UNIVERSITY OF MALAYA

Faculty of Computer Science & Information Technology

Group Assignment

Climate Dynamics: Assessing Trends and Implication Across Nations

---

Course Code & Name	WQD7009 BIG DATA ANALYTICS
Name	MUHAMMAD HAKIM BIN NASARUDDIN (23079722)
	BAIZID YALDRAM (23117259)
	MUHAMAD NOOR FAQEH BIN BAKAR (17165216)
	MUHAMAD RAFIQ IQBAL BIN SAMSUDIN (17206926)
	AHMAD DANIEL BIN MOHD SUPANDI (24064261)
Group Number	GROUP 13
Lecturer	DR. RIYAZ AHAMED ARIYALURAN HABEEB MOHAMED
Submission Date	27TH DECEMBER 2024

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Background

This assignment focuses on the development and implementation of a data lifecycle framework tailored to analyse climate change datasets. It consists of various stages which includes, data ingestion, pre-processing, analysis and visualization, utilizing cloud-based tools, particularly Google Cloud Platform (GCP). Key components of the framework involve:

1) **Data Ingestion**: Efficient storage and management using Google Cloud Storage (GCS).
2) **Data Pre-Processing**: Simplifying data transformation and cleaning through Cloud Data Fusion.
3) **Data Analysis**: Employing BigQuery for querying and deriving insights from large datasets.
4) **Data Visualization**: Utilizing Looker Studio to present data-driven insights interactively.

The framework is evaluated based on metrics like processing time, memory usage, and execution efficiency, ensuring its scalability and adaptability for addressing global climate challenges. This initiative integrates datasets related to climate change, CO2 emissions, and global weather, emphasizing a comprehensive and analytical approach to climate study.

1.2 Project Objectives

- To develop a data lifecycle framework for analyzing climate datasets that include different phases of data ingestion and storage, pre-processing, analyzing, and data visualization.
- To implement the proposed data lifecycle framework on Google Cloud Platform to demonstrate the usability and effectiveness of each tool selected.
- To evaluate the performance of using several key performance metrics such as processing time, elapsed time, slot time and memory used during the process.

CHAPTER 2: MEETING MINUTE REPORT

Date : 8th December 2024

Time : 5:00 PM

2.1 Agenda

1. Selection of dataset for project
2. Framework and architecture design
3. Data cleaning and preprocessing
4. Planning for data visualization and dashboard development
5. Scheduling future meetings

2.2 Key Discussions and Decisions

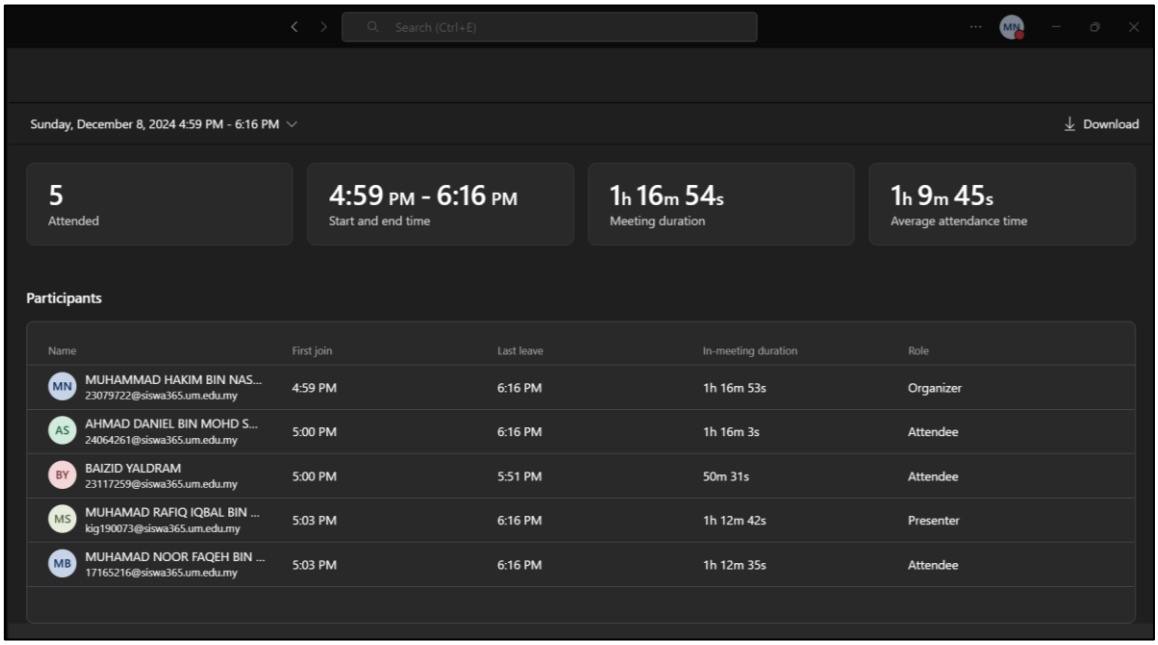
Dataset Selection	• Three datasets were considered, focusing on climate change, CO2 emissions, and global weather data. • Climate change data, including temperature, rainfall, and deforestation metrics, was finalized for the project.
Framework Selection	• The team will use Google Cloud Platform for architecture design. • Proposed components include Data ingestion tools (e.g., BigQuery, Apache Spark), Data storage and processing tools, Visualization tools like Power BI and Tableau.
Data Cleaning and Preprocessing	• Tasks include identifying missing or erroneous data and performing imputation or deletion where necessary. • Suggestions to integrate datasets and add meaningful columns to enhance prediction capabilities.
Visualization and Dashboard Development	• Initial focus on designing a basic dashboard. • Tools like Tableau and Power BI will be explored, leveraging free student licenses where applicable.

Scheduling and Coordination	• Weekly meetings planned for Thursdays (progress checkpoints). • Agreed to set reminders and use shared calendars for scheduling.

2.3 Action Items

1. Each member to review and familiarize themselves with the selected dataset
2. Finalize tools for ingestion, storage, and visualization
3. Begin data cleaning and preprocessing tasks
4. Draft visualization framework by next meeting
5. Setup shared Google Drive folder for collaboration
6. Schedule reminders for weekly meetings

2.4 Attendance



CHAPTER 3: DATA ARCHITECURE

3.1 Different Layers in Data Lifecycle Proces

The pipeline below showcases a comprehensive data lifecycle process or architecture diagram, as illustrated in ***Figure 1***.

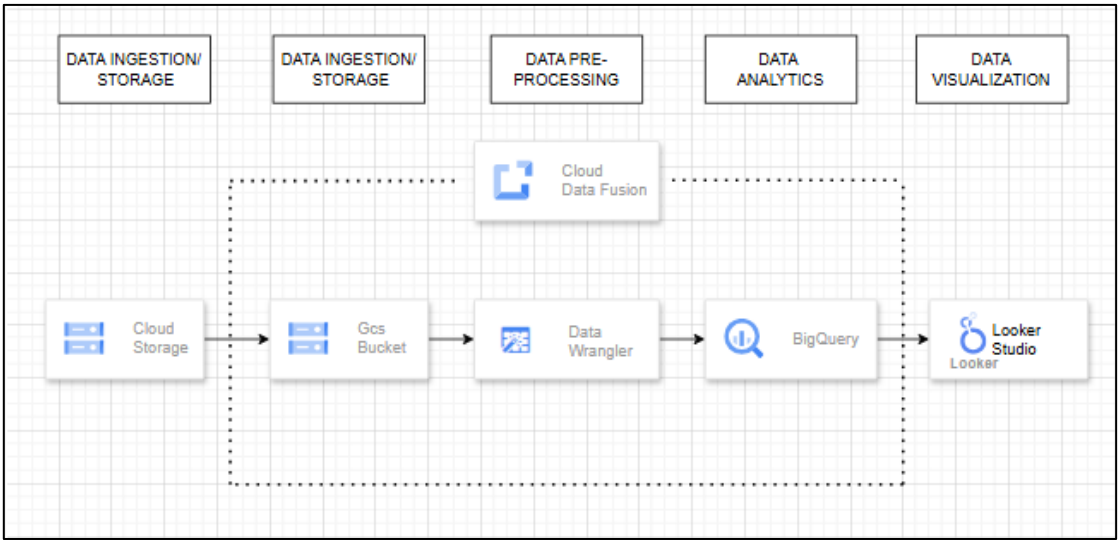


Figure 1: Proposed Data Architecture Diagram

The data lifecycle contains several stages which each of one playing a crucial role in ingesting, storage, processing, analyzing and visualizing effectively. The process is briefly explained below:

1. **Data Ingestion/Processing**:
    - Component: Cloud Storage and GCS Bucket
    - Purpose: The data firstly ingested through google cloud storage. The objective is to make sure that the raw data is securely stored and readily available for processing in the Google Cloud Platform. Data was stored in a specific container called GCS Bucket. This folder acts as a repository to organize and manage data within the GCS environment.
    - Communication: Ingestion in cloud storage enables seamless data transfer to other Google Cloud Tools and Environment.
2. **Data Preprocessing**:
    - Component: Cloud Data Fusion
    - Purpose: Cloud data fusion allows us to design ETL process (Extract, Transform, Load) pipelines by using a visual interface or drag-and-drop interface. This layer

handles the transformation and cleansing of raw data into a structured formats that are suitable for our analytics in BigQuery.

- Communication: Cloud data fusion can extract data from gcs bucket, transform using data wrangler tools and load it into data warehouse (BigQuery).

3. **Data Analytics**:
   - Component: Big Query
   - Purpose: Bigquery is a cloud data warehouse that is scalable and allows us to query and analyze the data process by the cloud data fusion. Insights can be derived from the large datasets through the use of SQL queries.
4. **Data Visualization**:
   - Component: Looker Studio
   - Purpose: This layer converts analytical results into a variety of visualization including graphs, chart, table, and dashboards. This visualizes insights and allows businesses to make more data driven decisions.
   - Communication: Data from BigQuery can be uploaded automatically onto Looker Studio allows seamless integration between these tools.

CHAPTER 4: TOOLS SELECTION JUSTIFICATION

In this section, we will justify the proposed framework of using cloud-based tools from Google Cloud Platform (GCP). Below are justifications for tools use for every stage in the proposed framework:

4.1 Data Ingestion

Google Cloud Storage (GCS) has been chosen as the primary data ingestion and storage tool in the cloud-based analytics framework due to its robust features and alignment with project requirements, as illustrated in ***Figure 2***.

Figure 2: Google Cloud Storage (GCS)

The following factors justify this selection:

1. **Relevance to the Dataset**: The CSV dataset needs an efficient storage solution. GCS is ideal for organizing files for easy access and effective data management.
2. **Integration with Other Tools**: GCS integrates well with Google Cloud tools like Data Fusion for ETL, BigQuery for analysis, and Looker Studio for visualization, streamlining workflows.
3. **High Performance and Redundancy**: GCS offers quick access and low-latency retrieval, with data replication across locations for redundancy and high availability.
4. **Backup, Archiving, and Disaster Recovery**: GCS provides robust data backup and archiving options and built-in disaster recovery features for data accessibility during disruptions.

4.2 Data Pre-Processing

In the data pre-processing stage, as shown in ***Figure 3***, Cloud Data Fusion is utilized to extract, transform and load data because of the following reasons:

Figure 3: Cloud Data Fusion

1. **Visual Interface**: Cloud data fusion is built as a user friendly, point-and-click or drag and drop interface for creating ETL data pipeline, as illustrated in ***Figure 4***. This reduces the hassle for coding, reduces complexity and allows non-technical users to also work in designing the ETL data pipeline.

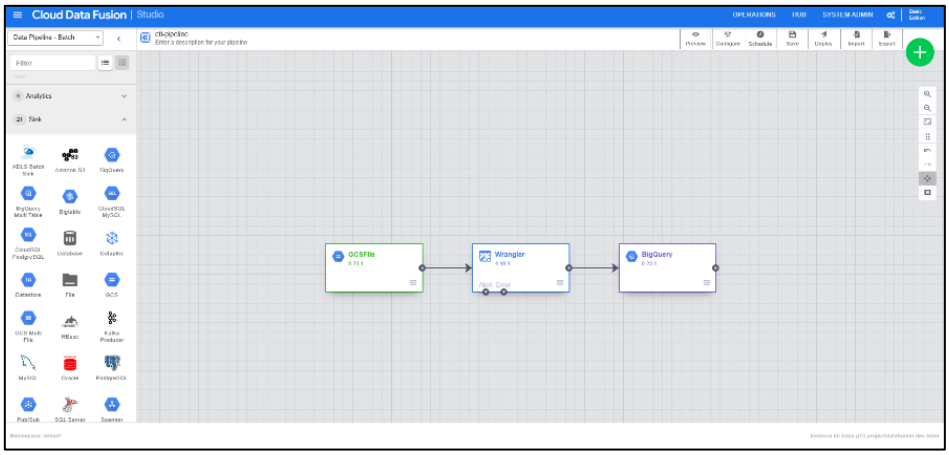


Figure 4: User-friendly Interface

2. **Scalability:** Dataproc Autoscaling in the backend allows the ETL pipeline to automatically manage cluster resources and scale worker VMs to help improve performance and scalability, as shown in ***Figure 5***. This makes it ideal for working either with small or large dataset without worrying about the infrastructure constraints.

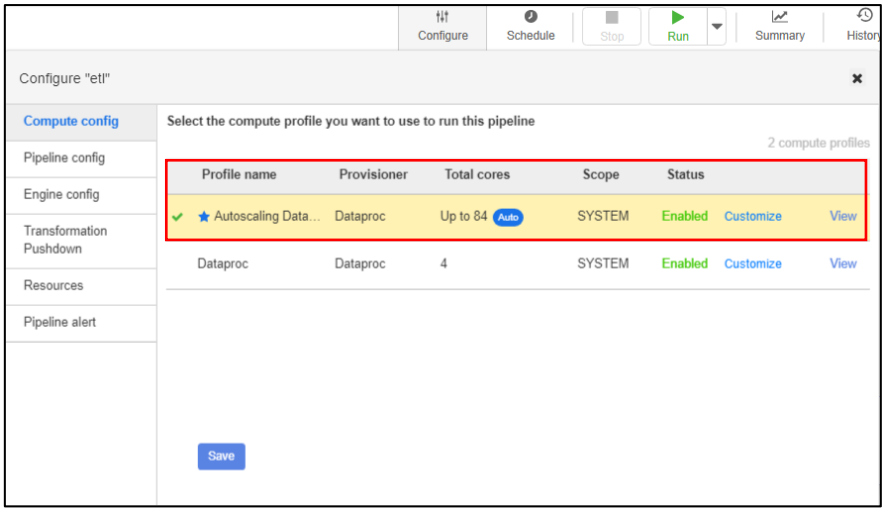


Figure 5: Scalability Function in Data Fusion

3. **Data Cleaning using Data Wrangler**: Cleaning and transforming data tasks are easier due to its interactive visual interface. Tasks like handling missing values, removing null values, data aggregation and filtering require no coding, and libraries make it relatively simple and suitable for beginners and non-technical users.

4.3 Google Big Query

Google Big Query is a powerful and fully managed data warehouse designed to handle large-scale data analytics efficiently, as illustrated in ***Figure 6***. It integrates seamlessly with the Google Cloud framework, making it an ideal choice for complex analyses.

