**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](https://myunt-my.sharepoint.com/:f:/g/personal/suprathikavangari_my_unt_edu/EgtirXwn6khLuZ-7rhpcsr0B2WE3UzFLx6uxkZ0lIJ6pXg?e=dSRloB)

**TABLE OF CONTENTS**

1. **Introduction**
   1. Problem Statement
   2. Importance
2. **Data Generation**
3. **Data Collection**
4. **Project Initialization in GCP**
   1. Infrastructure Setup
5. **Data Cleaning and Processing**
   1. Data Processing and Transformation using Open Refine.
6. **Data Storage**
   1. Storing data using Google Cloud Storage Buckets
7. **Data Management**
   1. Dataproc for provisioning Hadoop Infrastructure
   2. Querying in BigQuery
8. **Advanced-Data Analysis and Management**
   1. Querying in Apache Hive.
   2. Querying in Apache Spark.
9. **Key Findings and Conclusion**
   1. Summary of Key Findings
   2. Conclusion
10. **References**

**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**

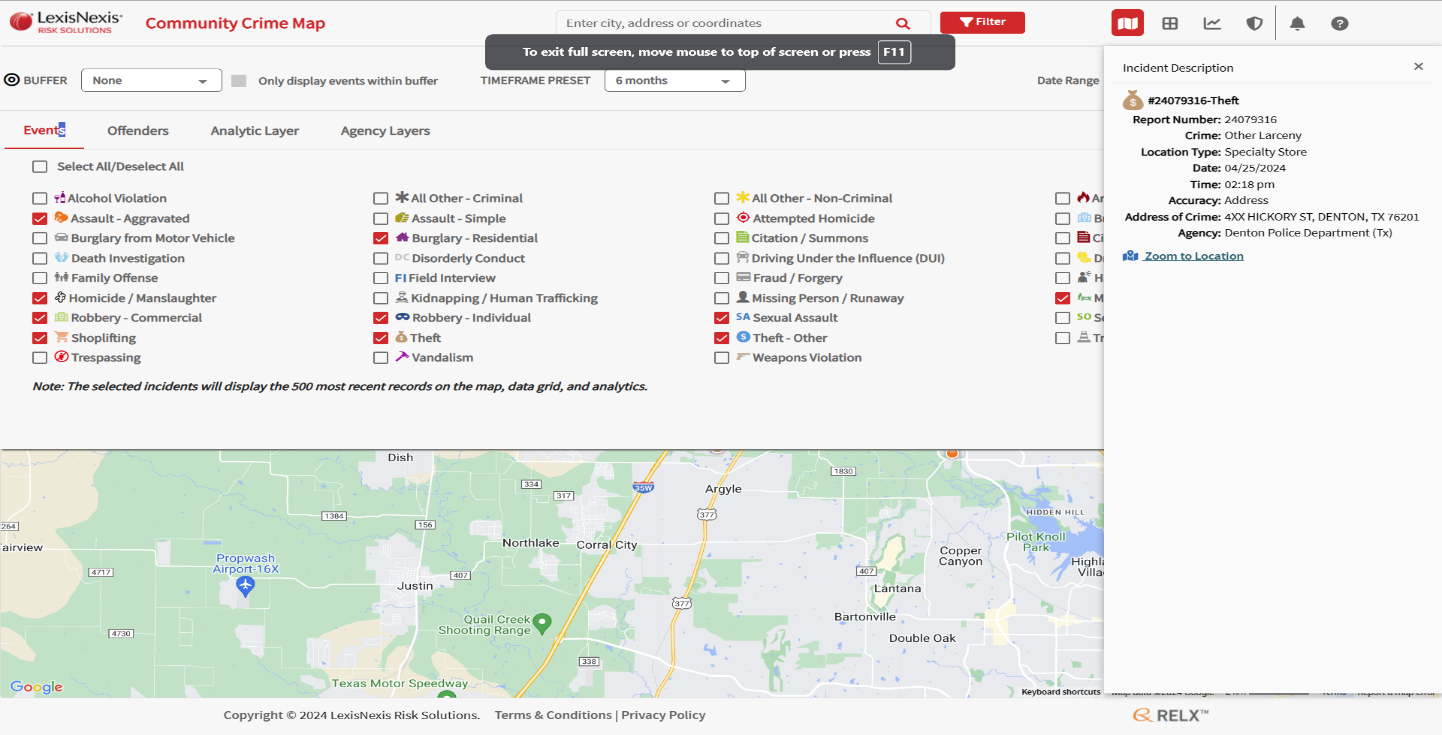
The process of Data Generation is a crucial beginning phase in our project. The first step in our data generation approach was to identify relevant data sources that could shed light on the potential theft patterns.

