

Crime and Safety in Residential 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 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.

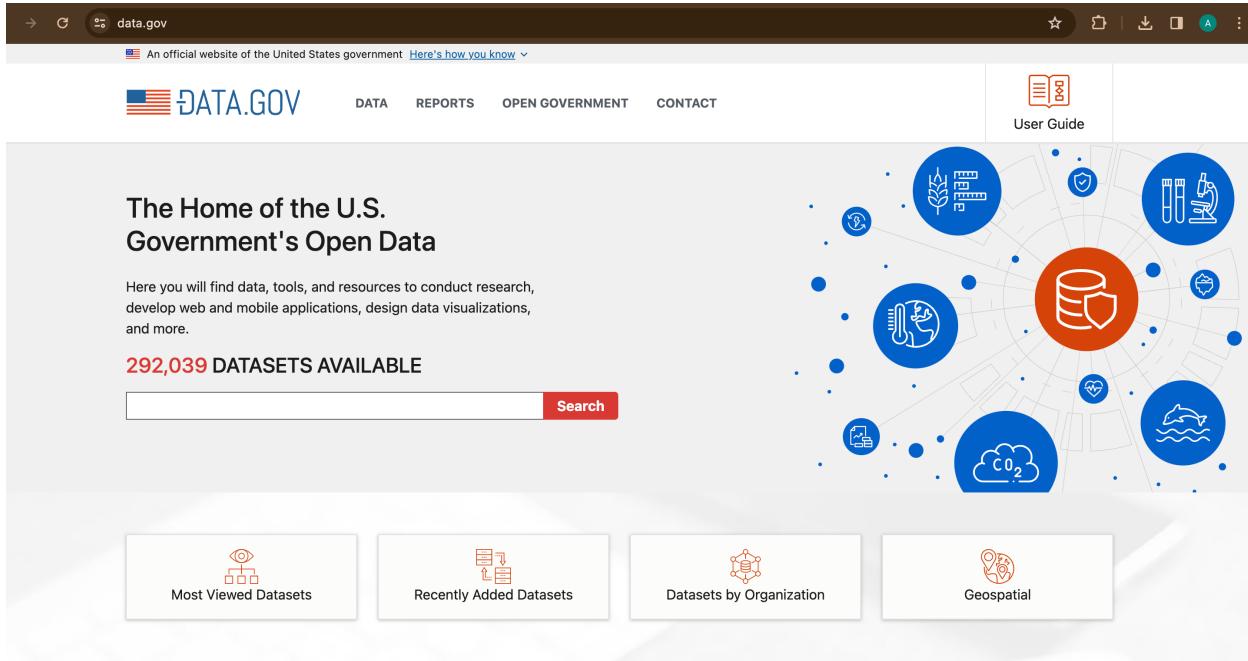
The Data lifecycle includes Generation where raw data is generated from repositories such as data.gov, Collection, Processing of Static and Streaming Data.

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.

An official website of the United States government [Here's how you know](#)

 DATA.GOV DATA REPORTS OPEN GOVERNMENT CONTACT  User Guide

DATA CATALOG  / Datasets Organizations 

 / City of Los Angeles / data.lacity.org

City of Los Angeles
There is no description for this organization

Topics
Local Government

Publisher
data.lacity.org

Contact

LAPD OpenData

Crime Data from 2010 to 2019
Metadata Updated: February 24, 2024

This dataset reflects incidents of crime in the City of Los Angeles from 2010 - 2019. This data is transcribed from original crime reports that are typed on paper and therefore there may be some inaccuracies within the data. Some location fields with missing data are noted as (0°, 0°). Address fields are only provided to the nearest hundred block in order to maintain privacy. This data is as accurate as the data in the database. Please note questions or concerns in the comments.

Access & Use Information

Public: This dataset is intended for public access and use.
Non-Federal: This dataset is covered by different Terms of Use than Data.gov.
License: See this page for license information.

Downloads & Resources

An official website of the United States government [Here's how you know](#)

 DATA REPORTS OPEN GOVERNMENT CONTACT  User Guide

DATA CATALOG  / Datasets Organizations 

 / City of Los Angeles / data.lacity.org

City of Los Angeles
There is no description for this organization

Topics
Local Government

Publisher
data.lacity.org

Contact

LAPD OpenData

Crime Data from 2020 to Present
Metadata Updated: March 8, 2024

Starting on March 7th, 2024, the Los Angeles Police Department (LAPD) will adopt a new Records Management System for reporting crimes and arrests. This new system is being implemented to comply with the FBI's mandate to collect NIBRS-only data (NIBRS — FBI). During this transition, users will temporarily see only incidents reported in the retiring system. However, the LAPD is actively working on generating new NIBRS datasets to ensure a smoother and more efficient reporting system.

**Update 1/18/2024 - LAPD is facing issues with posting the Crime data, but we are taking immediate action to resolve the problem. We understand the importance of providing reliable and up-to-date information and are committed to delivering it.

As we work through the issues, we have temporarily reduced our updates from weekly to bi-weekly to ensure that we provide accurate information. Our team is actively working to identify and resolve these issues promptly.

We apologize for any inconvenience this may cause and appreciate your understanding. Rest assured, we are doing everything we can to fix the problem and get back to providing weekly updates as soon as possible. **

This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020. This data is

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.

File Edit View Insert Runtime Tools Help Last edited on 6 March
Comment Share Connect | Colab

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]

```

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

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

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

↳ 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

```

1 #Verifying for null valued rows
2 df.isna().any()

```

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

```
[ ] 1 # checking the statistical data for each column
2 df.describe()
3 df.count()
```

Column	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA NAME	Crm Cd Desc
	1925662	1925662	1925662	1925662	1925662	1925662

streaming ☆

File Edit View Insert Runtime Tools Help Last edited on 6 March

+ Code + Text

Comment

Connect

```
[ ] Date Rptd False
DATE OCC False
TIME OCC False
AREA NAME False
{x} 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 781953
Date Rptd 781953
DATE OCC 781953
TIME OCC 781953
AREA NAME 781953
Crm Cd Desc 781953
Vict Age 781953
Vict Sex 781953
Vict Descent 781953
LAT 781953
LON 781953
dtype: int64
```

```
[ ] 1 # removing the values below one for the "Vict Age" column
2 df.drop(df[df['Vict Age'] < 1].index, axis=0, inplace=True)
3 df.count()
```

```
<> ipython-input-8-046a0b1081db>: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.drop(df[df['Vict Age'] < 1].index, axis=0, inplace=True)
DR_NO 674731
Date Rptd 674731
```

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
```