Figure 6: Big Query

1. **Scalability**: Big Query can dynamically scale to meet the demands of any workload, ensuring consistent performance regardless of data size.
2. **Fully Managed**: Big Query automatically handles storage, processing, and resource allocation, making the analysis part easier.
3. **Integration with Google Cloud**: Seamless integration with tools like Google Data Studio, Looker, Google Sheets, and AI/ML services (e.g., TensorFlow and Vertex AI). This enables end-to-end data workflows on a single platform.
4. **Real-Time Analytics**: Big Query's streaming API allows real-time data ingestion, which is crucial for modern applications like IoT, fraud detection, or user activity tracking.
5. **Cost-Effectiveness**: Big Query uses a pay-as-you-go pricing model, meaning you only pay for the queries you run and the storage you use. Its pricing structure makes it accessible for projects with varying resource demands.

4.4 Data Visualization

As shown in ***Figure 7***, Looker is a sophisticated, cloud-based business intelligence platform designed for interactivity and scalability of visualizing data, yet it integrates seamlessly with various databases and cloud platforms particularly BigQuery, predominantly making it an ideal choice for data workflows. The following are key reasons why Looker was chosen:

Figure 7: Looker Studio

1. **Seamless Integration**:
   - With its seamless connectivity, real-time data exploration is enabled without requiring manual intervention or scheduled updates even though in this project, real-time data is not implemented.
2. **Scalability and performance**:
   - Able to handle large datasets efficiently and smoothly, ensuring high performance even with complex data as it is designed for enterprise usage.
3. **Customizability and flexibility**:
   - Customized calculations, aggregations and derived fields are possible using Looker to tailor analytics specific demands.
   - On top of that, it supports a wide range of visualizations, turning it into a versatile tool for various reporting and storytelling needs.
4. **Ease of use**:
   - It has a user-friendly interface which then empowers technical and non-technical users to create and explore dashboards, which advanced users can utilize SQL-based queries for in-depth analysis.
5. **Actionable insights**:
   - Interactive filters, drill-down capabilities and alerts make it easier for insights deliverables and prompt action are possible, which is a vital criterion for decision-making in sustainability projects.

CHAPTER 5: MODEL IMPLEMENTATION

In this section, we outline the implementation of a framework on the Google Cloud Platform and detail the data lifecycle pipeline from ingestion to visualization.

5.1 Data Ingestion

Successful data migration is crucial for efficient storage, access, and analysis. Google Cloud Storage (GCS) facilitates fast and secure data transfers. The data ingestion process includes the following steps:

5.1.1 Creating a Project

For this assignment, a Google Cloud project named “bdaa-g13-project” was created, ***Figure 8***. Enabling the relevant API before using GCS ensures proper setup and accessibility within the project.

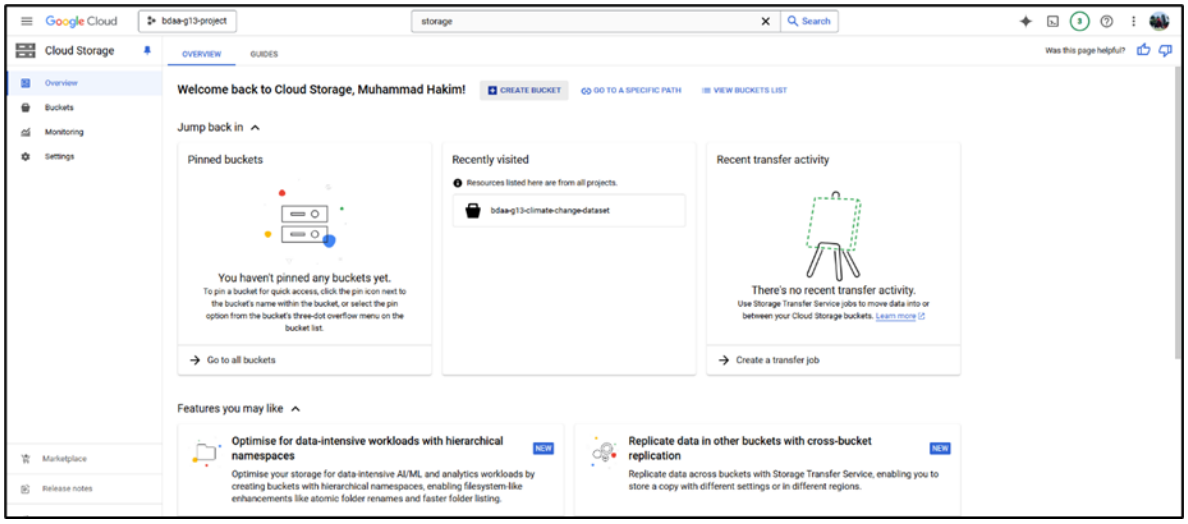


Figure 8: GCS Home Interface

5.1.2 Configuring and Creating a Bucket

The next step is to set up a project-specific storage bucket named “bkt-bdaa”, ***Figure 9***. During setup, a notification about restricted public access is popped up, ***Figure 10***. Clicking “CONFIRM” prevents public access. To allow it, the option “Enforce public access prevention on this bucket” can be unchecked.

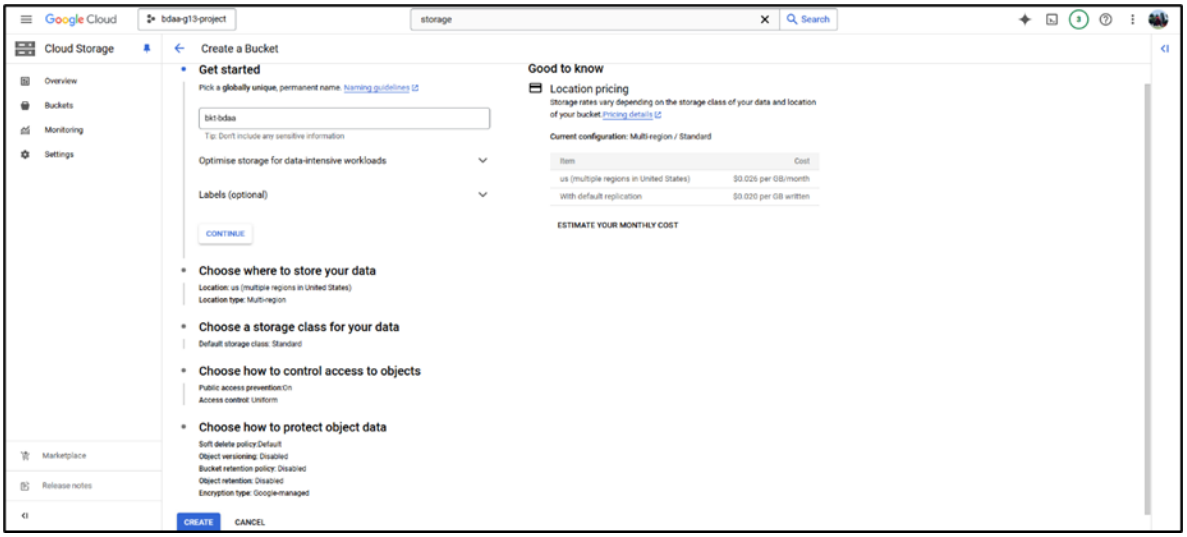


Figure 9 : GCS Bucket Configuration

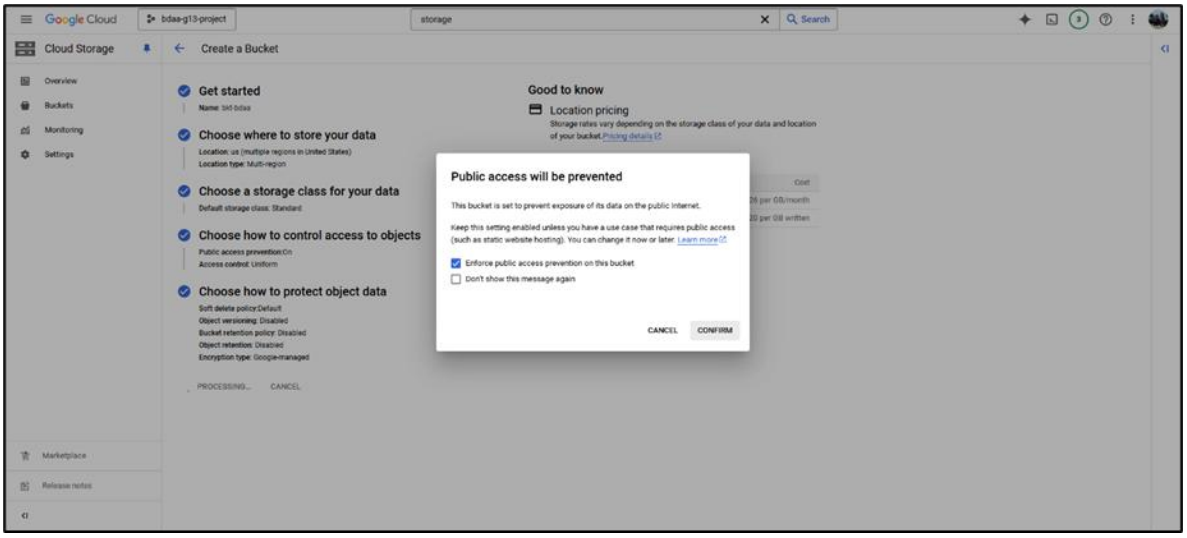


Figure 10: GCS Bucket Public Access Prevention Confirmation

5.1.3 Locating the Bucket

After creating the bucket, it appears in the “Folder browser” of the GCS interface. Clicking on the bucket reveals options to upload data files. In this case, the dataset is in a compact CSV format, allowing manual upload via the available option, as shown in ***Figure 11***.

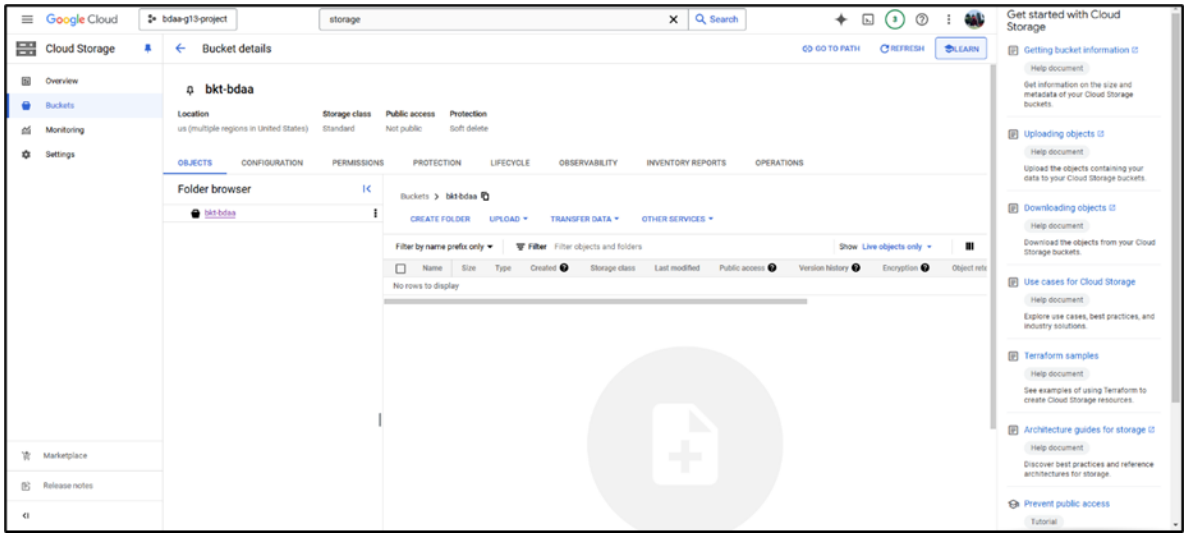


Figure 11: Created Bucket Called “bkt-bdaa” in Google Cloud Storage (GCS)

5.1.4 Uploading and Verifying Data

After uploading, the “climate-change-dataset.csv” will appear in the bucket's table section, displaying details like size, type, and upload time. As shown in ***Figure 12***, it is essential to open and check the file to ensure it is formatted and ready for further analysis.

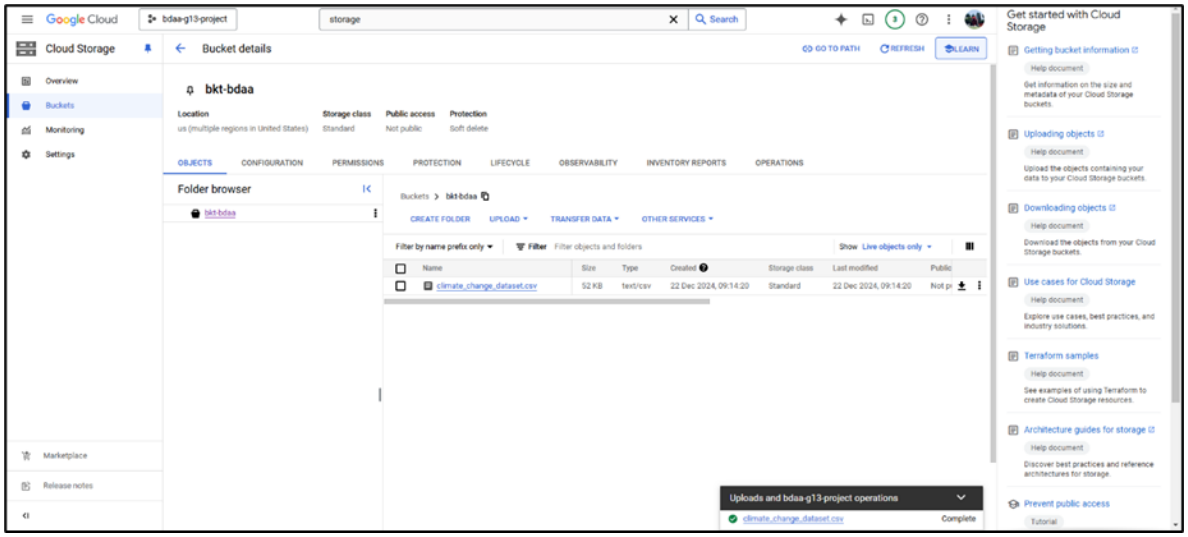


Figure 12: Uploaded Dataset in the Bucket Called “climate-change-dataset.csv”

5.2 Data Pre-Processing

After ingesting data into the google cloud environment, data pre-processing phase undergoes of ETL process which extract the data, transforming data, and loading data into data warehouses. Below is the step-by-step implementation of this phase.

5.2.1 Create a Data Fusion Instances

Firstly, to start building the ETL pipeline using cloud data fusion, the data Fusion instance was created using basic edition, as shown in ***Figure 13***. Time taken to create the instance and wait for it to start running was around 15 – 20 minutes. Instance can be viewed after the instance has been created.

