Crime and Safety in Residental Neighborhoods

Introduction

Our project mainly describes the data lifecycle of Data used in Crime and Safety in Residential Areas. The project focuses on addressing the rise in criminal activity in residential neighborhoods, with a specific emphasis on ensuring safety and security. We have taken the Static Dataset from https://catalog.data.gov/dataset/crime-data-from-2010-to-2019 and Streaming Dataset from https://catalog.data.gov/dataset/crime-data-from-2020-to-present. We have a started a brief analysis into Crime Statistics from 2010 to present using Static Data sets and Streaming Datasets from sources such as Data.gov.

Here the Use Case we have opted is Community: Residential Areas Issue: Crime and Safety; where crime and safety are the main concerns. We want to improve the overall safety of residential communities by exploring the aspects of this sector and offering practical insights to support crime prevention measures.

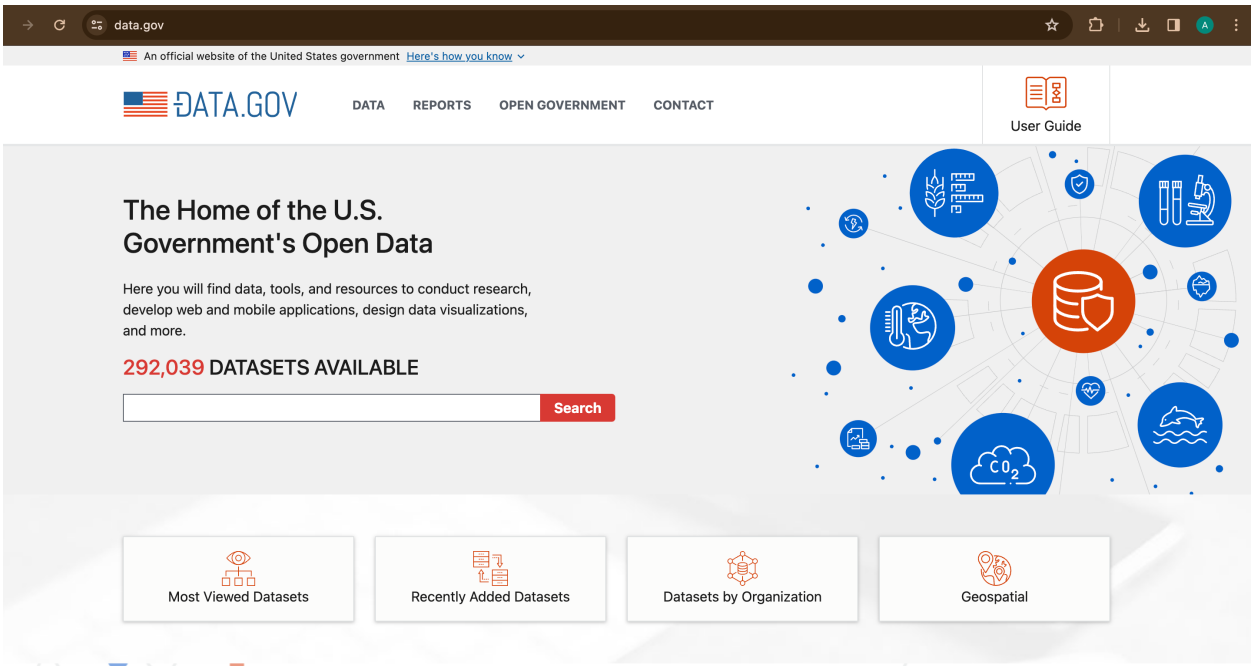
The Data lifecycle includes Generation where raw data is generated from repositories such as data.gov, Collection, Processing of Static and Streaming Data.

Tools and applications were used throughout the project to support various stages of the data lifecycle. Google Cloud Platform (GCP), which provides reliable infrastructure for processing and storing data. Large datasets could be stored with GCP Storage Bucket, and distributed data processing at scale was made possible by Dataproc's Hadoop architecture. Furthermore, Tools and applications such as Google Cloud Platform (GCP), GCP Storage Bucket, Dataproc, Hadoop Architecture, HDFS, HIVE, and Spark were utilized at different phases. The next steps for the data Science or Analyst teams involve performing queries in BigQuery, HIVE, and Spark to ensure data Quality. Spark developed as a powerhouse for high-performance analytics and processing, while HDFS and HIVE offered effective data management and querying capabilities.

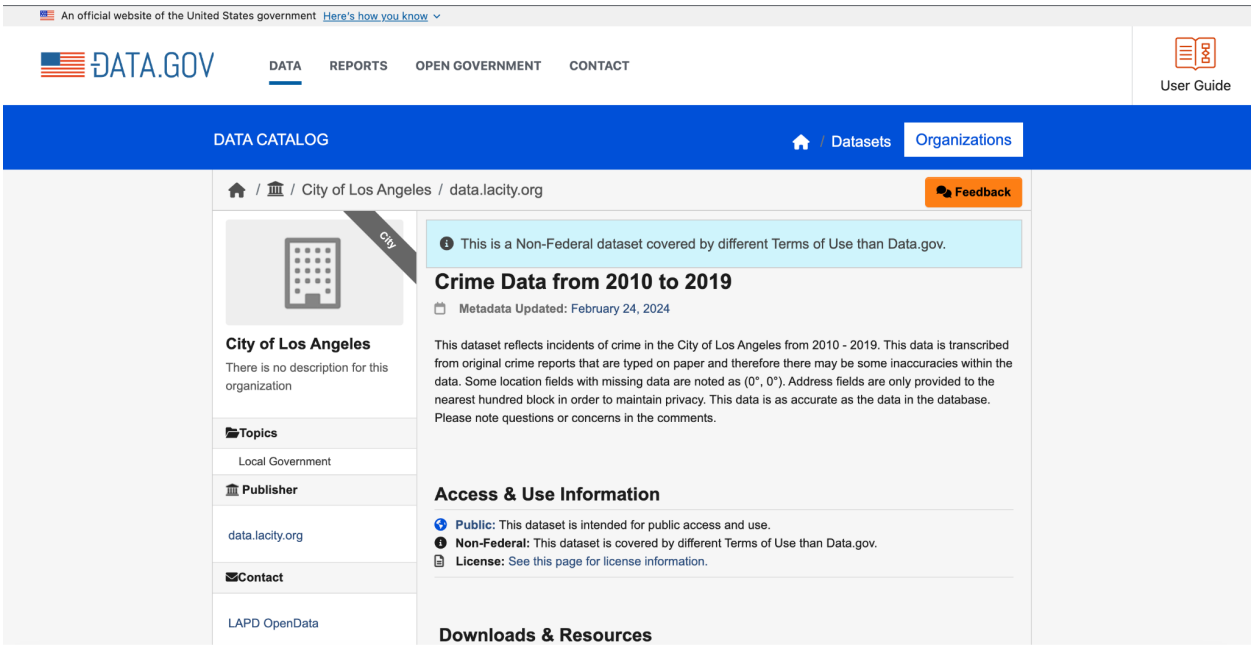
The emphasis of our project is shifting to ensure data integrity and extract relevant insights as we go into the next phase. Teams specializing in data science and analysis are prepared to explore the datasets more thoroughly and run complex queries using BigQuery, HIVE, and Spark. These searches will not only confirm the accuracy of the data but also reveal important trends and patterns that are essential for well-informed decision-making. In addition, cutting-edge analytical methods like predictive modeling and machine learning will be used to identify probable criminal hotspots and create proactive risk-reduction strategies.

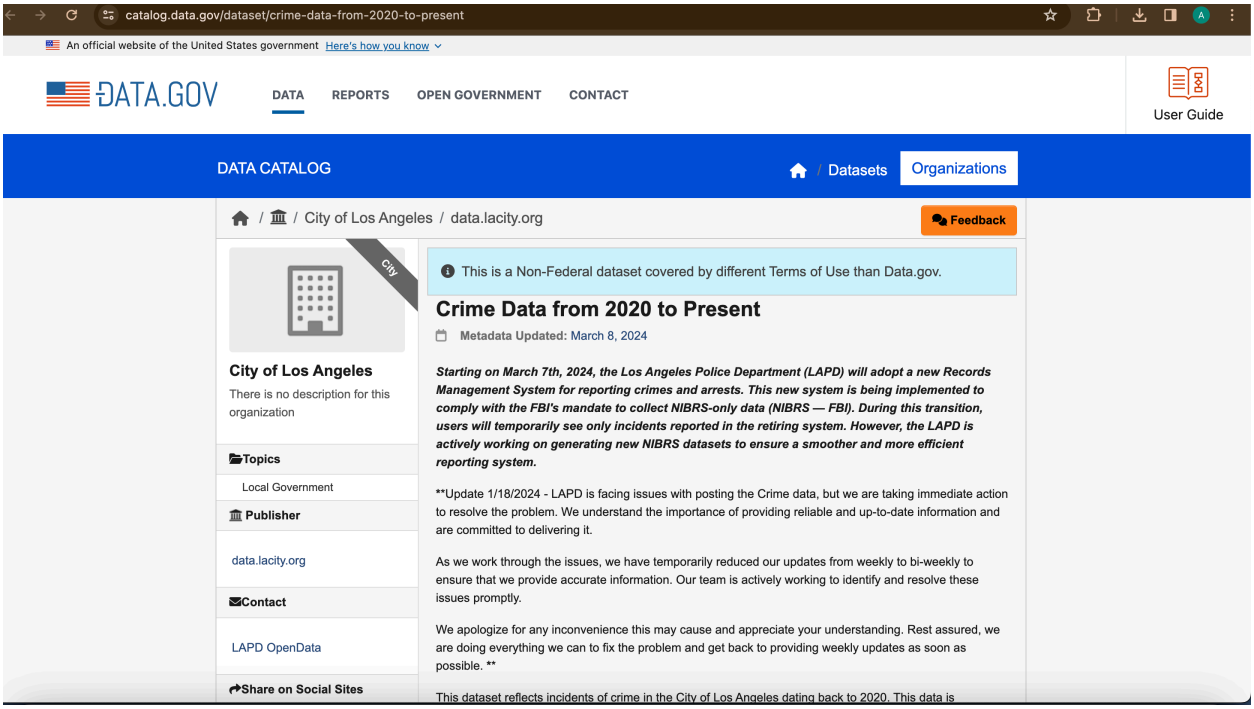
Step-1 Generation.

We have taken the raw Data from Data.gov repository that we have data from 2010 to 2019 for static Dataset and from 2020 to present as Streaming Dataset.