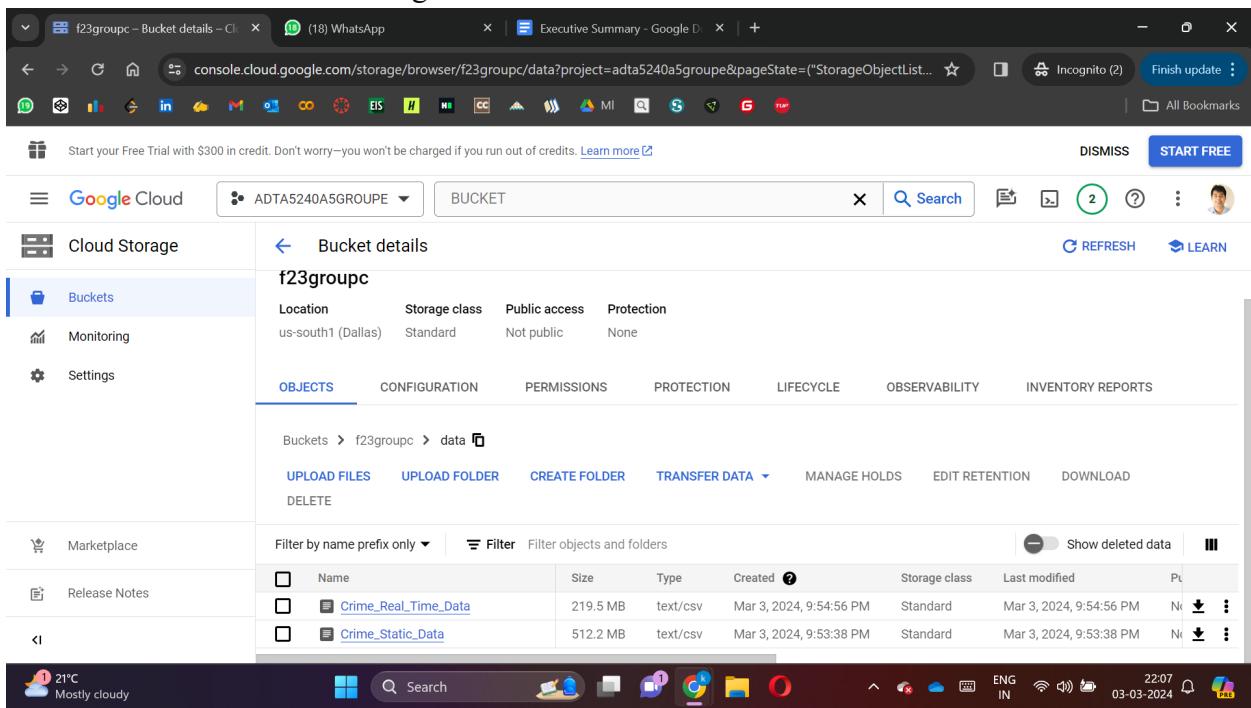
Step 4- Storage.

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.



The screenshot shows the Google Cloud Storage console for the 'f23groupc' bucket. The bucket details page is displayed, showing the following information:

Location	Storage class	Public access	Protection
us-south1 (Dallas)	Standard	Not public	None

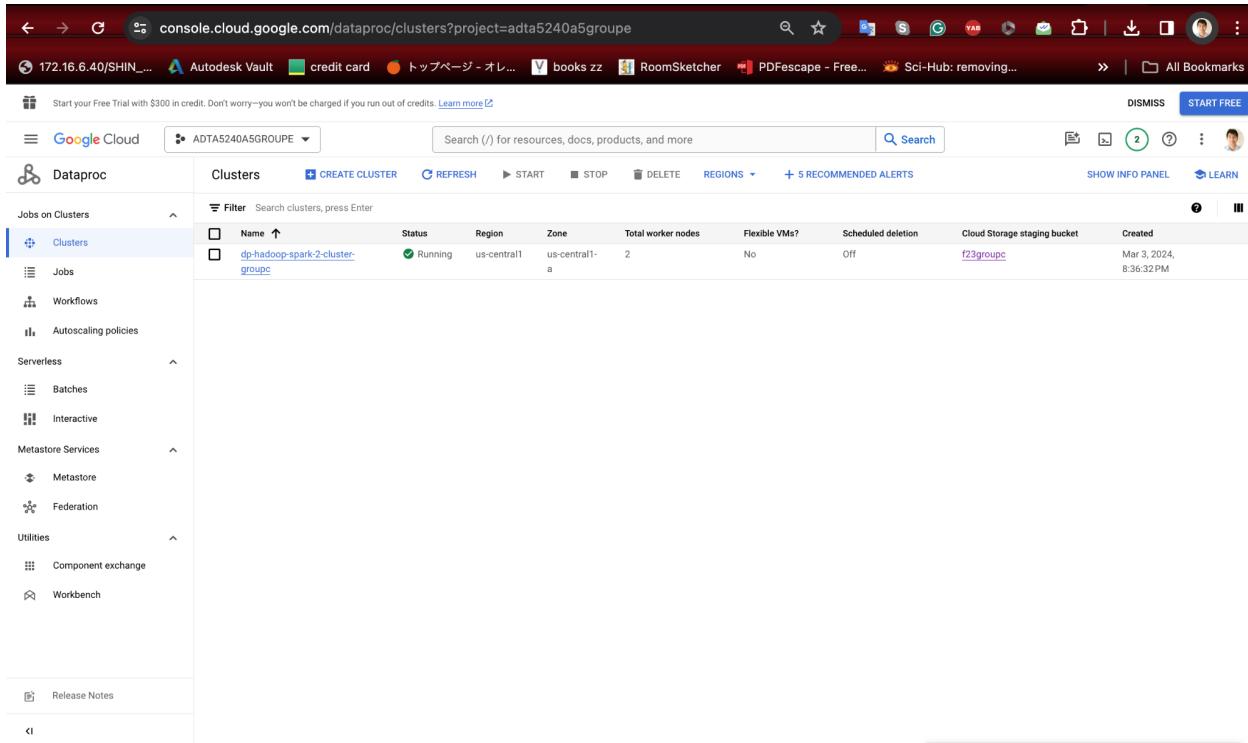
Below the bucket details, the 'OBJECTS' tab is selected, showing the contents of the 'data' folder:

Name	Size	Type	Created	Storage class	Last modified	Actions
Crime_Real_Time_Data	219.5 MB	text/csv	Mar 3, 2024, 9:54:56 PM	Standard	Mar 3, 2024, 9:54:56 PM	⋮
Crime_Static_Data	512.2 MB	text/csv	Mar 3, 2024, 9:53:38 PM	Standard	Mar 3, 2024, 9:53:38 PM	⋮

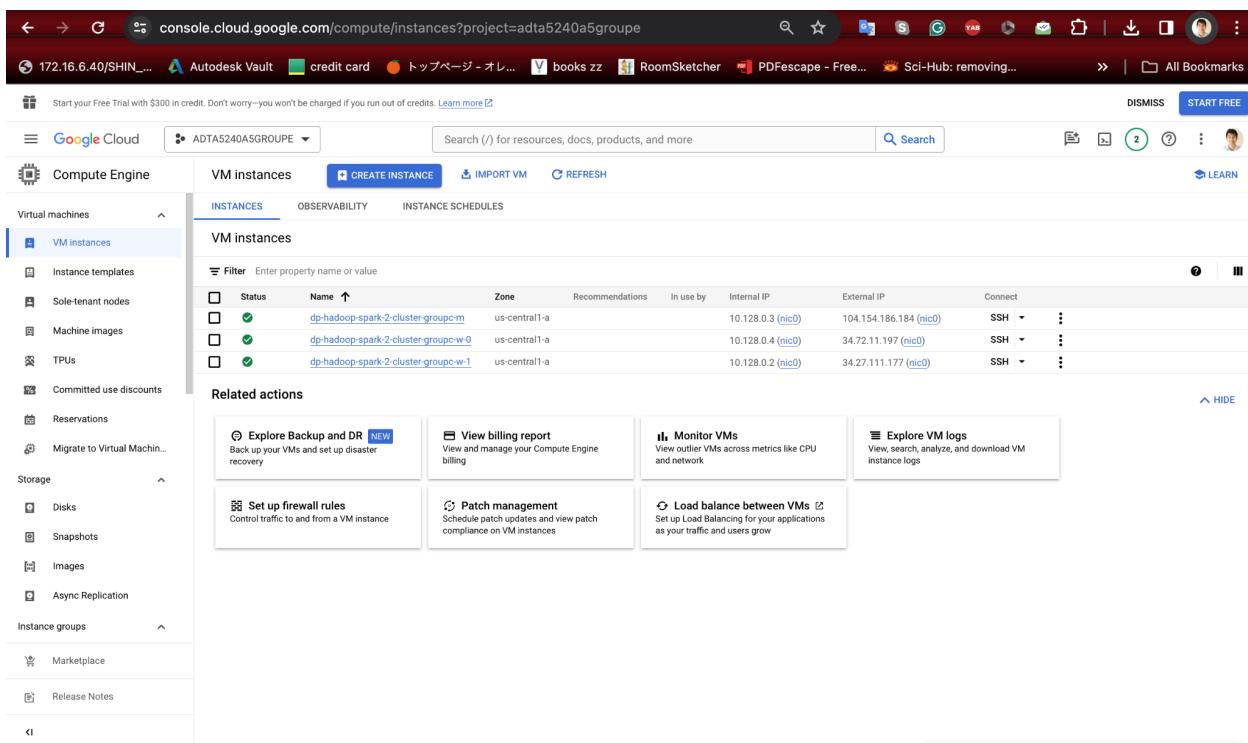
The console also includes a navigation bar with links for 'BUCKET', 'OBJECTS', 'CONFIGURATION', 'PERMISSIONS', 'PROTECTION', 'LIFECYCLE', 'OBSERVABILITY', and 'INVENTORY REPORTS'. The bottom of the screen shows the Windows taskbar with various pinned icons and system status indicators.