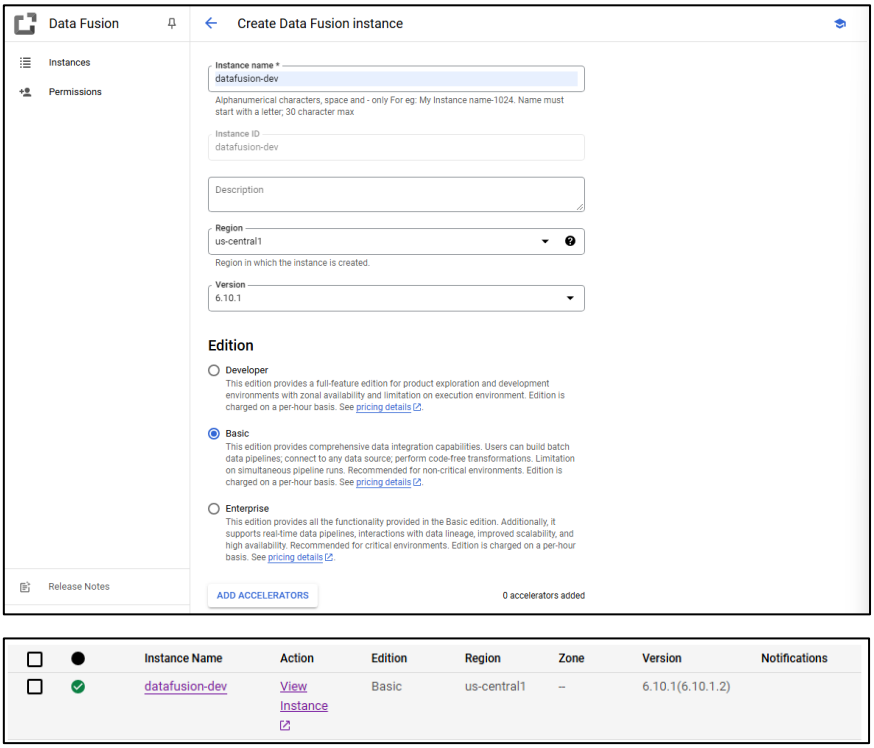


Figure 13: Data Fusion Instance Create

5.2.2 Choose task to perform

Next, the wrangling task was chosen to clean the dataset as in the ***Figure 14*** below.

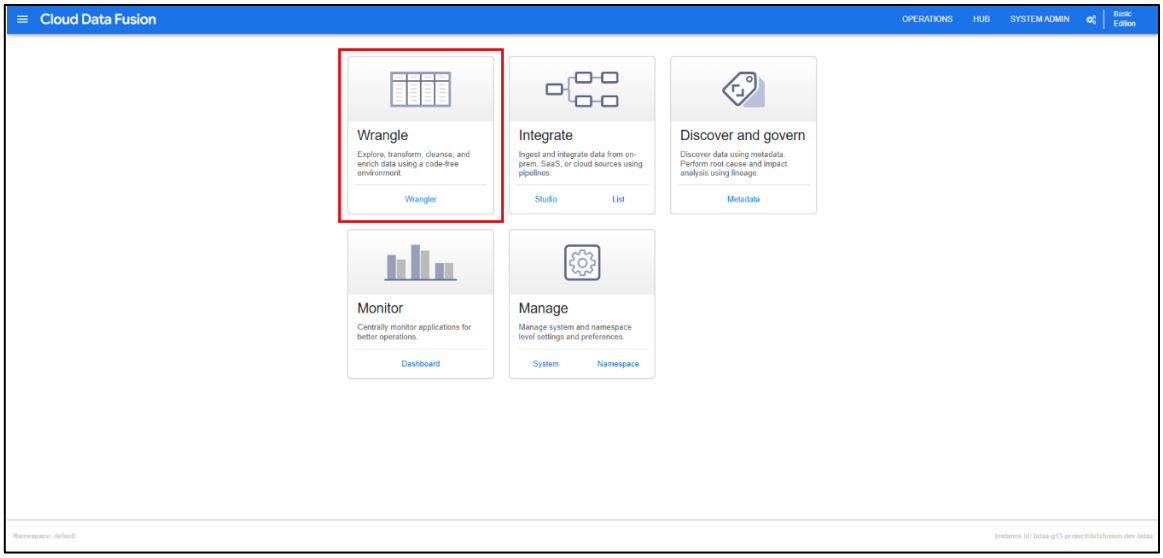


Figure 14: Data Wrangling in Data Fusion

5.2.3 Selecting Source for Dataset

Next, datasets from various databases, buckets, and data warehouses can be selected. In our case, the datasets that have been ingested on GCS bucket name ‘bkt-bdaa’ on previous section was selected, as shown in ***Figure 15***.

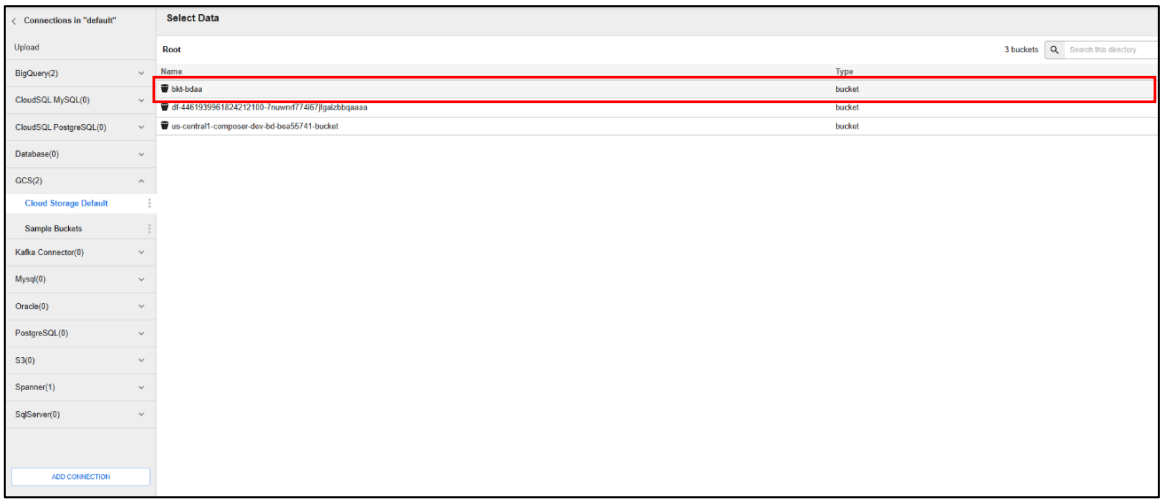


Figure 15: Selecting Datasets from Bucket Storage

5.2.4 Cleaning the Dataset

Next, the data selected can be transformed in the data wrangler, as illustrated in ***Figure 16a***. Processes like renaming, changing data type, removing null values and replacing null values

can be done without any coding acquire. In our case, the column type and column name have been transformed to improve consistency and to improve data accuracy during analysis. All transformation histories can be checked on the right side of the tab. After all transformation is done, the next step to proceed is to create a pipeline.

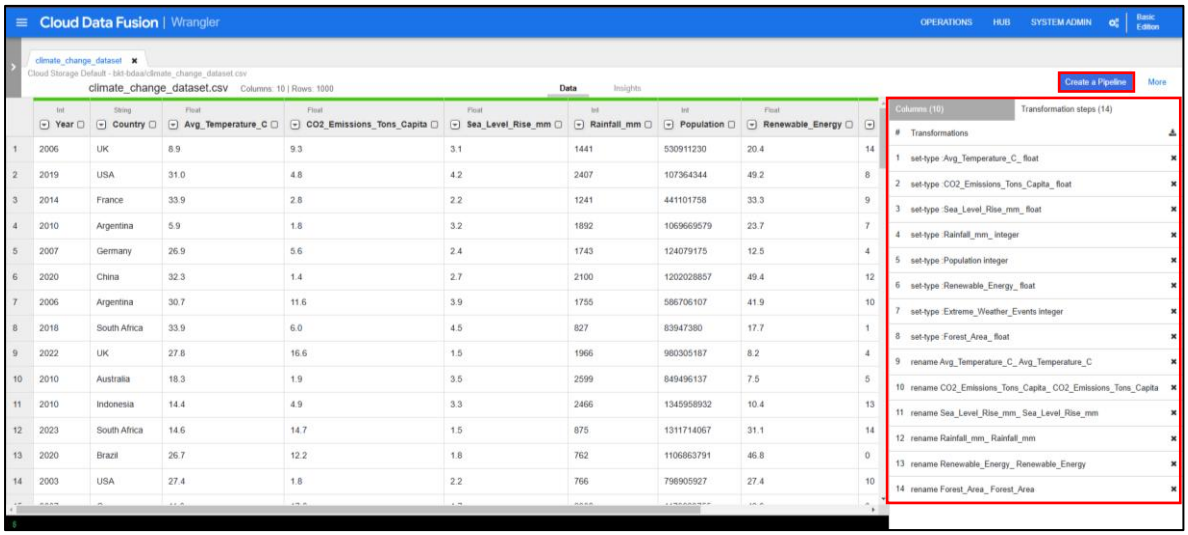


Figure 16a: Cleaning Data using Data Wrangler

5.2.5 Creating an ETL Data Pipeline

Next, batch processing is selected because our dataset is not a real-time dataset, as shown in ***Figure 16b***.

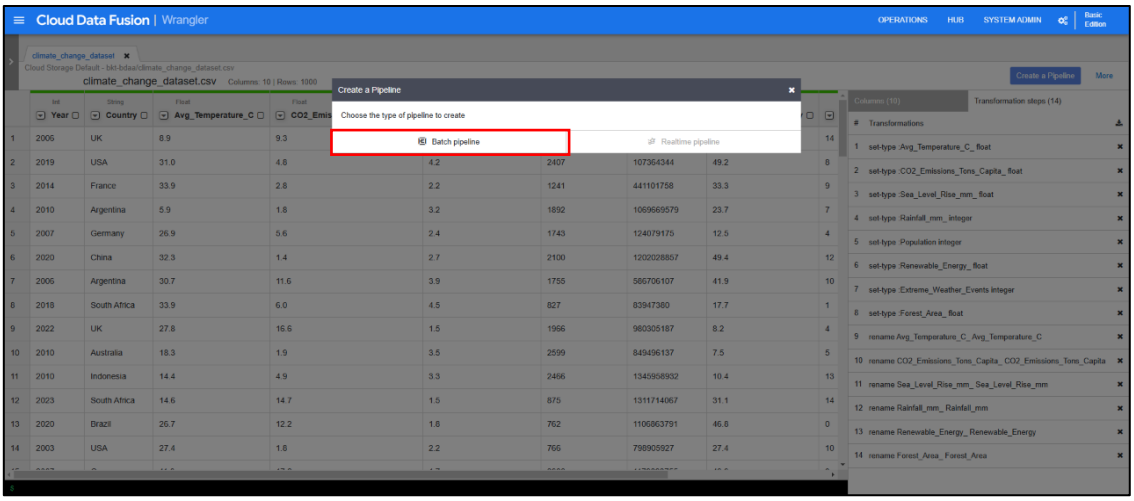


Figure 16b: Choosing Batch Processing

Based on ***Figure 17***, the overall data pipeline was completed connecting the data wrangling process into a BigQuery data warehouse.

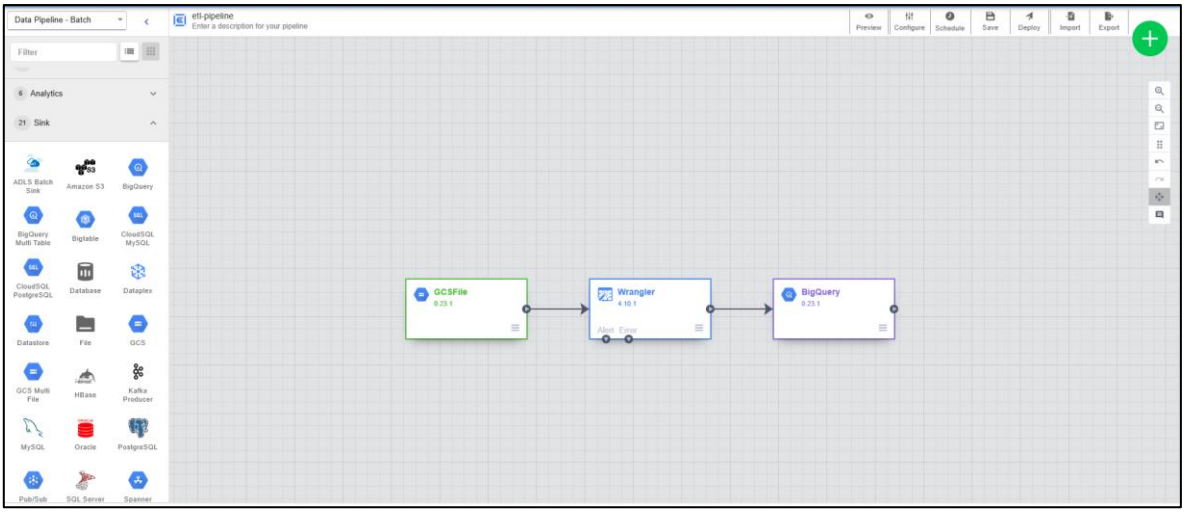


Figure 17: ETL Data Pipeline

Beforehand, the dataset table in BigQuery needs to be created first in order to connect these two processes, as shown in ***Figure 18***.

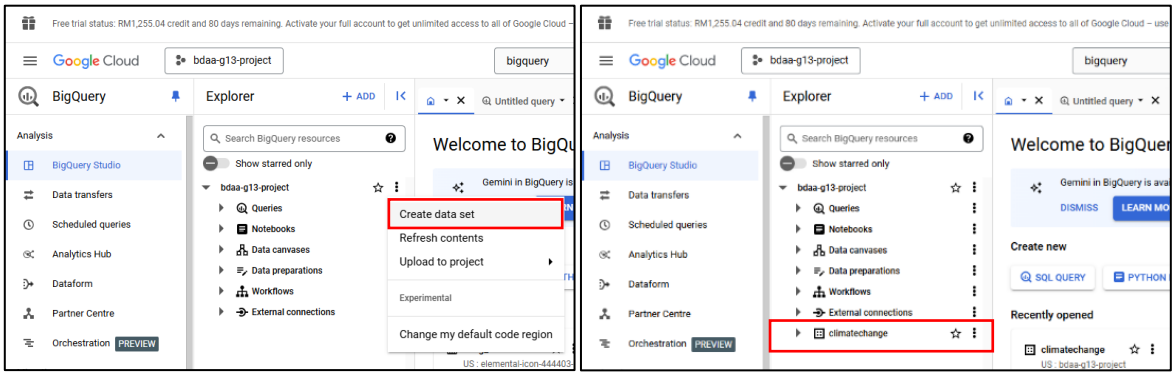


Figure 18: Creating Dataset on BigQuery

In order to connect the clean dataset that has been done by the data wrangler, Big Query properties in the data pipeline need to be set up to select the directory of the BigQuery table that we have created before as in the ***Figure 19*** below. Also, the name of the table to be transferred to BigQuery is set as 'cc_data'. After all Basic and advanced setup are completed. The BigQuery properties are validated to ensure it is correct.