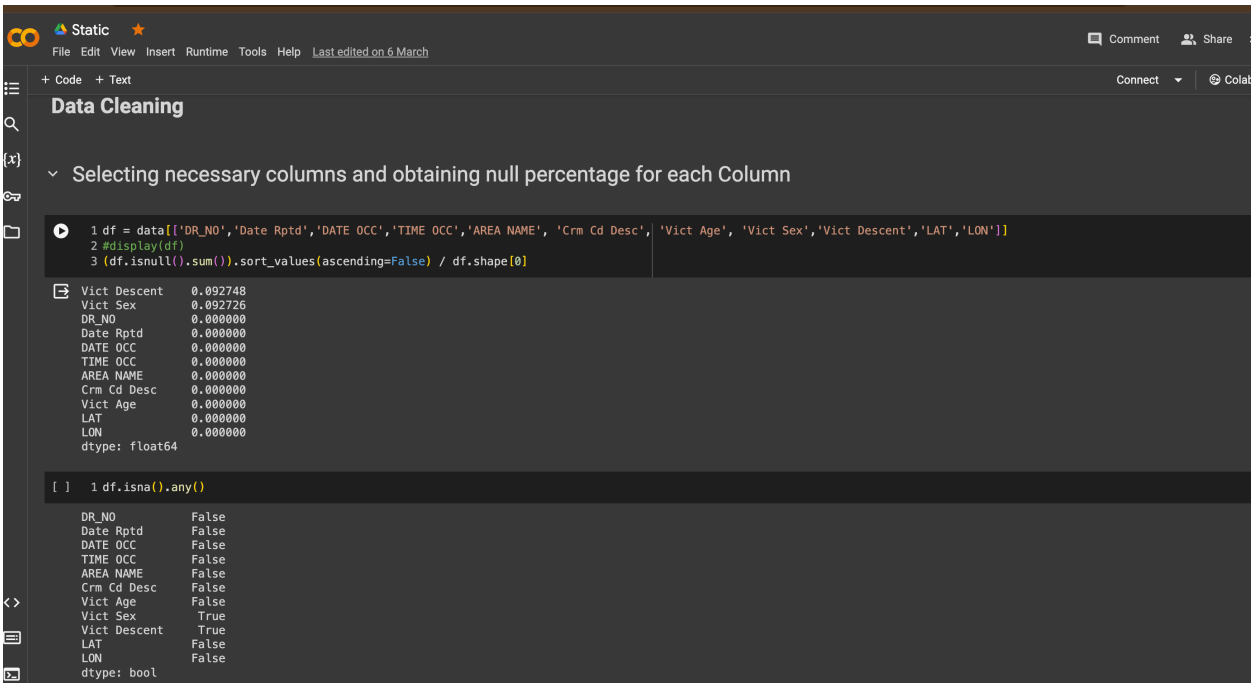
Step-2 Collection.

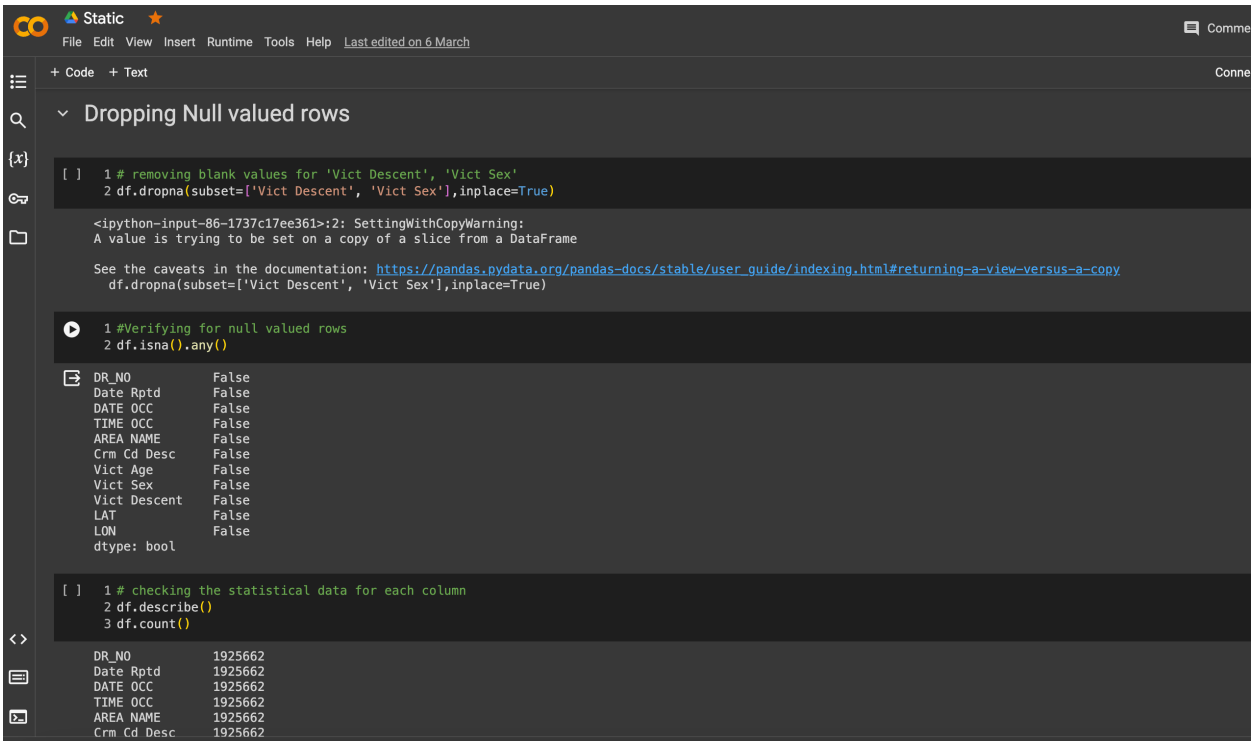


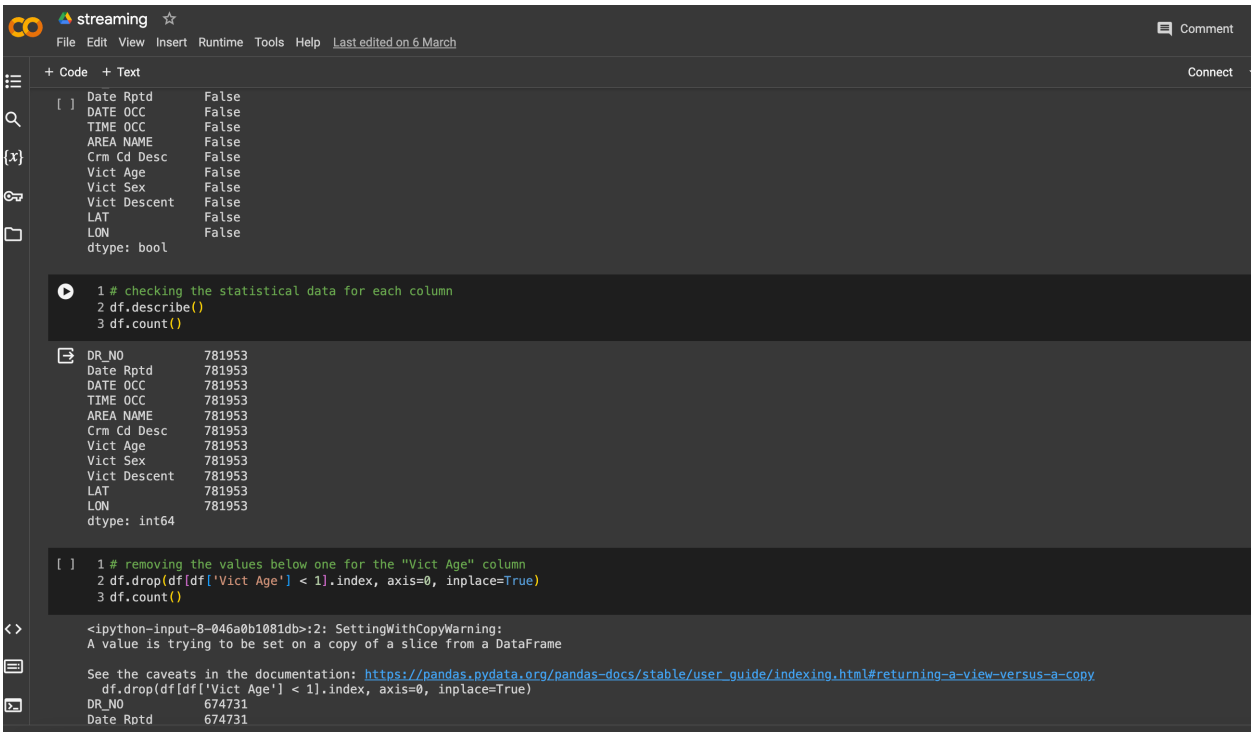


Static Dataset from https://catalog.data.gov/dataset/crime-data-from-2010-to-2019 and Streaming Dataset from https://catalog.data.gov/dataset/crime-data-from-2020-to-present which mainly have the data of crime in the city of Los Angeles.

Step 3- Processing.


Static
File Edit View Insert Runtime Tools Help Last edited on 6 March
Comment
Share
+ Code + Text
Connect
Data Cleaning
Selecting necessary columns and obtaining null percentage for each Column
1 df = data[['DR_NO','Date Rptd','DATE OCC','TIME OCC','AREA NAME', 'Crm Cd Desc', 'Vict Age', 'Vict Sex','Vict Descent','LAT','LON']]
2 #display(df)
3 (df.isnull().sum()).sort_values(ascending=False) / df.shape[0]
Vict Descent 0.092748
Vict Sex 0.092726
DR_NO 0.000000
Date Rptd 0.000000
DATE OCC 0.000000
TIME OCC 0.000000
AREA NAME 0.000000
Crm Cd Desc 0.000000
Vict Age 0.000000
LAT 0.000000
LON 0.000000
dtype: float64
1 df.isna().any()
DR_NO False
Date Rptd False
DATE OCC False
TIME OCC False
AREA NAME False
Crm Cd Desc False
Vict Age False
Vict Sex True
Vict Descent True
LAT False
LON False
dtype: bool


Static
File Edit View Insert Runtime Tools Help Last edited on 6 March
+ Code + Text
Dropping Null valued rows
1 # removing blank values for 'Vict Descent', 'Vict Sex'
2 df.dropna(subset=['Vict Descent', 'Vict Sex'],inplace=True)
<ipython-input-86-1737c17ee361>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
df.dropna(subset=['Vict Descent', 'Vict Sex'],inplace=True)
1 #Verifying for null valued rows
2 df.isna().any()
DR_NO False
Date Rptd False
DATE OCC False
TIME OCC False
AREA NAME False
Crm Cd Desc False
Vict Age False
Vict Sex False
Vict Descent False
LAT False
LON False
dtype: bool
1 # checking the statistical data for each column
2 df.describe()
3 df.count()
DR_NO 1925662
Date Rptd 1925662
DATE OCC 1925662
TIME OCC 1925662
AREA NAME 1925662
Crm Cd Desc 1925662



Data Cleaning

Selecting necessary columns and obtaining null percentage for each Column

```
1 df = data[['DR_NO','Date Rptd','DATE OCC','TIME OCC','AREA NAME', 'Crm Cd Desc', 'Vict Age', 'Vict Sex','Vict Descent','LAT','LON']]
2 #display(df)
3 (df.isnull().sum()).sort_values(ascending=False) / df.shape[0]
```

```
Vict Descent    0.092748
Vict Sex        0.092726
DR_NO           0.000000
Date Rptd       0.000000
DATE OCC        0.000000
TIME OCC        0.000000
AREA NAME       0.000000
Crm Cd Desc     0.000000
Vict Age        0.000000
LAT             0.000000
LON             0.000000
dtype: float64
```

```
1 df.isna().any()
```

```
DR_NO           False
Date Rptd       False
DATE OCC        False
TIME OCC        False
AREA NAME       False
Crm Cd Desc     False
Vict Age        False
Vict Sex         True
Vict Descent     True
LAT             False
LON             False
dtype: bool
```

