**ADTA 5240**-**Harvesting, Storing, and Retrieving Data**

**Professor - Schenita Floyd**

A green and black logo

Description automatically generated

**Lack of public space for Teenagers**

**Group-7**

**Team Members:**

1. **Sai Prasanth Mantha**
2. **Saiveeravenkata Ramakrishnachowdary Obinni**
3. **Srikanth Peethani**
4. **Charan Rekala**

**To:** North Dallas Community Development Network (NDCDN)

**Cc:** Collin Allred, North Dallas Community Leaders

**Bcc:** Prasanth, Ramakrishna, Srikanth, Charan

**Subject:** Addressing the Lack of Public Space for Teenagers: Data Insights and March Update

Dear Colleagues,

After weeks of hard work, the team of data analysis is proud of going a huge step towards solving the pressing problem: a shortage of adolescent-friendly public areas in North Dallas. We have collected, cleaned, and tabulated data from two key sources-community parks and recreation centers-to give a complete set of datasets reflecting both the availability and use of the public spaces at this day and age.

The first datasets one contains information about amenities in parks, including cultural centers, gyms, and outdoor sports fields. The second dataset covers location, accessibility features, and the operating status of fitness centers, recreation centers, and pools. These two together represent infrastructure for both teens and the greater community.

This data has been cleaned and saved in an affordable and secured platform for further planning, analysis, and visualization. Highlights include:

* Finding the Gaps: Where there is no access, or limited access, to public recreational spaces.
* Actual Utilization Patterns: Allocated usage data that enables better resource utilization.
* Improvement Opportunities: Information on underutilized spaces that could be renovated or repurposed to promote teen engagement.

We have also provided the link to the data sets and some screenshot views of key findings to your further use in subsequent research phases and strategy development. Those tools will help you derive concrete ways of how the standard and accessibility of public areas for young people in North Dallas can be improved.

Please do not hesitate to contact us for further assistance or in case of any queries. Together, we can give our younger citizens the gift of a more involved and integrated community when we all do our part.

Sincerely,  
Prasanth, Ramakrishna, Srikanth, Charan

**PROJECT IMPLEMENTATION**

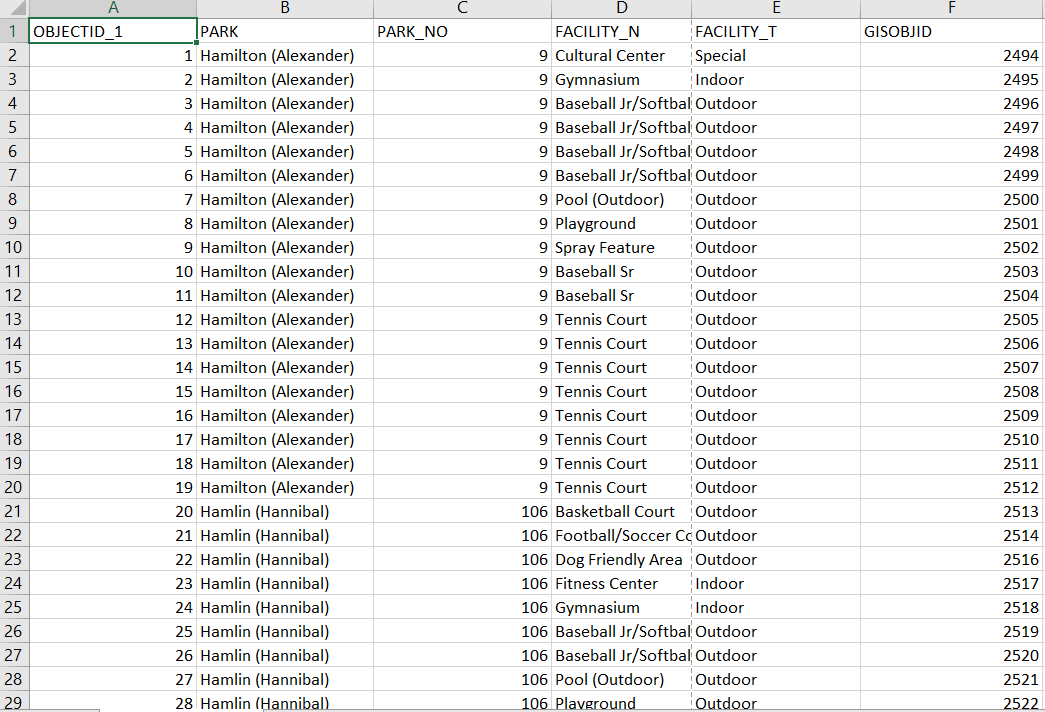
**Introduction**

The availability and accessibility of recreation facilities play a vital role in fostering community well-being and engagement. This report focuses on analyzing recreation facilities data to understand patterns, trends, and gaps in the distribution and availability of such spaces. By leveraging advanced data processing tools such as Google Cloud Platform (GCP), Hive, Spark, and OpenRefine, we aim to derive actionable insights that can contribute to optimizing the planning and utilization of these facilities.

**Data Generation**

**Parks - Chicago Park District Facilities (Static)**

* **Source**: The Chicago Park District Facilities dataset, available on the City of Chicago's open data platform.
* **Content**: This dataset provides comprehensive information about park facilities within the Chicago Park District, including locations, types of amenities available, and their respective details.
* **Frequency**: The dataset is updated regularly to ensure that it reflects current information about park facilities.
* **Utility**: The dataset is useful for understanding the distribution and accessibility of recreational facilities in Chicago, planning community events, and analysing urban green space availability.

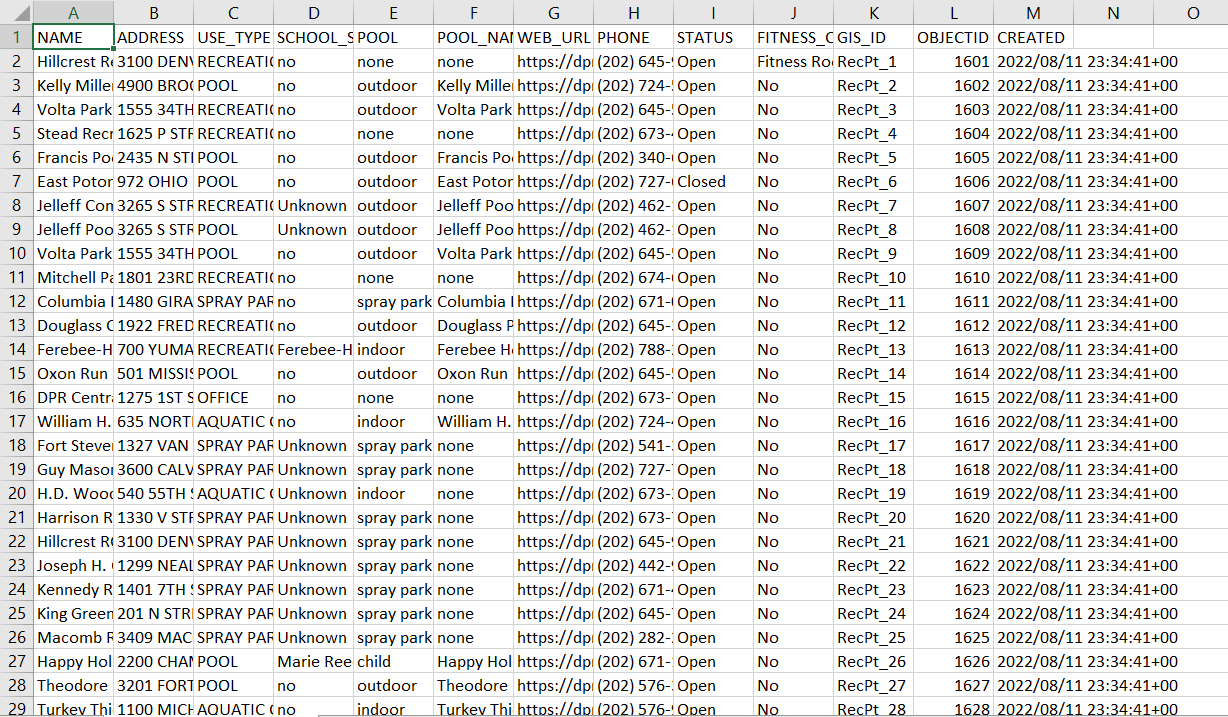
****

**Screenshot Owner :** Ramakrishna Chowdary

**Link:** [**https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Facilities-current-/**](https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Facilities-current-/5yyk-qt9y)

**Washington, D.C. Open Data - Real-Time Geospatial Dataset Explorer(Streaming)**

* **Source**: Washington, D.C.'s Open Data platform, providing a real-time geospatial dataset accessible through an interactive map explorer.
* **Content**: This dataset includes dynamic geographic data layers that feature real-time updates on locations, attributes, and spatial relationships within Washington, D.C. Users can explore real-time changes and updates in various elements, such as infrastructure, public services, or natural features, using the interactive interface.
* **Frequency**: The dataset is updated in real-time to ensure it reflects the most current geospatial information available.
* **Utility**: This real-time tool is indispensable for monitoring ongoing urban development projects, managing civic planning activities, conducting real-time spatial analysis, and enhancing community engagement by offering an up-to-date visualization of data in a map-based format.