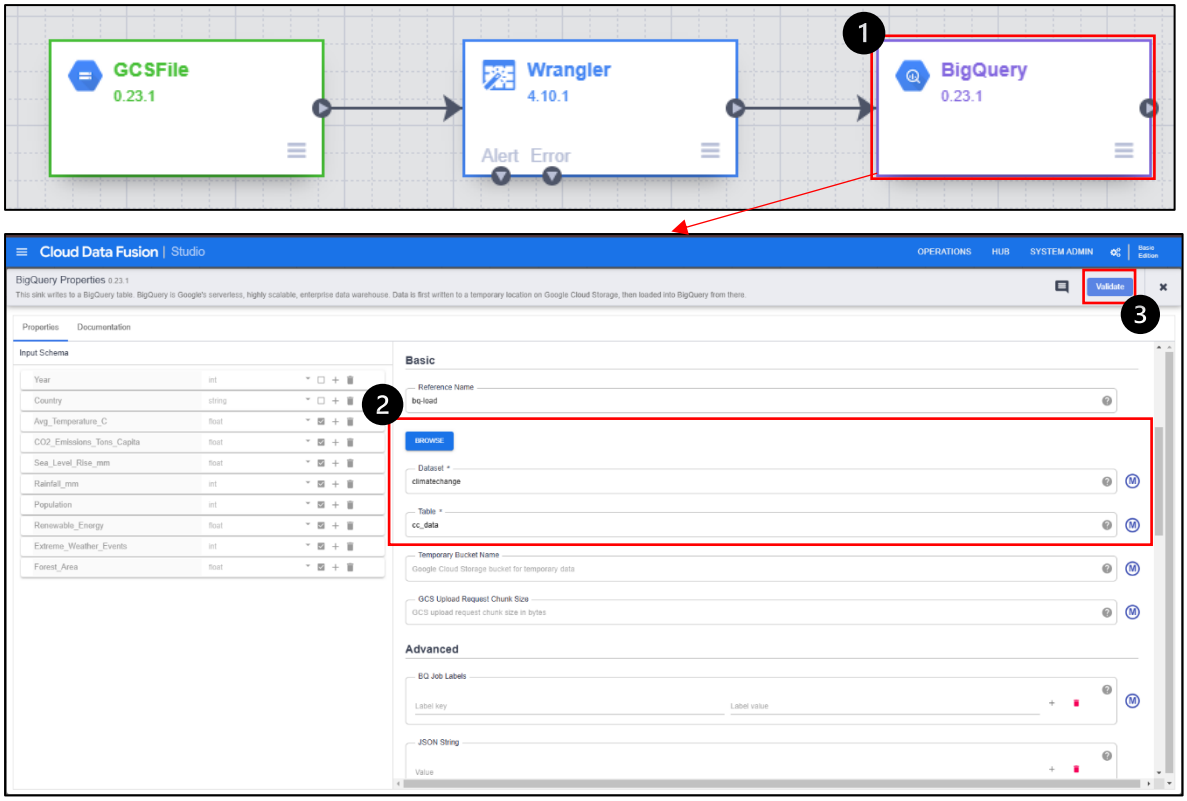


Figure 19: Big Query Property Setup

5.2.6 Deploying and Running the ETL data pipeline

After all the data pipeline has been correctly set up. We can start deploying the pipeline to start the overall ETL data pipeline process, as illustrated in ***Figure 20***.

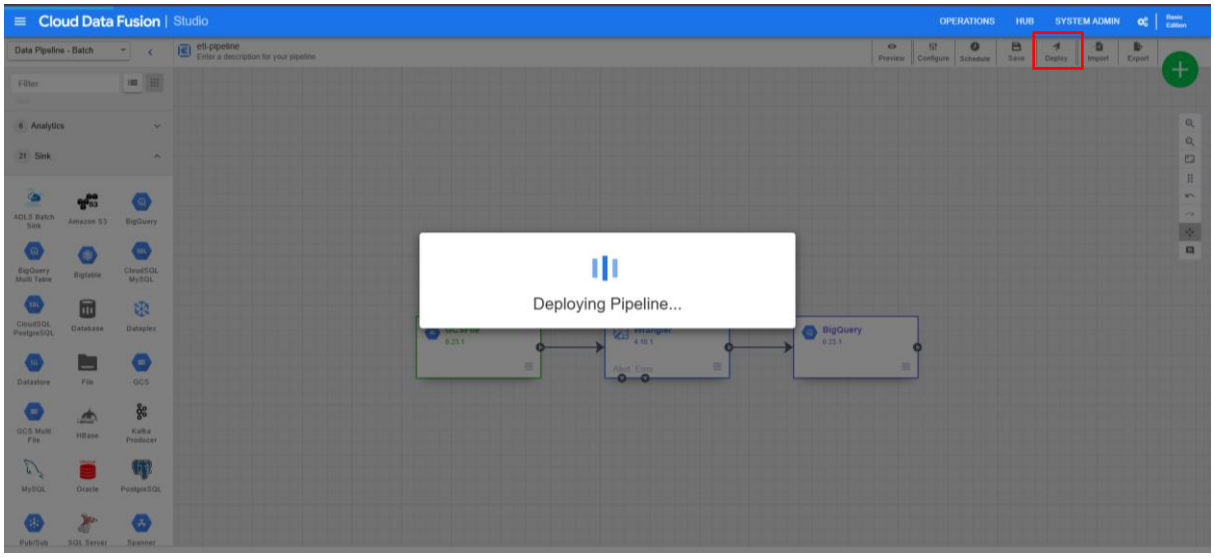


Figure 20: Deploying ETL Data Pipeline

After the pipeline is deployed, we can run the pipeline, as illustrated in ***Figure 21***, and wait for the status to be succeed. Any errors during the pipeline running can be checked in the logs.

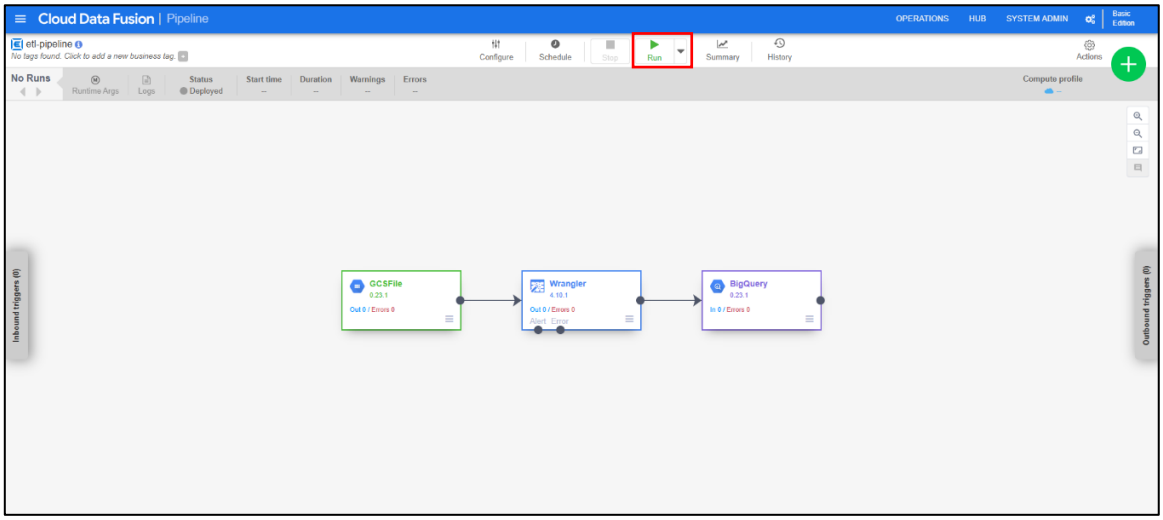


Figure 21: Running the Pipeline

The pipeline status will change from provisioning, starting, running, and succeeding if there are no errors during the process, as shown in ***Figure 22a***.

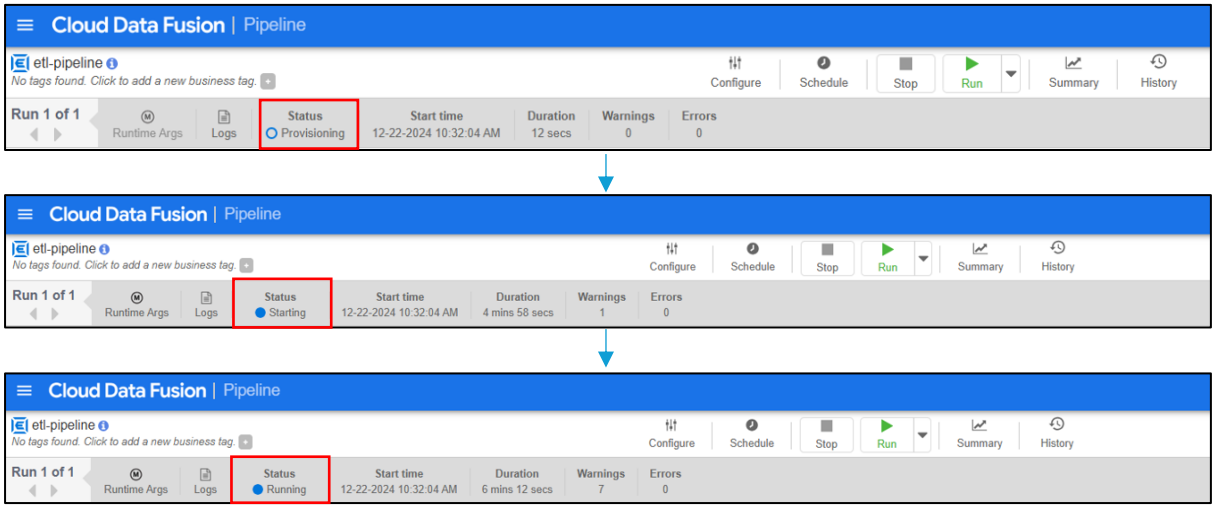


Figure 22a: Status in Cloud Data Fusion Execution

If there is no errors occur, as shown in ***Figure 22b***, it will change the status to succeeded and all history during the processing can be check through the logs as in the image below.

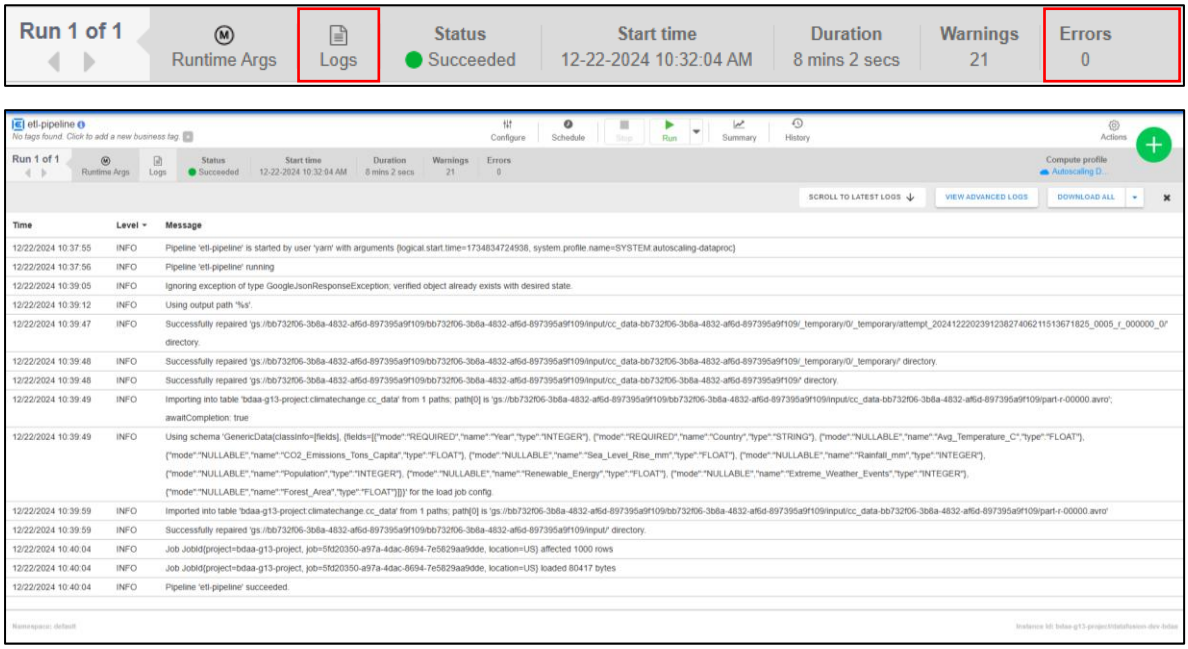


Figure 22b: Process Log

Lastly, the data can be checked on BigQuery table, the process was successfully transferring the clean datasets onto the BigQuery data warehouse as in the ***Figure 23*** below.

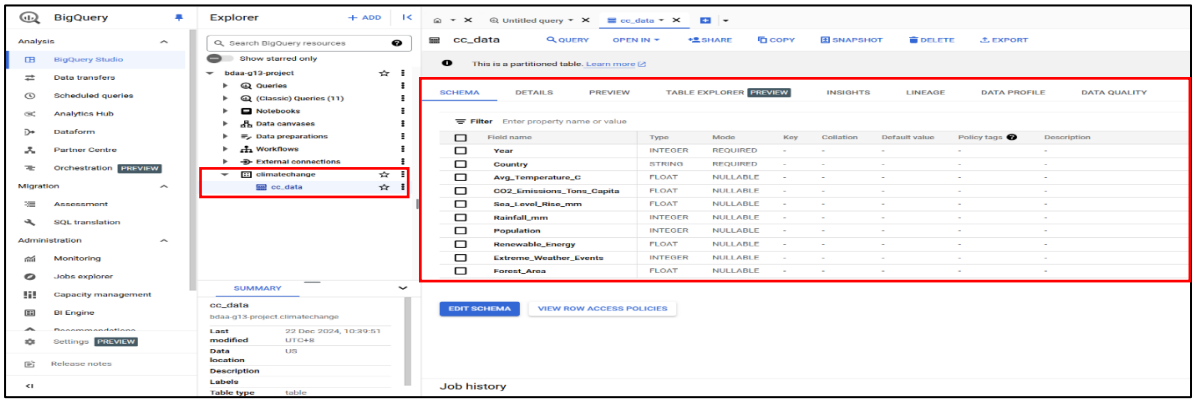


Figure 23: Transferred Dataset into BigQuery

5.3 Data Analytics

After successfully loading the clean dataset, as shown ***Figure* 24** from the data Pre-processing stage into Big Query, we displayed how our dataset looks like in the BigQuery table.

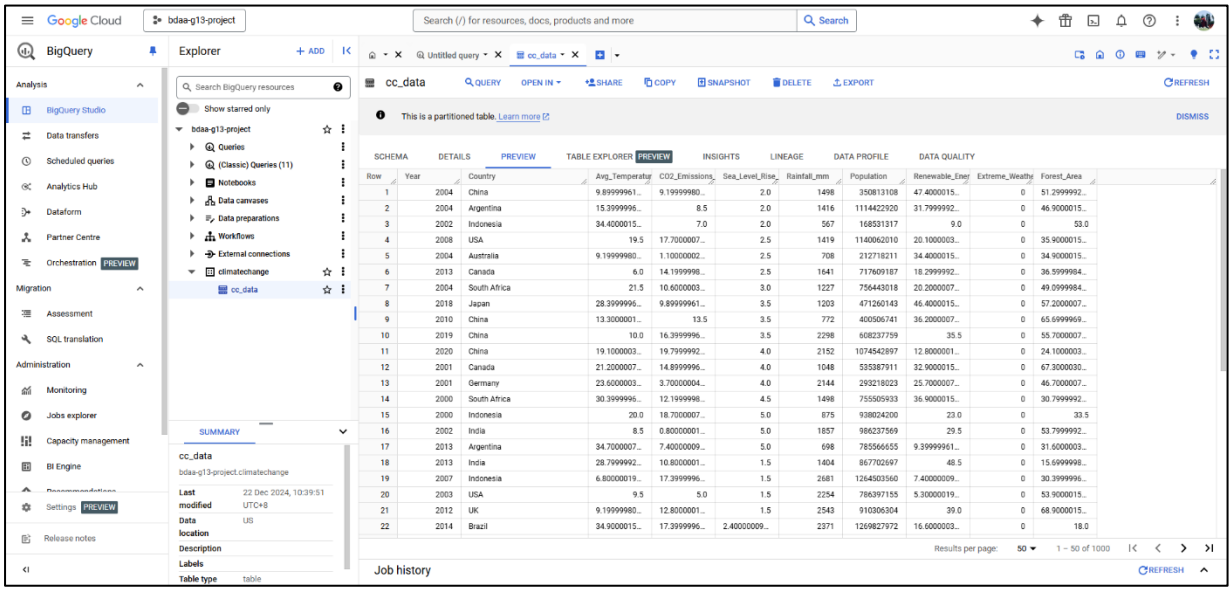


Figure 24: Loaded Dataset

Here are some of the meaningful queries related to our dataset:

5.3.1 Average Temperature by Year:

Query 1: Calculates the average temperature for each year, ***Figure 25a***.