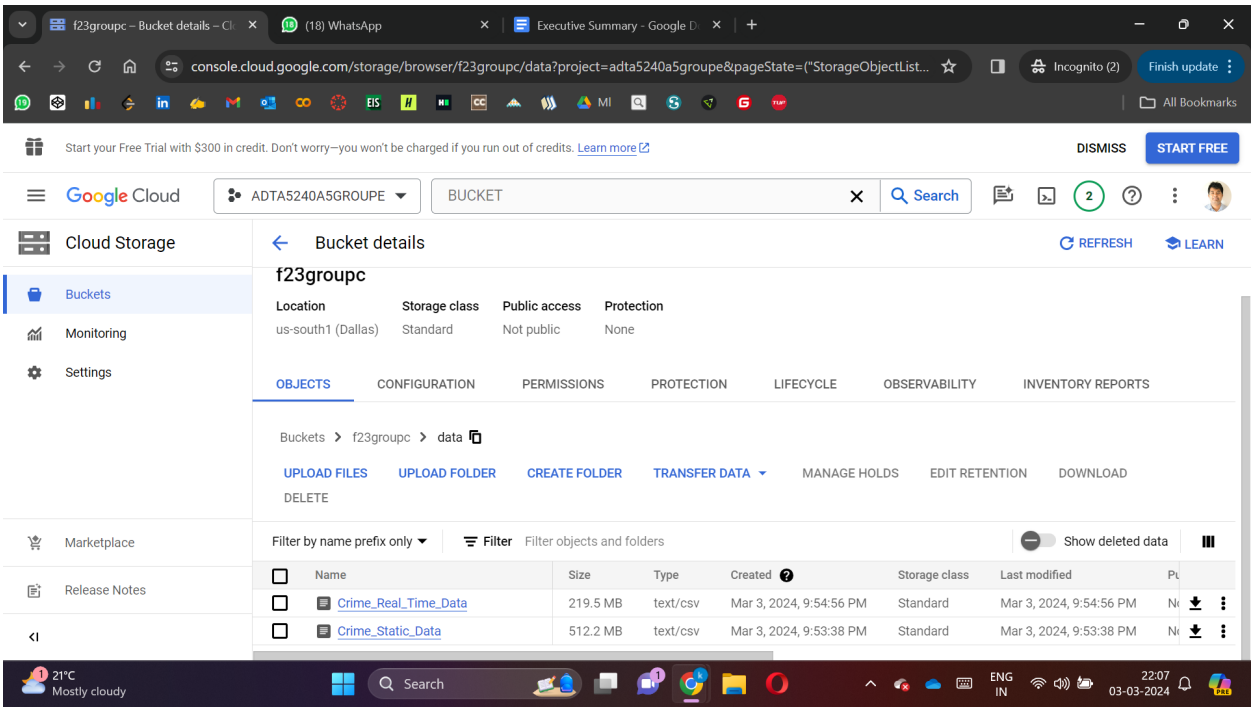
Step 4- Storage.

Google Cloud Platform

We have access to various tools the Google Cloud Platform (GCP) provides to create a solution that efficiently addresses the rise in criminal activity.

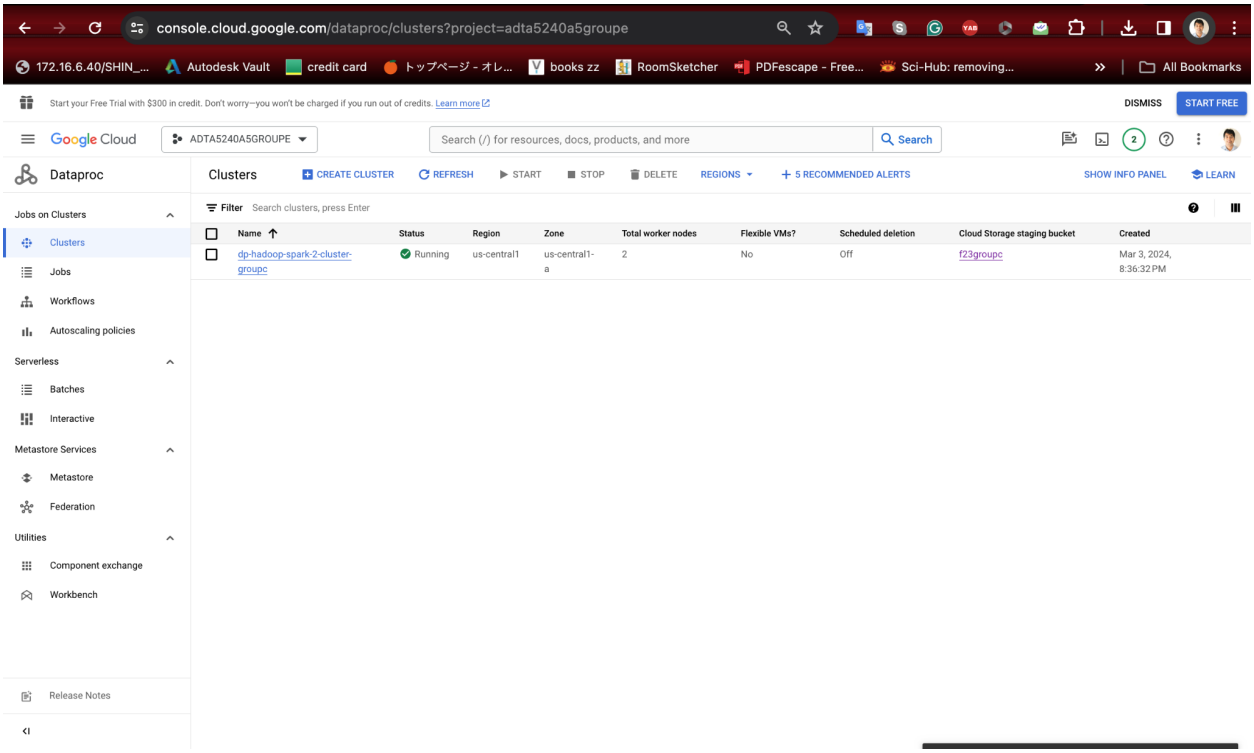
GCP Storage Bucket

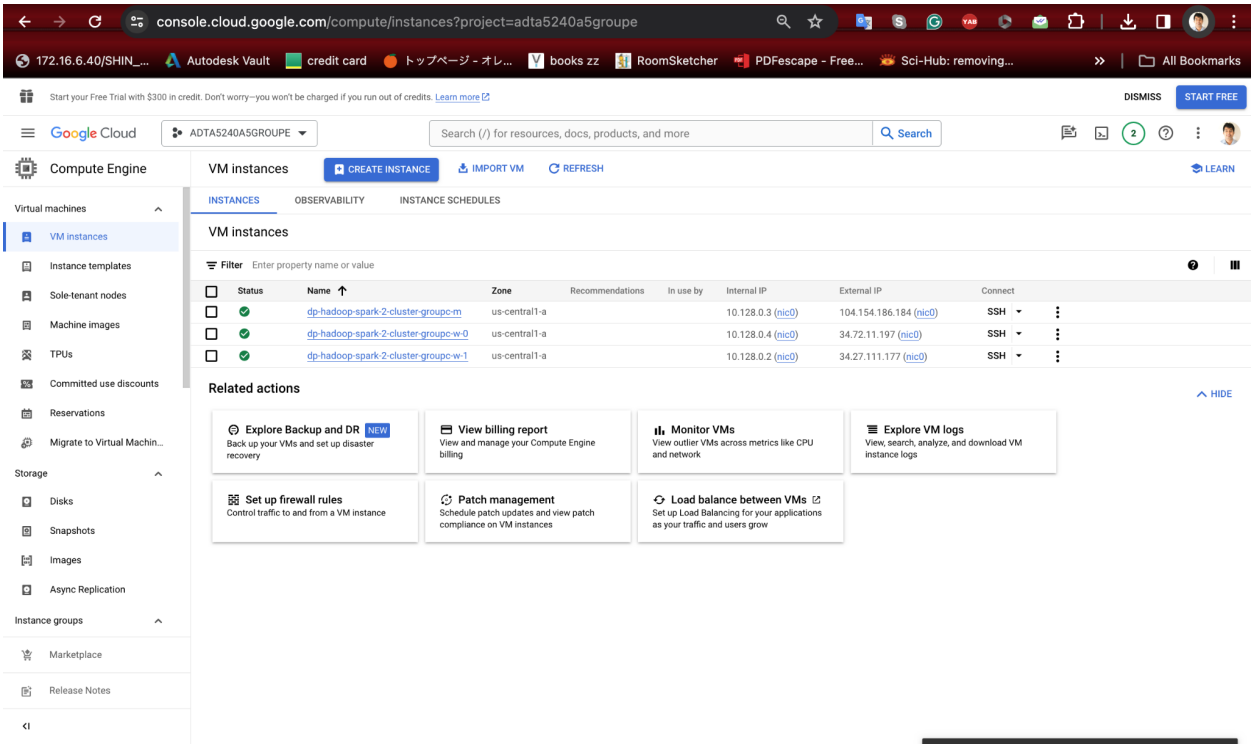
In our Google Cloud Platform (GCP) project, we established a new directory and uploaded the sanitized dataset to a GCP storage bucket.



Dataproc

We created a Hadoop architecture that has been implemented using Google Dataproc.