****

**Screenshot Owner :** Ramakrishna Chowdary

**Link:** [**https://opendata.dc.gov/datasets/Recreation\_Facilities**](https://opendata.dc.gov/datasets/7122c1c815314588abe5c1864da8a355_3/explore?location=39.246960%2C-76.077342%2C7.24)

**Data Collection**

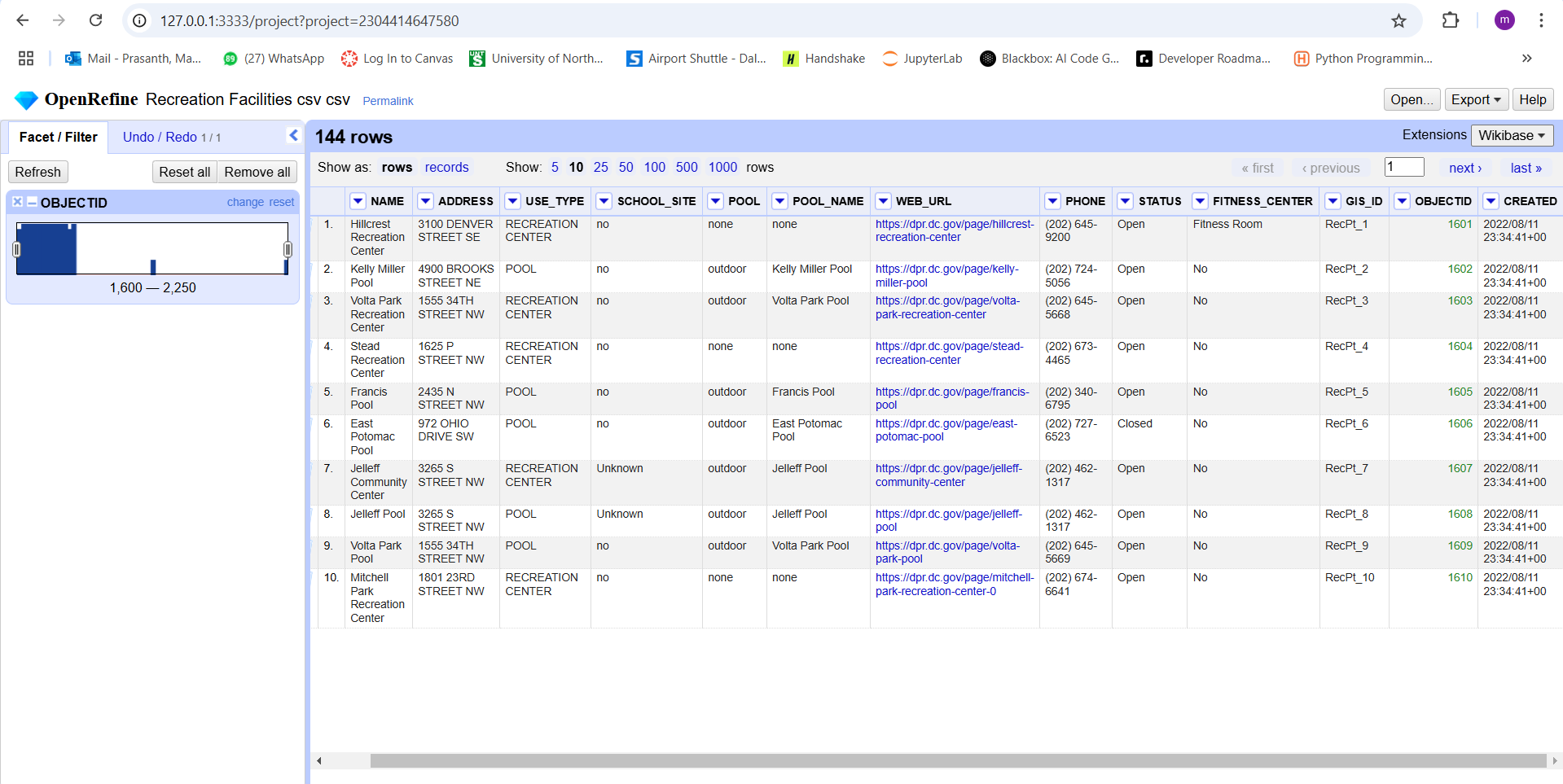
**Real-Time Dataset:** The real-time dataset is continuously collected and updated from the online government database through the use of automated scripts. These scripts ensure that the data remains current and accessible by regularly scraping and uploading new information. This process allows for a seamless flow of up-to-date data for real-time analysis**.**

**Static Dataset :** The dataset, sourced from Chicago's Open Data platform, provides a snapshot of park facilities, including locations and amenities. It is static, reflecting a specific point in time, and is periodically updated for analysis and planning**.**

**Data Processing**

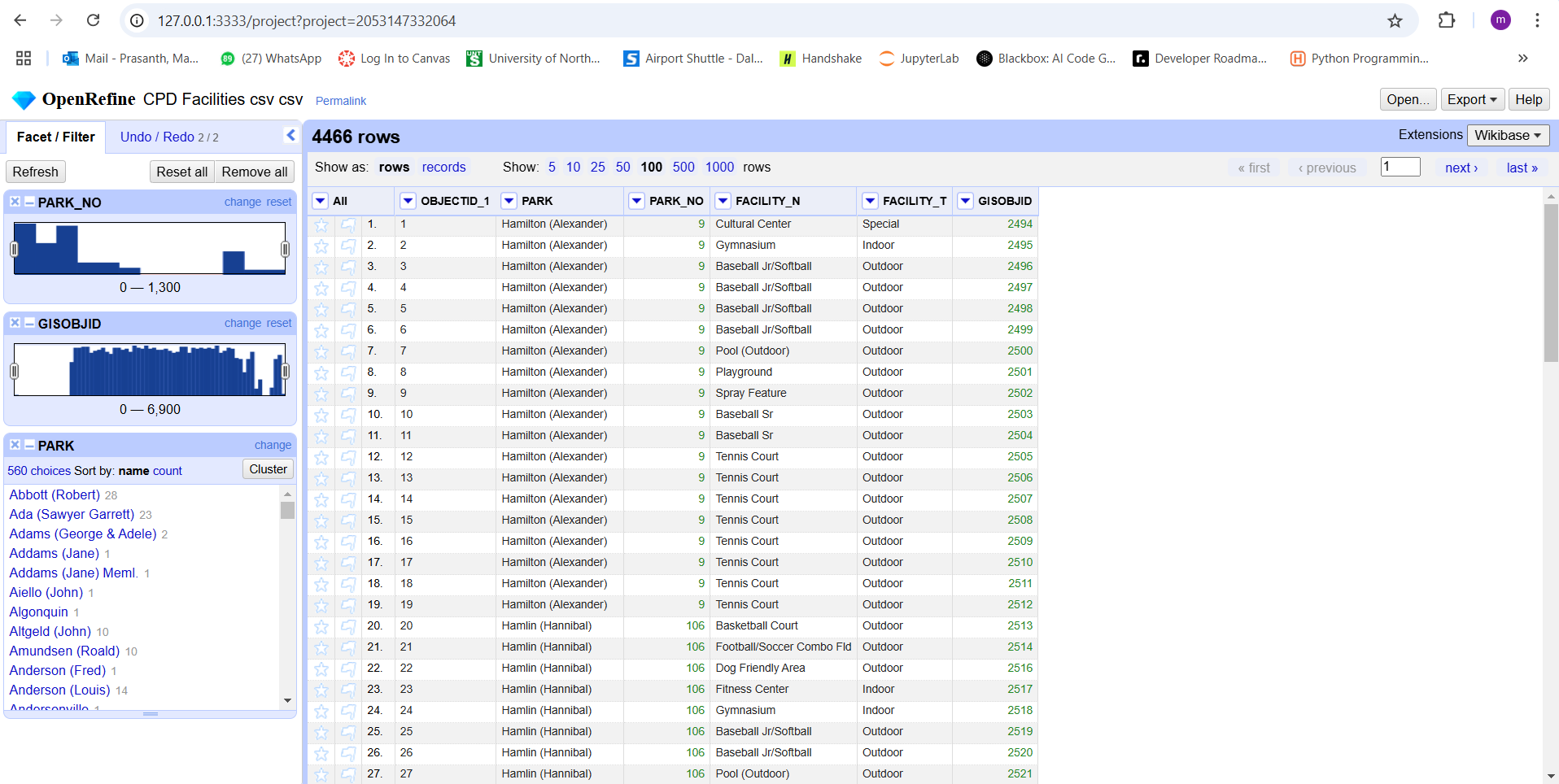
We could not afford GCP DataPrep by Trifacta, so we will use OpenRefine, which is free. So, in our real-time dataset, we discovered that the "OBJECTID" were stored in text format, which is not suitable for numerical analysis. To address this, we applied the function value.toNumber() to convert the zip codes into numeric values, enabling further analysis and better handling of the data.

Some entries in the **School\_Site** column were blank, indicating missing data. To ensure consistency and avoid gaps during analysis, these blank entries were replaced with the value "Unknown." This makes it clear that the status of the school site is not available, rather than leaving it ambiguous.



**Screenshot Owner :** Srikanth Peethani

In another dataset, the **PARK\_NO** and **GISOBJID** columns were originally stored as text values. To make them more useful for analysis, we converted them into numeric values in “**Open Refine**”. This is important because it allows us to perform numerical operations, such as sorting, filtering, or calculating statistics, which wouldn’t be possible if they remained as text.



