A Project Report

for

NEIGHBORHOOD THEFTS IN SUMMER

A Data Engineering Approach to Address Theft Patterns in the Neighborhood.

by

Group -9

Our Data Engineering Team - Respective Roles and Contributions:

- Harish Inavolu
    - Contributor to Data Cleaning, Processing, and Transformation.
    - Co-contributor in Data Storage.
    - Co-contributor in Presentation Preparation.
- Suprathika Vangari
    - Contributor to Data Storage.
    - Big Query.
    - Spark SQL.
    - Presentation and Project Report Preparation.
- Udaya Adepu
    - Contributor to Hive SQL.
    - Documented Analysis using Hive SQL.
- Vishala Vadla
    - Contributor in Cluster provisioning using Hadoop Infrastructure.
    - Co-contributor in Data Collection.
    - Documented Hadoop Infrastructure Process.
- Sai Likitha Uppala
    - Co-contributor in documenting Analysis using Spark SQL.
- Rajeshwari Chiraboina
    - Co-contributor in documenting Analysis using Spark SQL.

Keywords: Neighborhood thefts, summer, data engineering, data lifecycle, Google Cloud Platform, GCP, Hadoop, Dataproc, BigQuery, Hive, Spark, data analysis, crime prevention, community safety.

Subject: Neighborhood Thefts in Summer - Project Update

To: Denton Local Authority, Denton Police Department

Cc: Denton Local Authority Leaders, Denton Police Department

Bcc: Harish Inavolu, Suprathika Vangari, Udaya Adepu, Vishala Vadla, Sai Likitha Uppala, Rajeshwari Chiraboina.

Dear Colleagues,

I hope this email finds you well. Over the past two weeks, we as a data engineering team have successfully implemented the infrastructure needed to support the next phase of this project. We identified and collected quality data from several data sources.

Upon collecting the data, we processed the data using Open Refine, ensuring it was refined and ready for further. I am writing to provide an update on our project "Neighborhood Thefts in Summer" which we have been collaborating with and working together on our analysis.

We stored the processed data securely in a Google Cloud Platform (GCP) Storage Bucket, providing us with the flexibility and security we need while staying within our limited budget. We also set up a Big Query instance to manage the data, allowing us to run queries and gain deeper insights from the processed data.

Moving further, we leveraged the Hadoop ecosystem, including Spark and Hive SQL, to perform advanced analysis on the datasets. This led us to uncover valuable hidden patterns and trends that can inform our efforts to address the neighborhood theft issue in Denton.

I attached a few screenshots that should provide an overview of the work we have carried out so far. These visuals should give you a better understanding of the progress we have made so far and about the next steps in the data lifecycle for this project.

If you have questions or need more information, please contact us. We are committed to ensuring the success of this project and look forward to collaborating with you further.

Cordially,

TEAM NTS

Project Files – ADTA 5240 Group9

TABLE OF CONTENTS

INTRODUCTION

Our project aims to address the growing issue of neighborhood thefts during the summer months in Denton. With longer days and increased vacations, homes become more vulnerable, providing opportunities for criminals to target the community or a neighborhood. By understanding the underlying factors leading to an increase in thefts, this project looks to empower local authorities and community leaders. It further assists data analysts in implementing targeted solutions that increase safety while improving the quality of life for Denton residents.

1.1 Problem Statement:

The city of Denton has experienced a huge increase in neighborhood thefts during the summer, thereby leading to a decline in community safety and a sense of security among the residents. As a result, the local authorities and community leaders are seeking a data-driven approach to understand the underlying patterns, factors, and trends contributing to these thefts, to develop necessary strategies, and bring up a solution to address the issue**.**

1.2 Importance:

The importance of this project lies in utilizing the Google Cloud Platform while leveraging the potential of the platform's tools to uncover insights while developing strategies to address the pressing issue of neighborhood thefts. By leveraging data engineering techniques and the power of Google Cloud Platform (GCP), this project aims to provide data-driven solutions that can be implemented by local authorities and community leaders. This approach is prominent in terms of enhancing the overall safety and quality of life for the people of Denton.

2. Data Generation

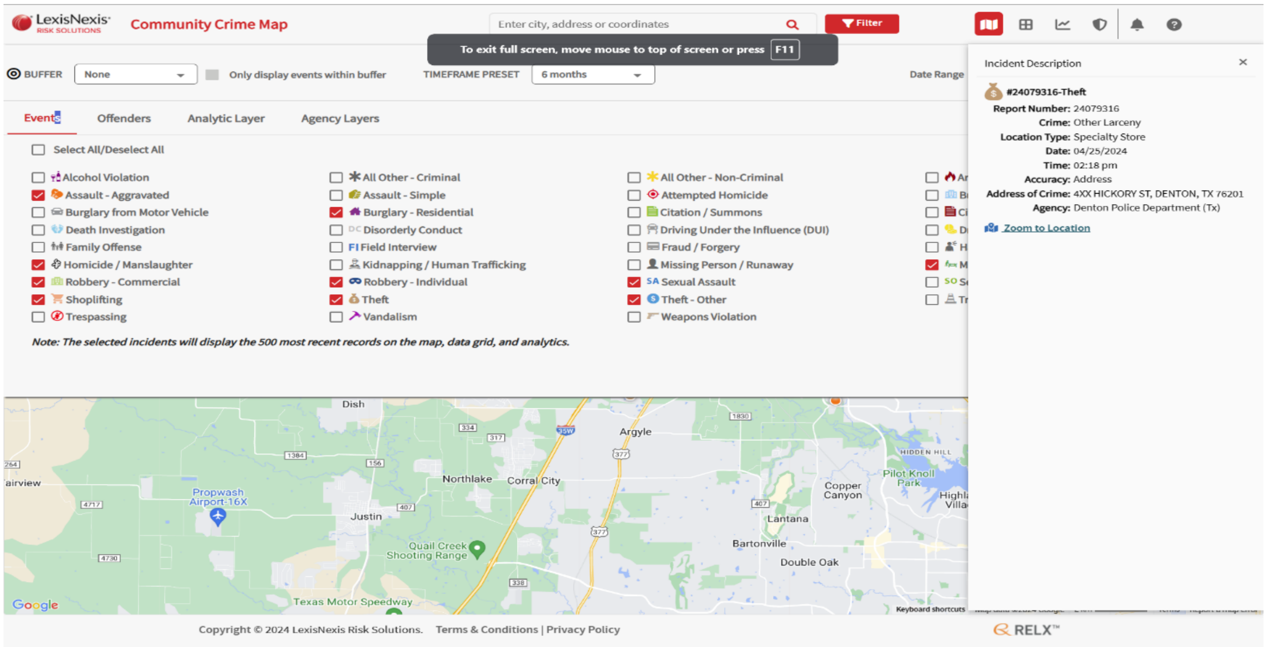
The process of Data Generation is a crucial beginning phase in our project. The first step in our data generation approach was to identify relevant data sources that could shed light on the potential theft patterns.

Community Crime Map: This interactive map provided a visual representation of reported crimes in the area, including thefts. By leveraging this resource, we could pinpoint the hotspots and analyze the spatial distribution of the incidents. There is an option to search for events by location, viewing results on the map, in a data grid or through analytics on the data for the location selected. We tried to customize the map with the crime data within Denton Local Region.

Syracuse Open Data Portal: This government-maintained data repository contained a huge amount of information on crime incidents within the city. This dataset is invaluable in understanding the nature and frequency of thefts in Denton.

Dallas Open Data: The data obtained from this source enabled us in recognizing the potential for drawing parallels and helped in identifying common factors contributing to summer thefts across the region. Police reports were extracted from the local department's internal records. It included details such as the type of theft, location, time, and any available suspect information.

