Chicago Crime

Big Data Project

Table of Contents

Architecture

Hadoop is an open-source distributed computing framework that allows users to store and process large datasets across clusters of commodity hardware. It is designed to be scalable, fault-tolerant, and efficient in handling big data. The Hadoop architecture consists of four key components: Hadoop Distributed File System (HDFS), Yet Another Resource Negotiator (YARN), MapReduce, and Hadoop Common.

Hadoop Distributed File System (HDFS): HDFS is the primary storage system for Hadoop. It is designed to store large files across multiple machines and provides reliable and scalable data storage. HDFS is composed of two types of nodes: NameNode, which manages the file system metadata, and DataNode, which stores data. HDFS provides a high level of fault tolerance by replicating data across different nodes in the cluster, ensuring that data is always available, even in the event of node failures.

Yet Another Resource Negotiator (YARN): YARN is a resource management framework that allows users to manage resources in a Hadoop cluster. It is responsible for allocating resources to various applications and coordinating the execution of tasks on a Hadoop cluster. YARN also provides a set of APIs that enable developers to write custom resource managers. YARN enables Hadoop to support a wide range of processing frameworks, such as Apache Spark, Apache Flink, and Apache Hive, among others.

MapReduce: MapReduce is a programming model used to process large datasets in parallel across a Hadoop cluster. It consists of two phases: map and reduce. The map phase processes input data and outputs key-value pairs, and the reduce phase aggregates the output of the map phase. MapReduce is highly scalable, and it can process terabytes or petabytes of data in a matter of hours or days.

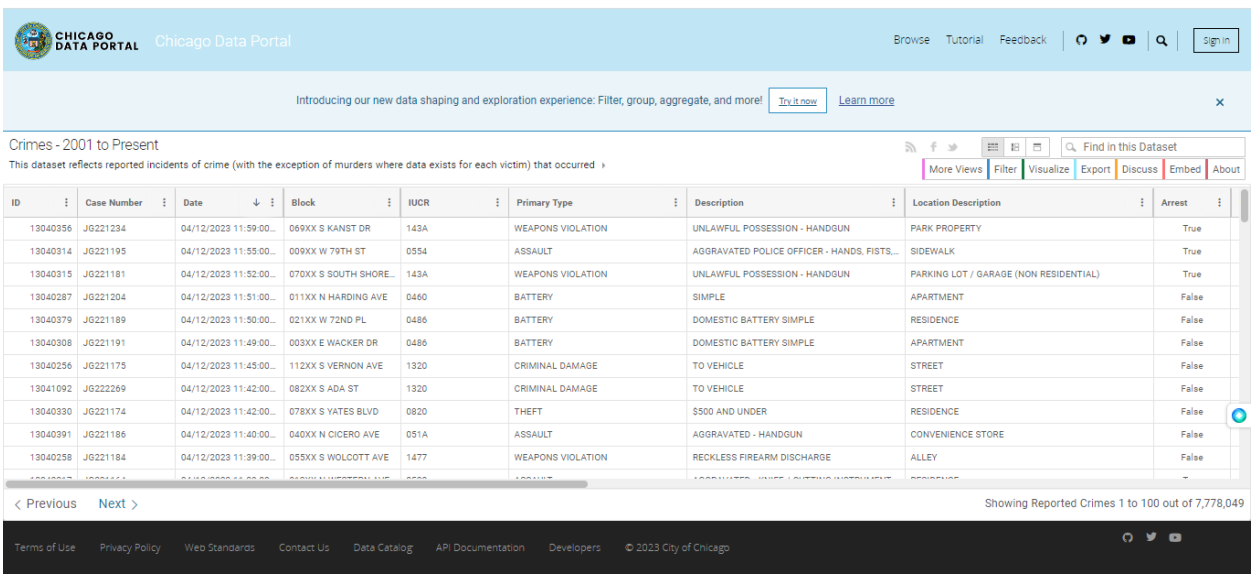
Hadoop Common: Hadoop Common is a set of common libraries and utilities used by all the components of the Hadoop ecosystem. It includes modules for configuration management, security, and other common functionalities. Hadoop Common provides a consistent platform for developers to build and deploy Hadoop applications.