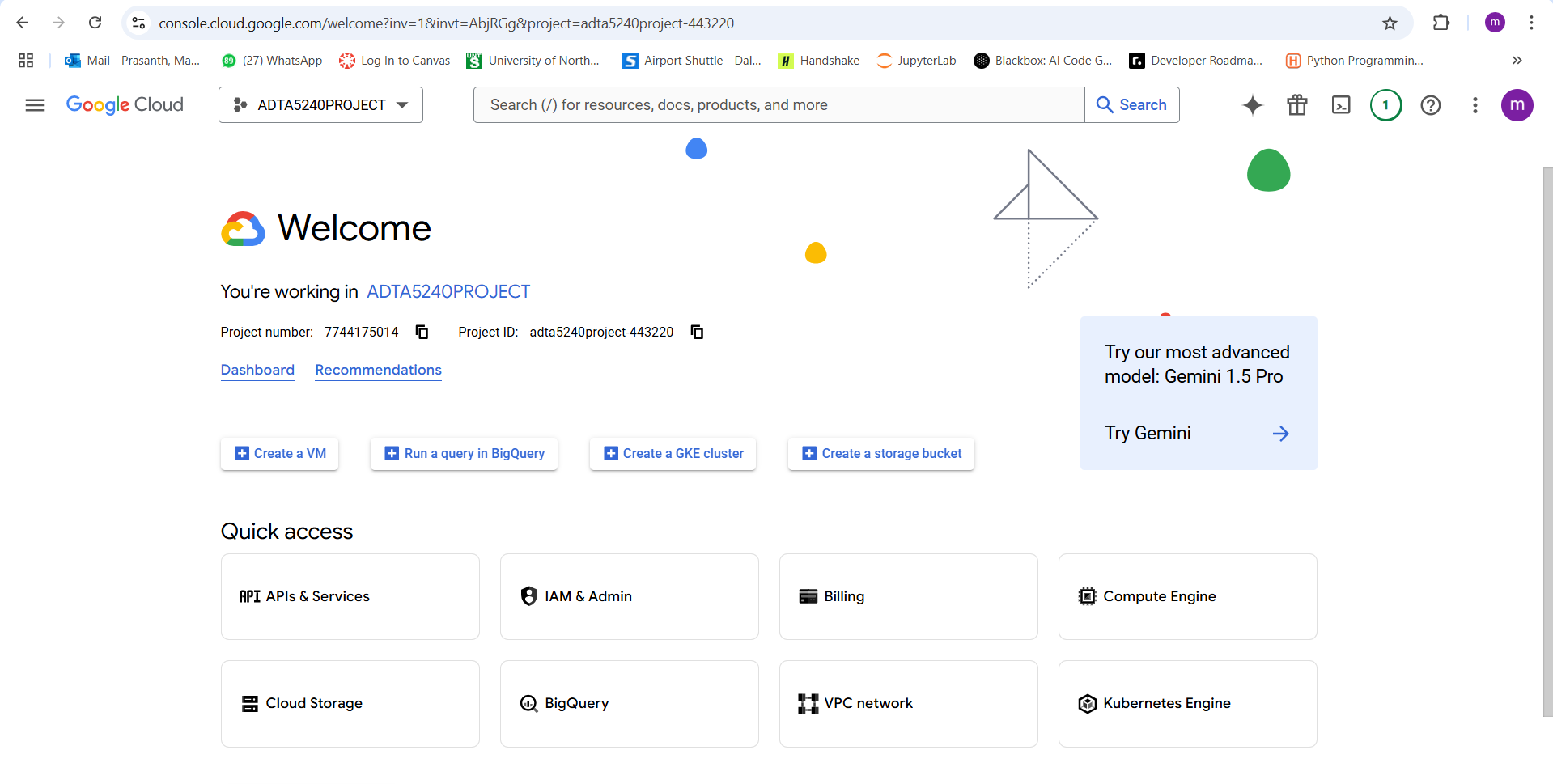
**Screenshot Owner :** Srikanth Peethani

**Data Storage**

To address the lack of public spaces for teenagers, we utilized **Google Cloud Platform (GCP)** to analyze the problem and develop data-driven solutions. GCP's suite of tools and global infrastructure enabled us to efficiently process and analyze data, helping us better understand the gaps in public facilities for teenagers.

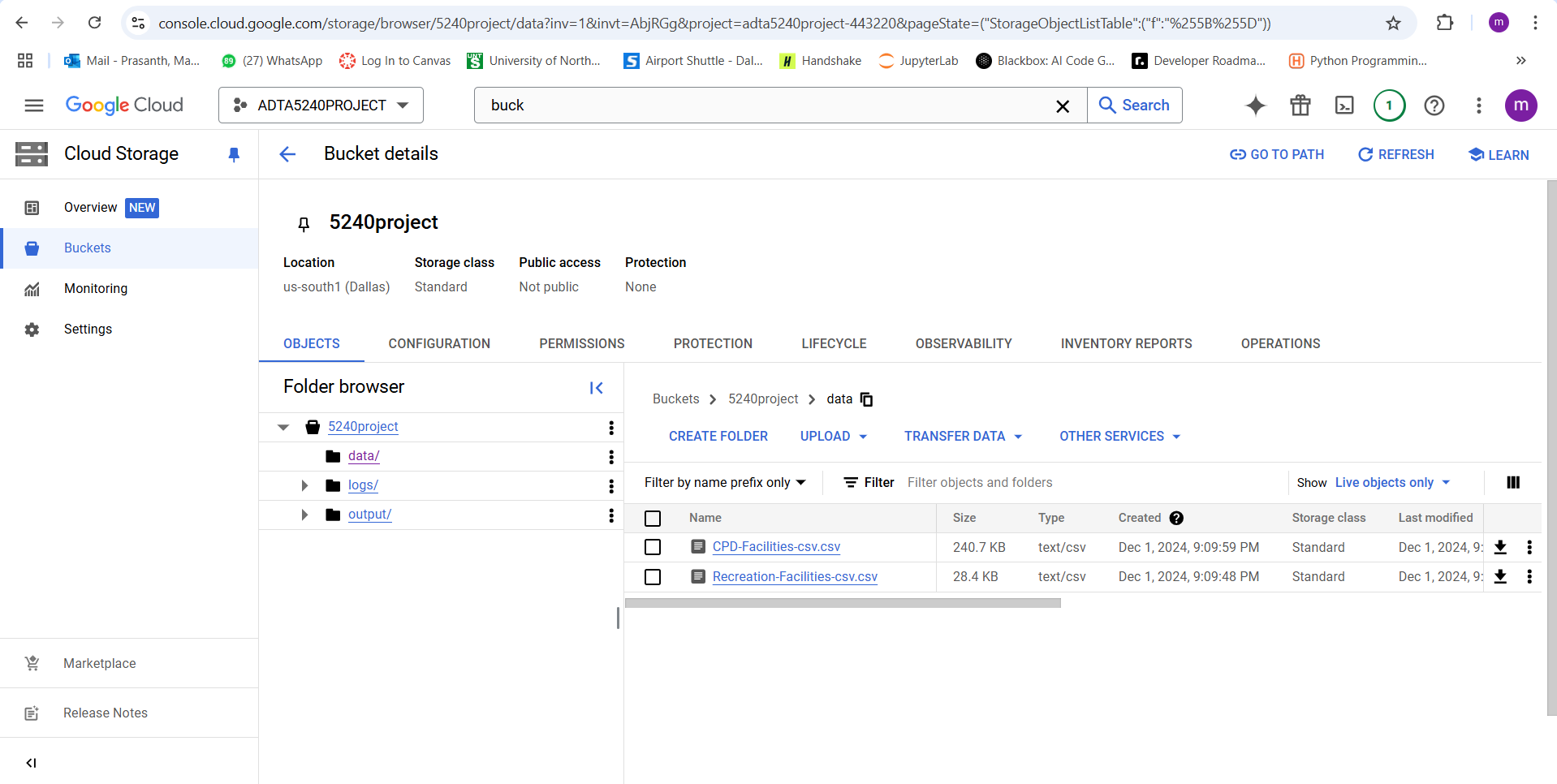
We created a dedicated project in GCP to organize our resources and streamline the collaboration process. Using **Google Dataproc**, a managed service for running Apache Hadoop and Spark, we built a scalable data processing infrastructure. This allowed us to analyze large datasets, including demographic information, facility usage data, and geographic distributions of existing public spaces.