Screenshot - One of the three data sources - Community Crime map data:


LexisNexis RISK SOLUTIONS
Community Crime Map
Enter city, address or coordinates
Filter
To exit full screen, move mouse to top of screen or press F11
Dish
Argyle
Northlake
Corral City
Justin
Propwash Airport-16X
Quail Creek Shooting Range
Texas Motor Speedway
Copper Canyon
Lantana
Bartonville
Double Oak
Pilot Knoll Park
HIDDEN HILL
airview
Google
Keyboard shortcuts
Copyright © 2024 LexisNexis Risk Solutions. Terms & Conditions | Privacy Policy
RELX™

◎ BUFFER: None | □ Only display events within buffer | TIMEFRAME PRESET: 6 months | Date Range

Events | Offenders | Analytic Layer | Agency Layers

□ Select All/Deselect All

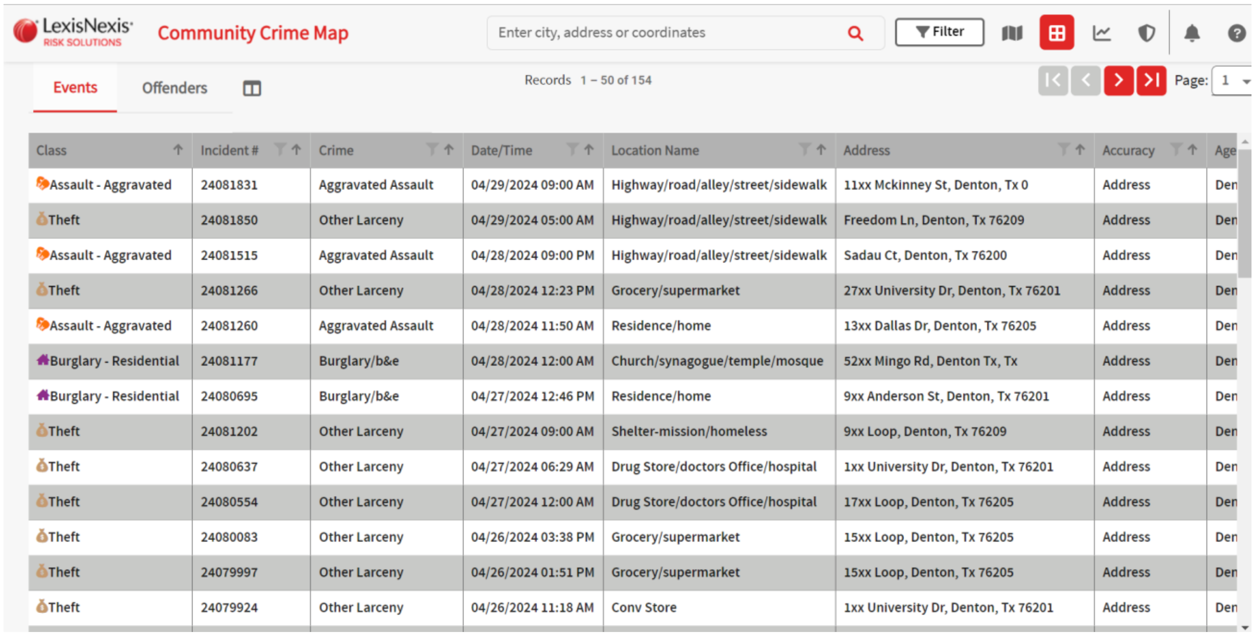
- □ Alcohol Violation
- ☑ Assault - Aggravated
- □ Burglary from Motor Vehicle
- □ Death Investigation
- □ Family Offense
- ☑ Homicide / Manslaughter
- ☑ Robbery - Commercial
- ☑ Shoplifting
- □ Trespassing
- □ All Other - Criminal
- □ Assault - Simple
- ☑ Burglary - Residential
- □ DC Disorderly Conduct
- □ FI Field Interview
- □ Kidnapping / Human Trafficking
- ☑ Robbery - Individual
- ☑ Theft
- □ Vandalism
- □ All Other - Non-Criminal
- □ Attempted Homicide
- □ Citation / Summons
- □ Driving Under the Influence (DUI)
- □ Fraud / Forgery
- □ Missing Person / Runaway
- ☑ SA Sexual Assault
- ☑ Theft - Other
- □ Weapons Violation
- □ Ar
- □ Bi
- □ Ci
- □ Di
- □ H
- ☑ M
- □ SO Se
- □ Tr

Note: The selected incidents will display the 500 most recent records on the map, data grid, and analytics.

Incident Description ×

#24079316-Theft
Report Number: 24079316
Crime: Other Larceny
Location Type: Specialty Store
Date: 04/25/2024
Time: 02:18 pm
Accuracy: Address
Address of Crime: 4XX HICKORY ST, DENTON, TX 76201
Agency: Denton Police Department (Tx)
Zoom to Location

Screenshot Owner: Harish Inavolu



Class	Incident #	Crime	Date/Time	Location Name	Address	Accuracy	Age
Assault - Aggravated	24081831	Aggravated Assault	04/29/2024 09:00 AM	Highway/road/alley/street/sidewalk	11xx Mckinney St, Denton, Tx 0	Address	Den
Theft	24081850	Other Larceny	04/29/2024 05:00 AM	Highway/road/alley/street/sidewalk	Freedom Ln, Denton, Tx 76209	Address	Den
Assault - Aggravated	24081515	Aggravated Assault	04/28/2024 09:00 PM	Highway/road/alley/street/sidewalk	Sadau Ct, Denton, Tx 76200	Address	Den
Theft	24081266	Other Larceny	04/28/2024 12:23 PM	Grocery/supermarket	27xx University Dr, Denton, Tx 76201	Address	Den
Assault - Aggravated	24081260	Aggravated Assault	04/28/2024 11:50 AM	Residence/home	13xx Dallas Dr, Denton, Tx 76205	Address	Den
Burglary - Residential	24081177	Burglary/b&e	04/28/2024 12:00 AM	Church/synagogue/temple/mosque	52xx Mingo Rd, Denton Tx, Tx	Address	Den
Burglary - Residential	24080695	Burglary/b&e	04/27/2024 12:46 PM	Residence/home	9xx Anderson St, Denton, Tx 76201	Address	Den
Theft	24081202	Other Larceny	04/27/2024 09:00 AM	Shelter-mission/homeless	9xx Loop, Denton, Tx 76209	Address	Den
Theft	24080637	Other Larceny	04/27/2024 06:29 AM	Drug Store/doctors Office/hospital	1xx University Dr, Denton, Tx 76201	Address	Den
Theft	24080554	Other Larceny	04/27/2024 12:00 AM	Drug Store/doctors Office/hospital	17xx Loop, Denton, Tx 76205	Address	Den
Theft	24080083	Other Larceny	04/26/2024 03:38 PM	Grocery/supermarket	15xx Loop, Denton, Tx 76205	Address	Den
Theft	24079997	Other Larceny	04/26/2024 01:51 PM	Grocery/supermarket	15xx Loop, Denton, Tx 76205	Address	Den
Theft	24079924	Other Larceny	04/26/2024 11:18 AM	Conv Store	1xx University Dr, Denton, Tx 76201	Address	Den

Screenshot Owner: Harish Inavolu

3. Data Collection

For the Community Crime Map, we employed a data collection tool **Listly** to extract the reported crime incidents. The Syracuse Open Data Portal and Dallas Open Data repositories provided structured datasets that you could directly download and integrate into your analysis pipeline.

Overall, the datasets contained a vast amount of information, including descriptions of incidents, crime categories, and other relevant details that focus on the nature and circumstances surrounding the thefts.

4. Project Initialization in Google Cloud Console

Prior to processing the data that we have collected; we accessed the Google Cloud Console and created a new project named “PROJECT-GROUP9-NTHEFTS” to proceed with the initial step of setting up a cloud storage bucket to store the processed data.