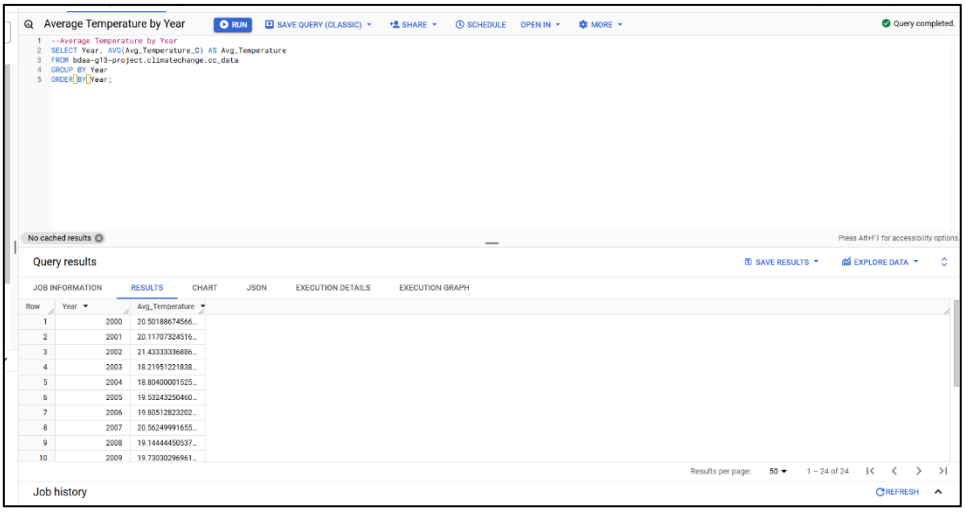


Figure 25a: Query 1

Insight: Tracks global temperature trends over time, highlighting periods of warming or cooling, ***Figure 25b***.

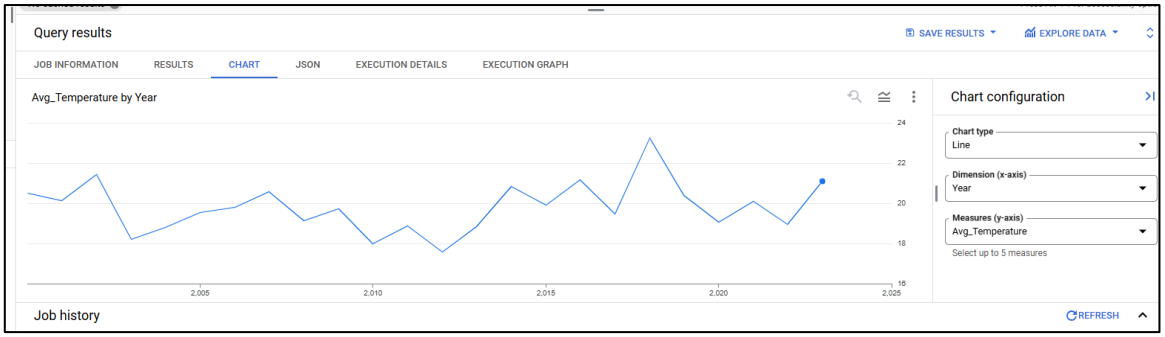


Figure 25b: Chart from Query 1

5.3.2 Countries with the Most Extreme Weather Events:

Query 2: Sums the total number of extreme weather events per country, ***Figure 26a***.

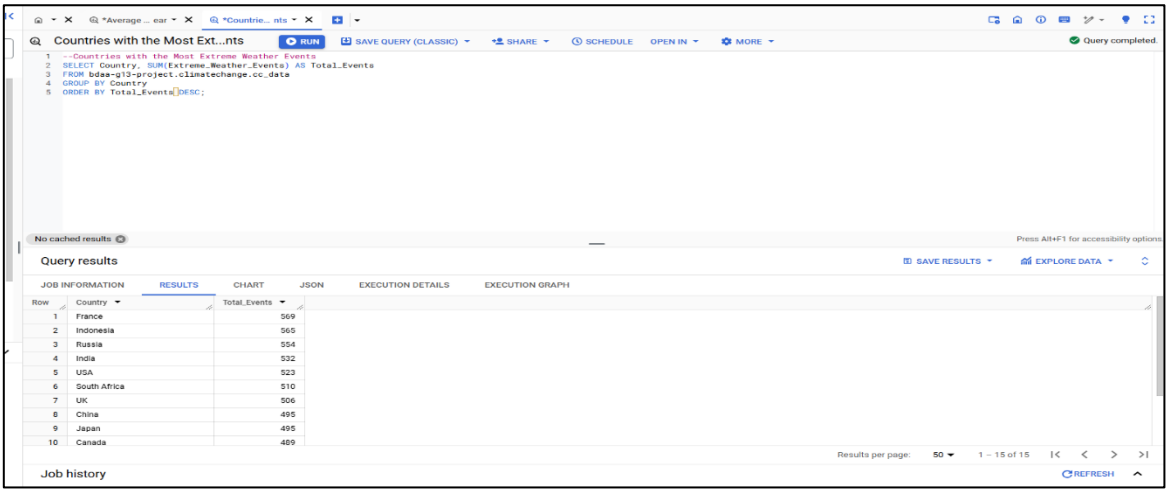


Figure 26a: Query 2

Insight: Identifies countries most affected by severe weather, useful for disaster management planning, ***Figure 26b***.

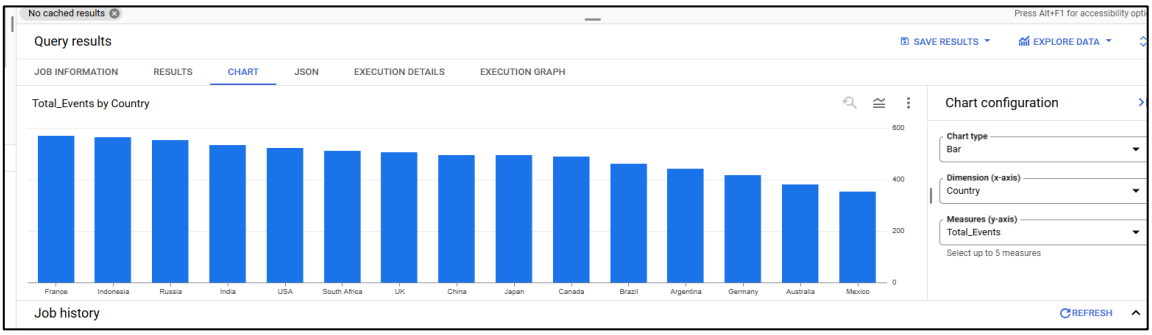


Figure 26b: Chart from Query 2

5.3.3 Forest Area Trends Over Time by Country:

Query 3: Computes the average forest area by country and year, ***Figure 27a***.

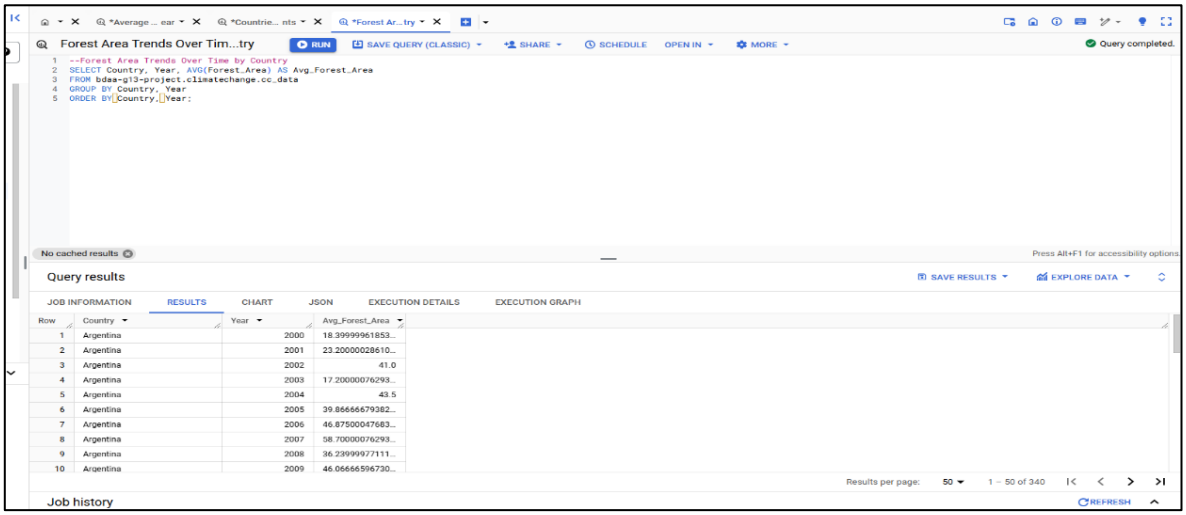


Figure 27a: Query 3

Insight: Examines deforestation or reforestation trends across regions, ***Figure 27b***.

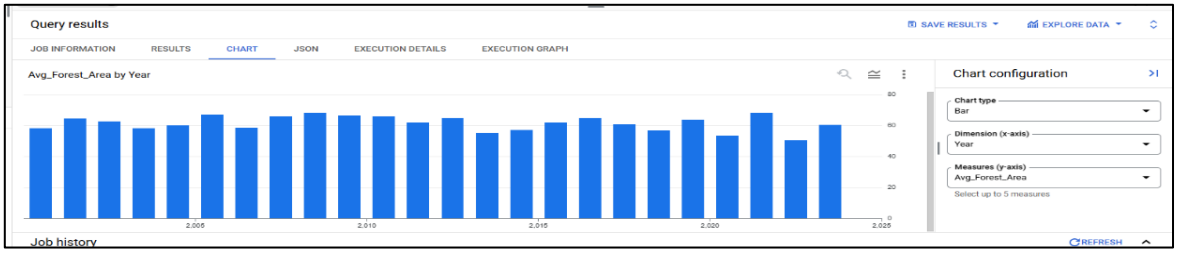


Figure 27b: Chart from Query 3

5.3.4 Relationship Between Renewable Energy Usage and CO2 Emissions:

Query 4: Analyses the relationship between renewable energy usage and per capita CO2 emissions, ***Figure 28a***.

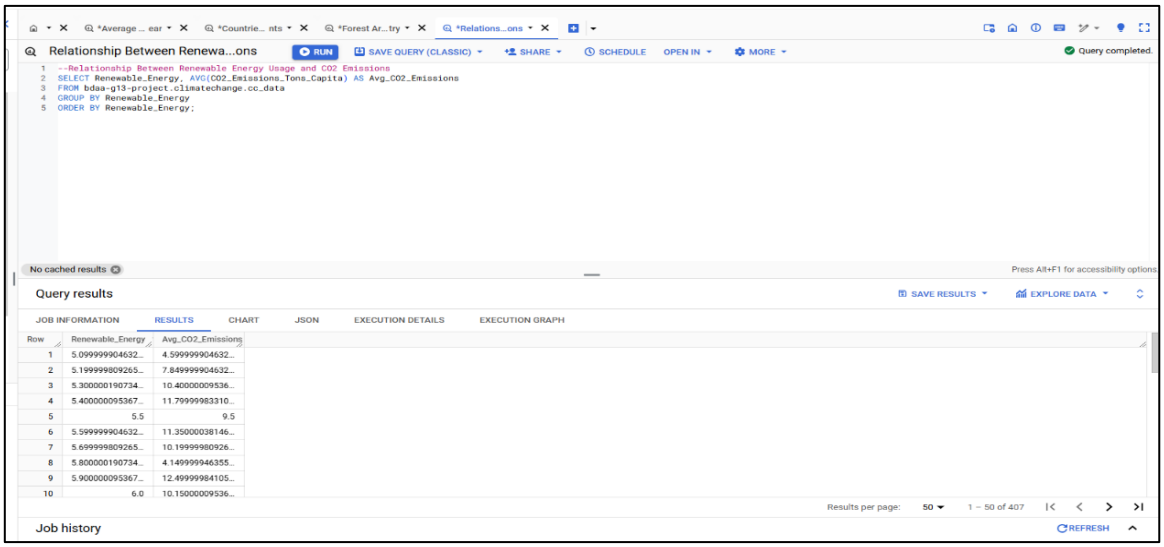


Figure 28a: Query 4

Insight: Evaluates how renewable energy adoption impacts carbon footprints, ***Figure 28b***.

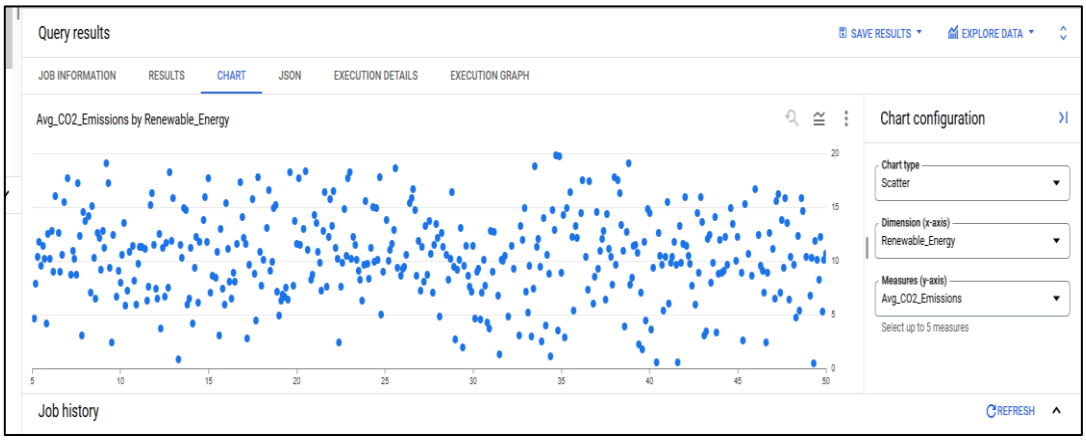


Figure 28b: Chart from Query 4

5.3.5 Sea Level Rise and Extreme Weather Events Correlation:

Query 5: Aggregates average sea level rise and weather events by year, ***Figure 29a***.