Overall, the Hadoop architecture provides a scalable and fault-tolerant solution for processing and managing big data. By utilizing HDFS for reliable and scalable storage, YARN for resource management, MapReduce for parallel processing, and Hadoop Common for common functionalities, Hadoop enables efficient and effective processing of large datasets. With its robust architecture and open-source nature, Hadoop has become a popular choice for businesses and organizations looking to harness the power of big data.

Data Description :

As a dataset, the Chicago crime data contains information on reported incidents of crime in the city from 2001 to the present date. The data is updated daily and includes records of crimes such as homicide, theft, assault, and drug offenses, among others. Each record includes information such as the date and time the crime was reported, the location of the crime, the type of crime committed, and the FBI crime classification. It contains over 7 million records and can be used to analyze crime trends and patterns in the city over time.

- The dataset is obtained from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.
- It includes the following:
  - 29 columns of which 15-20 columns are used for our analysis.
  - 7,778,049 instances.
  - Missing and redundant values which we have cleaned.



ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest
13040356	JG221234	04/12/2023 11:59:00...	069XX S KANST DR	143A	WEAPONS VIOLATION	UNLAWFUL POSSESSION - HANDGUN	PARK PROPERTY	True
13040314	JG221195	04/12/2023 11:55:00...	009XX W 79TH ST	0554	ASSAULT	AGGRAVATED POLICE OFFICER - HANDS, FISTS,...	SIDEWALK	True
13040315	JG221181	04/12/2023 11:52:00...	070XX S SOUTH SHORE...	143A	WEAPONS VIOLATION	UNLAWFUL POSSESSION - HANDGUN	PARKING LOT / GARAGE (NON RESIDENTIAL)	True
13040287	JG221204	04/12/2023 11:51:00...	011XX N HARDING AVE	0460	BATTERY	SIMPLE	APARTMENT	False
13040379	JG221189	04/12/2023 11:50:00...	021XX W 72ND PL	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False
13040308	JG221191	04/12/2023 11:49:00...	003XX E WACKER DR	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	False
13040256	JG221175	04/12/2023 11:45:00...	112XX S VERNON AVE	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	False
13041092	JG222269	04/12/2023 11:42:00...	082XX S ADA ST	1320	CRIMINAL DAMAGE	TO VEHICLE	STREET	False
13040330	JG221174	04/12/2023 11:42:00...	078XX S YATES BLVD	0820	THEFT	$500 AND UNDER	RESIDENCE	False
13040391	JG221186	04/12/2023 11:40:00...	040XX N CICERO AVE	051A	ASSAULT	AGGRAVATED - HANDGUN	CONVENIENCE STORE	False
13040258	JG221184	04/12/2023 11:39:00...	055XX S WOLCOTT AVE	1477	WEAPONS VIOLATION	RECKLESS FIREARM DISCHARGE	ALLEY	False

Data Cleaning :
Python was used to clean the data through Panda library.

Column Reduction-

- X coordinate
- Y coordinate
- Latitude
- Longitude

Data Transformation-

- Date formatting.
- Handling Null values by dropping unnecessary rows.
- Replacing null with appropriate values.

Column Name	Description	Type
ID	Unique identifier for the record.	Number
Case Number	The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	Plain Text
Date	Date when the incident occurred. this is sometimes a best estimate.	Date & Time
Block	The partially redacted address where the incident occurred, placing it on the same block as the actual address.	Plain Text
IUCR	The Illinois Unifrom Crime Reporting code. This is directly linked to the Primary Type and Description. See the list of IUCR codes at https://data.cityofchicago.org/d/c7ck-438e.	Plain Text
Primary Type	The primary description of the IUCR code.	Plain Text
Description	The secondary description of the IUCR code, a subcategory of the primary description.	Plain Text
Location Description	Description of the location where the incident occurred.	Plain Text
Arrest	Indicates whether an arrest was made.	Checkbox
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	Checkbox
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts. See the beats at https://data.cityofchicago.org/d/aerh-rz74.	Plain Text
District	Indicates the police district where the incident occurred. See the districts at https://data.cityofchicago.org/d/fthy-xz3r.	Plain Text
Ward	The ward (City Council district) where the incident occurred. See the wards at https://data.cityofchicago.org/d/sp34-6z76.	Number
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas. See the community areas at https://data.cityofchicago.org/d/cauq-8yn6.	Plain Text
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS). See the Chicago Police Department listing of these classifications at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.	Plain Text
Year	Year the incident occurred.	Number
Updated On	Date and time the record was last updated.	Date & Time
Location	The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.	Location