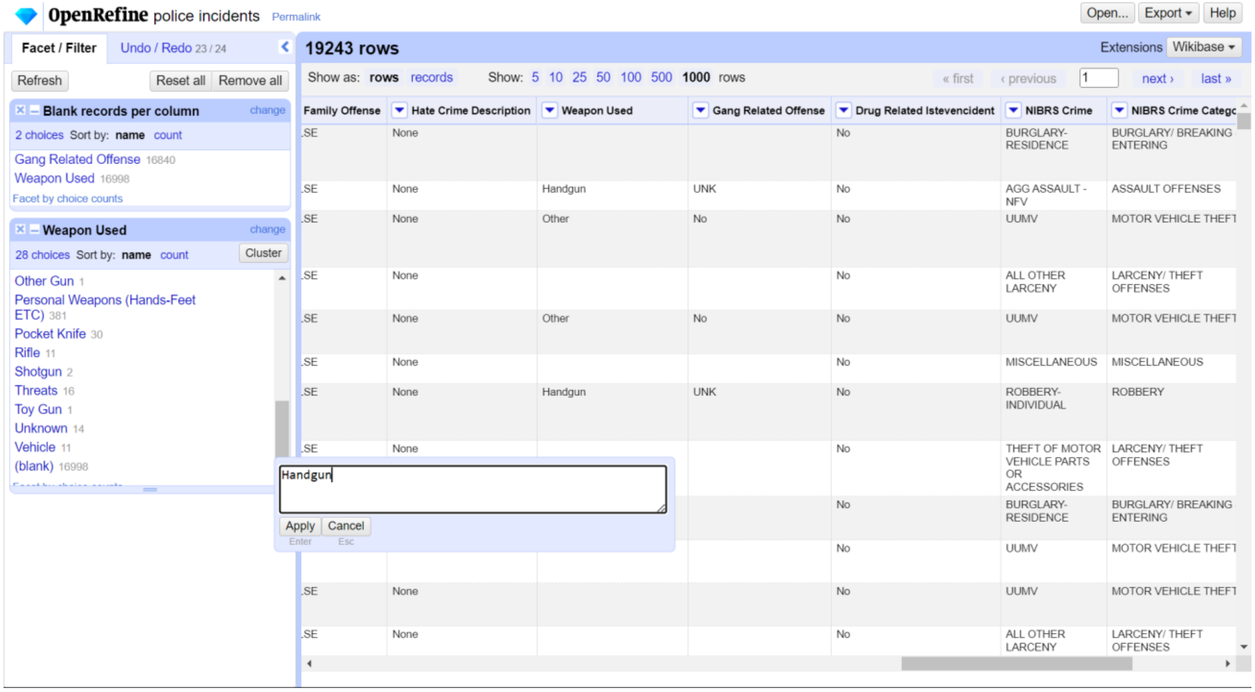
5. Data Processing

5.1 Data Cleaning, Preprocessing and Transformation using Open Refine.

Upon collecting the data from the sources, we utilized **OpenRefine,** a powerful data cleaning and transformation tool, to refine the datasets. We removed duplicates, standardized data formats, handled missing values, and resolved any inconsistencies to ensure high-quality data for further analysis. Overall, we implemented following techniques:

- Find and fix inconsistencies in the data, such as spelling variations or capitalization issues.
- Split and combine columns to extract more granular information from the data.
- Apply custom transformations and functions to standardize the data format.
- Cluster similar values together to resolve duplicates and inconsistencies.

Screenshot 1: Replacing empty values with most occurred values using Open Refine.



19243 rows

Blank records per column

2 choices Sort by: name count

Gang Related Offense 16840
Weapon Used 16998

Facet by choice counts

Weapon Used

28 choices Sort by: name count

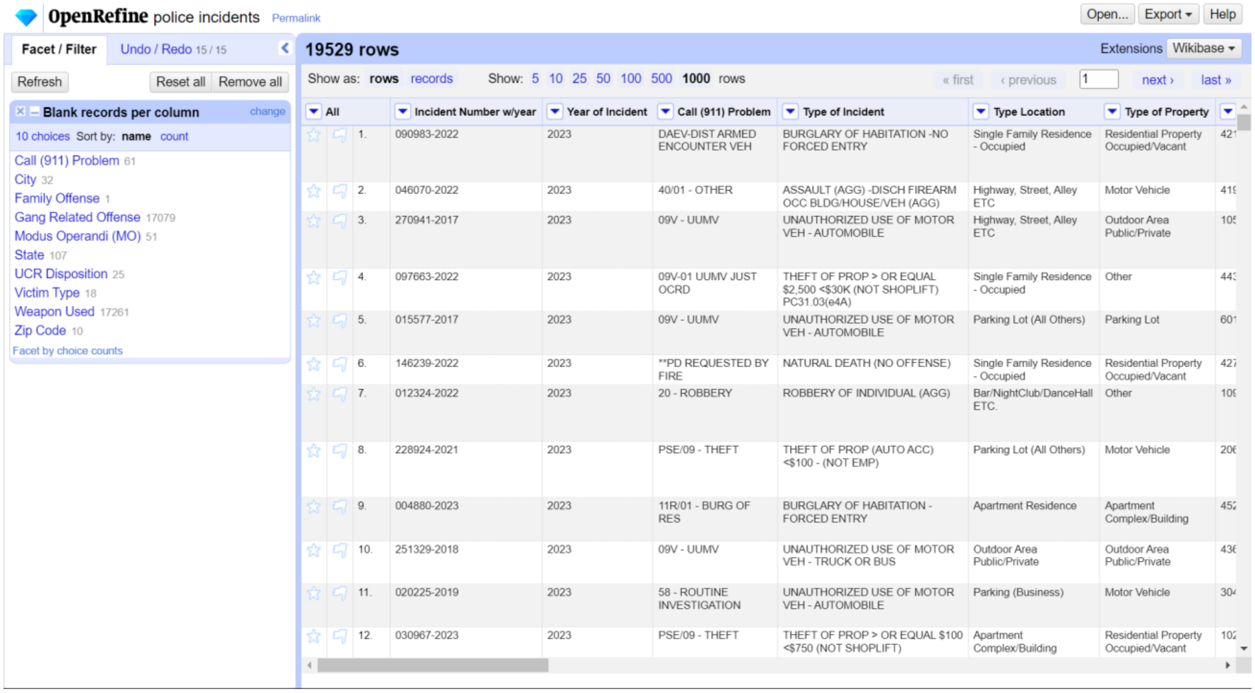
Other Gun 1
Personal Weapons (Hands-Feet ETC) 381
Pocket Knife 30
Rifle 11
Shotgun 2
Threats 16
Toy Gun 1
Unknown 14
Vehicle 11
(blank) 16998

Handgun

Family Offense	Hate Crime Description	Weapon Used	Gang Related Offense	Drug Related Istevencident	NIBRS Crime	NIBRS Crime Categc
.SE	None			No	BURGLARY-RESIDENCE	BURGLARY/ BREAKING ENTERING
.SE	None	Handgun	UNK	No	AGG ASSAULT - NFV	ASSAULT OFFENSES
.SE	None	Other	No	No	UUMV	MOTOR VEHICLE THEFT
.SE	None			No	ALL OTHER LARCENY	LARCENY/ THEFT OFFENSES
.SE	None	Other	No	No	UUMV	MOTOR VEHICLE THEFT
.SE	None			No	MISCELLANEOUS	MISCELLANEOUS
.SE	None	Handgun	UNK	No	ROBBERY-INDIVIDUAL	ROBBERY
.SE	None			No	THEFT OF MOTOR VEHICLE PARTS OR ACCESSORIES	LARCENY/ THEFT OFFENSES
				No	BURGLARY-RESIDENCE	BURGLARY/ BREAKING ENTERING
				No	UUMV	MOTOR VEHICLE THEFT
.SE	None			No	UUMV	MOTOR VEHICLE THEFT
.SE	None			No	ALL OTHER LARCENY	LARCENY/ THEFT OFFENSES

Screenshot Owner: Harish Inavolu

Screenshot 2: Selected blank facets in all columns to remove.