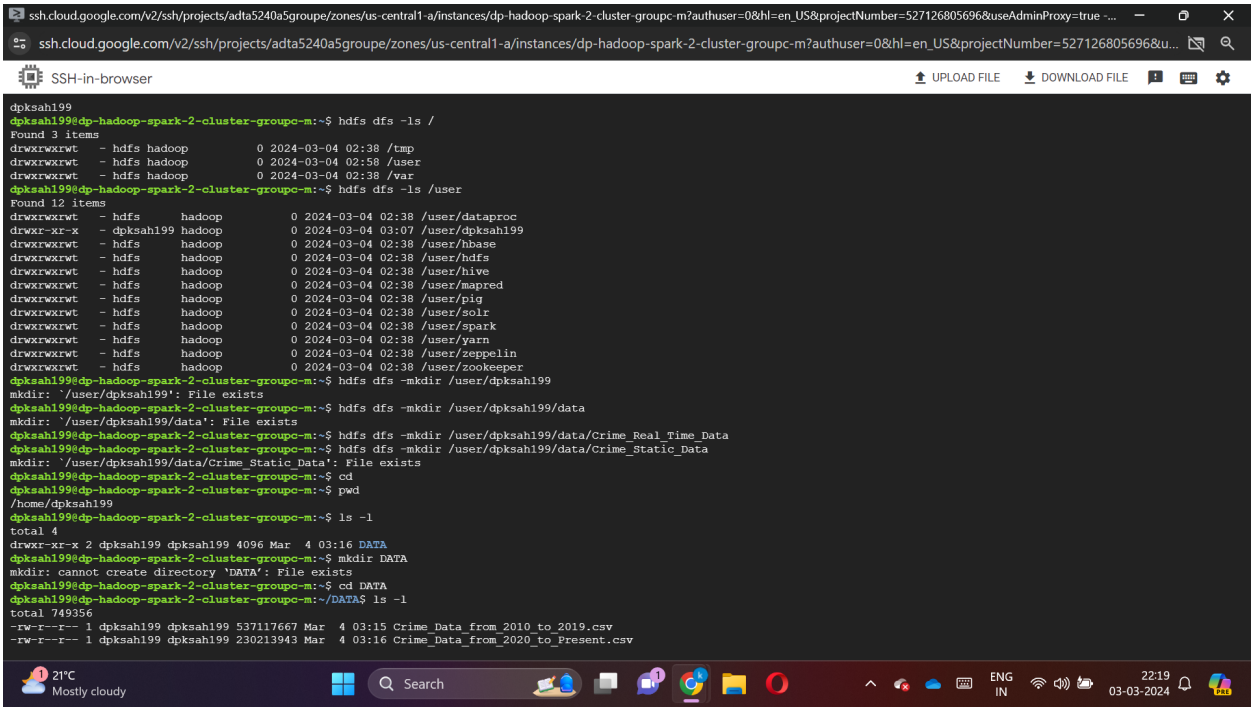


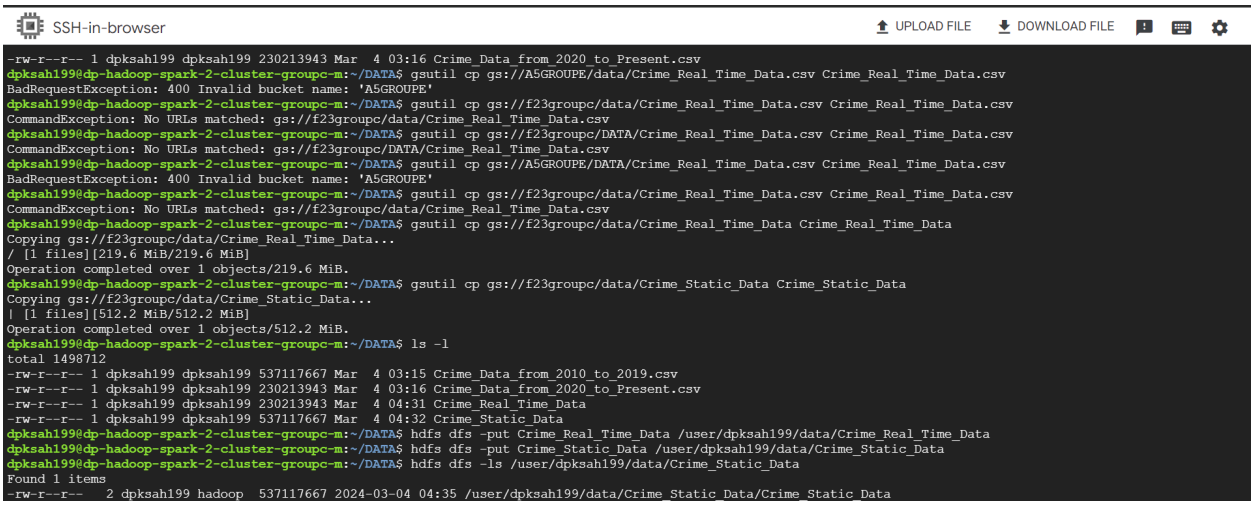


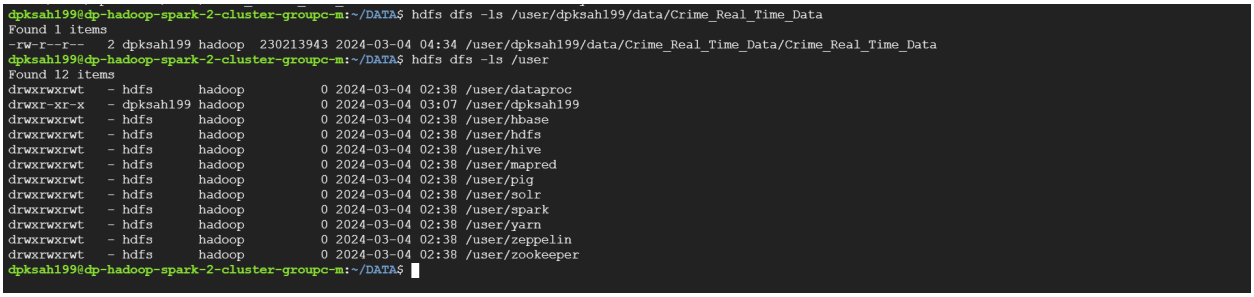
Step 5- Management

HDFS

Here, We have loaded the Data into the Manager Node and accordingly integrated the data into HDFS Echosystem.



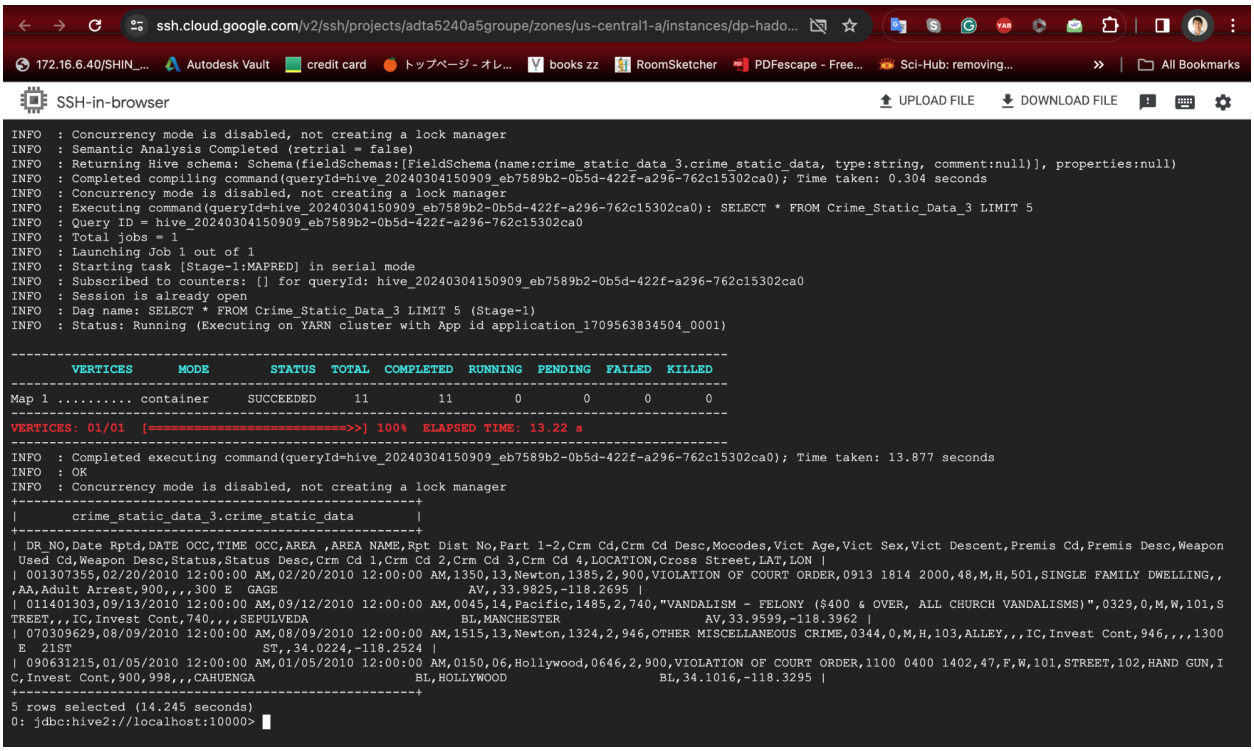




Step – 6 Analysis

HIVE on Static Data

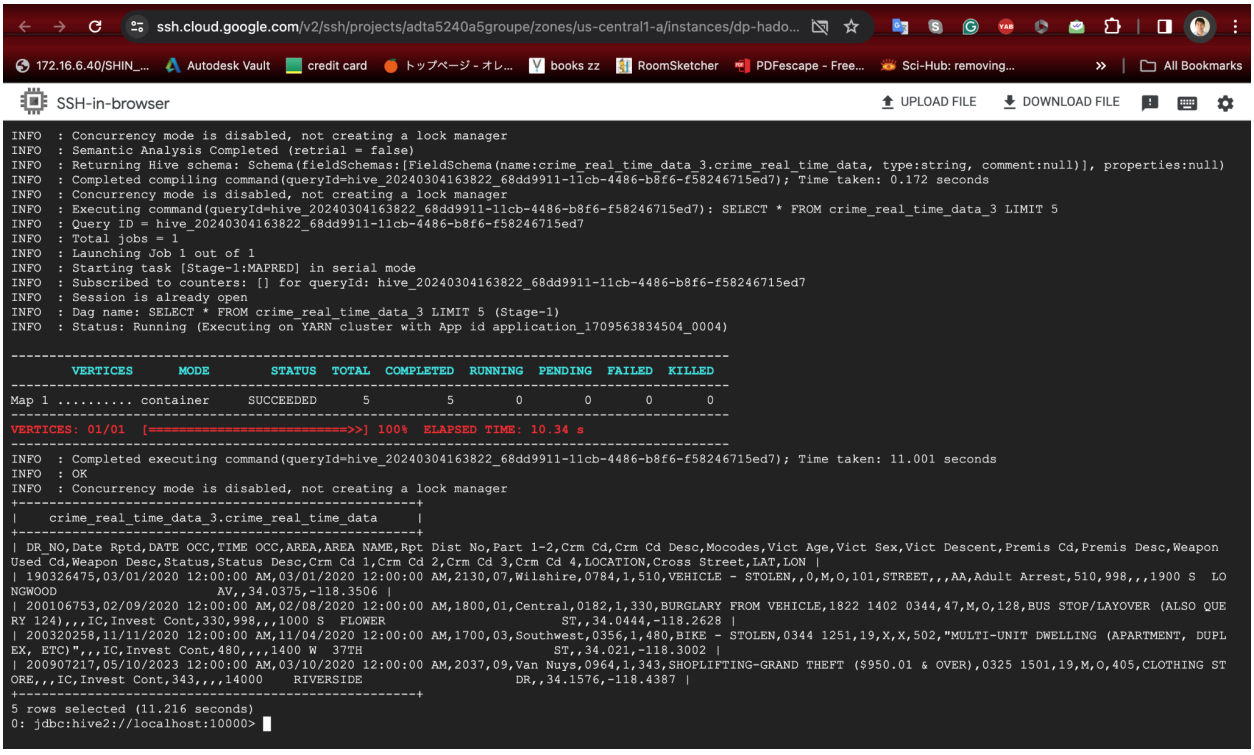
The below image shows the first five rows of static data for the crime.



We have Successfully extracted the data which needed from the dataset to verify it is a proper setup. The given query shows all the data of the table created.