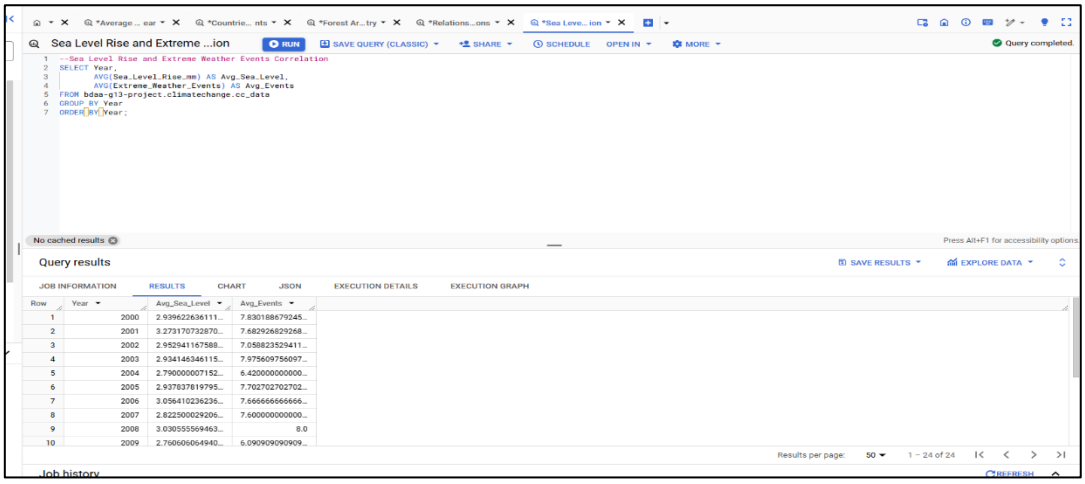


Figure 29a: Query 5

Insight: Investigates potential links between rising sea levels and increased weather extremity.

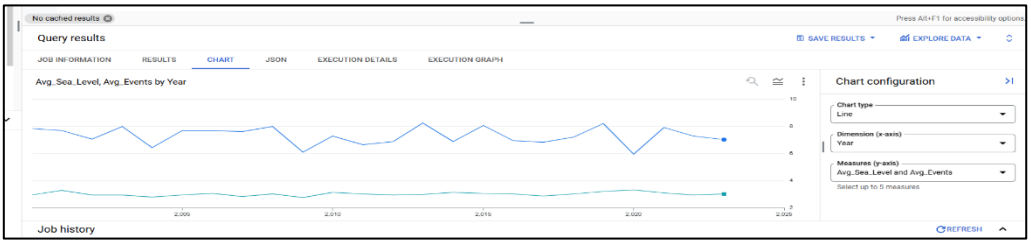


Figure 29b: Chart from Query 5

5.3.6 Top 10 Countries with the Highest CO2 Emissions:

Query 6: Ranks 10 countries by their CO2 emissions, ***Figure 30a***.

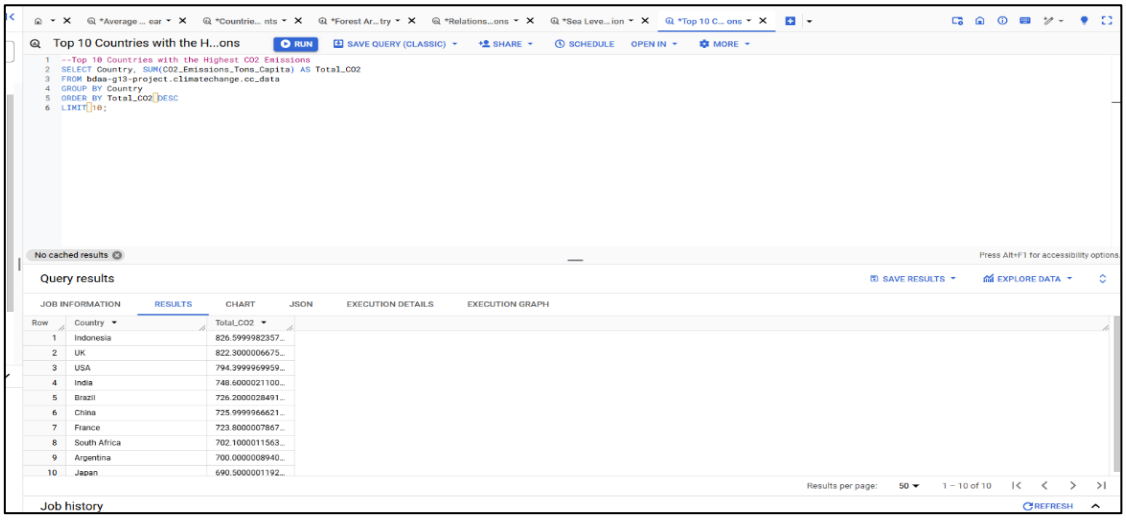


Figure 30a: Query 6

Insight: Identifies the largest contributors to global emissions for targeted environmental policies.

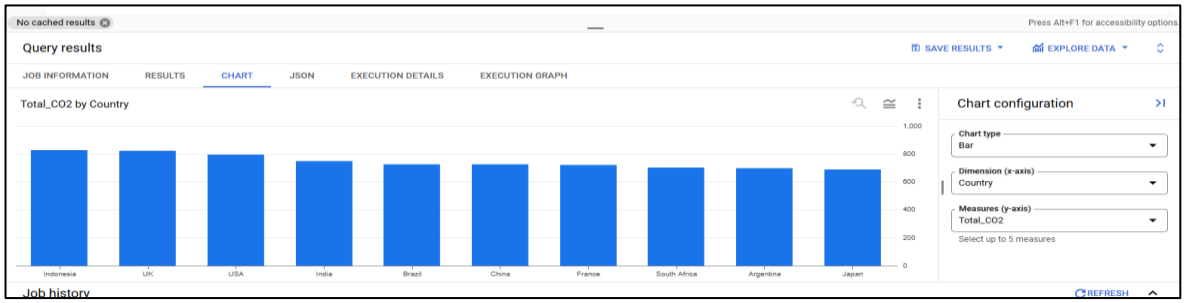


Figure 30b: Chart from Query 6

5.3.7. Top 10 Rainiest Countries:

Query 7: Ranks countries by average rainfall, ***Figure 31a***.

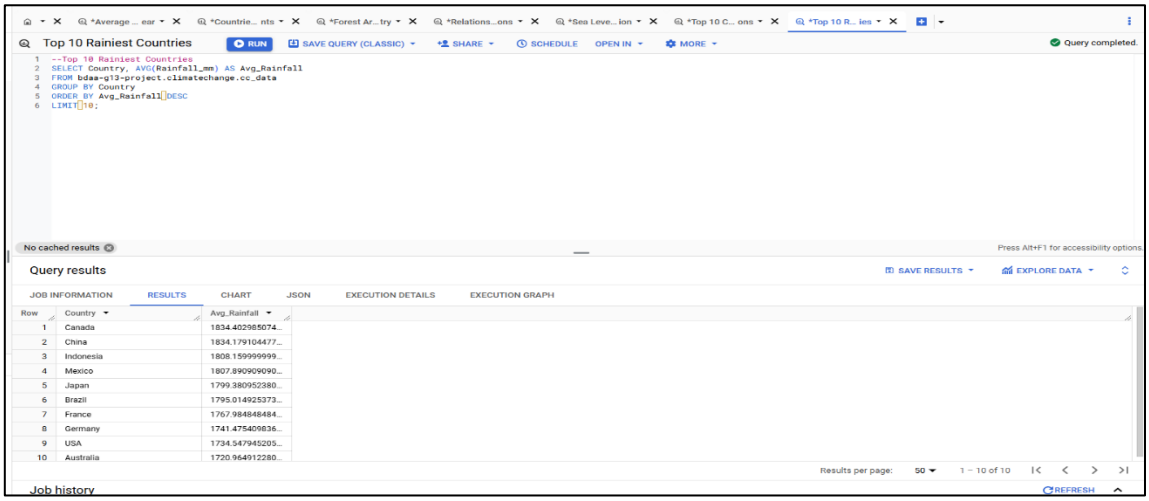


Figure 31a: Query 7

Insight: Highlights regions with the highest precipitation, critical for agriculture and water management, ***Figure 31b***.

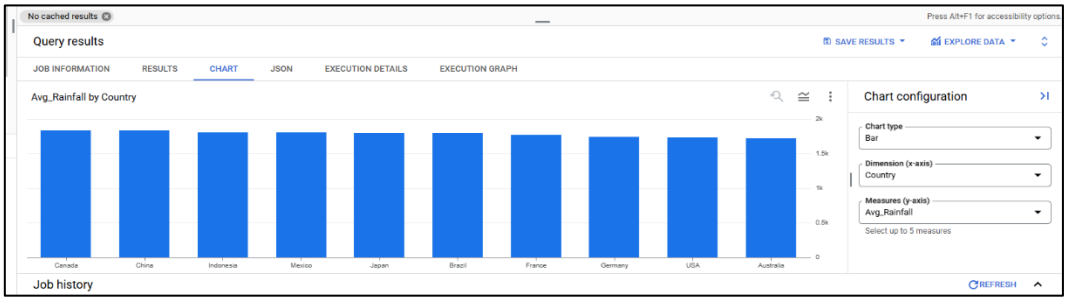


Figure 31b: Chart from Query 7

5.3.8 Yearly Change in CO2 Emissions:

Query 8: Analyses annual changes in CO2 emissions globally, ***Figure 32a***.

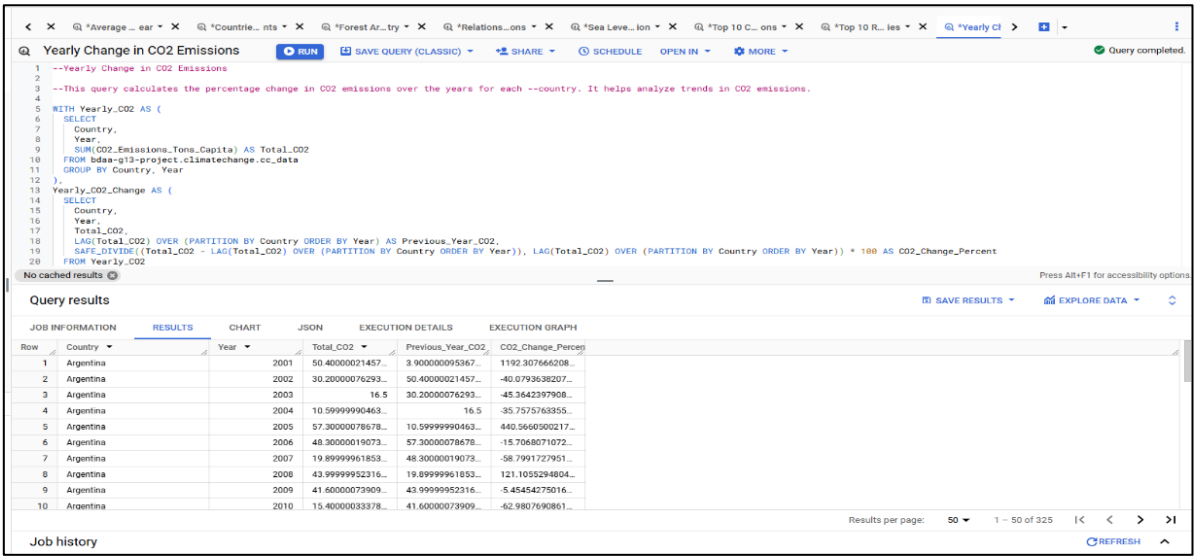


Figure 32a: Query 8

Insight: Tracks progress or setbacks in emission reduction efforts, ***Figure 32b***.

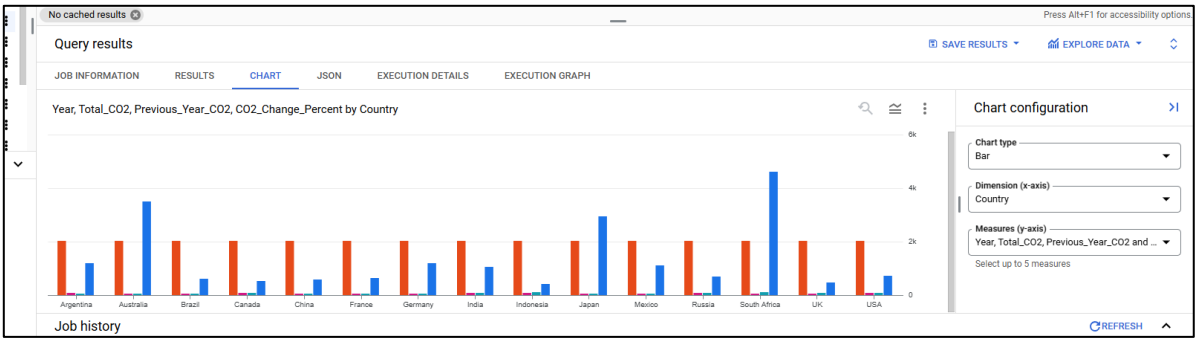
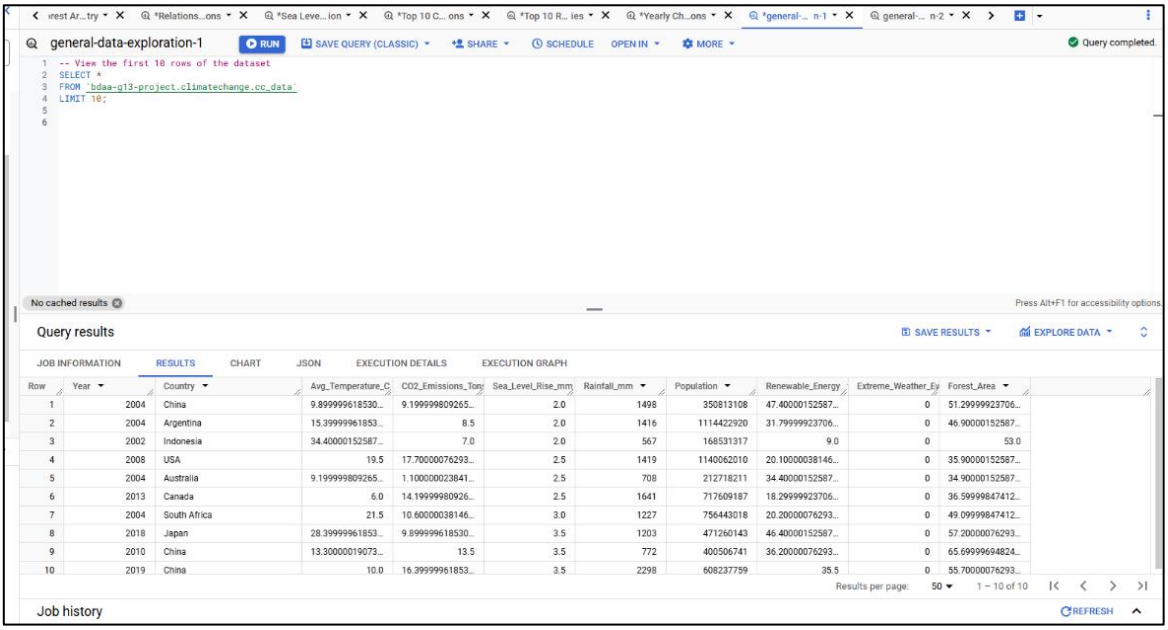


Figure 32b: Chart from Query 8

5.3.9 View the First 10 Rows of the Dataset:

Query 9: Displays a sample of the dataset, ***Figure 33a***.