**Project Name:** ADTA5240PROJECT



**Screenshot Owner :** Sai Prasanth Mantha

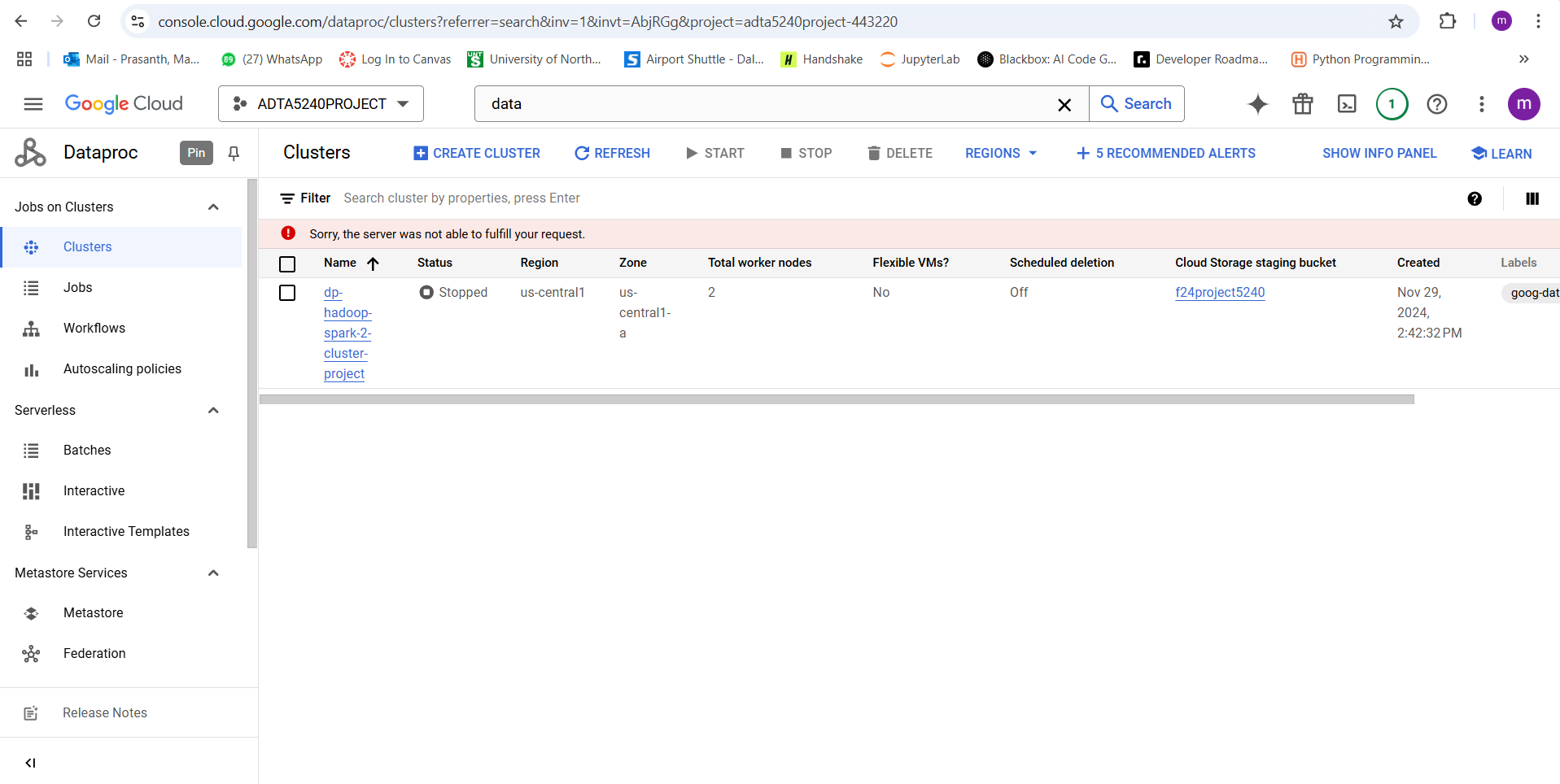
**Bucket Creation and Uploading Data:**



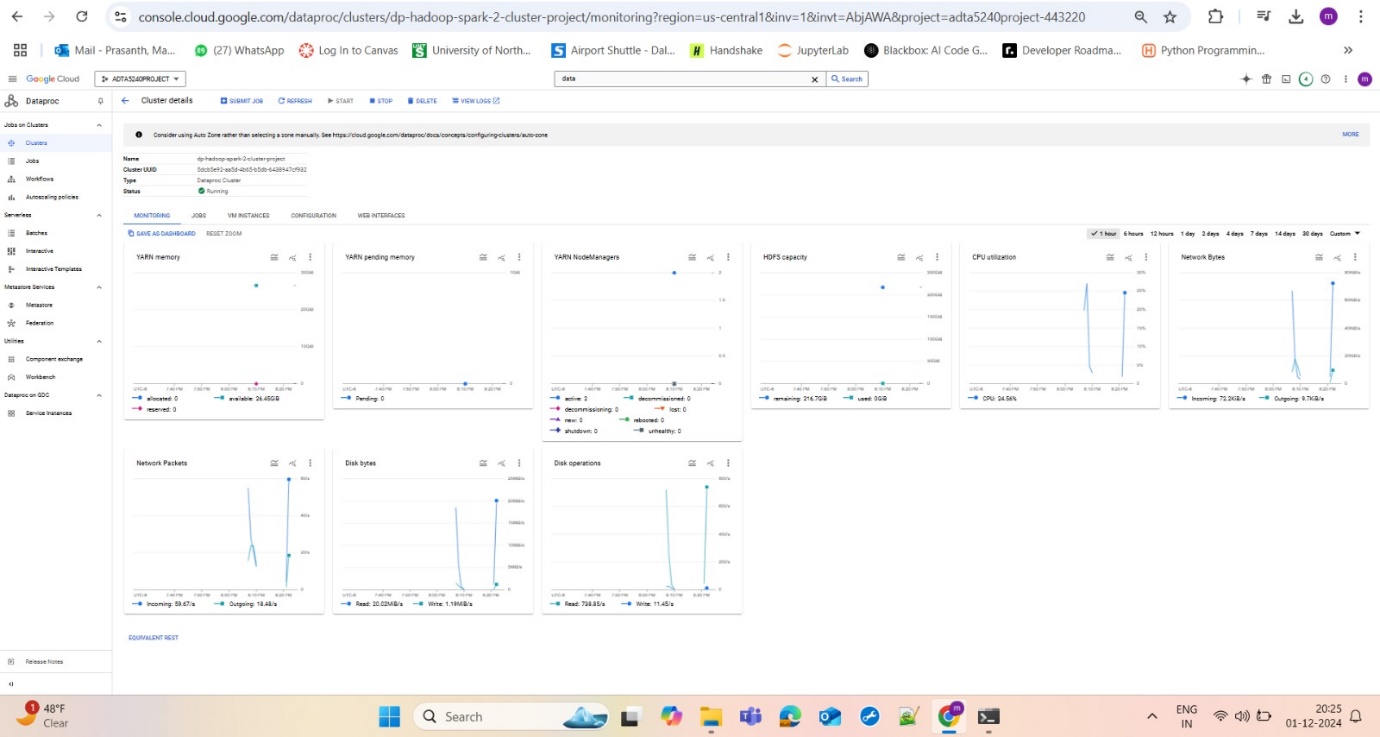
**Screenshot Owner :** Sai Prasanth Mantha

**Creating Clusters:**

Because clusters have so many advantages, they are a popular option in GCP. They provide fault tolerance, scalability, high availability, cost-effectiveness, security, and flexibility. Clusters lower latency, increase application performance, and adhere to rules by dividing tasks among several nodes. Additionally, they are adaptable to certain company requirements.

****

**Screenshot Owner :** Sai Prasanth Mantha



**Screenshot Owner :** Sai Prasanth Mantha

**DATA MANAGEMENT:**

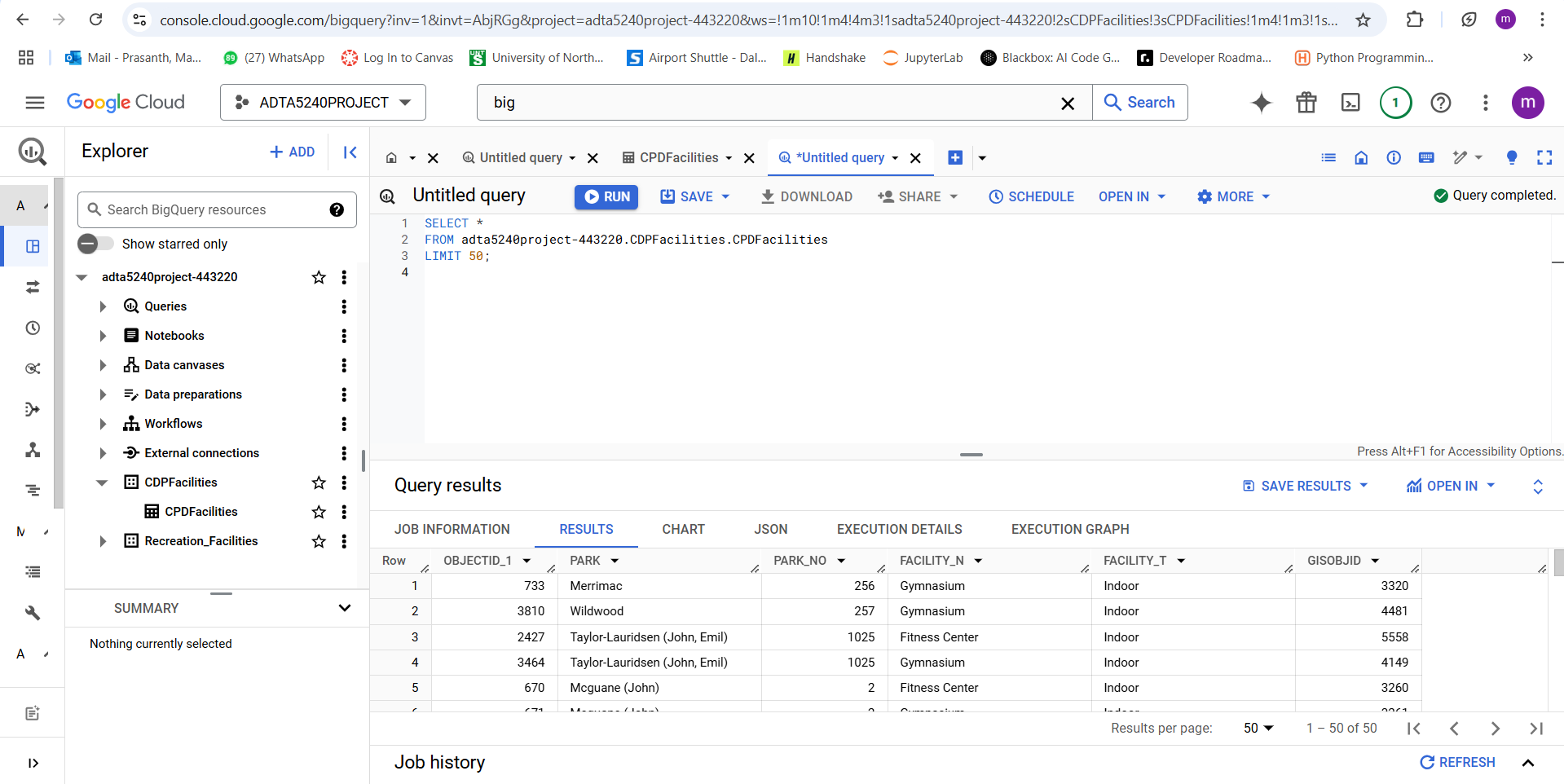
The collected datasets have been uploaded to Google Cloud Platform (GCP) and stored in the standard storage class. To ensure data security and compliance, access control and encryption measures have been implemented. This facilitates efficient analysis of the data using Big Query, Hive, and Spark, fostering informed decision-making, and maximizing productivity**.**

**DATA ANALYSIS:**

Data Analysis is done in BIG QUERY, HIVE AND SPARK.