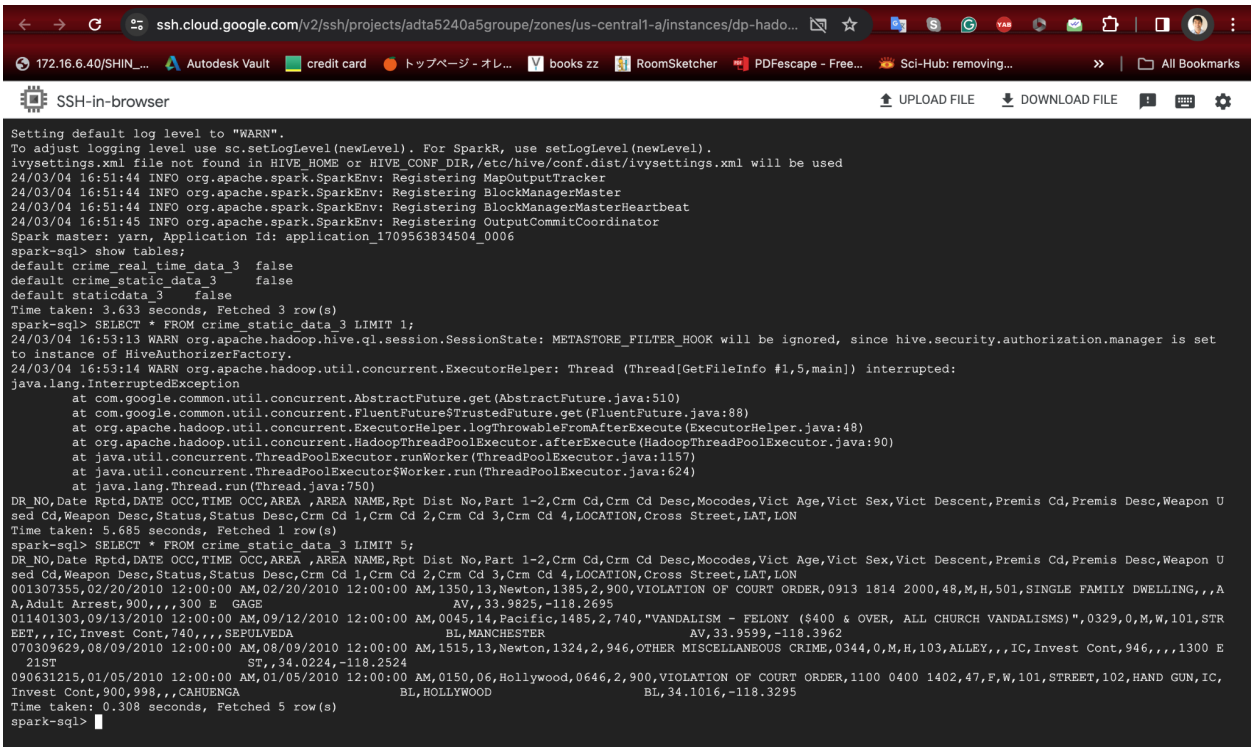
HIVE on Real-Time Data

The below image shows the first five rows of real-time for the crime.



SPARK on Static Data

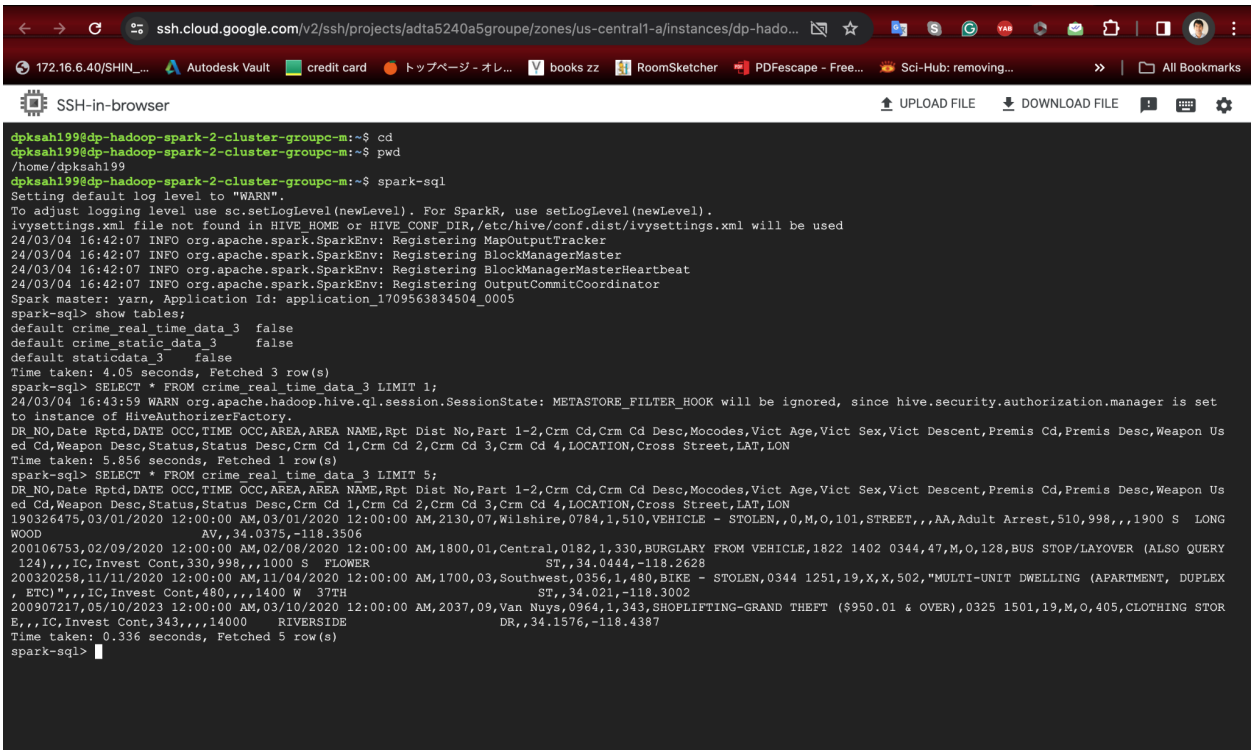
The below image shows the first five rows of static data for the crime.



For the same query in Hive, it took seconds 14.245, and here in Spark, it was seconds 0.308

SPARK on Real-Time Data

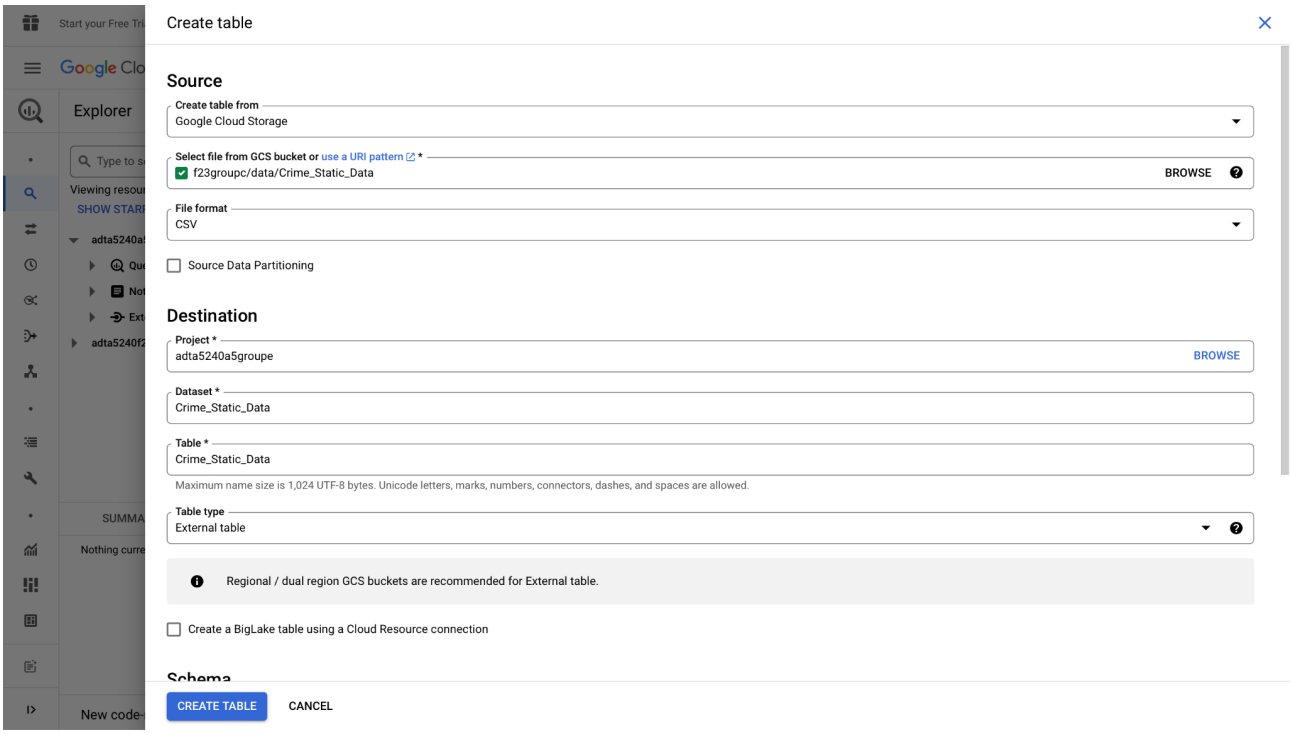
The below image shows the first five rows of real-time for the crime.