Row	Year	Country	Avg_Temperature_C	CO2_Emissions_Ton	Sea_Level_Rise_mm	Rainfall_mm	Population	Renewable_Energy	Extreme_Weather_Ey	Forest_Area
1	2004	China	9.899999618530...	9.199999809265...	2.0	1498	350813108	47.40000152587...	0	51.29999923706...
2	2004	Argentina	15.39999961853...	8.5	2.0	1416	1114422920	31.79999923706...	0	46.90000152587...
3	2002	Indonesia	34.40000152587...	7.0	2.0	567	168531317	9.0	0	53.0
4	2008	USA	19.5	17.70000076293...	2.5	1419	1140062010	20.10000038146...	0	35.90000152587...
5	2004	Australia	9.199999809265...	1.100000023841...	2.5	708	212718211	34.40000152587...	0	34.90000152587...
6	2013	Canada	6.0	14.19999980926...	2.5	1641	717609187	18.29999923706...	0	36.59999847412...
7	2004	South Africa	21.5	10.60000038146...	3.0	1227	756443018	20.20000076293...	0	49.09999847412...
8	2018	Japan	28.39999961853...	9.899999618530...	3.5	1203	471260143	46.40000152587...	0	57.20000076293...
9	2010	China	13.30000019073...	13.5	3.5	772	400506741	36.20000076293...	0	65.69999694824...
10	2019	China	10.0	16.39999961853...	3.5	2298	608237759	35.5	0	55.70000076293...

Results per page: 50 1 – 10 of 10

Figure 33a: Query 9

Insight: Helps understand data structure and content before deeper analysis.

5.4 Data Visualization

5.4.1 Connecting Looker with BigQuery

As the last step in the data processing pipeline which is the visualization part, as shown in ***Figure 34***, start by going to Looker Studio (https://lookerstudio.google.com), and click on the "+" button to create new visualization dashboard.

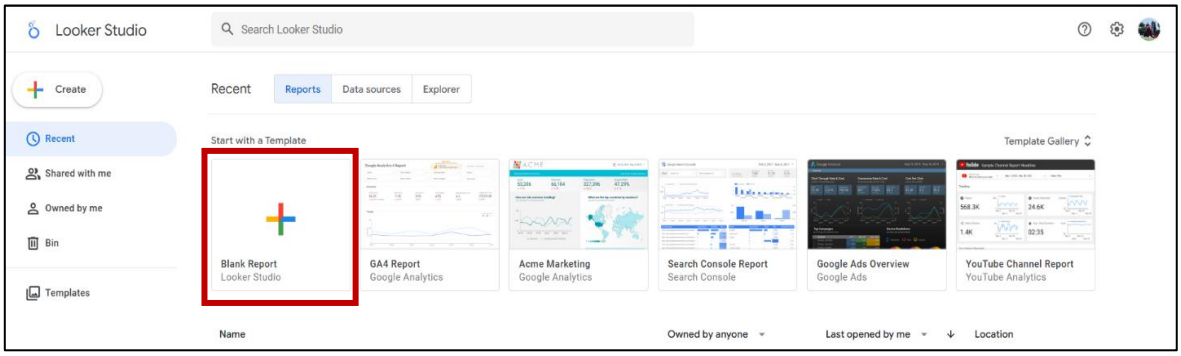


Figure 34: Create blank report

As we are using BigQuery to store the data, clicked on the BigQuery button, as illustrated in ***Figure 35***, access authorization would be needed afterwards by logging in to BigQuery account.

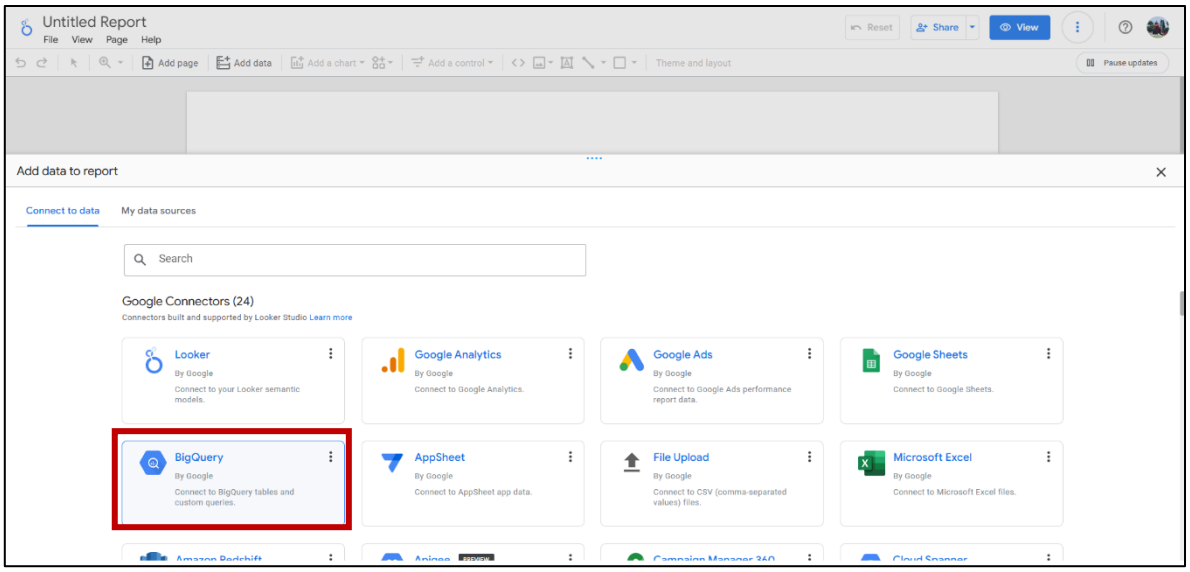


Figure 35: Choosing BigQuery as Dataset Source

Based on ***Figure 36***, we need to choose which “Project”, “Data set” and “Table” that we stored our data; in this case Project “bdaa-g13-project", Data set “climatechange” and Table “cc_data”.

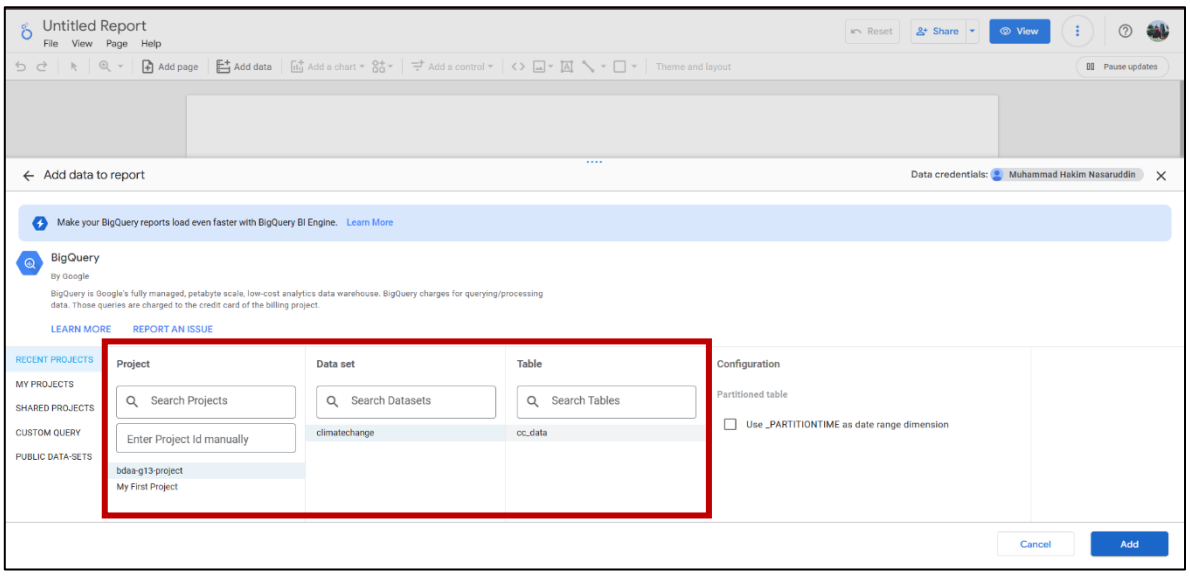


Figure 36: Configuring dataset for dashboard

5.4.2 Creating Dashboard in Looker

As illustrated in ***Figure 37***, to create a chart, start by clicking on" Add a chart" button. Choose" Bubble map" chart to create a fascinating map chart to consolidate the countries involved in this analysis.

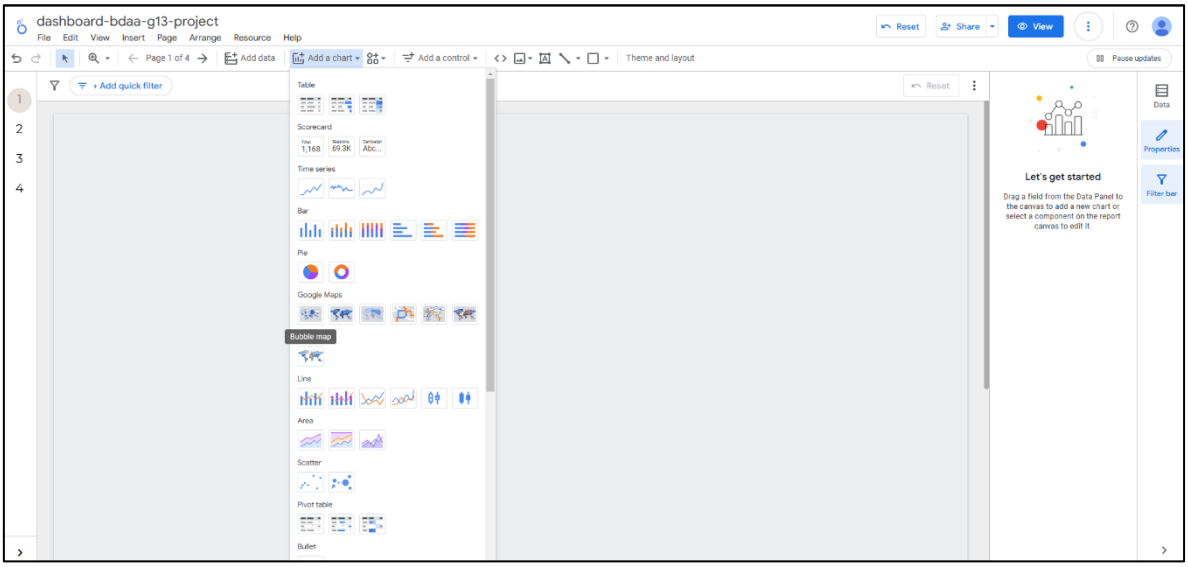


Figure 37: Add a chart in the dashboard

Drag and drop appropriate dimensions into respective field so that data can be visualized as per intended, as shown in ***Figure 38***. By dropping "Country" into Location Field, "Forest_Area" into Size field and Avg_Temperature_C into Color metric, map chart as shared below (XXX) can be obtained.

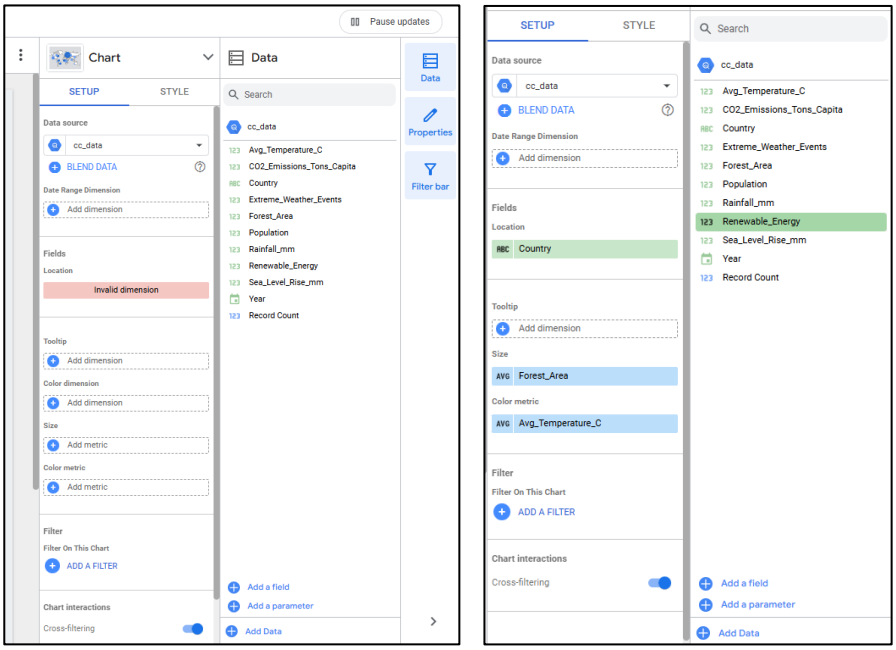


Figure 38: Add the corresponding dimensions into the charts

Based on ***Figure 39a***, **Figure *39b***, and ***Figure 39c***, please notice that the aggregation of the data chosen can be changed as we see fit by clicking on the button to the left side of the data column name, as in this case we chose to use Average so that we will be able to get data from each country per year by average for more inclusivity yet simple visualization.

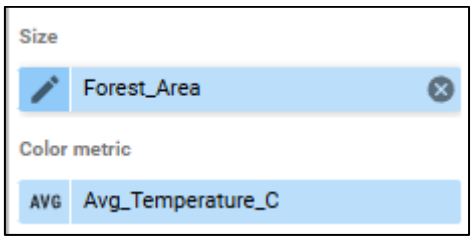


Figure 39a: "AVG" Aggregation for "Avg_Temperature_C"

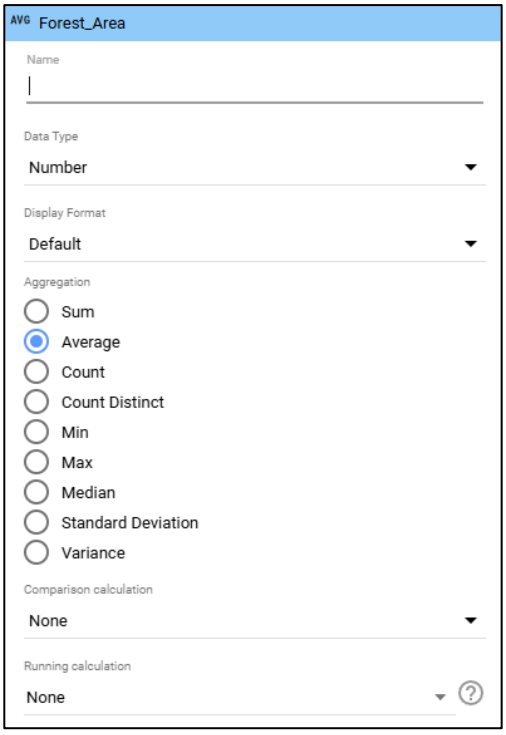


Figure 39b: Detailed configuration of "Aggregation" settings

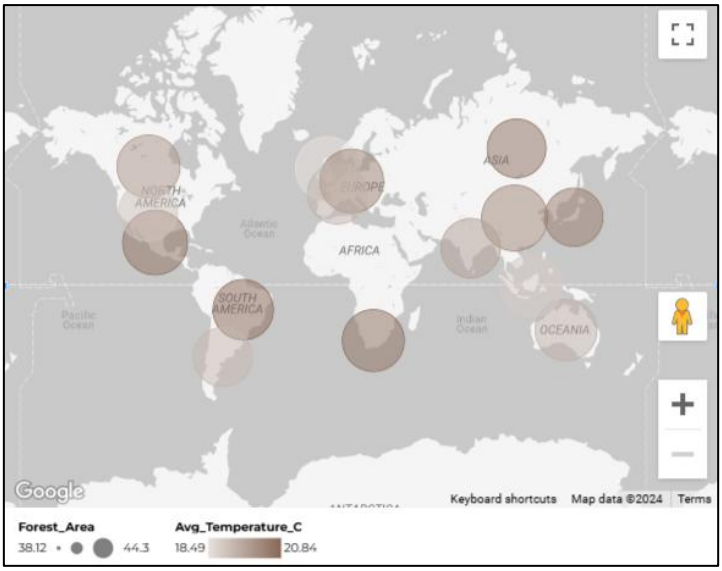


Figure 39c: Map chart

Based on ***Figure 40a*** and ***Figure 40b***, the design of the map chart can be customized to make it more appealing by going to the “Style” tab, where you can input the title of the chart, change the type of the map’s style, change the colour for the bubbles and various more.