19529 rows

Blank records per column change

10 choices Sort by: name count

Call (911) Problem	61
City	32
Family Offense	1
Gang Related Offense	17079
Modus Operandi (MO)	51
State	107
UCR Disposition	25
Victim Type	18
Weapon Used	17261
Zip Code	10

Facet by choice counts

All	Incident Number w/year	Year of Incident	Call (911) Problem	Type of Incident	Type Location	Type of Property	
1.	090983-2022	2023	DAEV-DIST ARMED ENCOUNTER VEH	BURGLARY OF HABITATION -NO FORCED ENTRY	Single Family Residence - Occupied	Residential Property Occupied/Vacant	421
2.	046070-2022	2023	40/01 - OTHER	ASSAULT (AGG) -DISCH FIREARM OCC BLDG/HOUSE/VEH (AGG)	Highway, Street, Alley ETC	Motor Vehicle	419
3.	270941-2017	2023	09V - UUMV	UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE	Highway, Street, Alley ETC	Outdoor Area Public/Private	105
4.	097663-2022	2023	09V-01 UUMV JUST OCRD	THEFT OF PROP > OR EQUAL $2,500 <$30K (NOT SHOPLIFT) PC31.03(e4A)	Single Family Residence - Occupied	Other	443
5.	015577-2017	2023	09V - UUMV	UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE	Parking Lot (All Others)	Parking Lot	601
6.	146239-2022	2023	**PD REQUESTED BY FIRE	NATURAL DEATH (NO OFFENSE)	Single Family Residence - Occupied	Residential Property Occupied/Vacant	427
7.	012324-2022	2023	20 - ROBBERY	ROBBERY OF INDIVIDUAL (AGG)	Bar/NightClub/DanceHall ETC.	Other	109
8.	228924-2021	2023	PSE/09 - THEFT	THEFT OF PROP (AUTO ACC) <$100 - (NOT EMP)	Parking Lot (All Others)	Motor Vehicle	206
9.	004880-2023	2023	11R/01 - BURG OF RES	BURGLARY OF HABITATION - FORCED ENTRY	Apartment Residence	Apartment Complex/Building	452
10.	251329-2018	2023	09V - UUMV	UNAUTHORIZED USE OF MOTOR VEH - TRUCK OR BUS	Outdoor Area Public/Private	Outdoor Area Public/Private	436
11.	020225-2019	2023	58 - ROUTINE INVESTIGATION	UNAUTHORIZED USE OF MOTOR VEH - AUTOMOBILE	Parking (Business)	Motor Vehicle	304
12.	030967-2023	2023	PSE/09 - THEFT	THEFT OF PROP > OR EQUAL $100 <$750 (NOT SHOPLIFT)	Apartment Complex/Building	Residential Property Occupied/Vacant	102

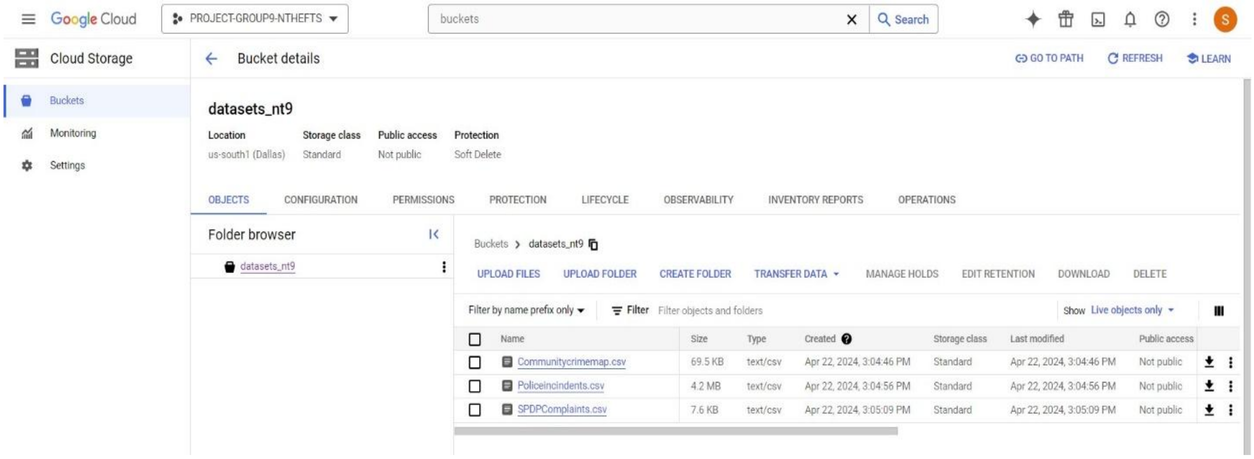
Screenshot Owner: Harish Inavolu

6. Data Storage

6.1 Google Cloud Storage Buckets

After processing the data using OpenRefine we recognized the need for a robust and scalable storage solution, and so we used Google Cloud Platform's Storage Bucket service to store our data. Google Cloud Storage provided a centralized and easily accessible repository for the processed data, helping seamless integration across various GCP's services and tools throughout our data engineering pipeline. Firstly, we set up a new project in the GCP Console. Secondly, we created a storage bucket named **datasets_nt9.** We created three subfolders namely, data, logs and output and uploaded our datasets into the data folder.

Screenshot 1: Cloud Storage Bucket with processed Data files



datasets_nt9

Location	Storage class	Public access	Protection
us-south1 (Dallas)	Standard	Not public	Soft Delete

□	Name	Size	Type	Created	Storage class	Last modified	Public access
□	Communitycrimemap.csv	69.5 KB	text/csv	Apr 22, 2024, 3:04:46 PM	Standard	Apr 22, 2024, 3:04:46 PM	Not public
□	Policeincindents.csv	4.2 MB	text/csv	Apr 22, 2024, 3:04:56 PM	Standard	Apr 22, 2024, 3:04:56 PM	Not public
□	SPDPComplaints.csv	7.6 KB	text/csv	Apr 22, 2024, 3:05:09 PM	Standard	Apr 22, 2024, 3:05:09 PM	Not public

Screenshot Owner: Harish Inavolu

7. Data Management

7.1 Dataproc for provisioning Hadoop Infrastructure

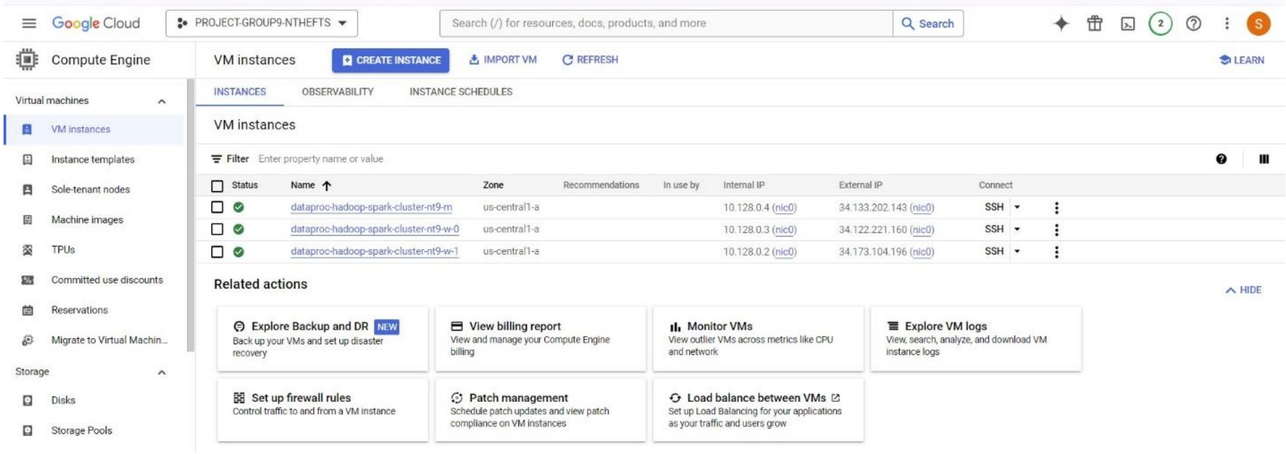
To begin with the management phase, a crucial step in our data engineering pipeline, we used Google Cloud's Dataproc service to set up the Hadoop ecosystem for managing and processing the neighborhood theft data. By using the Hadoop environment (HDFS (Hadoop Distributed File System)) managed by Dataproc, we were able to focus on the data engineering tasks. We enabled the Compute Engine API within GCP, through which we were able to access the powerful VM (virtual machines) instances served as a backbone of our data management infrastructure.

We configured the cluster with the following specifications:

- Cluster Name: “dataproc-Hadoop-spark-cluster"
- Region and Zone: us-central1
- Master Node: e2-standard-4 machine type with 128 GB disk size.
- Worker Nodes: two nodes with 128 GB disk size

To ensure seamless integration and management of this cluster, we enabled the Cloud Dataproc API. With the API in place, you gained access to a user-friendly interface and command-line tools, empowering you to effortlessly provision, configure, and manage the cluster resources.

Screenshot - The three dataproc cluster nodes.