Dataproc

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



The screenshot shows the Google Cloud Dataproc interface. On the left, a sidebar lists various services: Jobs on Clusters (Clusters, Jobs, Workflows, Autoscaling policies), Serverless (Batches, Interactive), Metastore Services (Metastore, Federation), Utilities (Component exchange, Workbench), and Release Notes. The main content area shows a table of clusters. One cluster is selected: 'dp-hadoop-spark-2-cluster-groupc', which is 'Running' in the 'us-central1' region with 2 worker nodes. The table includes columns for Name, Status, Region, Zone, Total worker nodes, Flexible VMs?, Scheduled deletion, Cloud Storage staging bucket, and Created date (Mar 3, 2024, 8:36:32 PM). The top navigation bar shows the URL 'console.cloud.google.com/dataproc/clusters?project=adta5240a5groupe' and the IP '172.16.6.40/SHIN...'. The top right has 'DISMISS' and 'START FREE' buttons.



The screenshot shows the Google Cloud Compute Engine interface. On the left, a sidebar lists Virtual machines (VM instances, Instance templates, Sole-tenant nodes, Machine images, TPUs, Committed use discounts, Reservations, Migrate to Virtual Machine...), Storage (Disks, Snapshots, Images, Async Replication), and Instance groups (Marketplace, Release Notes). The main content area shows a table of VM instances. Three instances are listed: 'dp-hadoop-spark-2-cluster-groupc-m' (Status: Running, Zone: us-central1-a, Internal IP: 10.128.0.3, External IP: 104.154.186.184), 'dp-hadoop-spark-2-cluster-groupc-w-0' (Status: Running, Zone: us-central1-a, Internal IP: 10.128.0.4, External IP: 34.72.11.197), and 'dp-hadoop-spark-2-cluster-groupc-w-1' (Status: Running, Zone: us-central1-a, Internal IP: 10.128.0.2, External IP: 34.27.111.177). The table includes columns for Status, Name, Zone, Recommendations, In use by, Internal IP, External IP, and Connect. Below the table, there are 'Related actions' cards: 'Explore Backup and DR' (NEW), 'View billing report', 'Monitor VMs', 'Explore VM logs', 'Set up firewall rules', 'Patch management', and 'Load balance between VMs'. The top navigation bar shows the URL 'console.cloud.google.com/compute/instances?project=adta5240a5groupe' and the IP '172.16.6.40/SHIN...'. The top right has 'DISMISS' and 'START FREE' buttons.

Step 5- Management

HDFS

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

```

SSH-in-browser
UPLOAD FILE DOWNLOAD FILE
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://A5GROUPE/data/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
BadRequestException: 400 Invalid bucket name: 'A5GROUPE'
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
CommandException: No URLs matched: gs://f23groupc/data/Crime_Real_Time_Data.csv
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/DATA/Crime_Real_Time_Data.csv Crime_Real_Time_Data.csv
CommandException: No URLs matched: gs://f23groupc/DATA/Crime_Real_Time_Data.csv
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Real_Time_Data Crime_Real_Time_Data
Copying gs://f23groupc/data/Crime_Real_Time_Data...
/ [1 files] [219.6 MiB/219.6 MiB]
Operation completed over 1 objects/219.6 MiB.
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ gsutil cp gs://f23groupc/data/Crime_Static_Data Crime_Static_Data
Copying gs://f23groupc/data/Crime_Static_Data...
/ [1 files] [512.2 MiB/512.2 MiB]
Operation completed over 1 objects/512.2 MiB.
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ ls -l
total 1498712
-rw-r--r-- 1 dpsah199 dpsah199 537117667 Mar  4 03:15 Crime_Data_from_2010_to_2019.csv
-rw-r--r-- 1 dpsah199 dpsah199 230213943 Mar  4 03:16 Crime_Data_from_2020_to_Present.csv
-rw-r--r-- 1 dpsah199 dpsah199 230213943 Mar  4 04:31 Crime_Real_Time_Data
-rw-r--r-- 1 dpsah199 dpsah199 537117667 Mar  4 04:32 Crime_Static_Data
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -put Crime_Real_Time_Data /user/dpsah199/data/Crime_Real_Time_Data
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -put Crime_Static_Data /user/dpsah199/data/Crime_Static_Data
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -ls /user/dpsah199/data/Crime_Static_Data
Found 1 items
-rw-r--r-- 2 dpsah199 hadoop 537117667 2024-03-04 04:35 /user/dpsah199/data/Crime_Static_Data/Crime_Static_Data

```

```

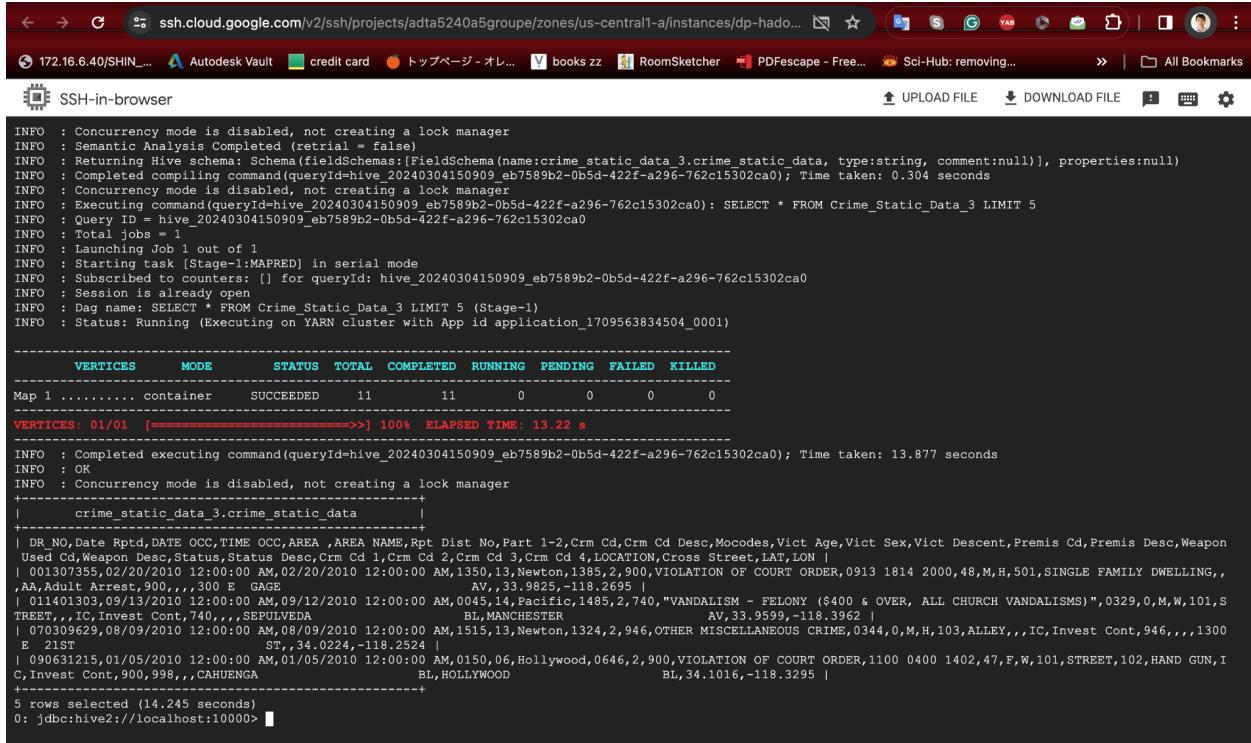
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -ls /user/dpsah199/data/Crime_Real_Time_Data
Found 1 items
-rw-r--r-- 2 dpsah199 hadoop 230213943 2024-03-04 04:34 /user/dpsah199/data/Crime_Real_Time_Data/Crime_Real_Time_Data
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ hdfs dfs -ls /user
Found 12 items
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/dataproc
drwxr-xr-x - dpsah199 hadoop      0 2024-03-04 03:07 /user/dpsah199
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/hbase
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/hdfs
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/hive
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/mapred
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/pig
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/solr
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/spark
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/yarn
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/zeppelin
drwxrwxrwt - hdfs  hadoop      0 2024-03-04 02:38 /user/zookeeper
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~/DATA$ 