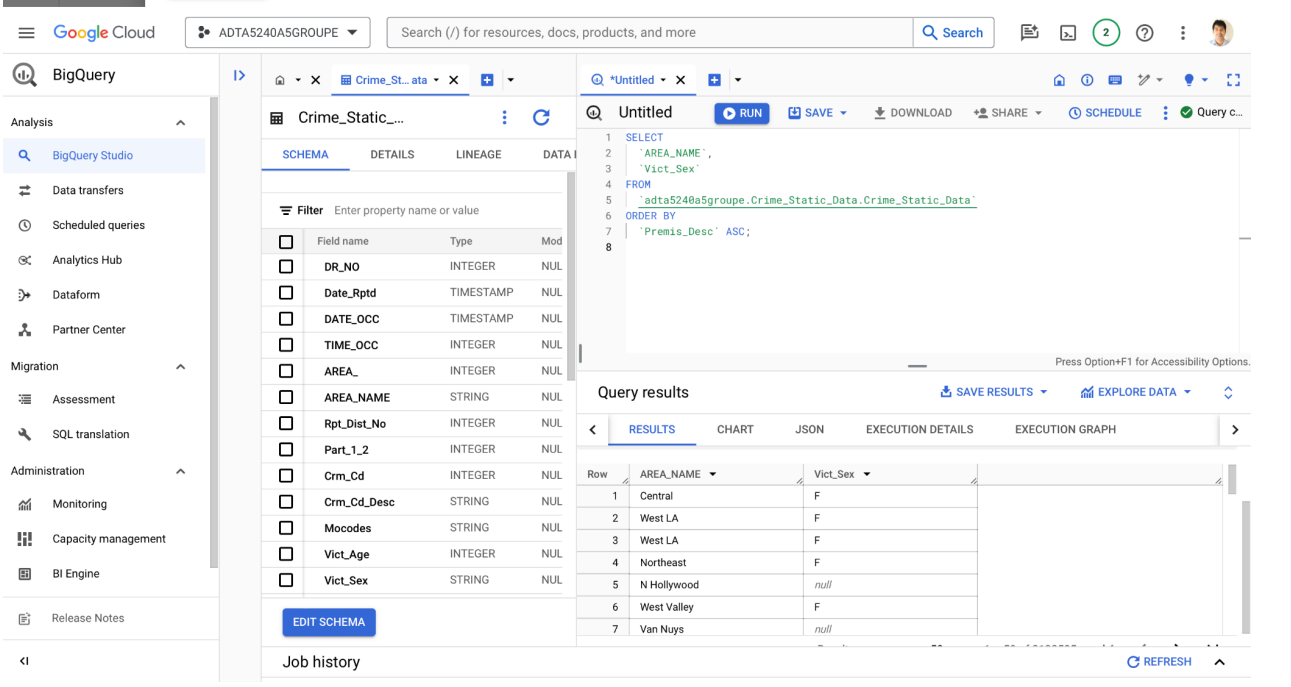


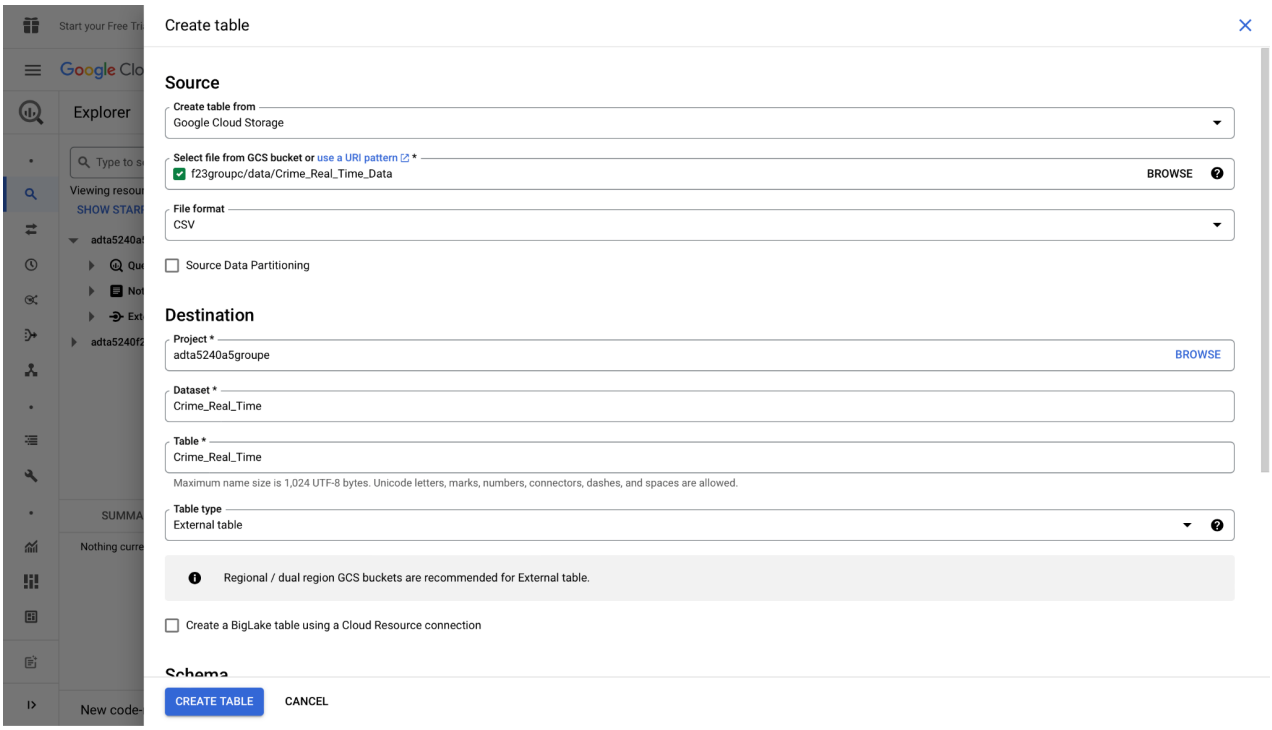
For the same query in Hive, it took seconds 11.216, and here in Spark, it was seconds 0.336.

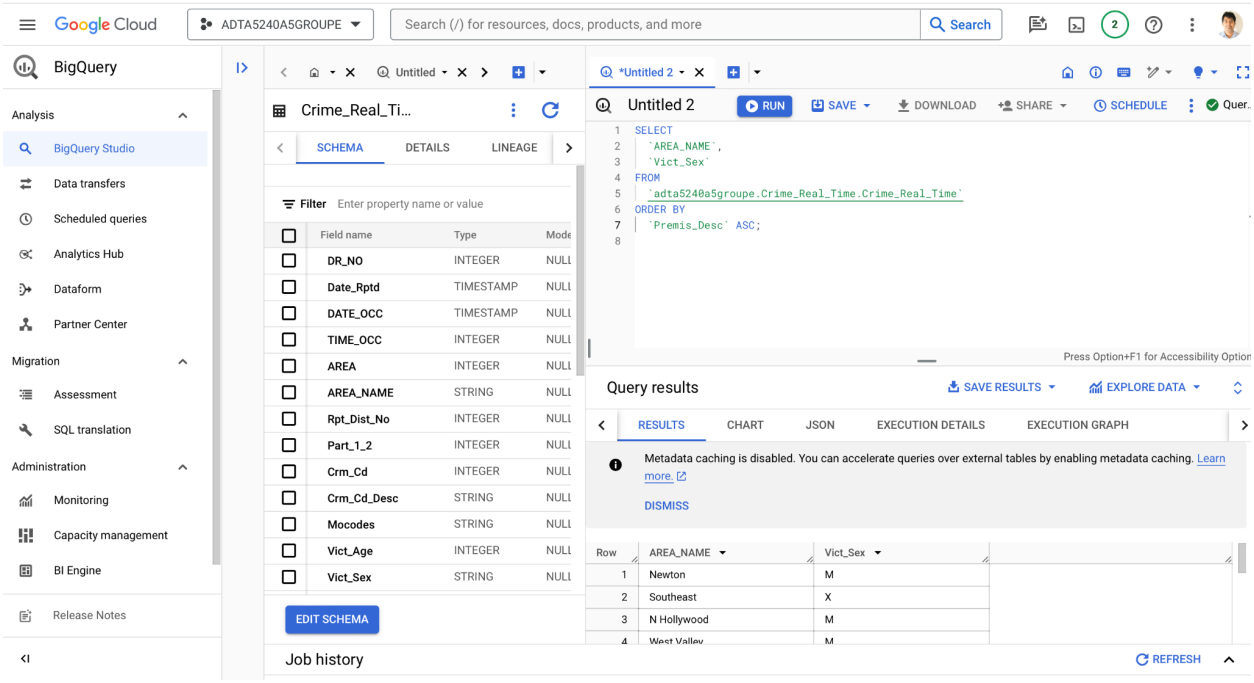
Big Query

Here, we have created the table in Big Query by defining the Dataset and Table by Crime_Static_Data and Crime_Real_Time dataset file that we had previously uploaded to our Google Cloud Storage. Also, we have used the Schema to auto detect the data as we have a large data set and would take so much time to enter the values manually.

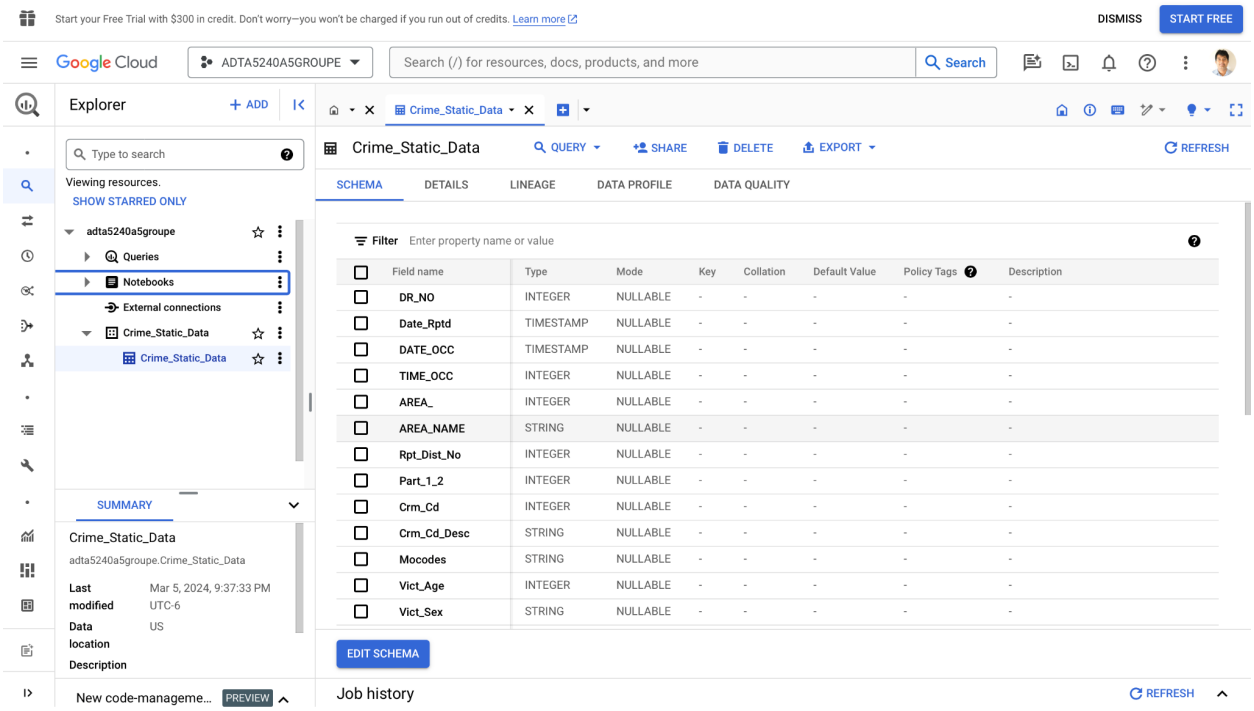

Create table
Source
Create table from
Google Cloud Storage
Select file from GCS bucket or use a URI pattern *
f23groupc/data/Crime_Static_Data
BROWSE
File format
CSV
Source Data Partitioning
Destination
Project *
adta5240a5groupe
BROWSE
Dataset *
Crime_Static_Data
Table *
Crime_Static_Data
Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.
Table type
External table
Regional / dual region GCS buckets are recommended for External table.
Create a BigLake table using a Cloud Resource connection
Schema
CREATE TABLE
CANCEL

