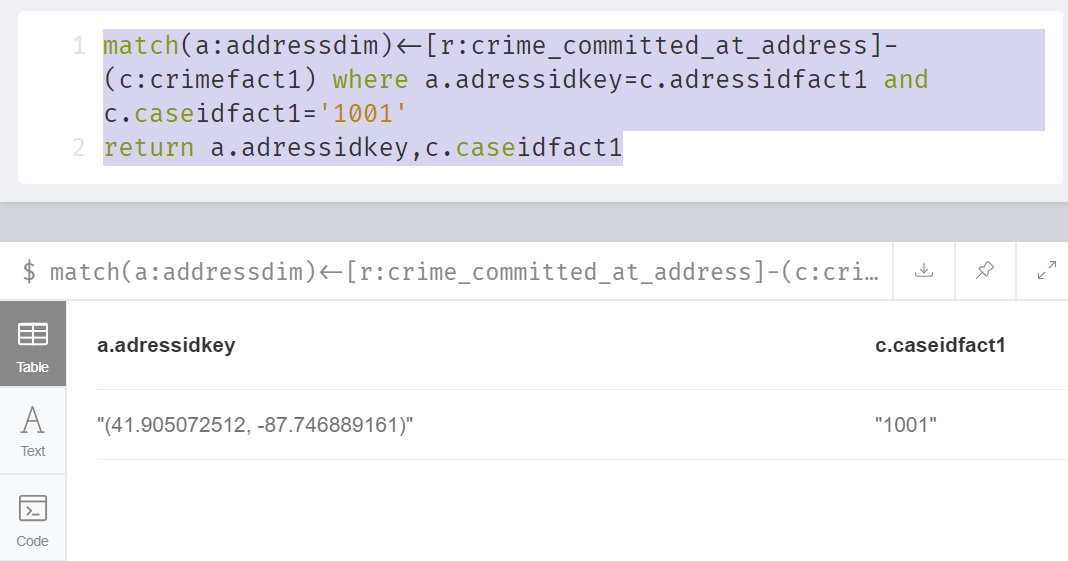
Result:

|  |  |
| --- | --- |
| case\_id\_fact | addressid |
| 1001 | (41.905072512, -87.746889161) |

Cypher Query:



Result:



1. **Example 3**-Find the total count of cases where primary type is ‘HOMICIDE’

SQL server query:



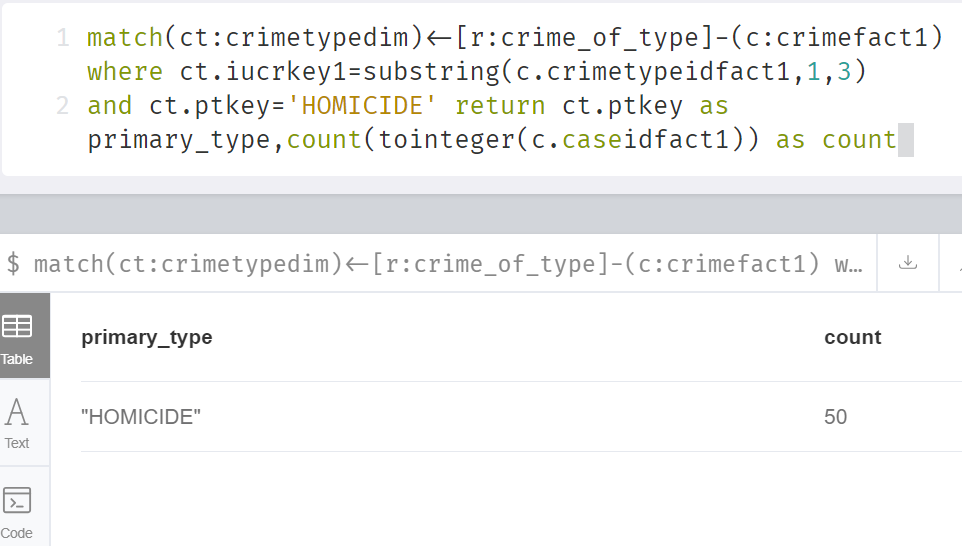
Result:

|  |  |
| --- | --- |
| PRIMARY\_TYPE | count |
| HOMICIDE | 50 |

Cypher Query:



Result:



# 7. Bibliography

1. Chicago Data Portal, Crimes 2001 to present

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>

2. Techtarget, Star Schema definition and benefits

<https://searchdatamanagement.techtarget.com/definition/star-schema>

3. Tutorialspoint , Neo4j Overview

<https://www.tutorialspoint.com/neo4j/neo4j_overview.htm>

1. Matillion,Data visualization crucial to your BI success

<https://www.matillion.com/resources/blog/data-visualization-crucial-to-your-bi-success/>

1. Code for R references.

<https://www.analyticalgaurav.com>

<https://rpubs.com/Ravinderkhatri/crimetrends>

<https://github.com/eugenividal/>

# Appendix A – SSIS Packages

**ORDER BY Customers.CustomerID;**

# Appendix B – VISUALIZATION CODES in R



# Appendix C– SSRS CODES

****

# Appendix D – NEO4j codes