**Big Query**

Uploaded the **Static Dataset** in Big Query.



**Screenshot Owner** : Sreekanth Peethani

**Query Analysis in Big Query**

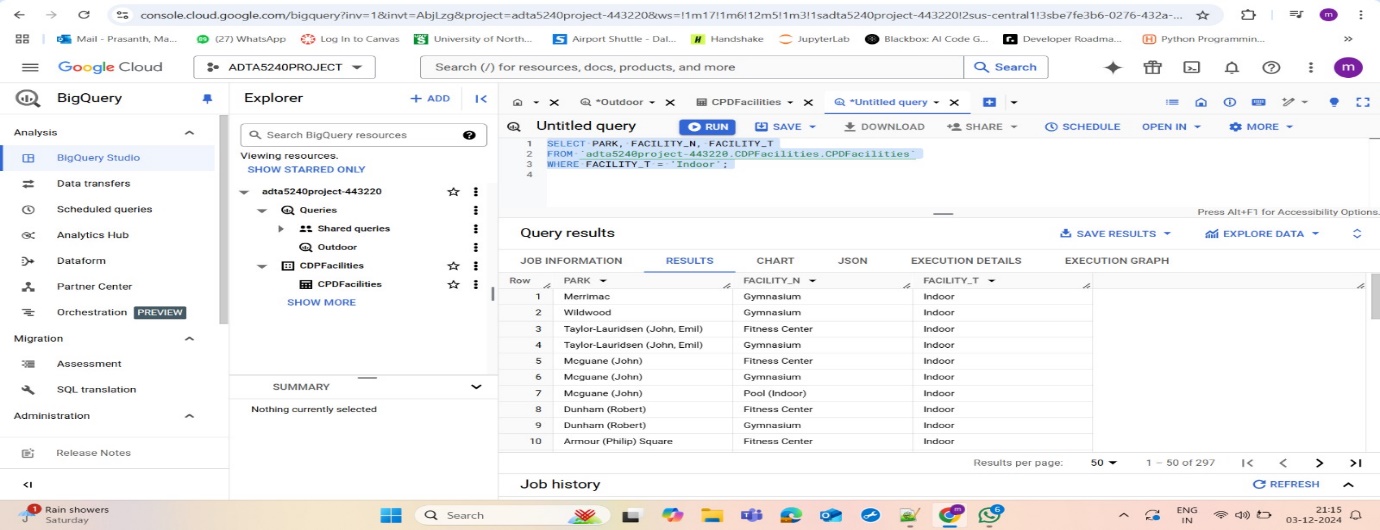
**QUERY 1:**

The below query could be used for data analysis to identify the parks and facilities that are categorized as indoor facilities. It retrieves the park names, facility names, and facility types to help identify the distribution and availability of indoor spaces in the dataset.

SELECT PARK, FACILITY\_N, FACILITY\_T

FROM adta5240project-443220.CDPFacilities.CPDFacilities

WHERE FACILITY\_T = 'Indoor';



**Screenshot Owner:** Ramakrishna Chowdary

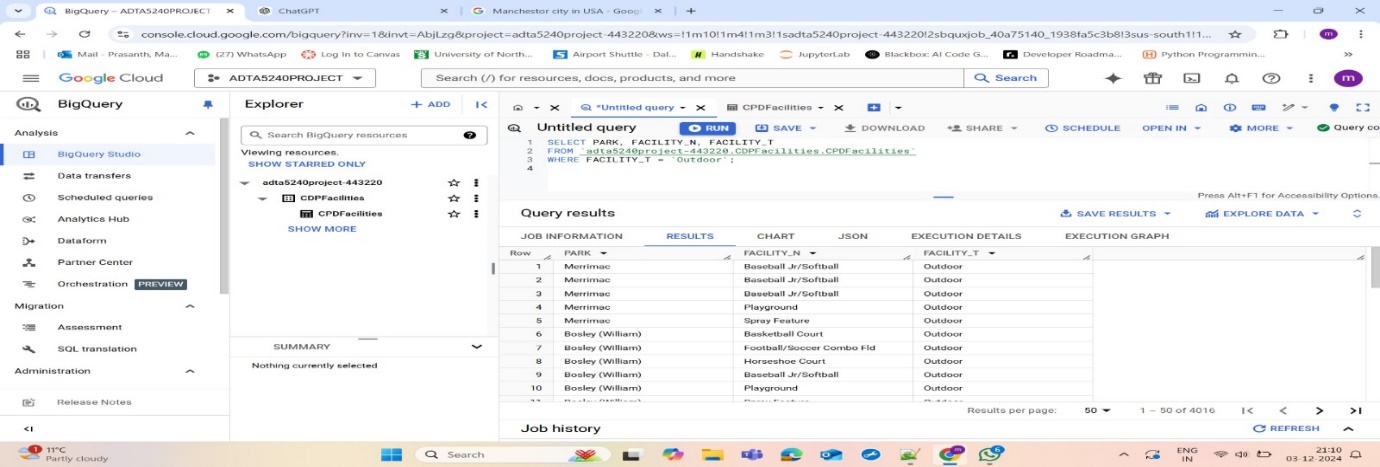
**Query 2:**

The below query could be used for data analysis to identify the parks and facilities that are categorized as outdoor facilities. It retrieves the park names, facility names, and facility types to help understand the distribution and availability of outdoor spaces in the dataset.

SELECT PARK, FACILITY\_N, FACILITY\_T

FROM adta5240project-443220.CDPFacilities.CPDFacilities

WHERE FACILITY\_T = 'Outdoor';



**Screenshot Owner :** Charan Rakela

**Query 3:**

The below query could be used for data analysis to identify the parks with the highest number of outdoor facilities. It counts the number of outdoor facilities for each park, organizes the data by park, and sorts the results in descending order to highlight parks with the most outdoor facilities.

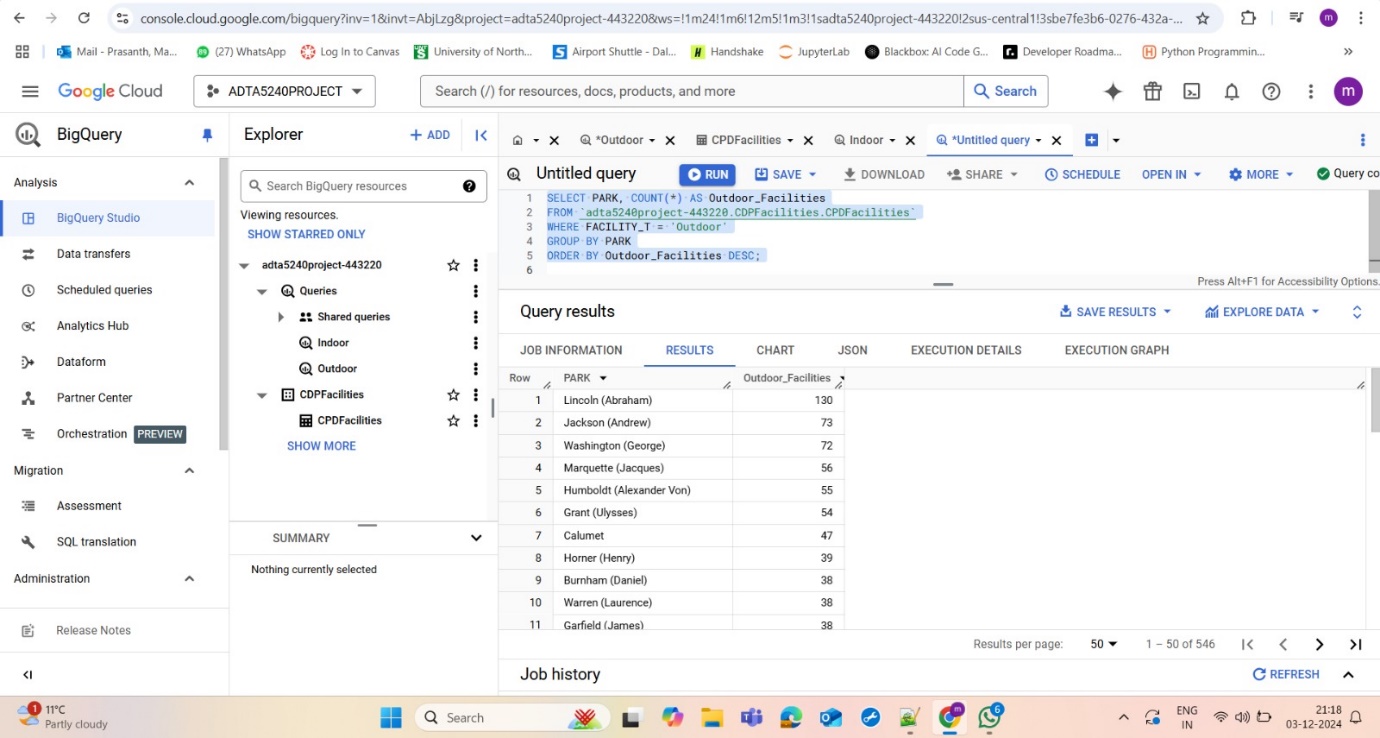
SELECT PARK, COUNT(\*) AS Outdoor\_Facilities