Business Questions :-

1. **What are the top 5 most common types of crimes in Chicago?**

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
      POSS: HEROIN(BRN/TAN)   SIDEWALK      True    False   1412    14      35      21      18      2015    02/10
/2018 03:50:01 PM
Time taken: 0.538 seconds, Fetched: 5 row(s)
hive> Select Primary_Type, COUNT(ID) AS Crime_Count
    > From crime1
    > GROUP BY Primary_Type
    > ORDER BY Crime_Count DESC
    > LIMIT 5;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diff
erent execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hitali_20230421134503_03b05b12-2e67-4566-b542-70badacc193b
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 6
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
```

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
2023-04-21 13:49:56,793 Stage-2 map = 0%,  reduce = 0%
2023-04-21 13:50:14,729 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 5.4 sec
2023-04-21 13:50:21,126 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 7.71 sec
MapReduce Total cumulative CPU time: 7 seconds 710 msec
Ended Job = job_1680035323433_0002
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 6   Cumulative CPU: 111.05 sec   HDFS Read: 1301664273 HDFS Write: 1904 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.71 sec   HDFS Read: 8868 HDFS Write: 232 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 58 seconds 760 msec
OK
THEFT   1638908
BATTERY 1420259
CRIMINAL DAMAGE 885403
NARCOTICS       747160
ASSAULT 505931
Time taken: 320.709 seconds, Fetched: 5 row(s)
hive>
```

Hive query calculates the top 5 most common types of crimes in Chicago. Let me explain the information in more detail:

- Theft - This is the most common type of crime in Chicago, with 1,638,908 reported incidents. Theft refers to the unlawful taking or attempted taking of property from another person or business without the use of force.

- Battery - This is the second most common type of crime in Chicago, with 1,420,259 reported incidents. Battery is a crime that involves the intentional use of force or violence against another person.

- Criminal Damage - This is the third most common type of crime in Chicago, with 885,403 reported incidents. Criminal Damage refers to the intentional damage or destruction of property belonging to another person or business.

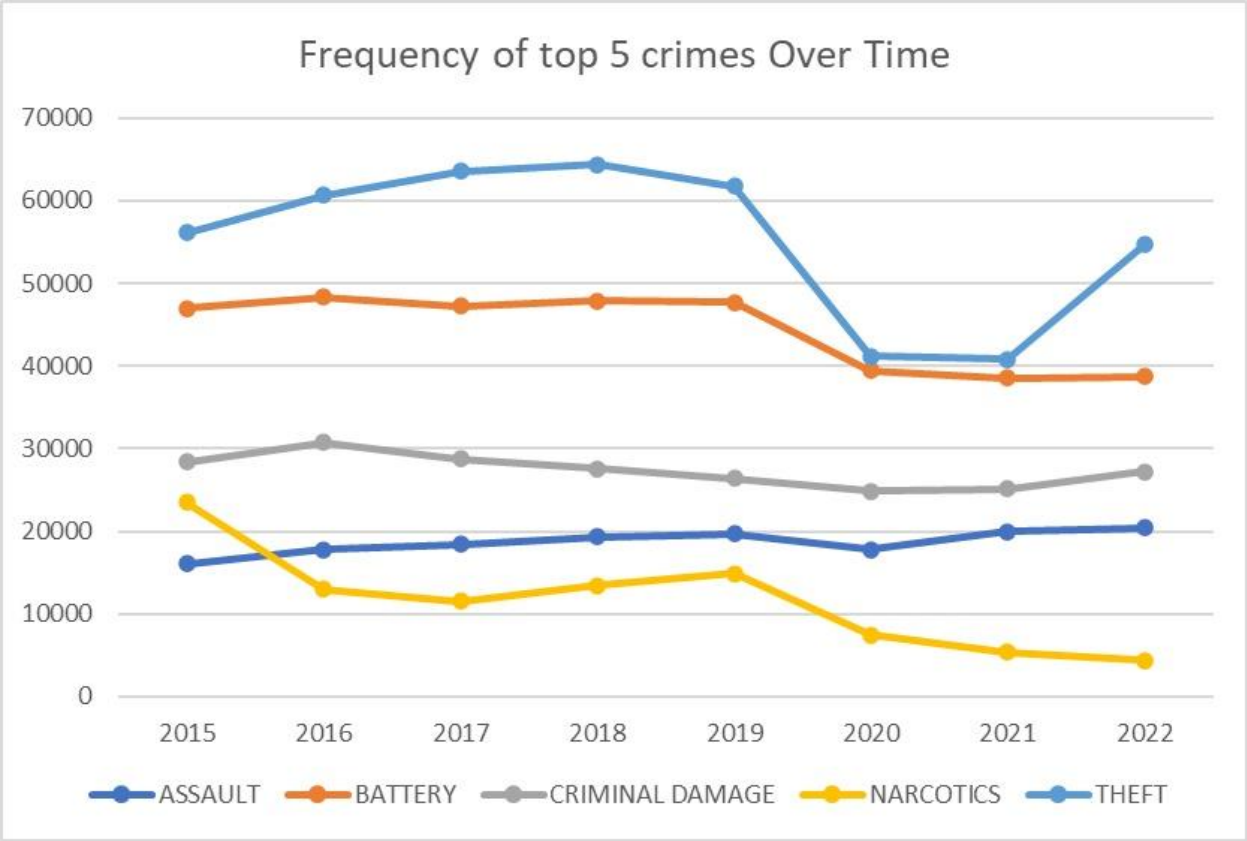
- Narcotics - This is the fourth most common type of crime in Chicago, with 747,160 reported incidents. Narcotics offenses include the possession, sale, or distribution of illegal drugs.