```

Step – 6 Analysis

HIVE on Static Data

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



```

INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldschemas:[FieldSchema(name:crime_static_data_3.crime_static_data, type:string, comment:null)], properties:null)
INFO : Completed compiling command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0); Time taken: 0.304 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0): SELECT * FROM Crime_Static_Data_3 LIMIT 5
INFO : Query ID = hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0
INFO : Total Jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0
INFO : Session is already open
INFO : Dag name: SELECT * FROM Crime_Static_Data_3 LIMIT 5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1709563834504_0001)

-----  

  VERTICES      MODE      STATUS TOTAL COMPLETED RUNNING PENDING FAILED KILLED  

Map 1 ..... container      SUCCEEDED      11      11      0      0      0      0  

-----  

  VERTICES: 01/01  [=====>>>] 100% ELAPSED TIME: 13.22 s  

-----  

INFO : Completed executing command(queryId=hive_20240304150909_eb7589b2-0b5d-422f-a296-762c15302ca0); Time taken: 13.877 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----  

| crime_static_data_3.crime_static_data |  

+-----  

| DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA ,AREA NAME,Rpt Dist No,Part 1=2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon  

Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON |  

| 001307355,02/20/2010 12:00:00 AM,02/20/2010 12:00:00 AM,1350,13,Newton,1385,2,900,VIOLATION OF COURT ORDER,0913 1814 2000,48,M,H,501,SINGLE FAMILY DWELLING,,  

,AA,Adult Arrest,900,,,300 E GAGE | AV,33,9825,-118,2695 |  

| 011401303,09/13/2010 12:00:00 AM,09/12/2010 12:00:00 AM,0045,14,Pacific,1485,2,740,"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)",0329,0,M,W,101,S  

TREET,,IC,Invest Cont,740,,,SEFULVEDA BB,MANCHESTER AV,33,9599,-118,3962 |  

| 070309629,08/09/2010 12:00:00 AM,08/09/2010 12:00:00 AM,1515,13,Newton,1324,2,946,OTHER MISCELLANEOUS CRIME,0344,0,M,H,103,ALLEY,,IC,Invest Cont,946,,,1300  

E 21ST ST,,34,0224,-118,2524 |  

| 090631215,01/05/2010 12:00:00 AM,01/05/2010 12:00:00 AM,0150,06,Hollywood,0646,2,900,VIOLATION OF COURT ORDER,1100 0400 1402,47,F,W,101,STREET,102,HAND GUN,I  

C,Invest Cont,900,998,,,CAHUENGA BL,HOLLYWOOD BL,34,1016,-118,3295 |  

+-----  

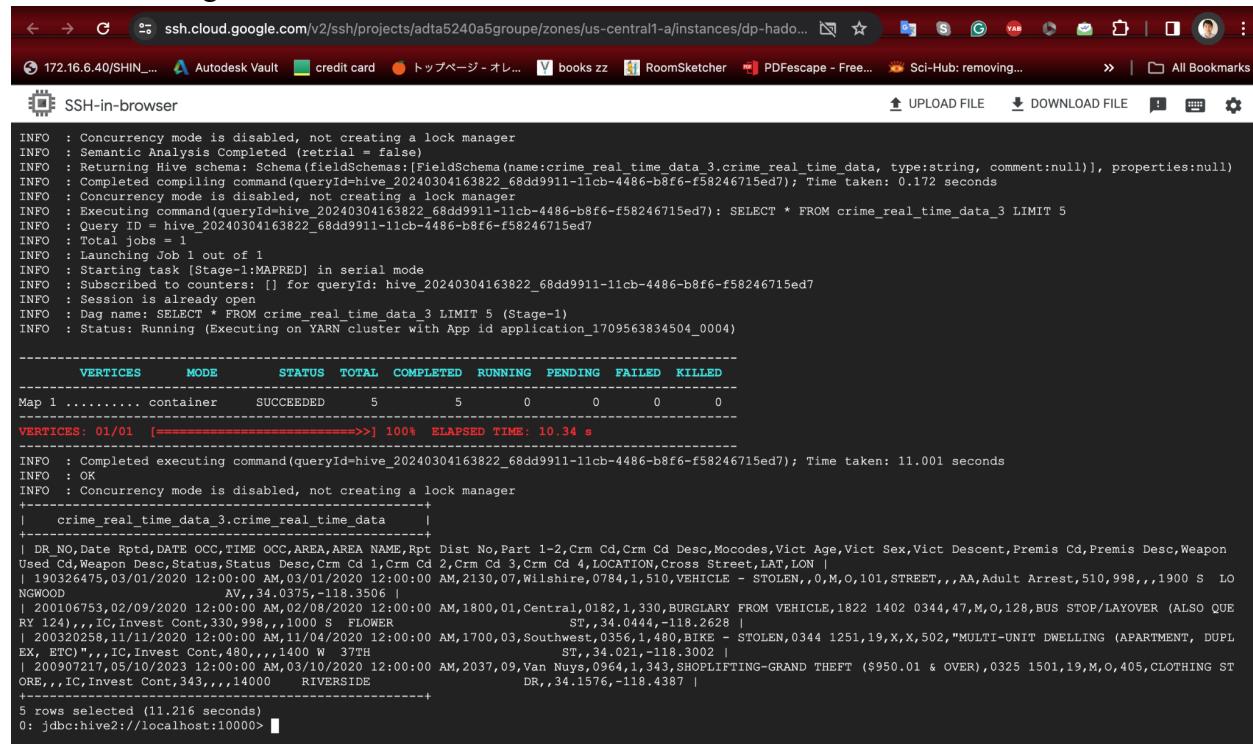
5 rows selected (14.245 seconds)
0: jdbc:hive2://localhost:10000> 

```

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.



ssh.cloud.google.com/v2/ssh/projects/adta5240a5groupe/zones/us-central1-a/instances/dp-hado... 172.16.6.40/SHIN... Autodesk Vault credit card トップページ - オレ... books zz RoomSketcher PDFescape - Free... Sci-Hub: removing... > | All Bookmarks

SSH-in-browser

```
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Semantic Analysis Completed (retryal = false)
INFO : Returning Hive schema: Schema(fieldSchemas:[FieldSchema{name:crime_real_time_data_3.crime_real_time_data, type:string, comment:null}), properties:null)
INFO : Completed compiling command(queryId:hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7); Time taken: 0.172 seconds
INFO : Concurrency mode is disabled, not creating a lock manager
INFO : Executing command(queryId:hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7): SELECT * FROM crime_real_time_data_3 LIMIT 5
INFO : Query ID = hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7
INFO : Total jobs = 1
INFO : Launching Job 1 out of 1
INFO : Starting task [Stage-1:MAPRED] in serial mode
INFO : Subscribed to counters: [] for queryId: hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7
INFO : Session is already open
INFO : Dag name: SELECT * FROM crime_real_time_data_3 LIMIT 5 (Stage-1)
INFO : Status: Running (Executing on YARN cluster with App id application_1709563834504_0004)

-----
```