VM instances

Status	Name ↑	Zone	Recommendations	In use by	Internal IP	External IP	Connect
✓	dataproc-hadoop-spark-cluster-nt9-m	us-central1-a			10.128.0.4 (nic0)	34.133.202.143 (nic0)	SSH
✓	dataproc-hadoop-spark-cluster-nt9-w-0	us-central1-a			10.128.0.3 (nic0)	34.122.221.160 (nic0)	SSH
✓	dataproc-hadoop-spark-cluster-nt9-w-1	us-central1-a			10.128.0.2 (nic0)	34.173.104.196 (nic0)	SSH

Related actions

Explore Backup and DR NEW
Back up your VMs and set up disaster recovery

View billing report
View and manage your Compute Engine billing

Monitor VMs
View outlier VMs across metrics like CPU and network

Explore VM logs
View, search, analyze, and download VM instance logs

Set up firewall rules
Control traffic to and from a VM instance

Patch management
Schedule patch updates and view patch compliance on VM instances

Load balance between VMs
Set up Load Balancing for your applications as your traffic and users grow

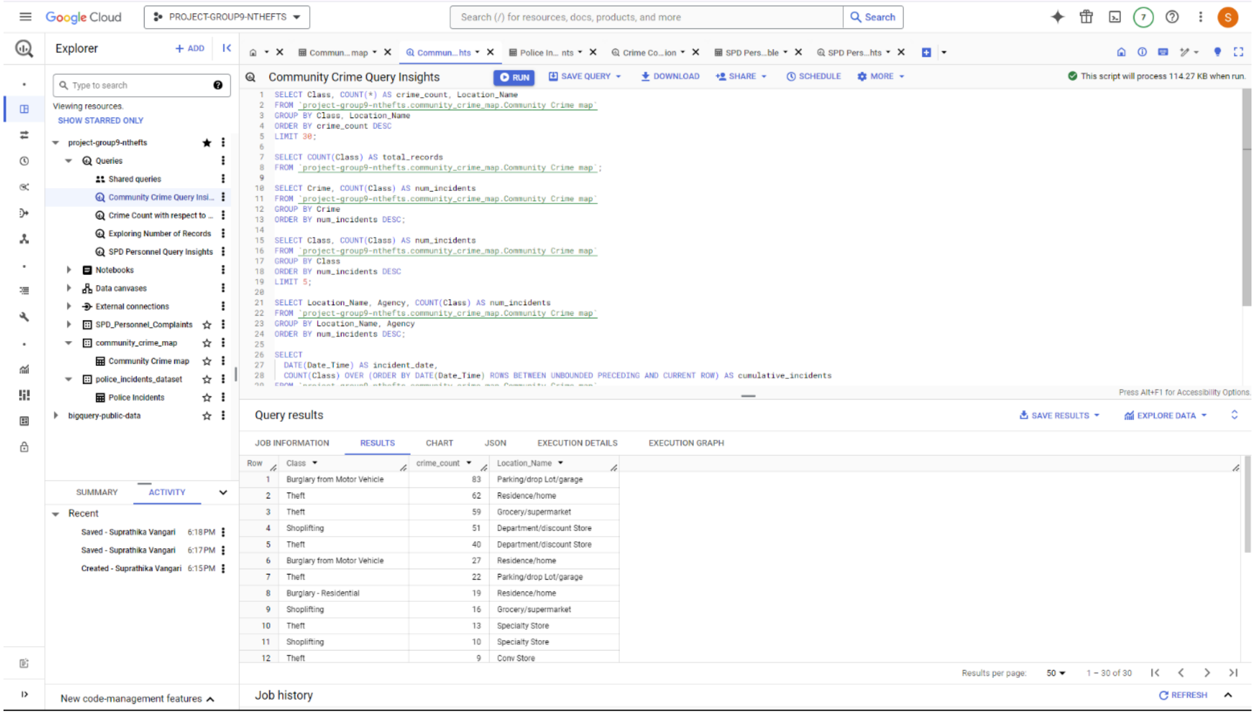
Screenshot Owner: Vishala Vadla

Once the cluster was created, we set up a secure connection to the manager node via an SSH terminal. This direct access allowed us to directly interact with the Hadoop and Spark ecosystems, execute data processing tasks, run queries, and perform advanced analytics.

7.2 Querying using BigQuery:

To manage the processed data, we used Google Cloud's BigQuery, a fully managed enterprise data warehouse. We created table schemas of the three datasets in BigQuery, while integrating it to the data stored in our Cloud Storage bucket. This will allow data analysts to run SQL queries on our data and gain insights into the neighborhood theft patterns to analyze factors such as theft hotspots, time patterns, and potential correlations with other variables.

Screenshot - Querying the Community Crime Map Data

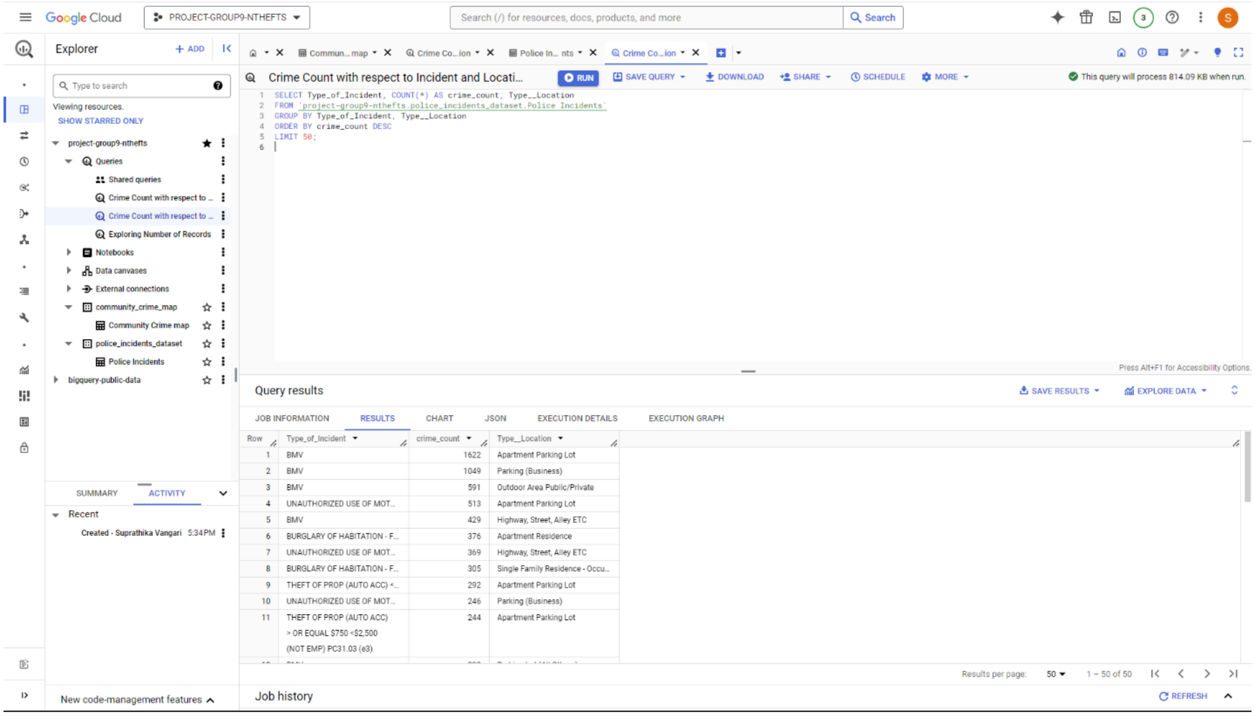


```sql
SELECT Class, COUNT(*) AS crime_count, Location_Name
FROM `project-group9-nthefts.community_crime_map.Community Crime map`
GROUP BY Class, Location_Name
ORDER BY crime_count DESC
LIMIT 30;

SELECT COUNT(Class) AS total_records
FROM `project-group9-nthefts.community_crime_map.Community Crime map`;

SELECT Crime, COUNT(Class) AS num_incidents
FROM `project-group9-nthefts.community_crime_map.Community Crime map`
GROUP BY Crime
ORDER BY num_incidents DESC;

SELECT Class, COUNT(Class) AS num_incidents
FROM `project-group9-nthefts.community_crime_map.Community Crime map`
GROUP BY Class
ORDER BY num_incidents DESC
LIMIT 5;

SELECT Location_Name, Agency, COUNT(Class) AS num_incidents
FROM `project-group9-nthefts.community_crime_map.Community Crime map`
GROUP BY Location_Name, Agency
ORDER BY num_incidents DESC;

SELECT
  DATE(Date_Time) AS incident_date,
  COUNT(Class) OVER (ORDER BY DATE(Date_Time) ROWS BETWEEN UNBOUNDED PRECEDING AND CURRENT ROW) AS cumulative_incidents
```