FROM adta5240project-443220.CDPFacilities.CPDFacilities

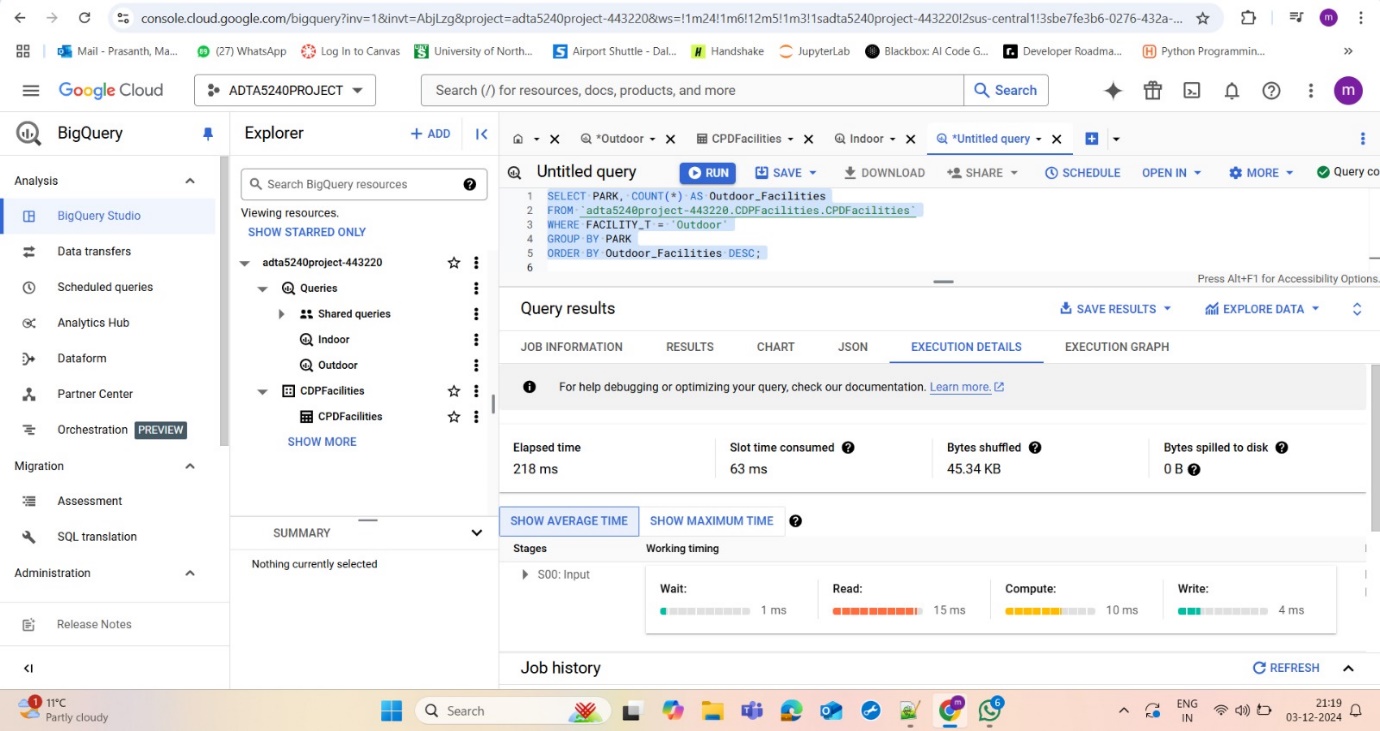
WHERE FACILITY\_T = 'Outdoor'

GROUP BY PARK

ORDER BY Outdoor\_Facilities DESC;



**Screenshot Owner :** Charan Rakela



**Screenshot Owner :** Charan Rakela

Time taken to run the query in Big Query is 218ms.

Now, we will run the same query in Hive & Spark.

**HADOOP ENVIRONMENT**:

Creation of Hadoop Environment for Hive and Spark.

**Query in Hive:**

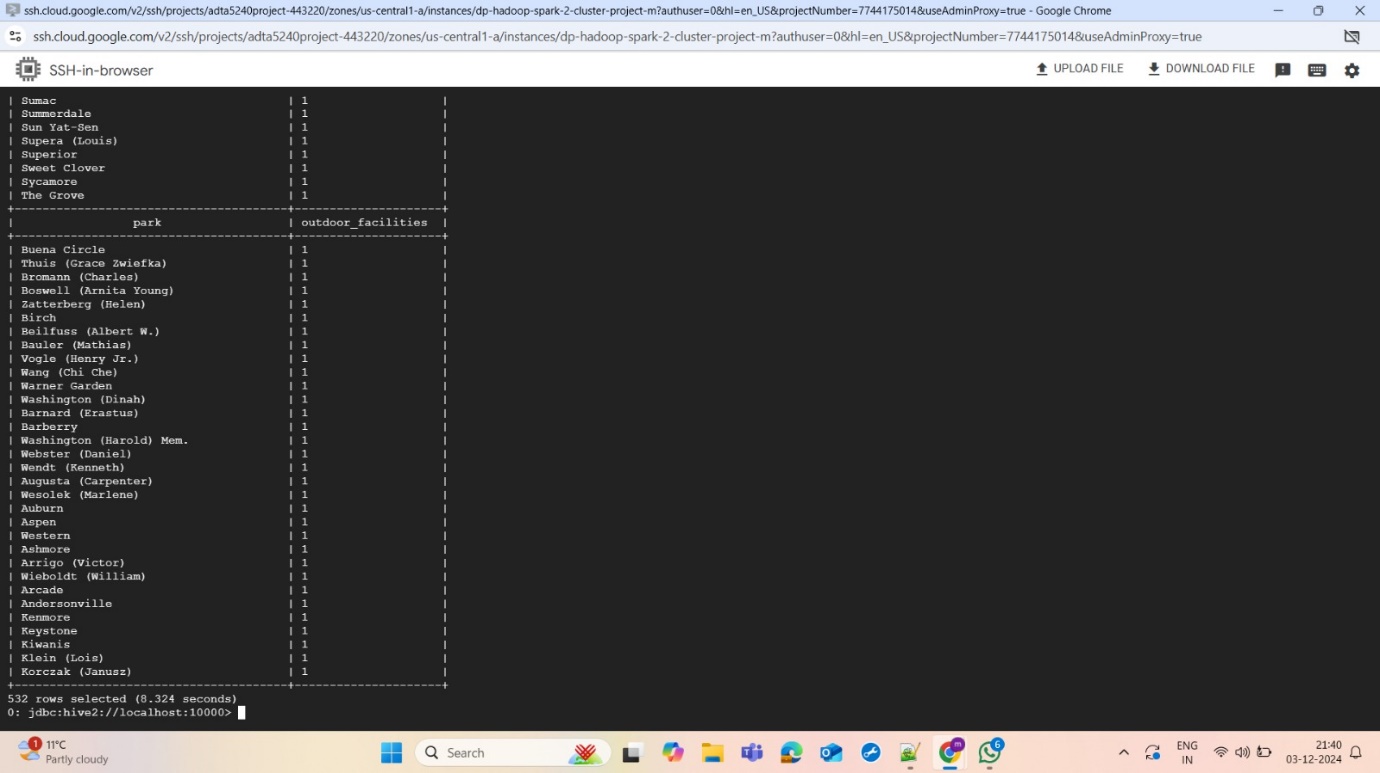
SELECT PARK, COUNT(\*) AS Outdoor\_Facilities

FROM CPDFacilities

WHERE FACILITY\_T = 'Outdoor'

GROUP BY PARK

ORDER BY Outdoor\_Facilities DESC;



**Screenshot Owner :** Sai Prasanth Mantha

Time Taken in Hive is 8.134 seconds.

**Query in Spark**

We have run the same query in spark for time analysis:

**QUERY:**

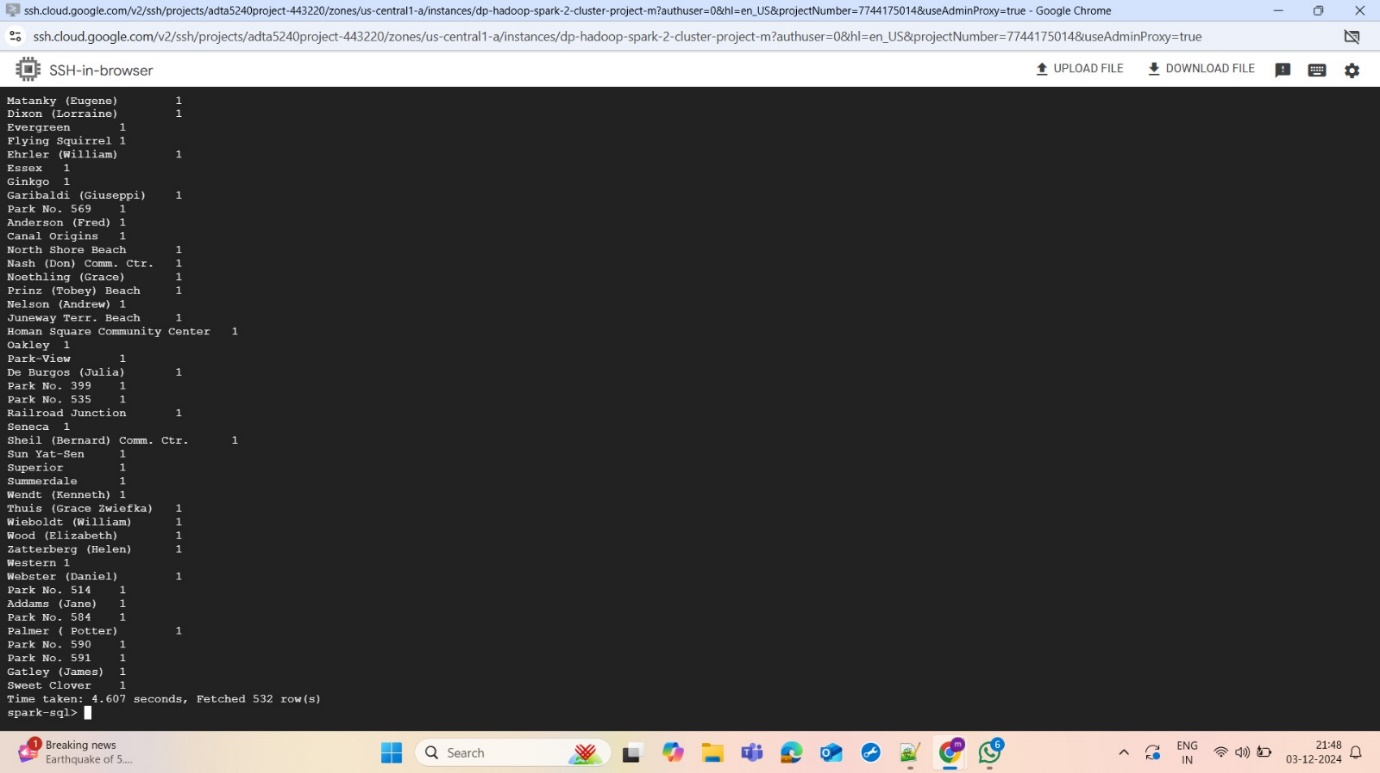
SELECT PARK, COUNT(\*) AS Outdoor\_Facilities

FROM CPDFacilities

WHERE FACILITY\_T = 'Outdoor'

GROUP BY PARK

ORDER BY Outdoor\_Facilities DESC;



**Screenshot Owner :** Charan Rakela

Time taken in Spark is 4.607 seconds.