**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:



Screenshot Owner: Harish Inavolu

A screenshot of a computer

Description automatically generated

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.

A screenshot of a computer

Description automatically generated  
Screenshot Owner: Harish Inavolu

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

A screenshot of a computer

Description automatically generated

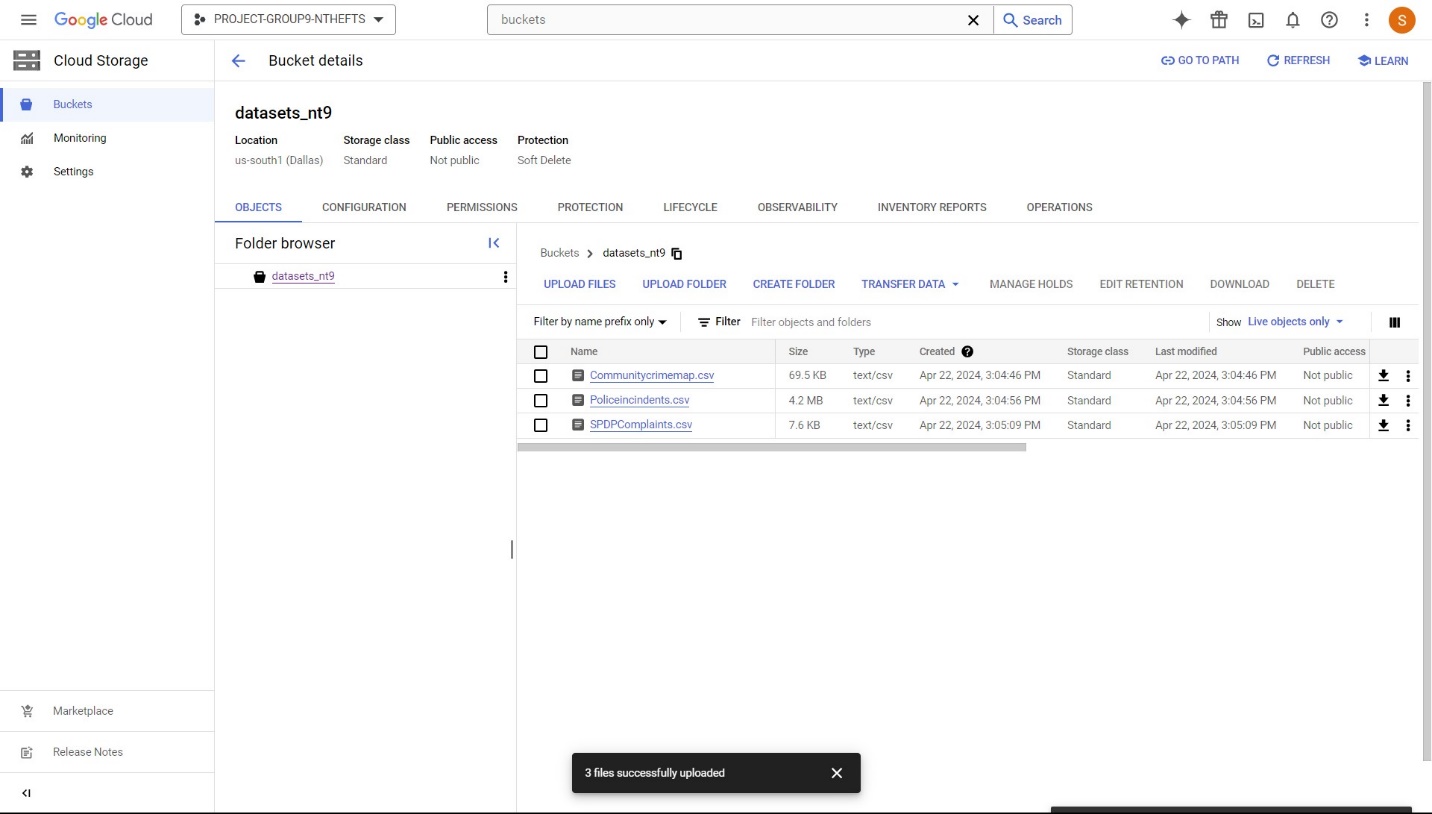
Screenshot Owner: Harish Inavolu

**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

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.