Row	Class	crime_count	Location_Name
1	Burglary from Motor Vehicle	83	Parking/drop Lot/garage
2	Theft	62	Residence/home
3	Theft	59	Grocery/supermarket
4	Shoplifting	51	Department/discount Store
5	Theft	40	Department/discount Store
6	Burglary from Motor Vehicle	27	Residence/home
7	Theft	22	Parking/drop Lot/garage
8	Burglary - Residential	19	Residence/home
9	Shoplifting	16	Grocery/supermarket
10	Theft	13	Specialty Store
11	Shoplifting	10	Specialty Store
12	Theft	9	Conv Store

Screenshot Owner: Suprathika Vangari

Screenshot - Querying the Police Incidents Dataset



```sql
SELECT Type_of_Incident, COUNT(*) AS crime_count, Type__Location
FROM `project-group9-nthefts.police_incidents_dataset.Police Incidents`
GROUP BY Type_of_Incident, Type__Location
ORDER BY crime_count DESC
LIMIT 50;
```

Row	Type_of_Incident	crime_count	Type__Location
1	BMV	1622	Apartment Parking Lot
2	BMV	1049	Parking (Business)
3	BMV	591	Outdoor Area Public/Private
4	UNAUTHORIZED USE OF MOT...	513	Apartment Parking Lot
5	BMV	429	Highway, Street, Alley ETC
6	BURGLARY OF HABITATION - F...	376	Apartment Residence
7	UNAUTHORIZED USE OF MOT...	369	Highway, Street, Alley ETC
8	BURGLARY OF HABITATION - F...	305	Single Family Residence - Occu...
9	THEFT OF PROP (AUTO ACC) <...	292	Apartment Parking Lot
10	UNAUTHORIZED USE OF MOT...	246	Parking (Business)
11	THEFT OF PROP (AUTO ACC) > OR EQUAL $750 <$2,500 (NOT EMP) PC31.03 (e3)	244	Apartment Parking Lot

Screenshot Owner: Suprathika Vangari

8. Advanced Analysis using Hive and Spark

In addition to BigQuery, we used Apache Spark and Apache Hive to perform advanced analysis on the datasets.

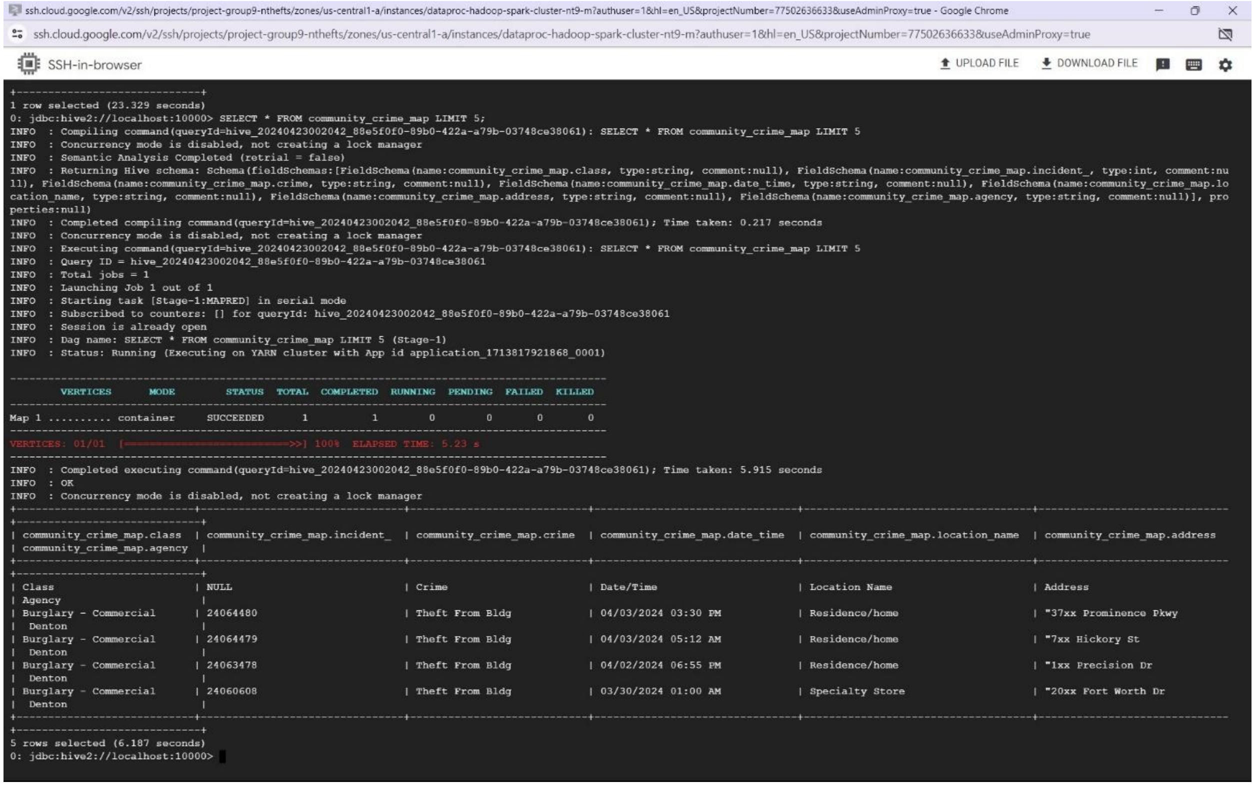
8.1 Querying in Hive

Using Hive, we defined a structured schema for the datasets and executed HQL (Hive Query Language) queries. We were able to perform complex data transformations using Hive's SQL-like interface.

Enabling Hive Access: After setting up the Dataproc cluster and moving the processed datasets to HDFS, the Hive environment was accessed through the beeline interface using the command **beeline -u jdbc: hive2://localhost:10000**

Creating Hive Tables: Hive tables were created to map the datasets stored in HDFS, allowing Hive to access and process the data in a manner like a relational database.