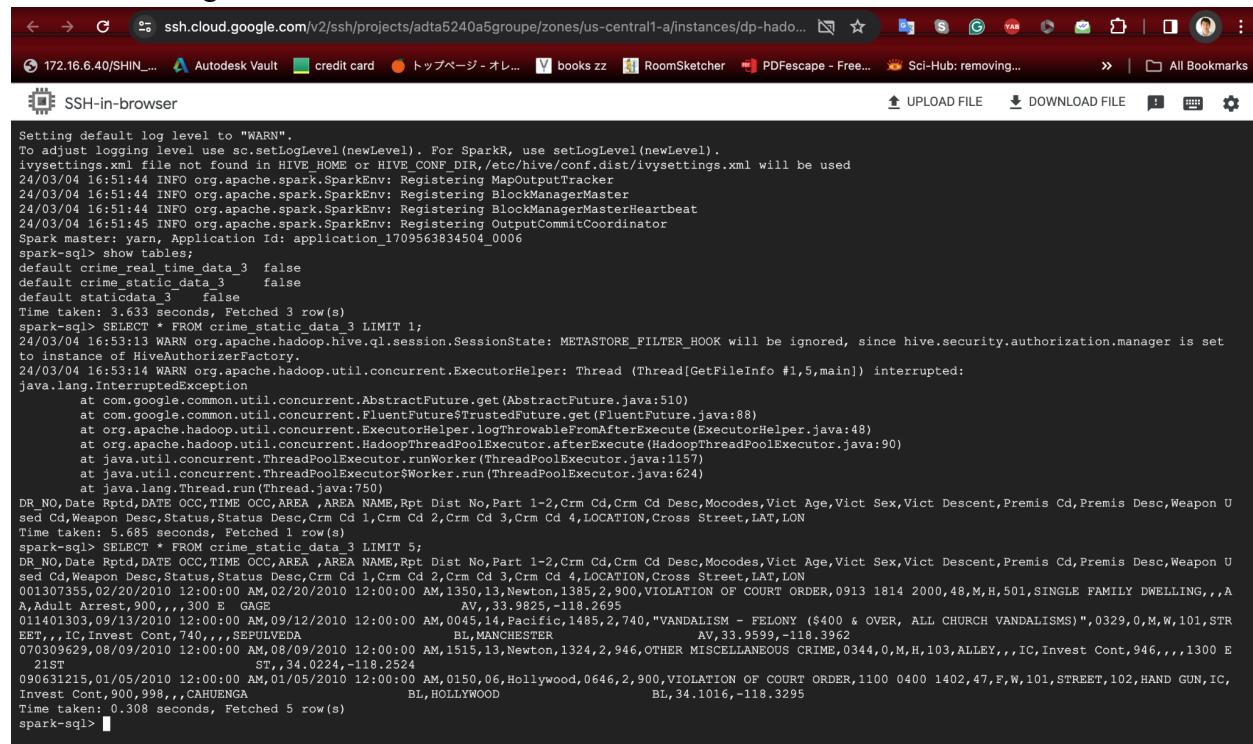
VERTICES	MODE	STATUS	TOTAL	COMPLETED	RUNNING	PENDING	FAILED	KILLED
Map 1	container	SUCCEEDED	5	5	0	0	0	0

```
VERTICES: 01/01 [=====>>>] 100% ELAPSED TIME: 10.34 s

INFO : Completed executing command(queryId:hive_20240304163822_68dd9911-11cb-4486-b8f6-f58246715ed7); Time taken: 11.001 seconds
INFO : OK
INFO : Concurrency mode is disabled, not creating a lock manager
+-----+
| crime_real_time_data_3.crime_real_time_data |
+-----+
| DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON |
| 190326475,03/01/2020 12:00:00 AM,03/01/2020 12:00:00 AM,2130,07,Wilshire,0784,1,510,VEHICLE - STOLEN,,0,M,0,101,STREET,,,AA,Adult Arrest,510,998,,,1900 S LO NGWOOD AV,,34.0375,-118.3506 |
| 200106753,02/09/2020 12:00:00 AM,02/08/2020 12:00:00 AM,1800,01,Central,0182,1,330,BURGLARY FROM VEHICLE,1822 1402 0344,47,M,0,128,BUS STOP/LAYOVER (ALSO QUREY 124),,IC,Invest Cont,330,998,,,1000 S FLOWER ST,,34.0444,-118.2628 |
| 200320258,11/11/2020 12:00:00 AM,11/04/2020 12:00:00 AM,1700,03,Southwest,0356,1,480,BIKE - STOLEN,0344 1251,19,X,X,502,"MULTI-UNIT DWELLING (APARTMENT, DUPL EX, ETC)"",,IC,Invest Cont,480,,,1400 W 37TH ST,,34.021,-118.3002 |
| 200907217,05/10/2023 12:00:00 AM,03/10/2020 12:00:00 AM,2037,09,Van Nuys,0964,1,343,SHOPLIFTING-GRAND THEFT ($950.01 & OVER),0325 1501,19,M,O,405,CLOTHING ST ORE,,,IC,Invest Cont,343,,,14000 RIVERSIDE DR,,34.1576,-118.4387 |
+-----+
5 rows selected (11.216 seconds)
0: jdbc:hive2://localhost:10000> ■
```

SPARK on Static Data

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



ssh.cloud.google.com/v2/ssh/projects/adta5240a5groupe/zones/us-central1-a/instances/dp-hado... 172.16.6.40/SHIN... Autodesk Vault credit card トップページ - オレ... books zz RoomSketcher PDFescape - Free... Sci-Hub: removing... > | All Bookmarks