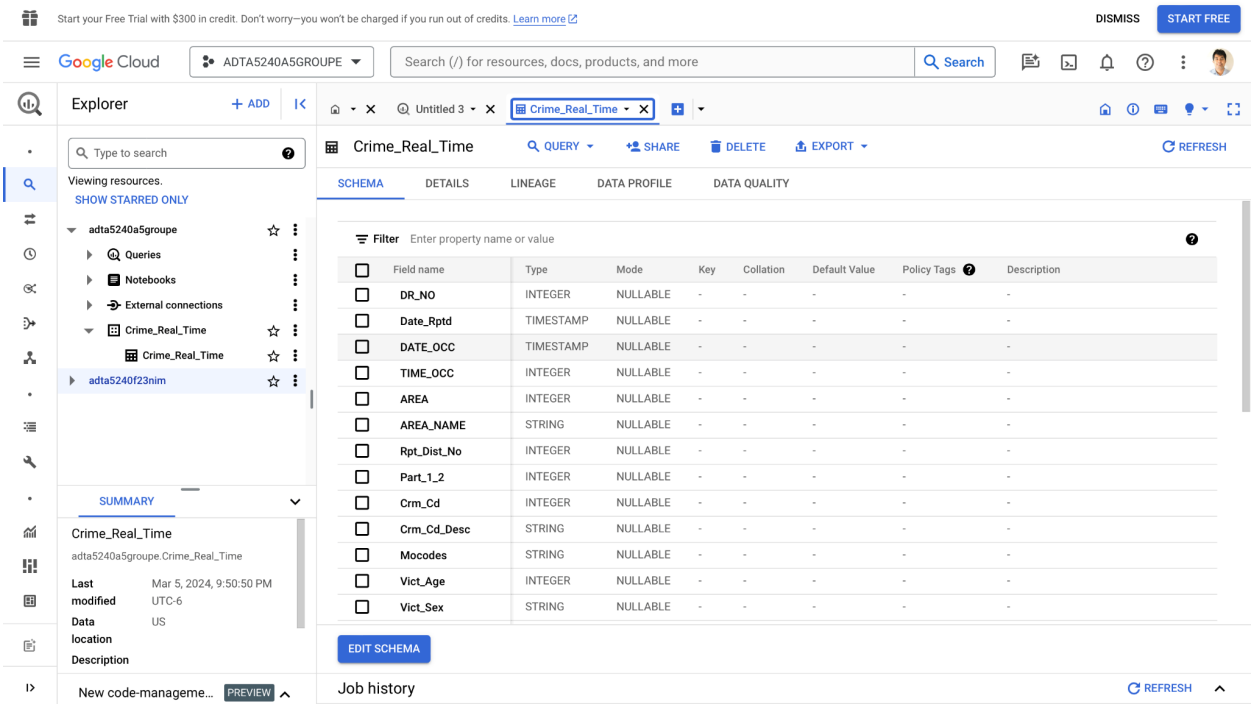

Google Cloud
ADTA5240A5GROUPE
Search (/) for resources, docs, products, and more
Search
BigQuery
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Migration
Assessment
SQL translation
Administration
Monitoring
Capacity management
BI Engine
Release Notes
Crime_Static_...
SCHEMA
DETAILS
LINEAGE
Filter
Enter property name or value
Field name
Type
DR_NO INTEGER
Date_Rptd TIMESTAMP
DATE_OCC TIMESTAMP
TIME_OCC INTEGER
AREA_ INTEGER
AREA_NAME STRING
Rpt_Dist_No INTEGER
Part_1_2 INTEGER
Crm_Cd INTEGER
Crm_Cd_Desc STRING
Mocodes STRING
Vict_Age INTEGER
Vict_Sex STRING
EDIT SCHEMA
Untitled
RUN
SAVE
DOWNLOAD
SHARE
SCHEDULE
SELECT
`AREA_NAME`,
`Vict_Sex`
FROM
`adta5240a5groupe.Crime_Static_Data.Crime_Static_Data`
ORDER BY
`Premis_Desc` ASC;
Press Option+F1 for Accessibility Options.
Query results
SAVE RESULTS
EXPLORE DATA
RESULTS
CHART
JSON
EXECUTION DETAILS
EXECUTION GRAPH
Row
AREA_NAME
Vict_Sex
1 Central F
2 West LA F
3 West LA F
4 Northeast F
5 N Hollywood null
6 West Valley F
7 Van Nuys null
Job history
REFRESH


Create table
Source
Create table from
Google Cloud Storage
Select file from GCS bucket or use a URI pattern *
f23groupc/data/Crime_Real_Time_Data
BROWSE
File format
CSV
Source Data Partitioning
Destination
Project *
adta5240a5groupe
BROWSE
Dataset *
Crime_Real_Time
Table *
Crime_Real_Time
Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.
Table type
External table
Regional / dual region GCS buckets are recommended for External table.
Create a BigLake table using a Cloud Resource connection
Schema
CREATE TABLE
CANCEL

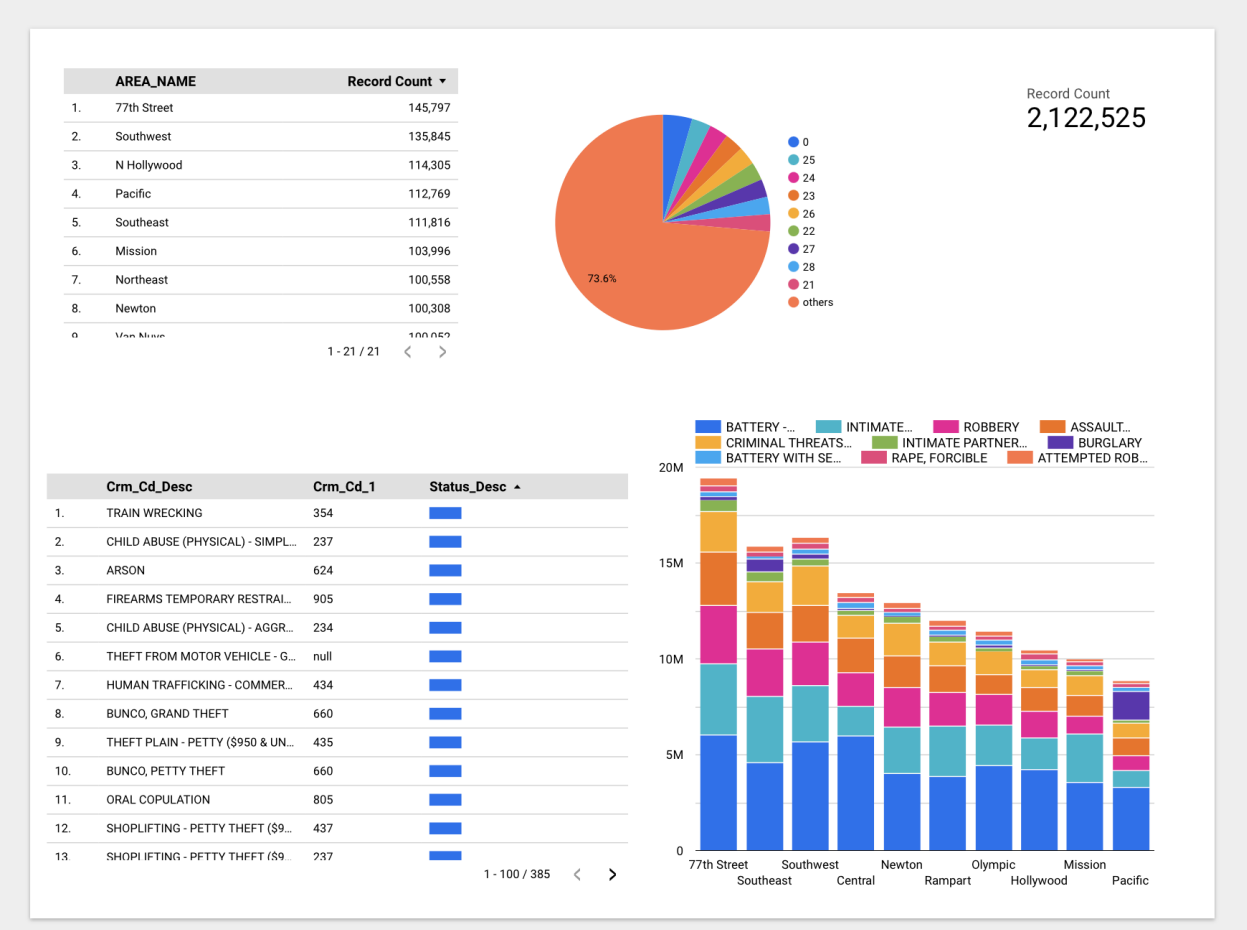

Google Cloud
ADTA5240A5GROUPE
Search (/) for resources, docs, products, and more
Search
BigQuery
Analysis
BigQuery Studio
Data transfers
Scheduled queries
Analytics Hub
Dataform
Partner Center
Migration
Assessment
SQL translation
Administration
Monitoring
Capacity management
BI Engine
Release Notes
Crime_Real_Ti...
SCHEMA
DETAILS
LINEAGE
Filter
Enter property name or value
Field name
Type
Mode
DR_NO INTEGER NULL
Date_Rptd TIMESTAMP NULL
DATE_OCC TIMESTAMP NULL
TIME_OCC INTEGER NULL
AREA INTEGER NULL
AREA_NAME STRING NULL
Rpt_Dist_No INTEGER NULL
Part_1_2 INTEGER NULL
Crm_Cd INTEGER NULL
Crm_Cd_Desc STRING NULL
Mocodes STRING NULL
Vict_Age INTEGER NULL
Vict_Sex STRING NULL
EDIT SCHEMA
Untitled 2
RUN
SAVE
DOWNLOAD
SHARE
SCHEDULE
SELECT
`AREA_NAME`,
`Vict_Sex`
FROM
`adta5240a5groupe.Crime_Real_Time.Crime_Real_Time`
ORDER BY
`Premis_Desc` ASC;
Press Option+F1 for Accessibility Option
Query results
SAVE RESULTS
EXPLORE DATA
RESULTS
CHART
JSON
EXECUTION DETAILS
EXECUTION GRAPH
Metadata caching is disabled. You can accelerate queries over external tables by enabling metadata caching. Learn more.
DISMISS
Row
AREA_NAME
Vict_Sex
1 Newton M
2 Southeast X
3 N Hollywood M
Job history
REFRESH

Schema


Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
ADTA5240A5GROUPE
Search (/) for resources, docs, products, and more
Search
Explorer
+ ADD
Type to search
Viewing resources.
SHOW STARRED ONLY
adta5240a5groupe
Queries
Notebooks
External connections
Crime_Static_Data
Crime_Static_Data
SUMMARY
Crime_Static_Data
adta5240a5groupe.Crime_Static_Data
Last modified
Mar 5, 2024, 9:37:33 PM UTC-6
Data location
US
Description
New code-manageme...
PREVIEW
Crime_Static_Data
QUERY
SHARE
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
LINEAGE
DATA PROFILE
DATA QUALITY
Filter
Enter property name or value
Field name
Type
Mode
Key
Collation
Default Value
Policy Tags
Description
DR_NO INTEGER NULLABLE
Date_Rptd TIMESTAMP NULLABLE
DATE_OCC TIMESTAMP NULLABLE
TIME_OCC INTEGER NULLABLE
AREA_ INTEGER NULLABLE
AREA_NAME STRING NULLABLE
Rpt_Dist_No INTEGER NULLABLE
Part_1_2 INTEGER NULLABLE
Crm_Cd INTEGER NULLABLE
Crm_Cd_Desc STRING NULLABLE
Mocodes STRING NULLABLE
Vict_Age INTEGER NULLABLE
Vict_Sex STRING NULLABLE
EDIT SCHEMA
Job history
REFRESH


Start your Free Trial with $300 in credit. Don't worry—you won't be charged if you run out of credits. Learn more
DISMISS
START FREE
Google Cloud
ADTA5240A5GROUPE
Search (/) for resources, docs, products, and more
Search
Explorer
+ ADD
Type to search
Viewing resources.
SHOW STARRED ONLY
adta5240a5groupe
Queries
Notebooks
External connections
Crime_Real_Time
Crime_Real_Time
adta5240f23nim
SUMMARY
Crime_Real_Time
adta5240a5groupe.Crime_Real_Time
Last modified
Mar 5, 2024, 9:50:50 PM UTC-6
Data location
US
Description
New code-manageme...
PREVIEW
Untitled 3
Crime_Real_Time
Crime_Real_Time
QUERY
SHARE
DELETE
EXPORT
REFRESH
SCHEMA
DETAILS
LINEAGE
DATA PROFILE
DATA QUALITY
Filter
Enter property name or value
Field name
Type
Mode
Key
Collation
Default Value
Policy Tags
Description
DR_NO INTEGER NULLABLE
Date_Rptd TIMESTAMP NULLABLE
DATE_OCC TIMESTAMP NULLABLE
TIME_OCC INTEGER NULLABLE
AREA INTEGER NULLABLE
AREA_NAME STRING NULLABLE
Rpt_Dist_No INTEGER NULLABLE
Part_1_2 INTEGER NULLABLE
Crm_Cd INTEGER NULLABLE
Crm_Cd_Desc STRING NULLABLE
Mocodes STRING NULLABLE
Vict_Age INTEGER NULLABLE
Vict_Sex STRING NULLABLE
EDIT SCHEMA
Job history
REFRESH