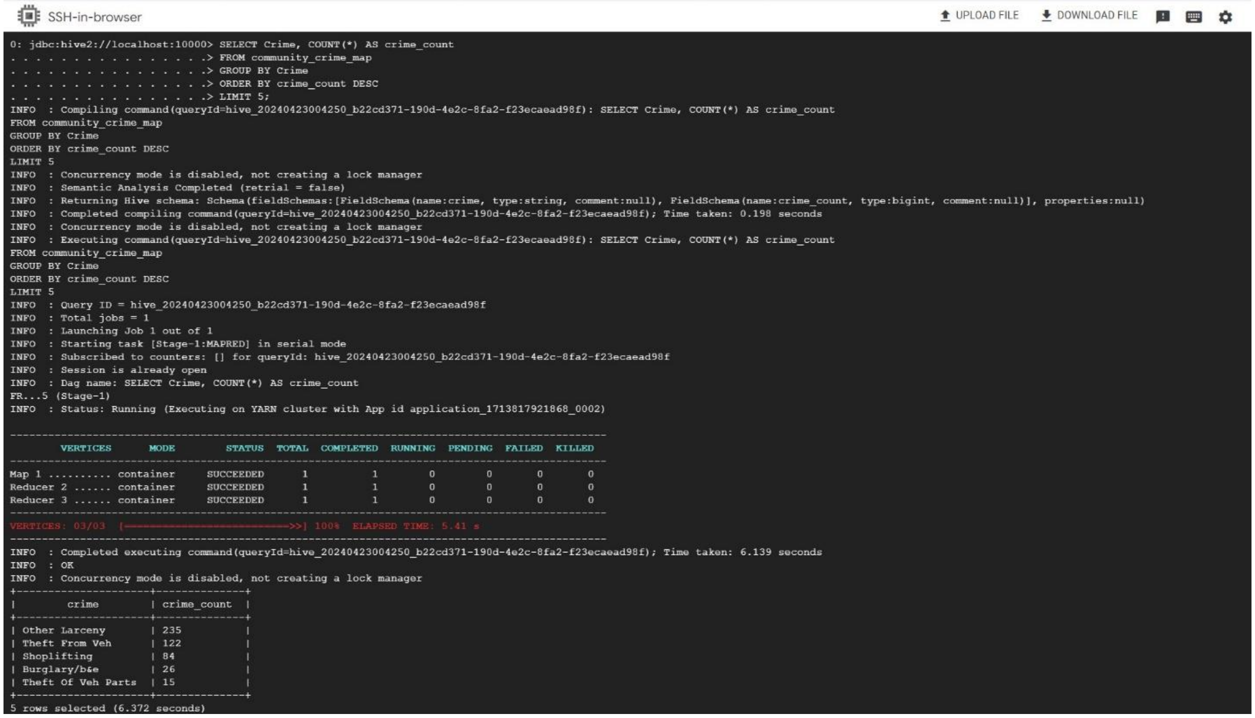
Screenshot - Creating Hive Tables


ssh.cloud.google.com/v2/ssh/projects/project-group9-nthefts/zones/us-central1-a/instances/dataproc-hadoop-spark-cluster-nt9-m?authuser=1&hl=en_US&projectNumber=77502636633&useAdminProxy=true - Google Chrome
ssh.cloud.google.com/v2/ssh/projects/project-group9-nthefts/zones/us-central1-a/instances/dataproc-hadoop-spark-cluster-nt9-m?authuser=1&hl=en_US&projectNumber=77502636633&useAdminProxy=true
SSH-in-browser
UPLOAD FILE
DOWNLOAD FILE

```
+-----------------------------+
1 row selected (23.329 seconds)
0: jdbc:hive2://localhost:10000> SELECT * FROM community_crime_map LIMIT 5;
INFO  : Compiling command(queryId=hive_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061): SELECT * FROM community_crime_map LIMIT 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:community_crime_map.class, type:string, comment:null), FieldSchema(name:community_crime_map.incident_, type:int, comment:nu
ll), FieldSchema(name:community_crime_map.crime, type:string, comment:null), FieldSchema(name:community_crime_map.date_time, type:string, comment:null), FieldSchema(name:community_crime_map.lo
cation_name, type:string, comment:null), FieldSchema(name:community_crime_map.address, type:string, comment:null), FieldSchema(name:community_crime_map.agency, type:string, comment:null)], pro
perties:null)
INFO  : Completed compiling command(queryId=hive_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061); Time taken: 0.217 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061): SELECT * FROM community_crime_map LIMIT 5
INFO  : Query ID = hive_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061
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_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061
INFO  : Session is already open
INFO  : Dag name: SELECT * FROM community_crime_map LIMIT 5 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1713817921868_0001)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 01/01  [==========================>>] 100%  ELAPSED TIME: 5.23 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240423002042_88e5f0f0-89b0-422a-a79b-03748ce38061); Time taken: 5.915 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+----------------------------+--------------------------------+----------------------------+--------------------------------+------------------------------------+------------------------------
+-----------------------------+
| community_crime_map.class  | community_crime_map.incident_  | community_crime_map.crime  | community_crime_map.date_time  | community_crime_map.location_name  | community_crime_map.address
| community_crime_map.agency  |
+----------------------------+--------------------------------+----------------------------+--------------------------------+------------------------------------+------------------------------
+-----------------------------+
| Class                      | NULL                           | Crime                      | Date/Time                      | Location Name                      | Address
| Agency                      |
| Burglary - Commercial      | 24064480                       | Theft From Bldg            | 04/03/2024 03:30 PM            | Residence/home                     | "37xx Prominence Pkwy
|  Denton                     |
| Burglary - Commercial      | 24064479                       | Theft From Bldg            | 04/03/2024 05:12 AM            | Residence/home                     | "7xx Hickory St
|  Denton                     |
| Burglary - Commercial      | 24063478                       | Theft From Bldg            | 04/02/2024 06:55 PM            | Residence/home                     | "1xx Precision Dr
|  Denton                     |
| Burglary - Commercial      | 24060608                       | Theft From Bldg            | 03/30/2024 01:00 AM            | Specialty Store                    | "20xx Fort Worth Dr
|  Denton                     |
+----------------------------+--------------------------------+----------------------------+--------------------------------+------------------------------------+------------------------------
+-----------------------------+
5 rows selected (6.187 seconds)
0: jdbc:hive2://localhost:10000>
```

Screenshot Owner: Udaya Adepu

Screenshot - Sample Hive Query and Results



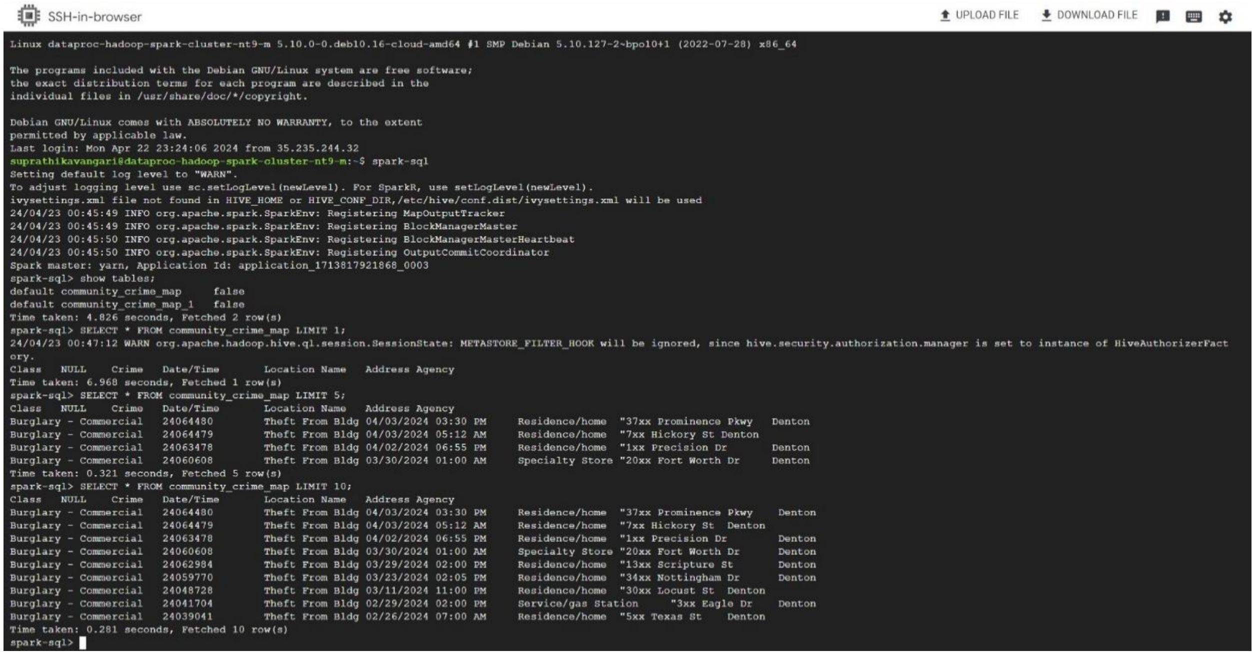
```
0: jdbc:hive2://localhost:10000> SELECT Crime, COUNT(*) AS crime_count
. . . . . . . . . . . . . . . . .> FROM community_crime_map
. . . . . . . . . . . . . . . . .> GROUP BY Crime
. . . . . . . . . . . . . . . . .> ORDER BY crime_count DESC
. . . . . . . . . . . . . . . . .> LIMIT 5;
INFO  : Compiling command(queryId=hive_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f): SELECT Crime, COUNT(*) AS crime_count
FROM community_crime_map
GROUP BY Crime
ORDER BY crime_count DESC
LIMIT 5
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Returning Hive schema: Schema(fieldSchemas:[FieldSchema(name:crime, type:string, comment:null), FieldSchema(name:crime_count, type:bigint, comment:null)], properties:null)
INFO  : Completed compiling command(queryId=hive_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f); Time taken: 0.198 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hive_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f): SELECT Crime, COUNT(*) AS crime_count
FROM community_crime_map
GROUP BY Crime
ORDER BY crime_count DESC
LIMIT 5
INFO  : Query ID = hive_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f
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_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f
INFO  : Session is already open
INFO  : Dag name: SELECT Crime, COUNT(*) AS crime_count
FR...5 (Stage-1)
INFO  : Status: Running (Executing on YARN cluster with App id application_1713817921868_0002)

----------------------------------------------------------------------------------------------
        VERTICES      MODE        STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
----------------------------------------------------------------------------------------------
Map 1 .......... container     SUCCEEDED      1          1        0        0       0       0
Reducer 2 ...... container     SUCCEEDED      1          1        0        0       0       0
Reducer 3 ...... container     SUCCEEDED      1          1        0        0       0       0
----------------------------------------------------------------------------------------------
VERTICES: 03/03  [==========================>>] 100%  ELAPSED TIME: 5.41 s
----------------------------------------------------------------------------------------------
INFO  : Completed executing command(queryId=hive_20240423004250_b22cd371-190d-4e2c-8fa2-f23ecaead98f); Time taken: 6.139 seconds
INFO  : OK
INFO  : Concurrency mode is disabled, not creating a lock manager
+--------------------+--------------+
|       crime        | crime_count  |
+--------------------+--------------+
| Other Larceny      | 235          |
| Theft From Veh     | 122          |
| Shoplifting        | 84           |
| Burglary/b&e       | 26           |
| Theft Of Veh Parts | 15           |
+--------------------+--------------+
5 rows selected (6.372 seconds)
```