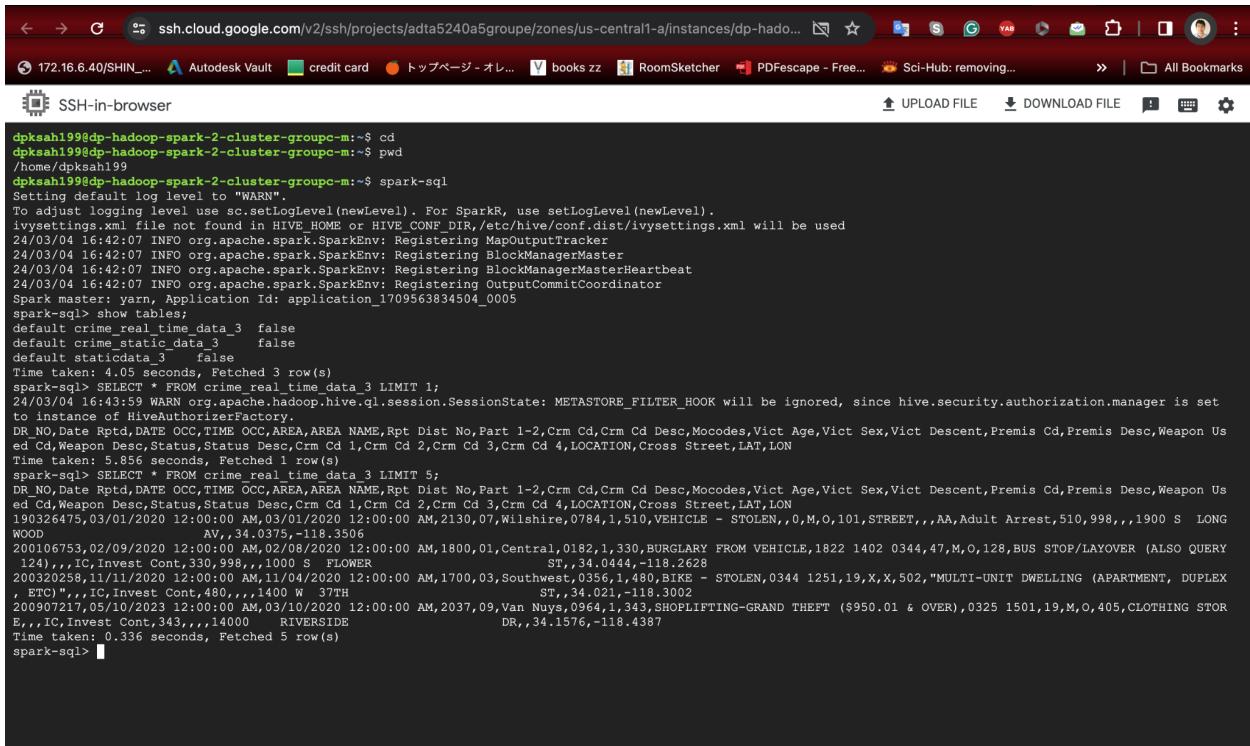
SSH-in-browser

```
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/ivysettings.xml will be used
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/03/04 16:51:44 INFO org.apache.spark.SparkEnv: Registering BlockManagerHeartbeat
24/03/04 16:51:45 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1709563834504_0006
spark-sql> show tables;
default_crime_real_time_data_3 false
default_crime_static_data_3 false
default_staticdata_3 false
Time taken: 3.633 seconds, Fetched 3 row(s)
spark-sql> SELECT * FROM crime_static_data_3 LIMIT 1;
24/03/04 16:53:13 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
24/03/04 16:53:14 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread (Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(FluentFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfterExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecute(HadoopThreadPoolExecutor.java:90)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:750)
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
Time taken: 5.685 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM crime_static_data_3 LIMIT 5;
DR_NO,Date Rptd,DATE OCC,TIME OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Used Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
001307355,02/20/2010 12:00:00 AM,02/20/2010 12:00:00 AM,1350,13,Newton,1385,2,900,VIOLATION OF COURT ORDER,0913 1814 2000,48,M,H,501,SINGLE FAMILY DWELLING,,,A,Adult Arrest,900,,,300 E GAGE AV,,33.9825,-118.2695
011401303,09/13/2010 12:00:00 AM,09/12/2010 12:00:00 AM,0045,14,Pacific,1485,2,740,"VANDALISM - FELONY ($400 & OVER, ALL CHURCH VANDALISMS)"",0329,0,M,W,101,STREET,,,IC,Invest Cont,740,,,SEPUVEDA BL,MANCHESTER AV,33.9599,-118.3962
0707309629,08/09/2010 12:00:00 AM,08/09/2010 12:00:00 AM,1515,13,Newton,1324,2,946,OTHER MISCELLANEOUS CRIME,0344,0,M,H,103,ALLEY,,,IC,Invest Cont,946,,,1300 E 21ST ST,,34.0224,-118.2524
090631215,01/05/2010 12:00:00 AM,01/05/2010 12:00:00 AM,0150,06,Hollywood,0646,2,900,VIOLATION OF COURT ORDER,1100 0400 1402,47,F,W,101,STREET,102,HAND GUN,IC,Invest Cont,900,998,,,CAHUENGA BL,HOLLYWOOD BL,34.1016,-118.3295
Time taken: 0.308 seconds, Fetched 5 row(s)
spark-sql> ■
```

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.



```
ssh.cloud.google.com/v2/ssh/projects/adta5240a5groupe/zones/us-central1-a/instances/dp-hado... 172.16.6.40/SHIN... Autodesk Vault credit card トップページ - オレ... books zz RoomSketcher PDFEscape - Free... Sci-Hub: removing... All Bookmarks
SSH-in-browser
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~$ cd
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~$ pwd
/home/dpsah199
dpsah199@dp-hadoop-spark-2-cluster-groupc-m:~$ spark-sql
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR, /etc/hive/conf.dist/ivysettings.xml will be used
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/03/04 16:42:07 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1709563834504_0005
spark-sql> show tables;
default crime_real_time_data_3 false
default crime_static_data_3 false
default staticdata_3 false
Time taken: 4.05 seconds, Fetched 3 row(s)
spark-sql> SELECT * FROM crime_real_time_data_3 LIMIT 1;
24/03/04 16:43:59 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
DR_NO,Date,Rptd,DATE,OCC,TIME,OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Us ed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
Time taken: 5.856 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM crime_real_time_data_3 LIMIT 5;
DR_NO,Date,Rptd,DATE,OCC,TIME,OCC,AREA,AREA NAME,Rpt Dist No,Part 1-2,Crm Cd,Crm Cd Desc,Mocodes,Vict Age,Vict Sex,Vict Descent,Premis Cd,Premis Desc,Weapon Us ed Cd,Weapon Desc,Status,Status Desc,Crm Cd 1,Crm Cd 2,Crm Cd 3,Crm Cd 4,LOCATION,Cross Street,LAT,LON
190326475,03/01/2020 12:00:00 AM,03/01/2020 12:00:00 AM,2130,07,Wilshire,0784,1,510,VEHICLE - STOLEN,,0,M,0,101,STREET,,,AA,Adult Arrest,510,998,,1900 S LONG WOOD
AV,,34.0375,-118.3506
200106753,02/09/2020 12:00:00 AM,02/08/2020 12:00:00 AM,1800,01,Central,0182,1,330,BURGLARY FROM VEHICLE,1822 1402 0344,47,M,0,128,BUS STOP/LAYOVER (ALSO QUERY 124),,IC,Invest Cont,330,998,,1000 S FLOWER
ST,,34.0444,-118.2628
200320258,11/11/2020 12:00:00 AM,11/04/2020 12:00:00 AM,1700,03,Southwest,0356,1,480,BIKE - STOLEN,0344 1251,19,X,X,502,"MULTI-UNIT DWELLING (APARTMENT, DUPLEX , ETC)"",,IC,Invest Cont,480,,1400 W 37TH
ST,,34.021,-118.3002
200907217,05/10/2023 12:00:00 AM,03/10/2020 12:00:00 AM,2037,09,Van Nuys,0964,1,343,SHOPLIFTING-GRAND THEFT ($950.01 & OVER),0325 1501,19,M,0,405,CLOTHING STOR E,,IC,Invest Cont,343,,14000 RIVERSIDE
DR,,34.1576,-118.4387
Time taken: 0.336 seconds, Fetched 5 row(s)
spark-sql> 
```

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.

Start your Free Trial

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 [BROWSE](#) [?](#)

Source Data Partitioning

Destination

Project [BROWSE](#)

Dataset [BROWSE](#)

Table Maximum name size is 1,024 UTF-8 bytes. Unicode letters, marks, numbers, connectors, dashes, and spaces are allowed.

Table type [BROWSE](#) [?](#)

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 [Data transfers](#) [Scheduled queries](#) [Analytics Hub](#) [Dataform](#) [Partner Center](#)

Migration [Assessment](#) [SQL translation](#)

Administration [Monitoring](#) [Capacity management](#) [BI Engine](#)

Release Notes

Crime_Static_Data

[SCHEMA](#) [DETAILS](#) [LINEAGE](#) [DATA](#)

Untitled

```

1 SELECT
2   `AREA_NAME`,
3   `Vict_Sex`
4 FROM
5   `adta5240a5groupe.Crime_Static_Data.Crime_Static_Data`
6 ORDER BY
7   `Premis_Desc` ASC;
8

```

Query results

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

[EDIT SCHEMA](#) [REFRESH](#)

Start your Free Trial

Google Cloud Explorer

Viewing resources

SHOW START

adta5240a5groupe

adta5240a5groupe

adta5240a5groupe

SUMMA

Nothing current

CREATE TABLE CANCEL

Google Cloud ADTA5240A5GROUPE Search (/) for resources, docs, products, and more

BigQuery Studio

Analysis

Data transfers

Scheduled queries

Analytics Hub

Dataform

Partner Center

Migration

Assessment

SQL translation

Administration

Monitoring

Capacity management

BI Engine

Release Notes

Job history

CREATE TABLE CANCEL

Source

Create table from Google Cloud Storage

Select file from GCS bucket or use a URI pattern f23group/data/Crime_Real_Time_Data BROWSE

File format CSV

Source Data Partitioning

Destination

Project adta5240a5groupe BROWSE

Dataset Crime_Real_Time

Table Crime_Real_Time

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

Untitled 2

RUN SAVE DOWNLOAD SHARE SCHEDULE Quer...