A screenshot of a computer

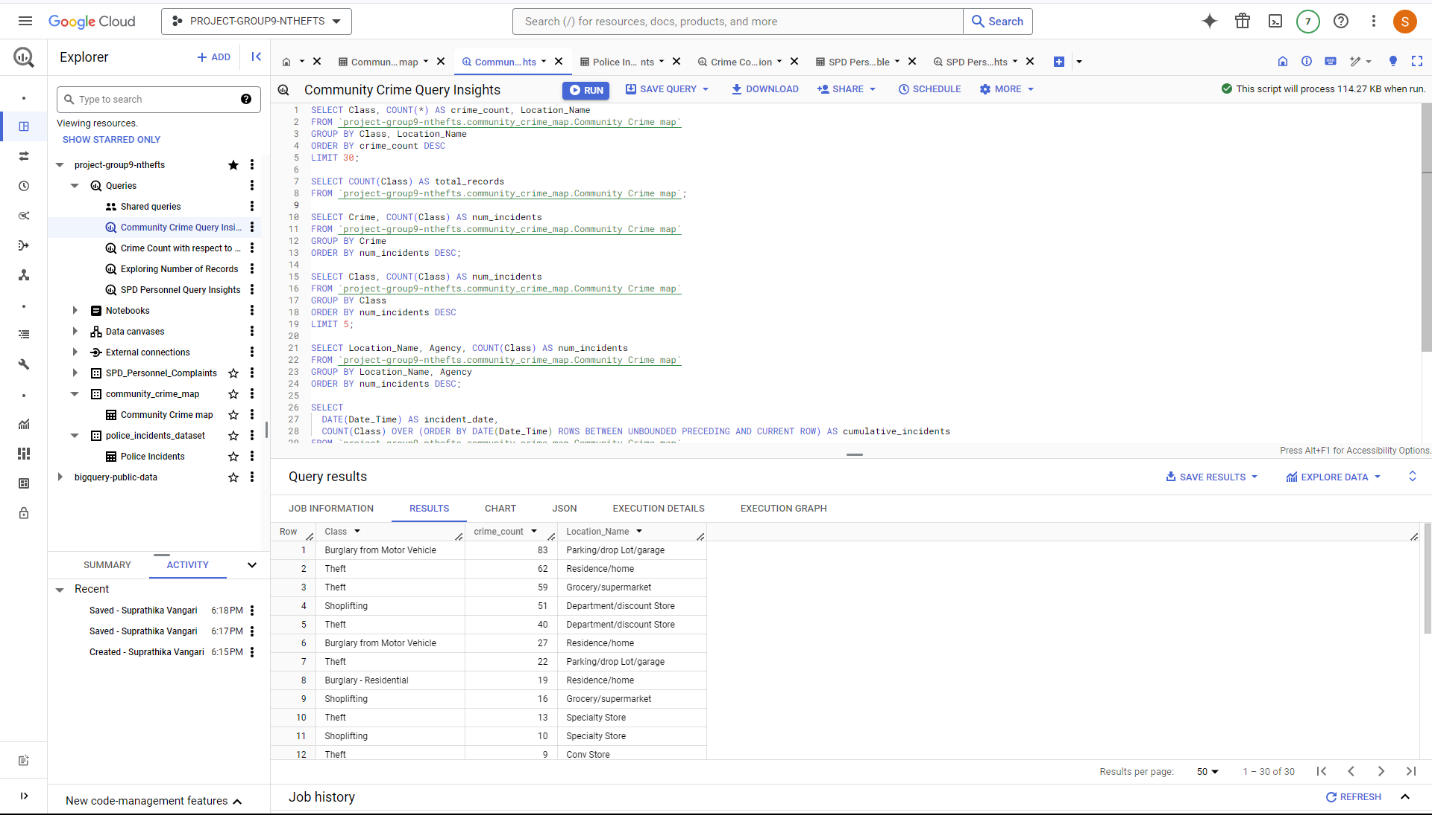
Description automatically generatedScreenshot Owner: Vishala Vadla

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.