- Assault - This is the fifth most common type of crime in Chicago, with 505,931 reported incidents. Assault is like battery but does not necessarily involve physical contact. It can also include the threat of violence or the use of a weapon.

2. **How have the frequencies of these crimes changed over the last 8 years?**

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
hive> select Year, Primary_Type, Count(ID) AS Crime_count
    > From crime1
    > WHERE (Year BETWEEN 2015 AND 2023) AND (Primary_Type = 'THEFT' OR Primary_Type = 'BATTERY' OR Primary_Type =
 'CRIMINAL DAMAGE' OR Primary_Type = 'NARCOTICS' OR Primary_Type = 'ASSAULT')
    > GROUP BY Year, Primary_Type
    > ORDER BY Year, Crime_count DESC;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diff
erent execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hitali_20230421141444_6cb13007-296f-4617-9c01-b662c6e16070
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 6
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1680035323433_0007, Tracking URL = http://hitali-VirtualBox:8088/proxy/application_168003532343
```

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 6   Cumulative CPU: 111.81 sec   HDFS Read: 1301677983 HDFS Write: 2043 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 7.68 sec   HDFS Read: 9494 HDFS Write: 1547 SUCCESS
Total MapReduce CPU Time Spent: 1 minutes 59 seconds 490 msec
OK
2015    THEFT   56125
2015    BATTERY 46965
2015    CRIMINAL DAMAGE 28380
2015    NARCOTICS       23509
2015    ASSAULT 16073
2016    THEFT   60652
2016    BATTERY 48345
2016    CRIMINAL DAMAGE 30705
2016    ASSAULT 17816
2016    NARCOTICS       13012
2017    THEFT   63551
2017    BATTERY 47271
2017    CRIMINAL DAMAGE 28747
2017    ASSAULT 18385
2017    NARCOTICS       11510
2018    THEFT   64394
2018    BATTERY 47878
2018    CRIMINAL DAMAGE 27582
2018    ASSAULT 19288
2018    NARCOTICS       13415
2019    THEFT   61750
2019    BATTERY 47672
2019    CRIMINAL DAMAGE 26427
2019    ASSAULT 19681
2019    NARCOTICS       14925
2020    THEFT   41164
2020    BATTERY 39489
2020    CRIMINAL DAMAGE 24840
2020    ASSAULT 17745
2020    NARCOTICS       7460
2021    THEFT   40766
2021    BATTERY 38585
2021    CRIMINAL DAMAGE 25091
2021    ASSAULT 19983
2021    NARCOTICS       5323
```



The table represents the frequency of the top 5 crimes in Chicago for the years 2015 to 2023. The crimes are categorized as Theft, Battery, Criminal Damage, Narcotics, and Assault. The data is plotted on a line graph to visualize the trends and changes in crime frequencies over time.