1 SELECT
2 'AREA_NAME',
3 'Vict_Sex'
4 FROM
5 'adta5240a5groupe.Crime_Real_Time.Crime_Real_Time'
6 ORDER BY
7 'Premis_Desc' ASC;

Press Option+F1 for Accessibility Options

Query results

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
4	West Valley	M

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 [I](#)

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-management... [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 [I](#)

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-management... [PREVIEW](#)

Crime_Real_Time

QUERY SHARE DELETE EXPORT REFRESH

SCHEMA DETAILS LINEAGE DATA PROFILE DATA QUALITY

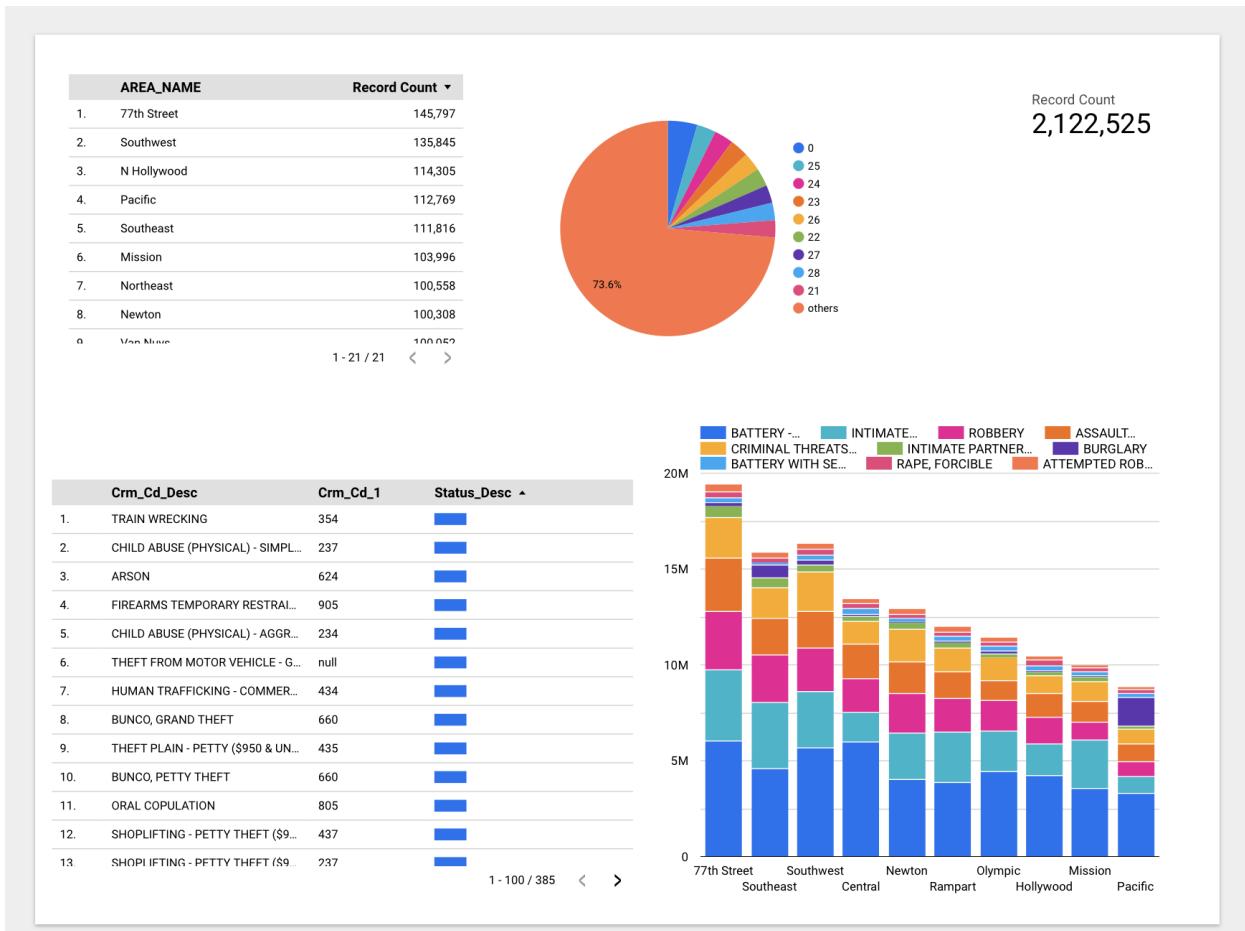
Filter Enter property name or value

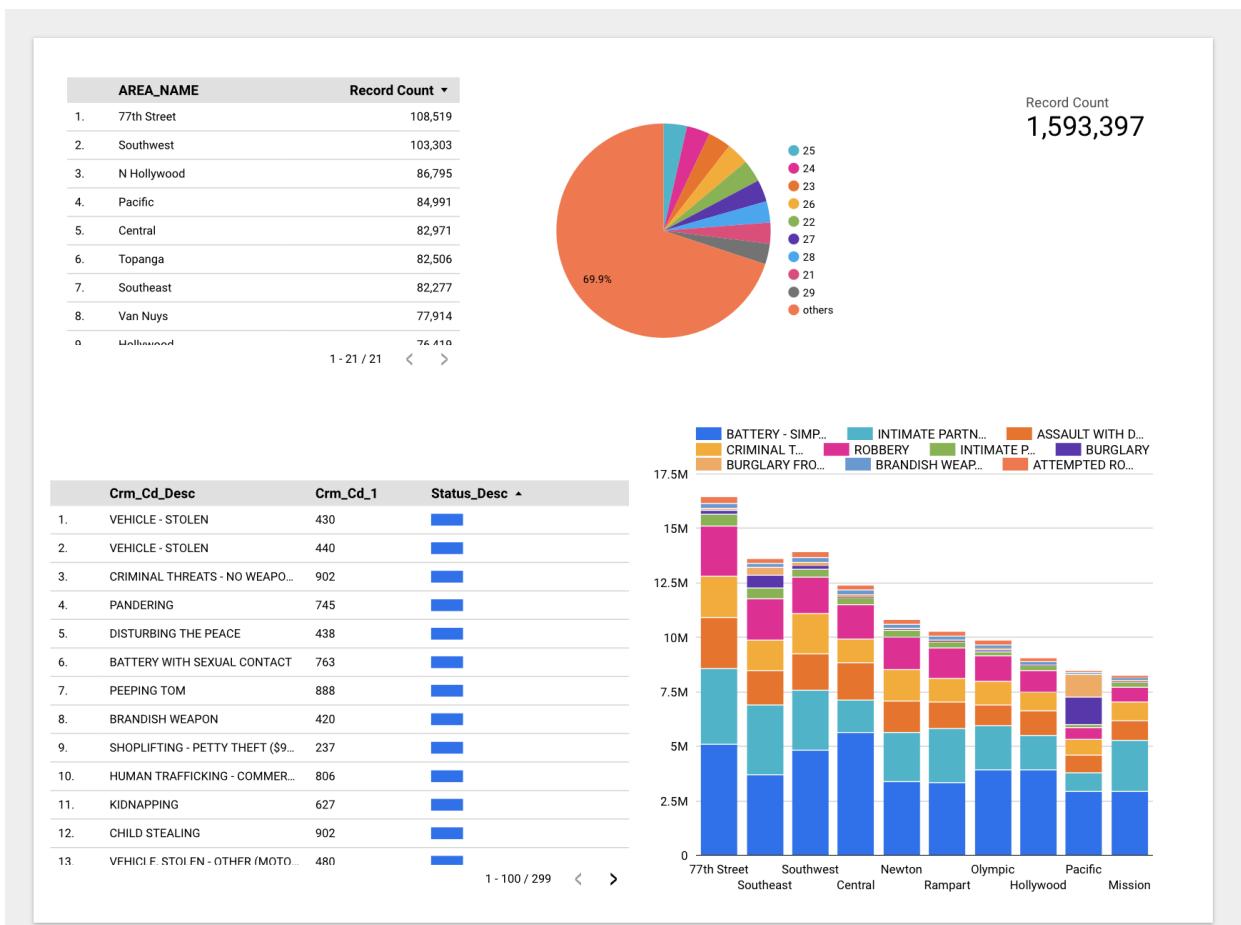
Field name	Type	Mode	Key	Collation	Default Value	Policy Tags	Description
DR_NO	INTEGER	NULLABLE	-	-	-	-	-
Date_Rptd	TIMESTAMP	NULLABLE	-	-	-	-	-
DATE_OCC	TIMESTAMP	NULLABLE	-	-	-	-	-
TIME_OCC	INTEGER	NULLABLE	-	-	-	-	-
AREA	INTEGER	NULLABLE	-	-	-	-	-
AREA_NAME	STRING	NULLABLE	-	-	-	-	-
Rpt_Dist_No	INTEGER	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