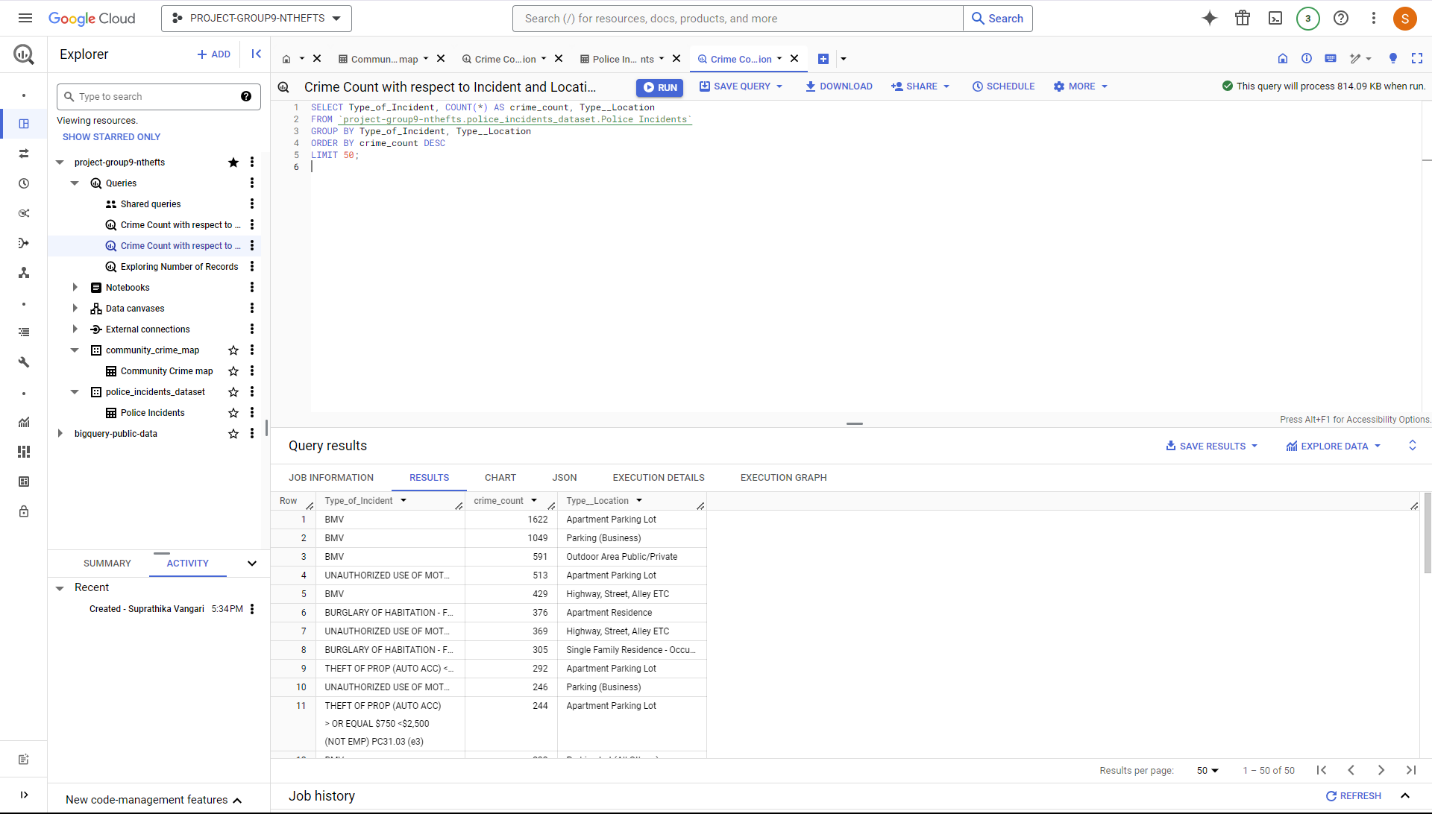
7.2 Querying using BigQuery:

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

Screenshot Owner: Suprathika Vangari

Screenshot - Querying the Police Incidents Dataset

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

A computer screen shot of a computer screen

Description automatically generatedScreenshot Owner: Udaya Adepu

Screenshot - Sample Hive Query and Results

A screenshot of a computer program

Description automatically generated

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.

A computer screen shot of a black screen

Description automatically generatedScreenshot - Sample Spark Query and Results

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**

1. Referred assignment instructions.
2. LinuxTrainingAcademy.com. (n.d.). LINUX COMMAND LINE CHEAT SHEET. <https://www.LinuxTrainingAcademy.com>
3. OpenRefine. (n.d.). Cleaning & Wrangling Data with OpenRefine. In OpenRefine User Guide (pp. 1–13).
4. Exploring facets | OpenRefine. (2024, April 29). <https://openrefine.org/docs/manual/facets>
5. Editor. (2021b, June 7). Hadoop vs Spark: Main Big Data Tools Explained. AltexSoft. <https://www.altexsoft.com/blog/hadoop-vs-spark/>
6. BigQuery documentation | Google Cloud. (n.d.). Google Cloud. https://cloud.google.com/bigquery/docs
7. All BigQuery code samples | Google Cloud. (n.d.). Google Cloud. <https://cloud.google.com/bigquery/docs/samples>
8. Police incidents | Dallas Open Data. (2024, April 30). <https://www.dallasopendata.com/Public-Safety/Police-Incidents/qv6i-rri7/about_data>
9. SPD Personnel Complaints (2021 to present). (n.d.). <https://data.syr.gov/datasets/5b53f40abbc54c559ca991c96816cc17_0/about>
10. LexisNexis® Community Crime Map. (n.d.). <https://communitycrimemap.com/>