Screenshot Owner: Udaya Adepu

8.2 Querying in Spark

On the other hand, Spark enabled us to develop custom queries to find and uncover relations between varied factors contributing to neighborhood thefts through the execution of queries.

Screenshot - Sample Spark Query and Results



```
Linux dataproc-hadoop-spark-cluster-nt9-m 5.10.0-0.deb10.16-cloud-amd64 #1 SMP Debian 5.10.127-2~bpo10+1 (2022-07-28) x86_64

The programs included with the Debian GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

Debian GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
Last login: Mon Apr 22 23:24:06 2024 from 35.235.244.32
suprathikavangari@dataproc-hadoop-spark-cluster-nt9-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
24/04/23 00:45:49 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
24/04/23 00:45:49 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
24/04/23 00:45:50 INFO org.apache.spark.SparkEnv: Registering BlockManagerMasterHeartbeat
24/04/23 00:45:50 INFO org.apache.spark.SparkEnv: Registering OutputCommitCoordinator
Spark master: yarn, Application Id: application_1713817921868_0003
spark-sql> show tables;
default community_crime_map     false
default community_crime_map_1   false
Time taken: 4.826 seconds, Fetched 2 row(s)
spark-sql> SELECT * FROM community_crime_map LIMIT 1;
24/04/23 00:47:12 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, since hive.security.authorization.manager is set to instance of HiveAuthorizerFact
ory.
Class   NULL    Crime   Date/Time       Location Name   Address Agency
Time taken: 6.968 seconds, Fetched 1 row(s)
spark-sql> SELECT * FROM community_crime_map LIMIT 5;
Class   NULL    Crime   Date/Time       Location Name   Address Agency
Burglary - Commercial   24064480        Theft From Bldg 04/03/2024 03:30 PM     Residence/home  "37xx Prominence Pkwy   Denton
Burglary - Commercial   24064479        Theft From Bldg 04/03/2024 05:12 AM     Residence/home  "7xx Hickory St Denton
Burglary - Commercial   24063478        Theft From Bldg 04/02/2024 06:55 PM     Residence/home  "1xx Precision Dr       Denton
Burglary - Commercial   24060608        Theft From Bldg 03/30/2024 01:00 AM     Specialty Store "20xx Fort Worth Dr     Denton
Time taken: 0.321 seconds, Fetched 5 row(s)
spark-sql> SELECT * FROM community_crime_map LIMIT 10;
Class   NULL    Crime   Date/Time       Location Name   Address Agency
Burglary - Commercial   24064480        Theft From Bldg 04/03/2024 03:30 PM     Residence/home  "37xx Prominence Pkwy    Denton
Burglary - Commercial   24064479        Theft From Bldg 04/03/2024 05:12 AM     Residence/home  "7xx Hickory St  Denton
Burglary - Commercial   24063478        Theft From Bldg 04/02/2024 06:55 PM     Residence/home  "1xx Precision Dr        Denton
Burglary - Commercial   24060608        Theft From Bldg 03/30/2024 01:00 AM     Specialty Store "20xx Fort Worth Dr      Denton
Burglary - Commercial   24062984        Theft From Bldg 03/29/2024 02:00 PM     Residence/home  "13xx Scripture St       Denton
Burglary - Commercial   24059770        Theft From Bldg 03/23/2024 02:05 PM     Residence/home  "34xx Nottingham Dr      Denton
Burglary - Commercial   24048728        Theft From Bldg 03/11/2024 11:00 PM     Residence/home  "30xx Locust St  Denton
Burglary - Commercial   24041704        Theft From Bldg 02/29/2024 02:00 PM     Service/gas Station     "3xx Eagle Dr    Denton
Burglary - Commercial   24039041        Theft From Bldg 02/26/2024 07:00 AM     Residence/home  "5xx Texas St    Denton
Time taken: 0.281 seconds, Fetched 10 row(s)
spark-sql> █
```

Screenshot Owner: Suprathika Vangari

We learned that Spark SQL could access the same Hive tables that were created earlier, as both Hive and Spark depend on the Hive Meta store part to understand the table data structure.

Comparison of Performances: We compared execution times for the same queries among Hive and Spark SQL. Our results prove that Spark SQL has better performance for certain types of queries. To conclude, Hive provided a familiar SQL-like interface for querying and processing the data, while Spark SQL offered improved performance and scalability for more complex queries and analyses.

9. Key findings and Conclusion

9.1 Summary of Key findings

By leveraging the power of Google Cloud Platform and employing best practices in data engineering, we have created a robust foundation for data analysts and scientists to uncover valuable insights and develop data-driven solutions. Integrated data sources serve as the basis for further exploration and discovery. The data processing, storage, and management techniques we have implemented ensure the reliability, accessibility, and scalability of the data, further empowering the data scientists to focus on advanced analysis.

9.2 Conclusion

In conclusion, we fulfilled our role as data engineers has been to set up a comprehensive and scalable data pipeline to support the analysis of neighborhood thefts.

Moving forward, we encourage the data analysts and scientists to fully utilize the data pipeline infrastructure we have established to delve deeper into finding patterns, trends, and correlations within the neighborhood theft data. By combining our data engineering expertise with their analytical skills and domain knowledge, we are confident that together, we can produce effective strategies to address the critical issue of neighborhood thefts.

10. REFERENCES

(i) Referred assignment instructions.

(ii) LinuxTrainingAcademy.com. (n.d.). LINUX COMMAND LINE CHEAT SHEET. https://www.LinuxTrainingAcademy.com

(iii) OpenRefine. (n.d.). Cleaning & Wrangling Data with OpenRefine. In OpenRefine User Guide (pp. 1–13).

(iv) Exploring facets | OpenRefine. (2024, April 29). https://openrefine.org/docs/manual/facets

(v) Editor. (2021b, June 7). Hadoop vs Spark: Main Big Data Tools Explained. AltexSoft. https://www.altexsoft.com/blog/hadoop-vs-spark/

(vi) BigQuery documentation | Google Cloud. (n.d.). Google Cloud. https://cloud.google.com/bigquery/docs

(vii) All BigQuery code samples | Google Cloud. (n.d.). Google Cloud. https://cloud.google.com/bigquery/docs/samples

(viii) Police incidents | Dallas Open Data. (2024, April 30). https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7/about_data

(ix) SPD Personnel Complaints (2021 to present). (n.d.). https://data.syr.gov/datasets/5b53f40abbc54c559ca991c96816cc17_0/about

(x) LexisNexis® Community Crime Map. (n.d.). https://communitycrimemap.com/