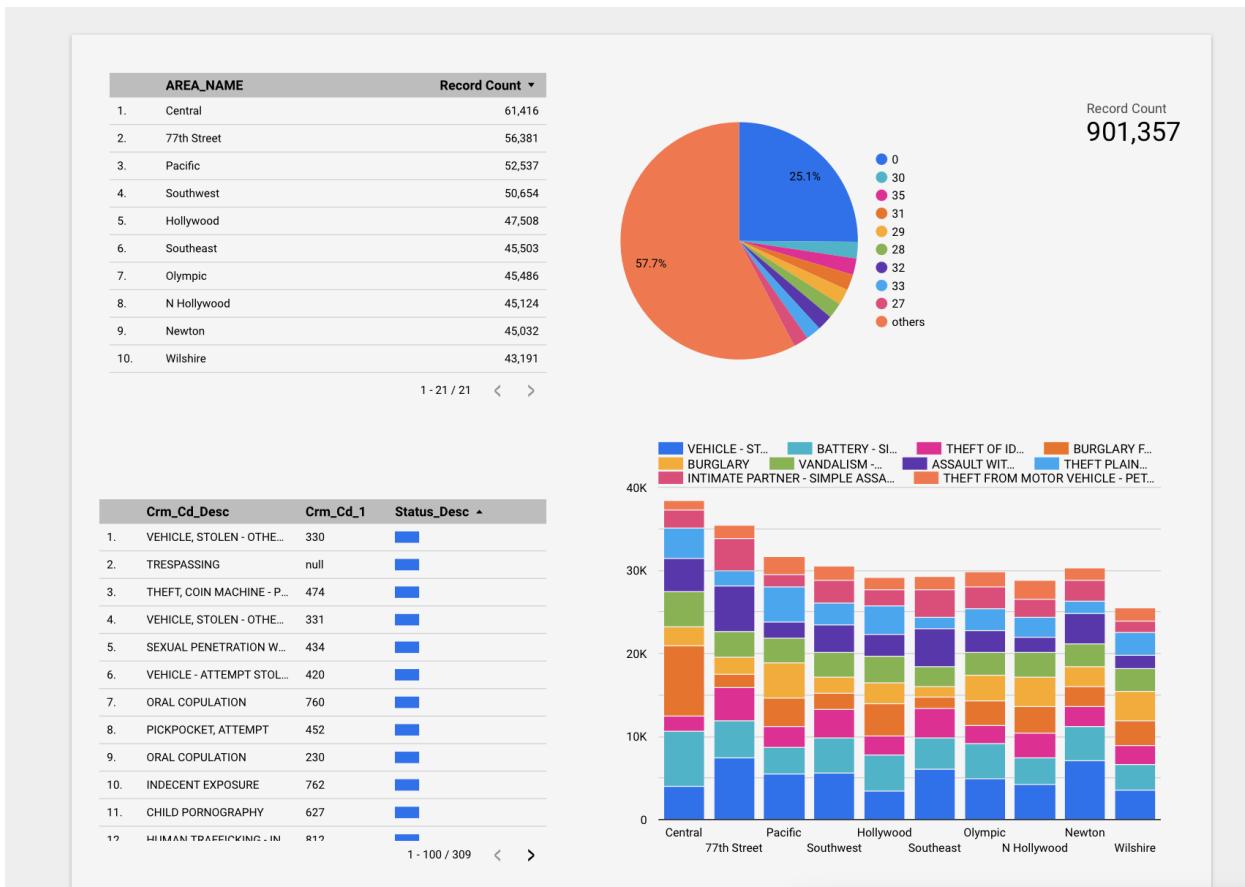
Job history [REFRESH](#)

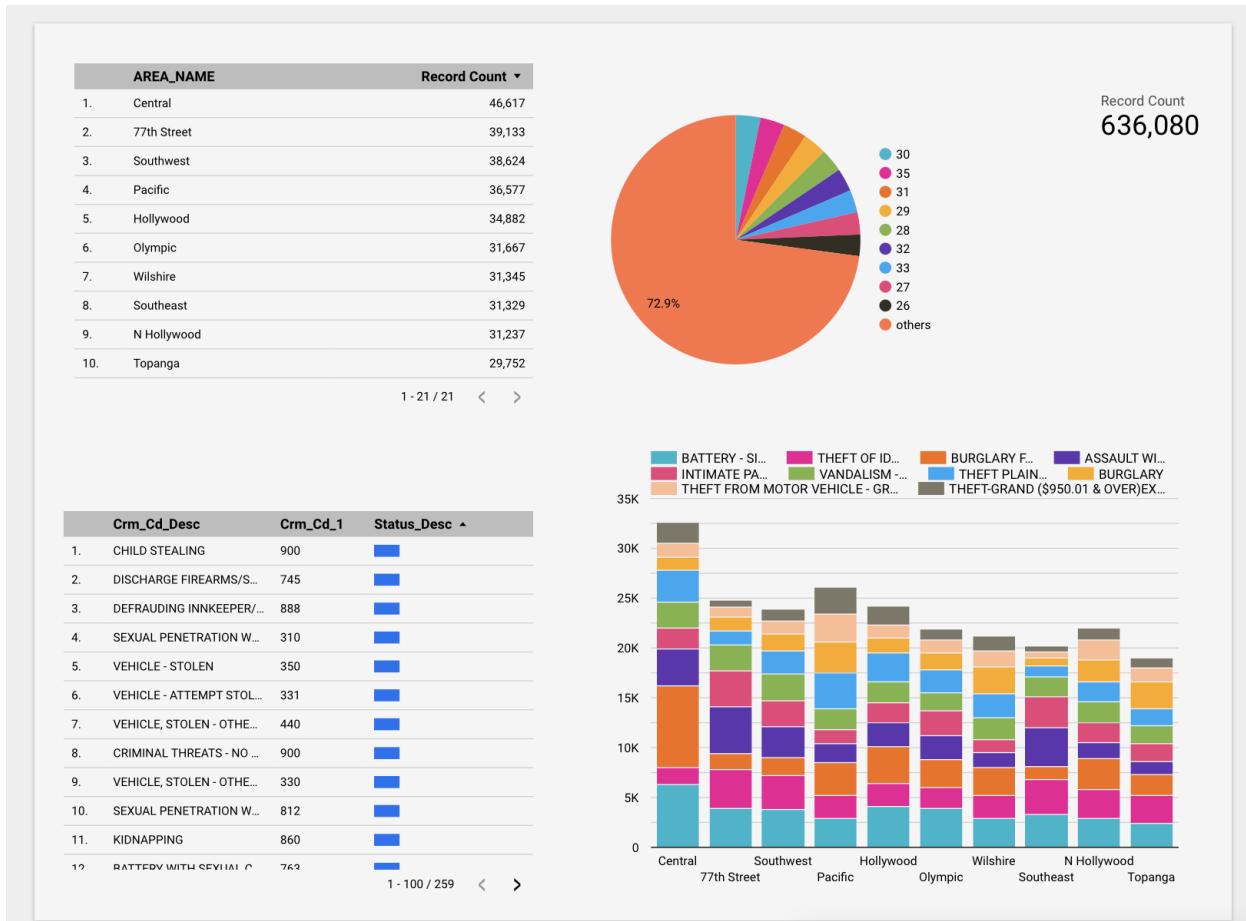
Step 7 -Visualization





Real time data





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 ☺-----