Step 7 -Visualization



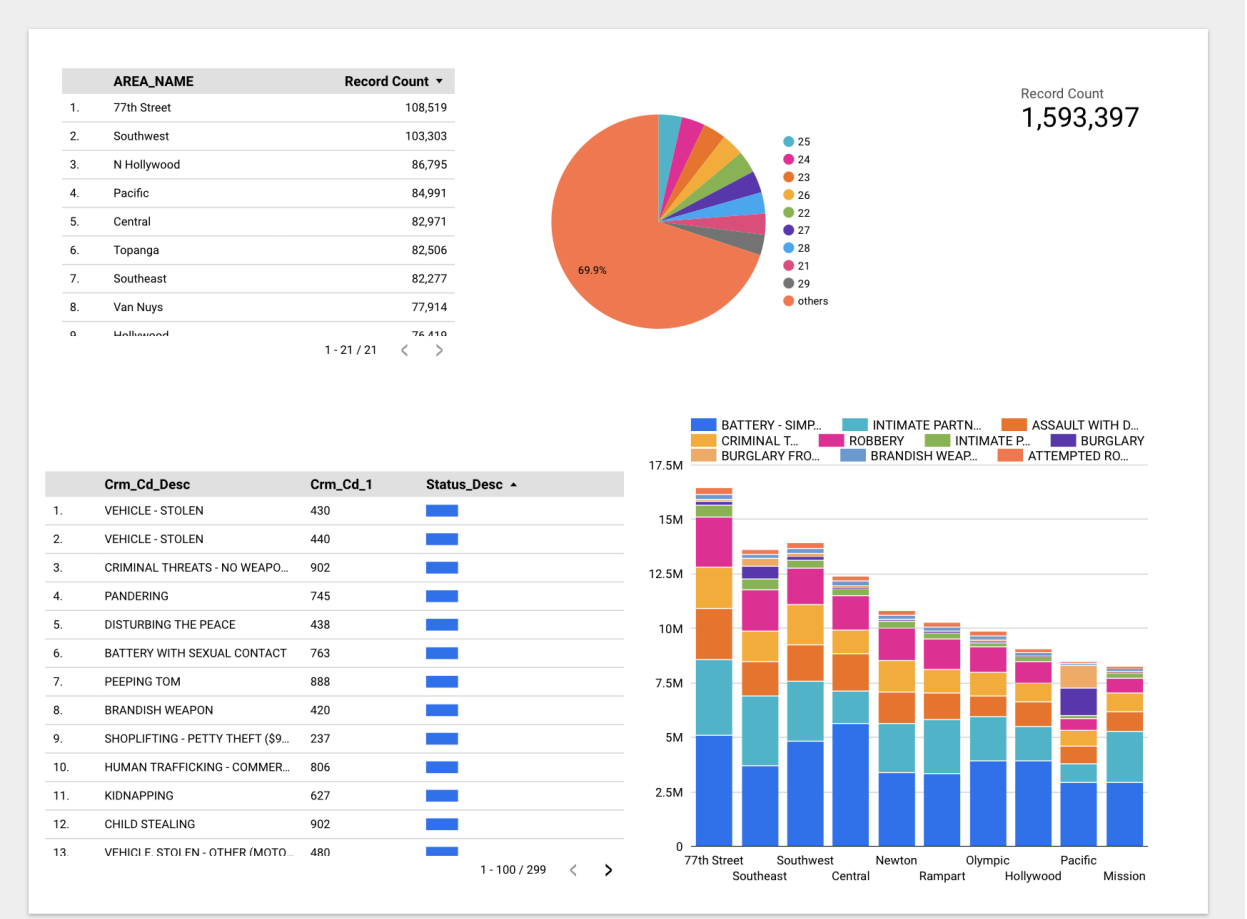
	AREA_NAME	Record Count
1.	77th Street	145,797
2.	Southwest	135,845
3.	N Hollywood	114,305
4.	Pacific	112,769
5.	Southeast	111,816
6.	Mission	103,996
7.	Northeast	100,558
8.	Newton	100,308
9.	Van Nuys	100,052

1 - 21 / 21

Record Count
2,122,525

	Crm_Cd_Desc	Crm_Cd_1	Status_Desc
1.	TRAIN WRECKING	354	
2.	CHILD ABUSE (PHYSICAL) - SIMPL...	237	
3.	ARSON	624	
4.	FIREARMS TEMPORARY RESTRAI...	905	
5.	CHILD ABUSE (PHYSICAL) - AGGR...	234	
6.	THEFT FROM MOTOR VEHICLE - G...	null	
7.	HUMAN TRAFFICKING - COMMER...	434	
8.	BUNCO, GRAND THEFT	660	
9.	THEFT PLAIN - PETTY ($950 & UN...	435	
10.	BUNCO, PETTY THEFT	660	
11.	ORAL COPULATION	805	
12.	SHOPLIFTING - PETTY THEFT ($9...	437	
13.	SHOPLIFTING - PETTY THEFT ($9...	237	

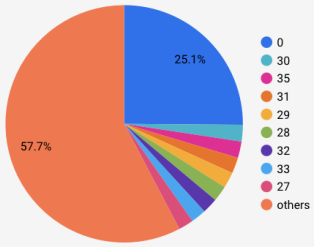
1 - 100 / 385



Real time data

	AREA_NAME	Record Count ▾
1.	Central	61,416
2.	77th Street	56,381
3.	Pacific	52,537
4.	Southwest	50,654
5.	Hollywood	47,508
6.	Southeast	45,503
7.	Olympic	45,486
8.	N Hollywood	45,124
9.	Newton	45,032
10.	Wilshire	43,191

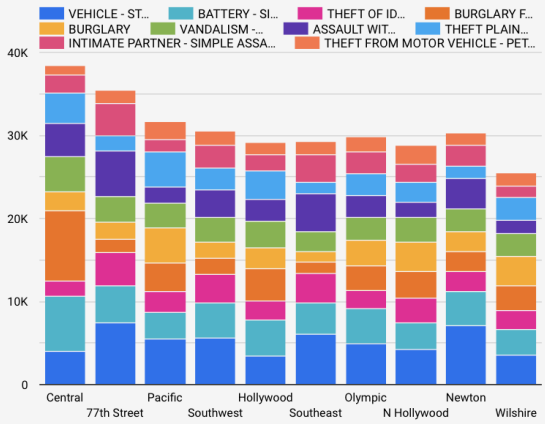
1 - 21 / 21

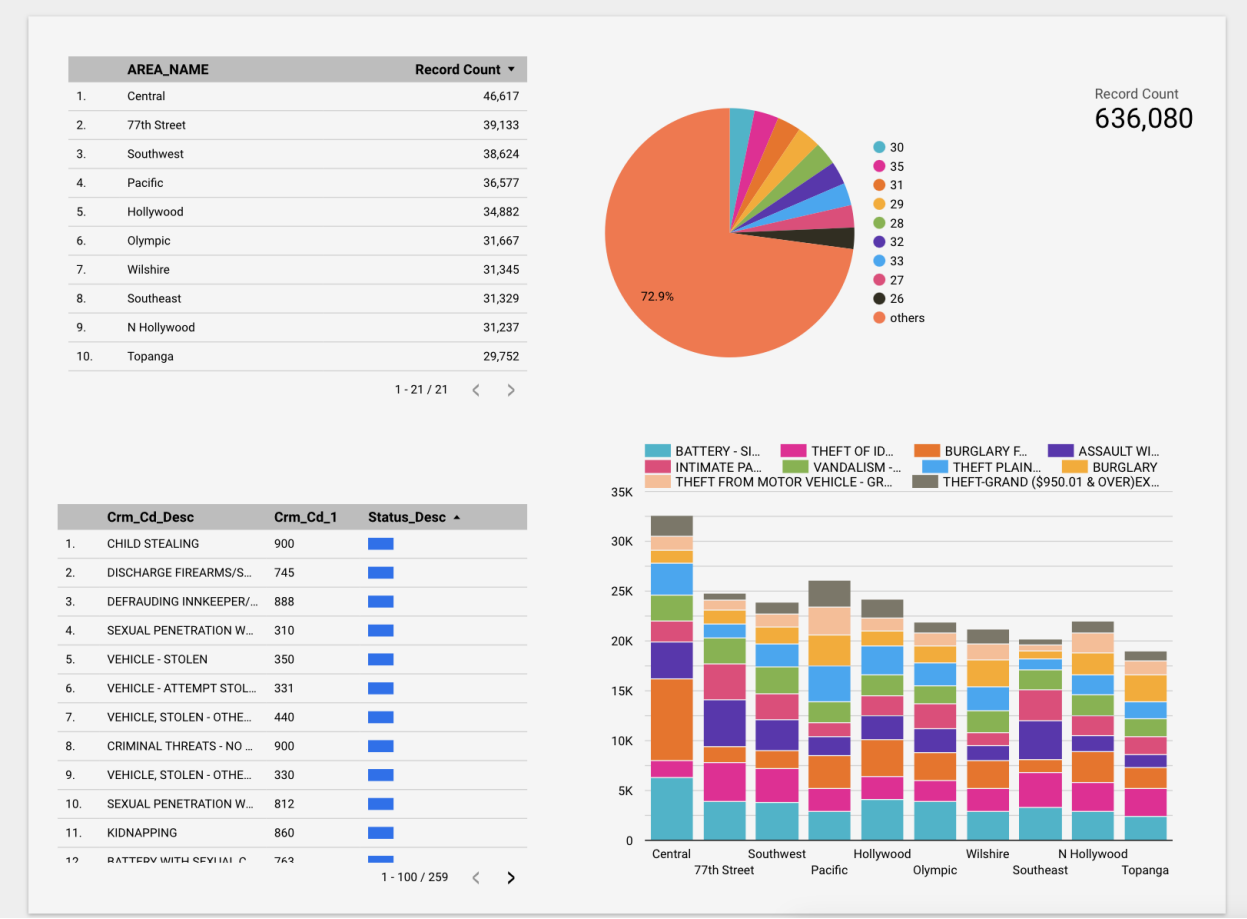


Record Count
901,357

	Crm_Cd_Desc	Crm_Cd_1	Status_Desc ▲
1.	VEHICLE, STOLEN - OTHE...	330	
2.	TRESPASSING	null	
3.	THEFT, COIN MACHINE - P...	474	
4.	VEHICLE, STOLEN - OTHE...	331	
5.	SEXUAL PENETRATION W...	434	
6.	VEHICLE - ATTEMPT STOL...	420	
7.	ORAL COPULATION	760	
8.	PICKPOCKET, ATTEMPT	452	
9.	ORAL COPULATION	230	
10.	INDECENT EXPOSURE	762	
11.	CHILD PORNOGRAPHY	627	
12.	HUMAN TRAFFICKING - IN...	812	

1 - 100 / 309





Step 8 - Interpretation

By this Project we came to know that the utilization of data science fosters proactive rather than reactive measures in addressing public security concerns. Through continuous analysis and adaptation, data-driven strategies can lead to sustainable improvements in public safety outcomes.

----------Thanks for the time and consideration ☺----------