The X-axis of the graph represents the years from 2015 to 2023, while the Y-axis represents the frequency of crimes reported. Each crime category is represented by a different color line on the graph. The legend on the graph shows which color line represents each crime category.

From the graph, we can observe that Theft has consistently been the most common crime reported in Chicago over the years. However, its frequency has decreased since 2016. Battery is the second most common crime reported, and its frequency has remained relatively stable over the years.

Criminal Damage and Narcotics have both decreased in frequency since 2015, while Assault has increased slightly. In 2020, there was a significant decrease in the frequency of all crime categories, likely due to the COVID-19 pandemic.

The graph provides a clear visual representation of the trends and changes in crime frequencies over time. This information can be helpful for businesses, residents, and law enforcement agencies to understand the patterns and trends of crime in Chicago and to take appropriate measures to prevent and combat crime in the city.

3. **What areas of the city have the highest crime rates?**

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
Kill Command = /usr/share/hadoop//bin/hadoop job  -kill job_1680035323433_0014
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2023-04-21 14:45:52,718 Stage-2 map = 0%,  reduce = 0%
2023-04-21 14:46:24,782 Stage-2 map = 100%,  reduce = 0%, Cumulative CPU 12.08 sec
2023-04-21 14:46:42,148 Stage-2 map = 100%,  reduce = 100%, Cumulative CPU 18.33 sec
MapReduce Total cumulative CPU time: 18 seconds 330 msec
Ended Job = job_1680035323433_0014
MapReduce Jobs Launched:
Stage-Stage-1: Map: 5  Reduce: 6   Cumulative CPU: 148.67 sec   HDFS Read: 1301668948 HDFS Write: 7954 SUCCESS
Stage-Stage-2: Map: 1  Reduce: 1   Cumulative CPU: 18.33 sec   HDFS Read: 14903 HDFS Write: 306 SUCCESS
Total MapReduce CPU Time Spent: 2 minutes 47 seconds 0 msec
OK
25      449549
8       266120
43      229088
23      217376
28      208751
24      206636
29      204259
67      200066
71      196337
49      181067
Time taken: 309.016 seconds, Fetched: 10 row(s)
hive>
```

```
hitali@hitali-VirtualBox: /usr/lib/hive/bin
hive> select Community_Area, count(ID) as Crime_Count
    > from crime1
    > where Community_Area IS NOT NULL AND Community_Area > 0
    > group by Community_Area
    > order by Crime_Count DESC
    > limit 10;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a diff
erent execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hitali_20230421144138_6ad37723-6927-424e-ae00-f9169f354373
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 6
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1680035323433_0013, Tracking URL = http://hitali-VirtualBox:8088/proxy/application_168003532343
3_0013/
Kill Command = /usr/share/hadoop//bin/hadoop job  -kill job_1680035323433_0013
Hadoop job information for Stage-1: number of mappers: 5; number of reducers: 6
2023-04-21 14:42:09,086 Stage-1 map = 0%,  reduce = 0%
```

Reference Table :-

Area Code	Area
5	North Center
25	Austin
8	Near North Side
43	South Shore
23	Humboldt Park
28	Near West Side
24	West Town
29	North Lawndale
67	West Englewood
71	Auburn Gresham
49	Roseland

Conclusion

The Chicago Police Department can implement a number of initiatives to reduce crime by identifying and targeting high-crime areas, using technology to identify potential crime hotspots, and improving community policing efforts. The city has seen increased efforts to engage communities in crime prevention and intervention programs, including efforts to address root causes of crime such as poverty and lack of access to education and healthcare. In recent years, Illinois has implemented a number of criminal justice reforms aimed at reducing incarceration rates and promoting alternatives to incarceration, such as drug treatment programs.

Based on these findings, it is recommended that law enforcement agencies focus on increasing their presence in the high-crime areas to deter criminal activities. Additionally, it is essential to investigate the reasons behind the decrease in the frequency of certain types of crimes, such as narcotics, and determine if the same strategies can be applied to reduce the frequency of other types of crimes. Furthermore, it is essential to continue monitoring crime rates and make adjustments to strategies as necessary to ensure public safety.