**STREAMING DATA SET- (REAL TIME) ANALYSIS:**

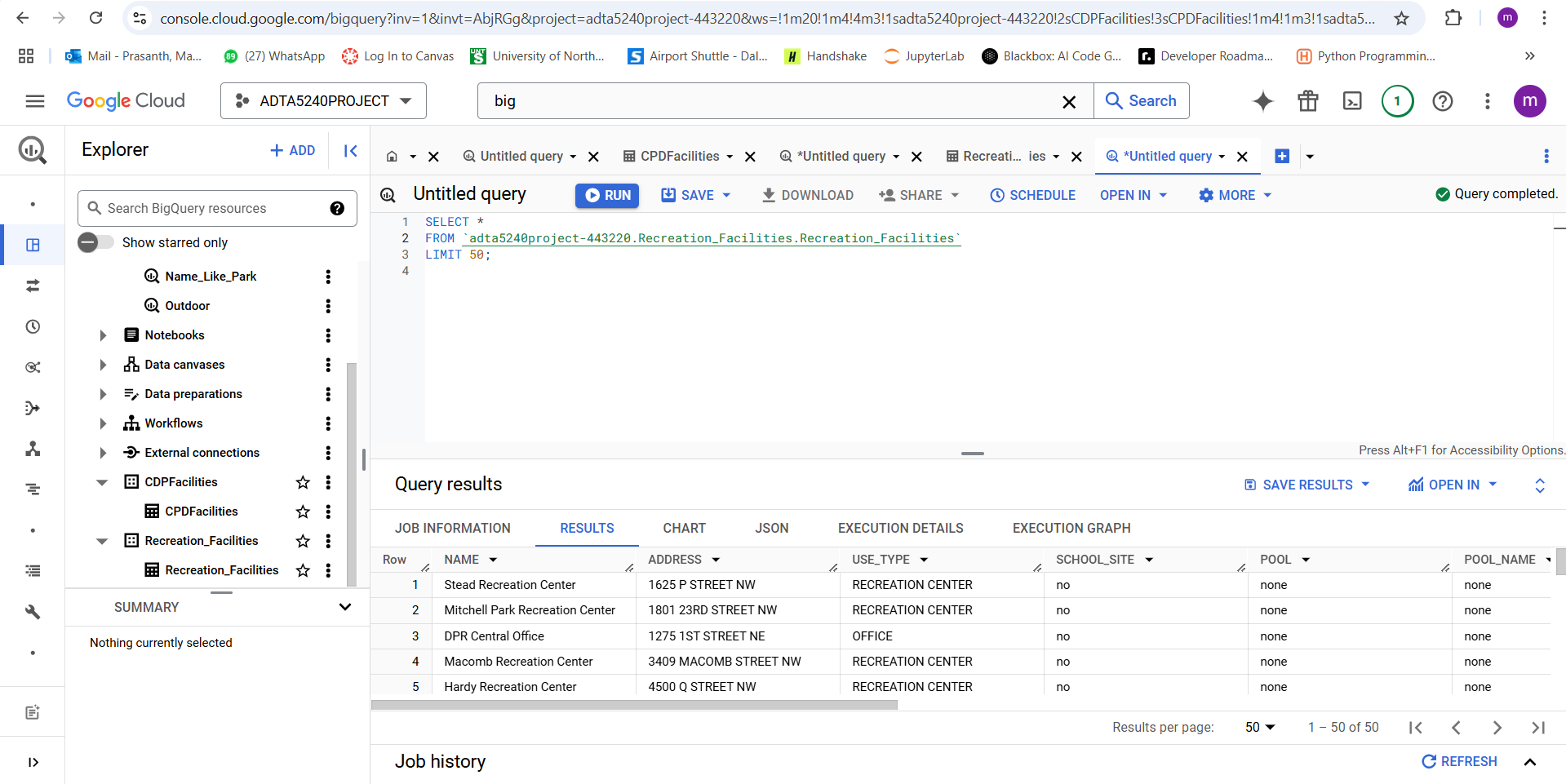
**BigQuery:**

Upload the **Streaming Dataset** in Big Query

SELECT \*

FROM `adta5240project-443220.Recreation\_Facilities.Recreation\_Facilities`

LIMIT 50;



**Screenshot Owner :** Ramakrishna Chowdary

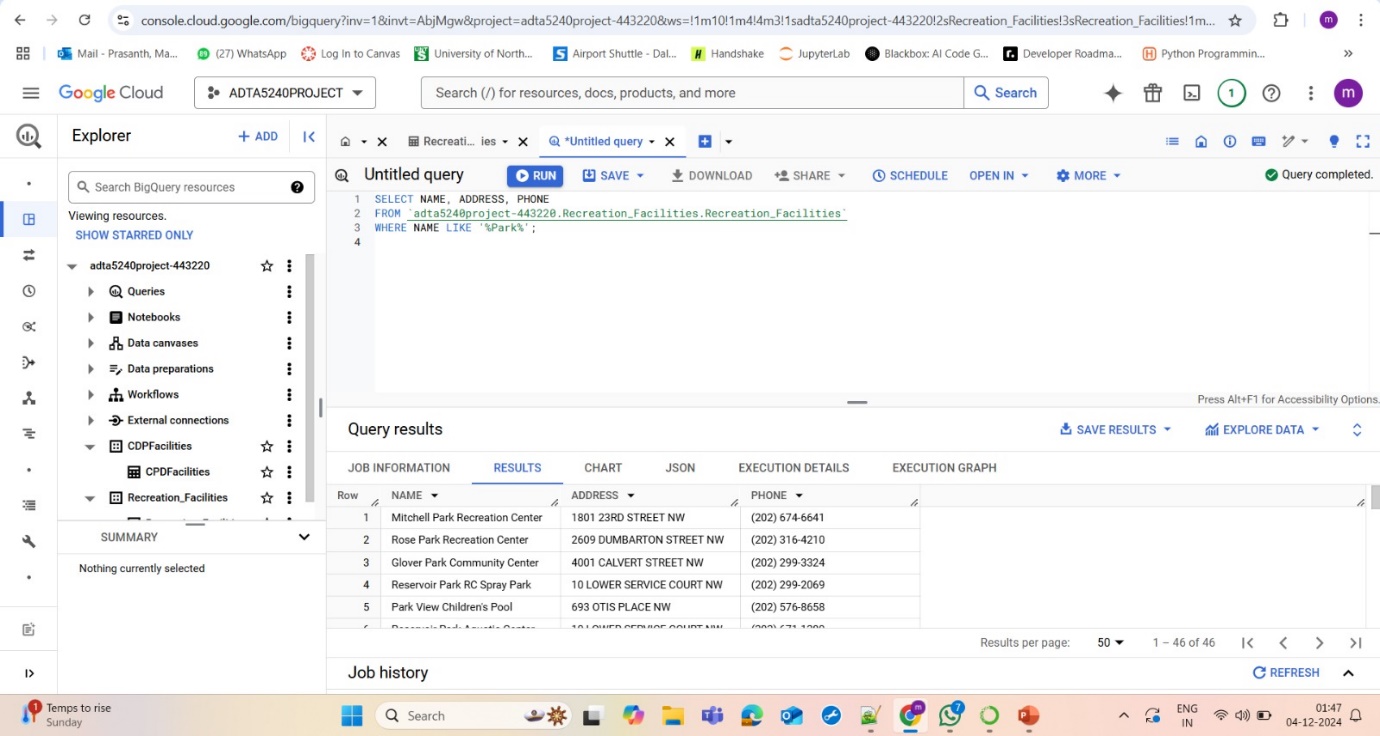
**Query 2:**

The below query is designed to fetch details (name, address, and phone) of all facilities whose names include the word "Park". It’s useful for identifying park-related recreation facilities**.**

SELECT NAME, ADDRESS, PHONE

FROM adta5240project-443220.Recreation\_Facilities.Recreation\_Facilities

WHERE NAME LIKE '%Park%';



**Screenshot Owner :** Sreekantha Peethani

**HADOOP ENVIRONMENT**:

Creation of Hadoop Environment for Hive and Spark.

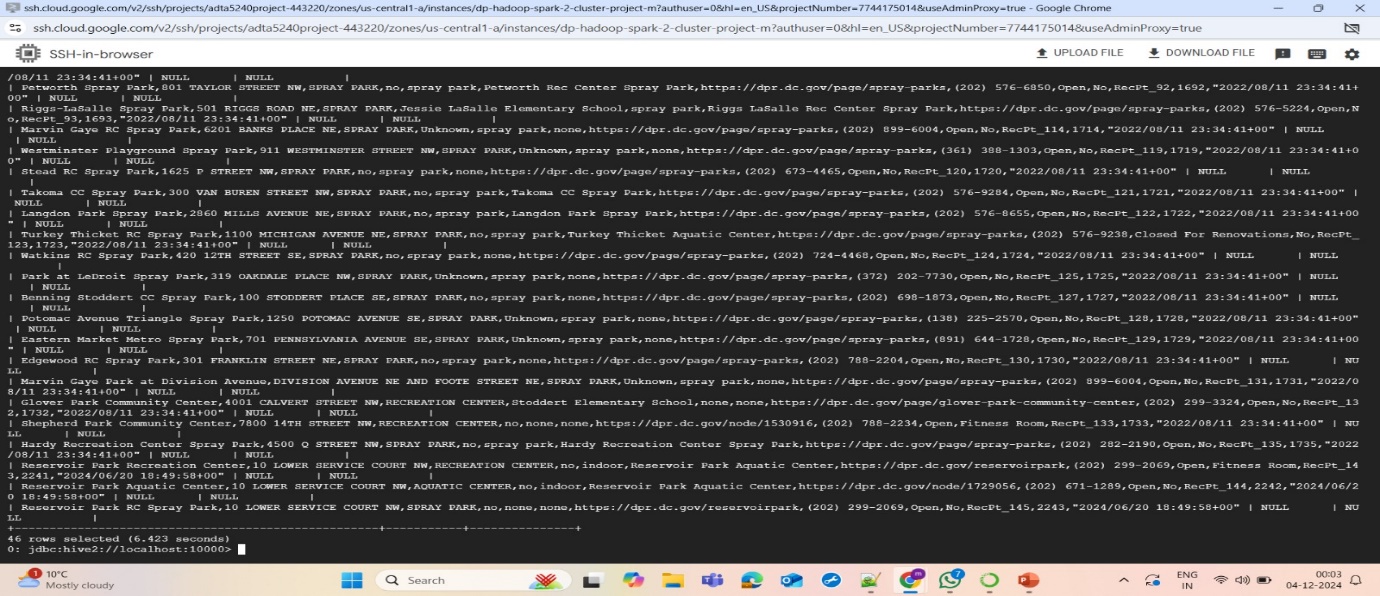
**Query in Hive**

We have ran the same Query in Hive:

SELECT facility\_name, location, contact\_info

FROM recreation\_facilities\_8

WHERE facility\_name LIKE '%Park%';



**Screenshot Owner :** Sai Prasanth Mantha

Time Elapsed: 6.423s

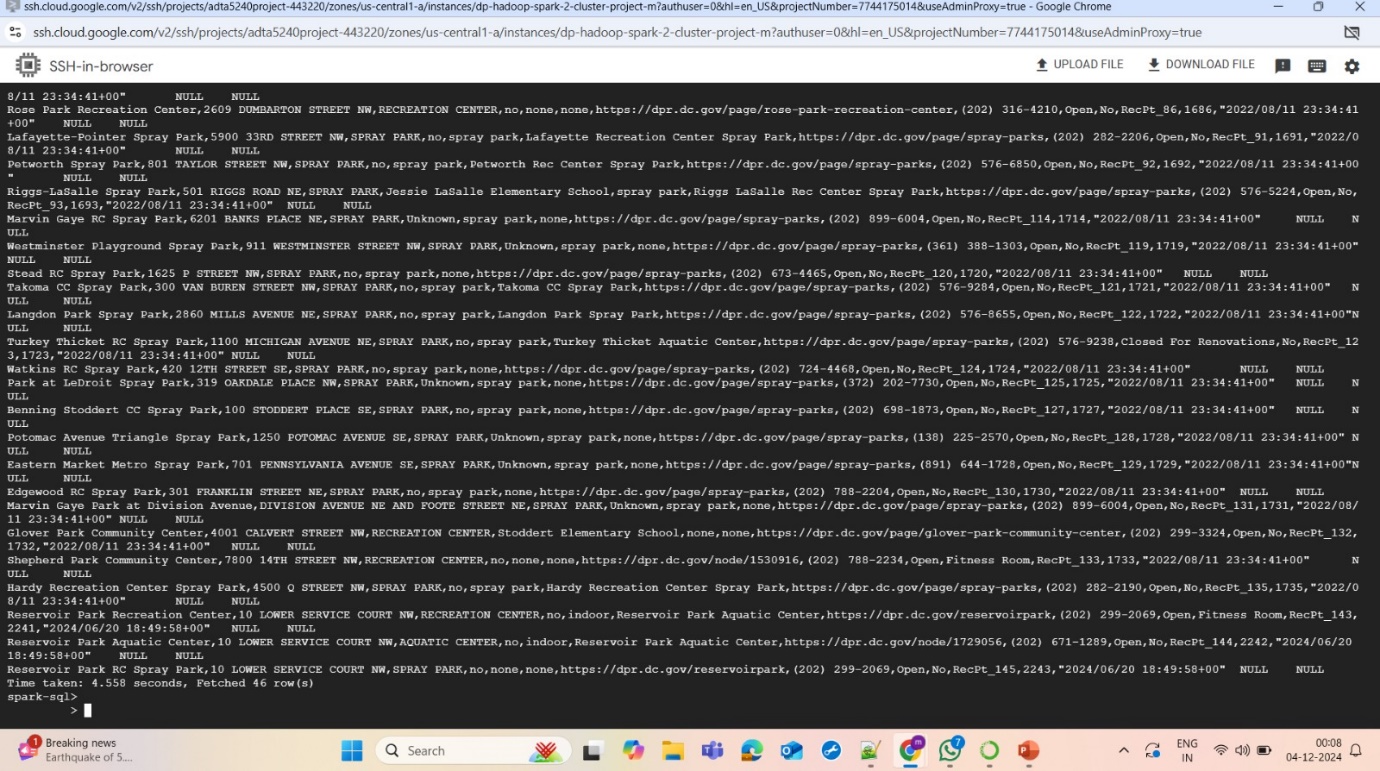
**Query in Spark**

We have run the same query in spark for time analysis:

SELECT facility\_name, location, contact\_info

FROM recreation\_facilities\_8

WHERE facility\_name LIKE '%Park%';



**Screenshot Owner :** Sai Prasanth Mantha

Time Elapsed:4.558s

We hope this report highlights our initial efforts to support the data scientists and analysts in addressing the lack of public spaces for teenagers. By leveraging data and tools, we aim to provide meaningful insights that pave the way for actionable solutions. Please feel free to reach out if you require further assistance or collaboration from our data engineering team. We are committed to working together to create inclusive and engaging spaces for the youth in our communities.

**Meeting Notes**

Date: November 25, 2023, and November 26, 2023  
Start Time: 10:00 AM  
End Time: 1:00 PM  
Note-taker: Ramakrishna  
Attendees: Prasanth, Ramakrishna, Srikanth, Charan

Meeting Summary

Day 1: November 20, 2023

* The team began by searching for datasets across various approved websites and finalized two datasets: a static dataset and a real-time dataset.
* The static dataset initially appeared complex but was later understood after careful examination.
* Data processing was conducted using OpenRefine, and it was observed that neither dataset contained missing or null values.
* Efforts were made to minimize the complexity of the static dataset for better usability and analysis.

Day 2: November 25, 2023

* Focus shifted to implementing queries on Google Cloud Platform (GCP) to analyze the issue and measure optimization solutions.
* Discussions included the efficiency of Hive and Spark for running queries on large datasets and optimizing performance.
* The team also distributed individual responsibilities to ensure smooth progress on subsequent action items and the final deliverables.

Decisions Made

1. From the static dataset, it was decided to analyze the distribution and frequency of recreation facilities across various parks to identify trends and patterns, focusing on key attributes like facility types and their availability in different locations.
2. The team decided to enhance the visualization aspect of the project to make the results more insightful and presentable.
3. Real-time datasets with timestamps were prioritized for running queries in Spark and Hive to calculate time differences and improve real-time analysis.
4. For future recommendations, the team discussed identifying larger and more complex datasets for additional case studies on frequent power outages and their impacts.

Final Changes

Meeting Date: December 2nd , 2023

* During the final review, it was confirmed that no null values were present in the datasets, so no imputation was required in OpenRefine.
* It was observed that some column in the real-time dataset and static dataset was originally in text format, so the team converted it to numerical values for better compatibility with analysis tools.

**Action Items:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Action Item # | Description | Owner | Due Date | Status |
| 1 | Data Set Collections | Charan | Due by Nov 20th | Completed |
| 2 | Documentation of Screenshots | Prasanth | Due by Nov 30th | Completed |
| 3 | Meeting notes draft and Open Refine Analysis | Ramakrishna | Due by Nov 24th | Completed |
| 4 | Executive Summary | Srikanth | Due by Dec 2nd | Completed |
| 5 | GCP, Hive and Spark Impletation | Prasanth | Due by Nov 29th | Completed |

**Refrences:**

* Google. (n.d.). *Tutorials | Big Query | google cloud*. Google. <https://cloud.google.com/bigquery/docs/introduction>
* IIT Madras - B.S. Degree Programme. (2022, January 6). *Prepare the data: Cleaning with Open Refine* [Video]. YouTube. https://www.youtube.com/watch?v=cX\_2MkShlJk
* *Hive tutorial*. (n.d.). <https://www.tutorialspoint.com/hive/index.htm>
* *Apache Spark Tutorial - Javatpoint*. (n.d.). www.javatpoint.com. <https://www.javatpoint.com/apache-spark-tutorial>
* Phases of data life cycle  
  <https://online.hbs.edu/blog/post/data-life-cycle>.
* Static Dataset Link: [https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Facilities](https://data.cityofchicago.org/Parks-Recreation/Parks-Chicago-Park-District-Facilities-current-/5yyk-qt9y)
* Streaming Dataset Link: [https://opendata.dc.gov/datasets/Recreation-Facilities](https://opendata.dc.gov/datasets/7122c1c815314588abe5c1864da8a355_3/explore?location=39.246960%2C-76.077342%2C7.24)