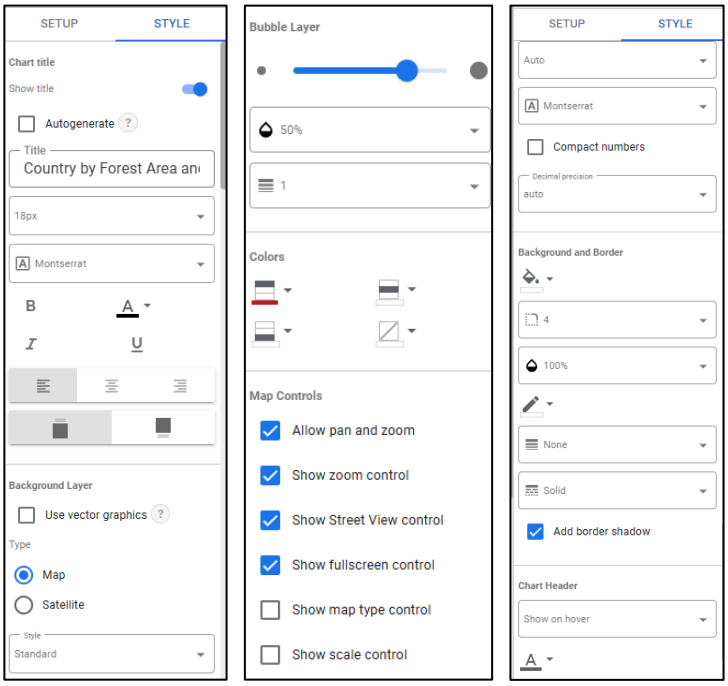


Figure 40a: “Style” settings of a chart

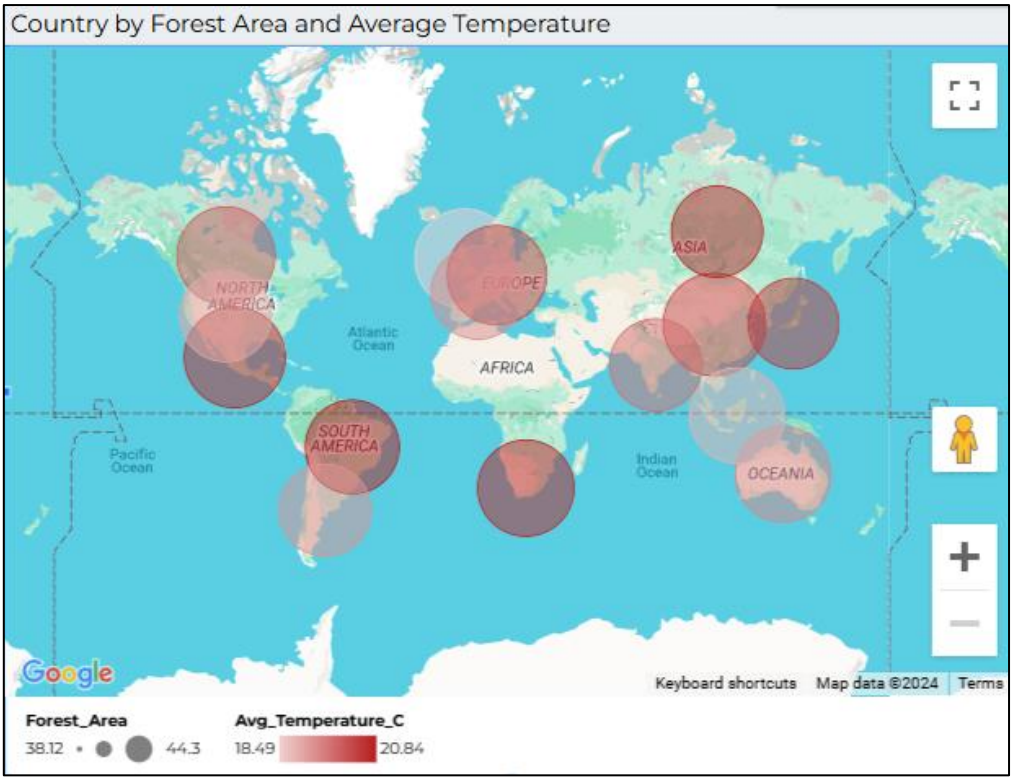


Figure 40b: End result of the customization done

5.4.3 Insights form the Dashboard Created

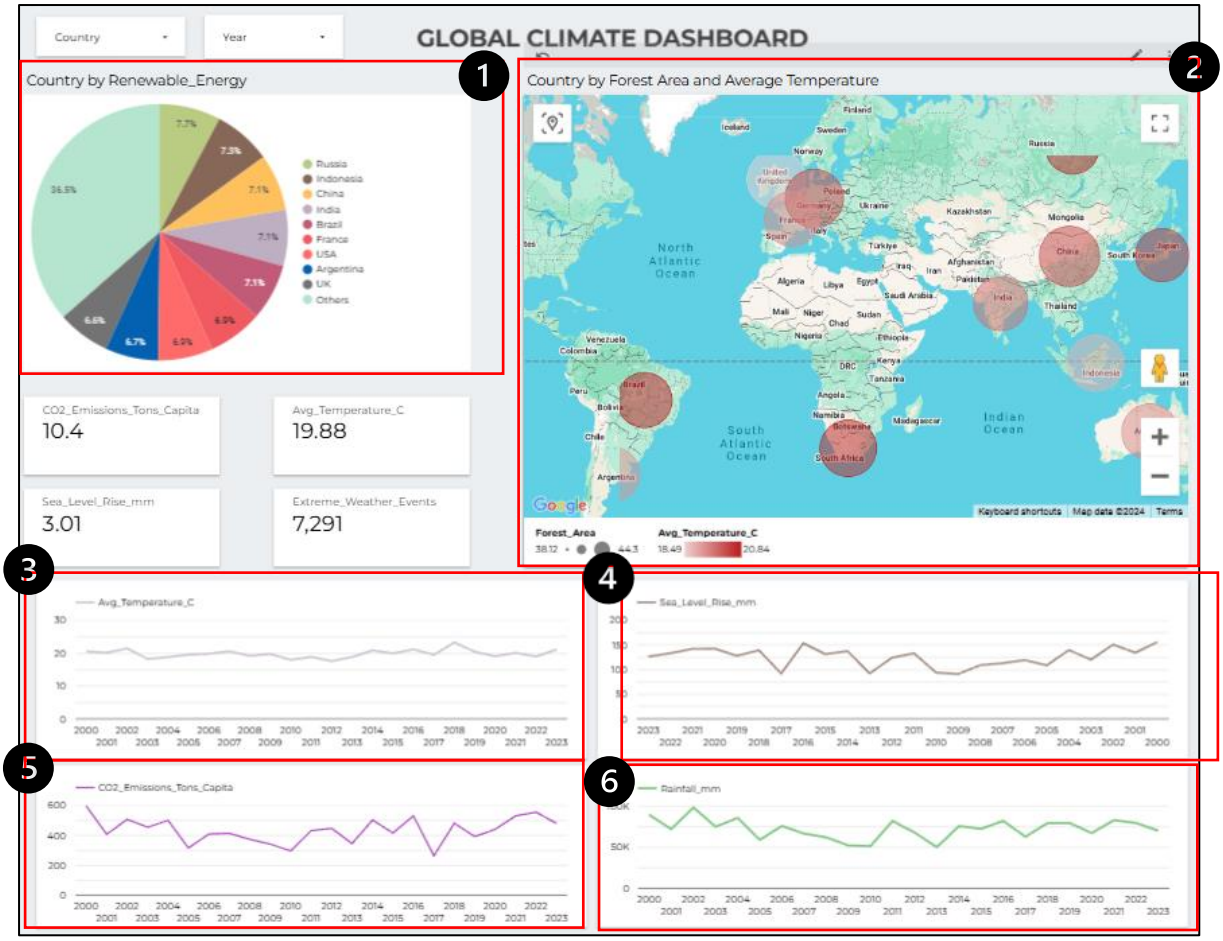


Figure 41: Dashboard created

5.4.3.1 Country by Renewable Energy

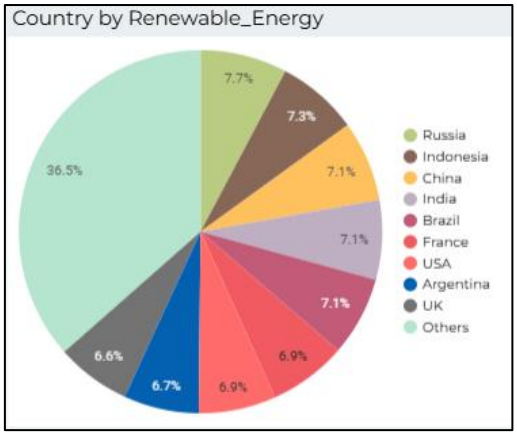


Figure 42: Pie chart of Country by Renewable Energy

Based on ***Figure 42***:

- The chart shows the percentage of renewable energy usage by country.
- Russia leads with the highest renewable energy adoption, followed by Indonesia and China, indicating significant investments in green energy.
- Other countries, including the USA and UK, have smaller shares.

5.4.3.2 Geo Chart: Country by Forest Area and Average Temperature

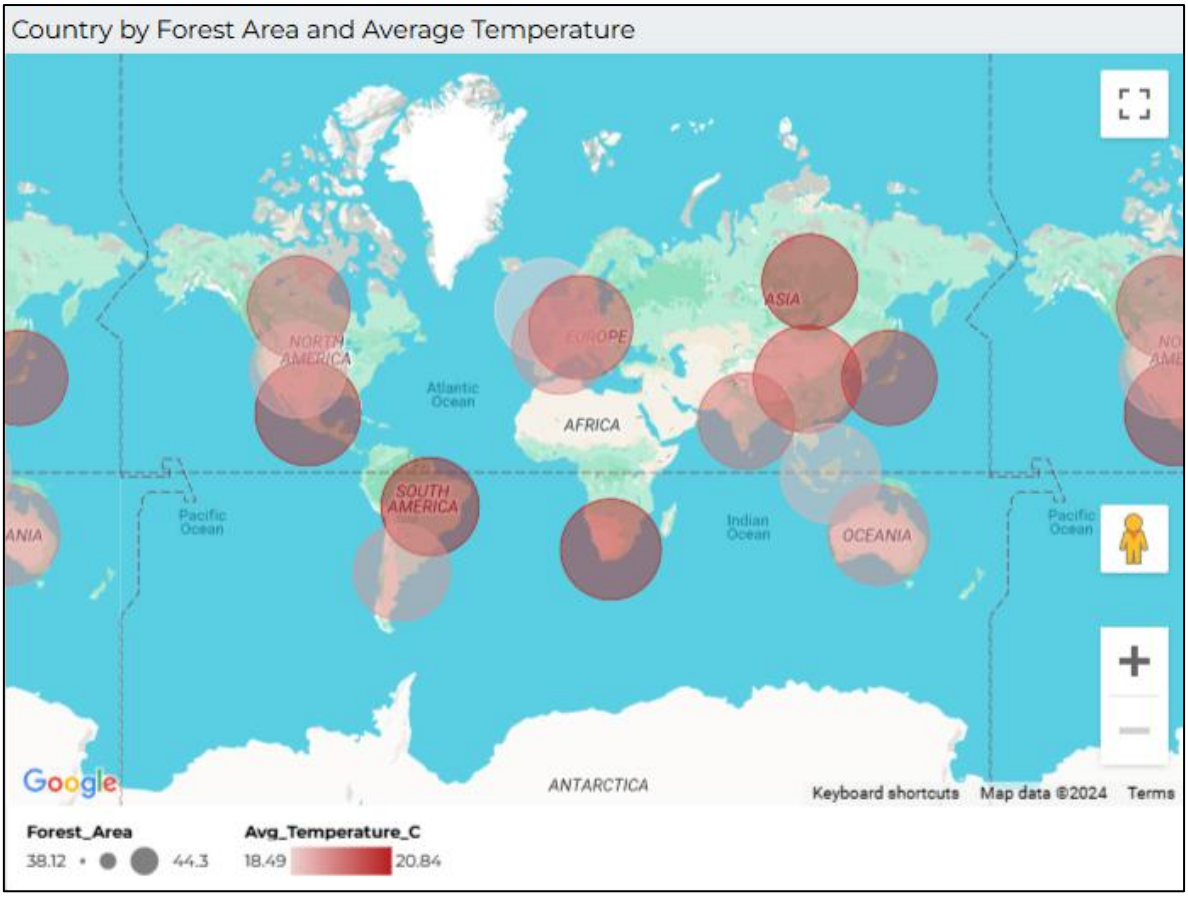


Figure 43: Geo chart of Country by Forest Area and Average Temperature

Based on ***Figure 43***, the map visualizes forest coverage and temperature variation by country:

- Countries like **Brazil** and **Russia** exhibit larger forest areas, critical for global climate regulation.
- Countries in regions such as South Asia and Africa reflect higher average temperatures, emphasizing the impacts of global warming.

5.4.3.3 Line Chart: Average Temperature Over Time

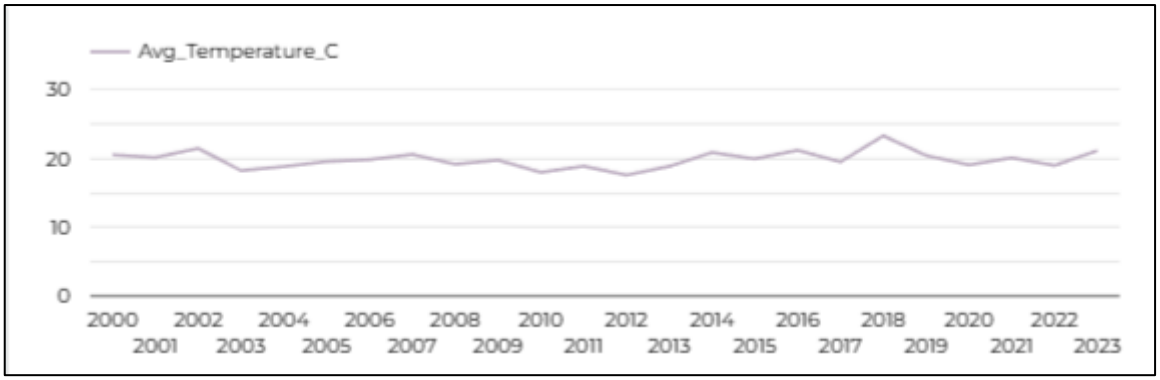


Figure 44: Line chart of Average Temperature over Time

Based on ***Figure 44***:

- The chart highlights a steady increase in global average temperatures over the years, signalling ongoing climate change.
- This trend reinforces the need for immediate measures to curb greenhouse gas emissions.

5.4.3.4 Line Chart: Sea Level Rise Over Time

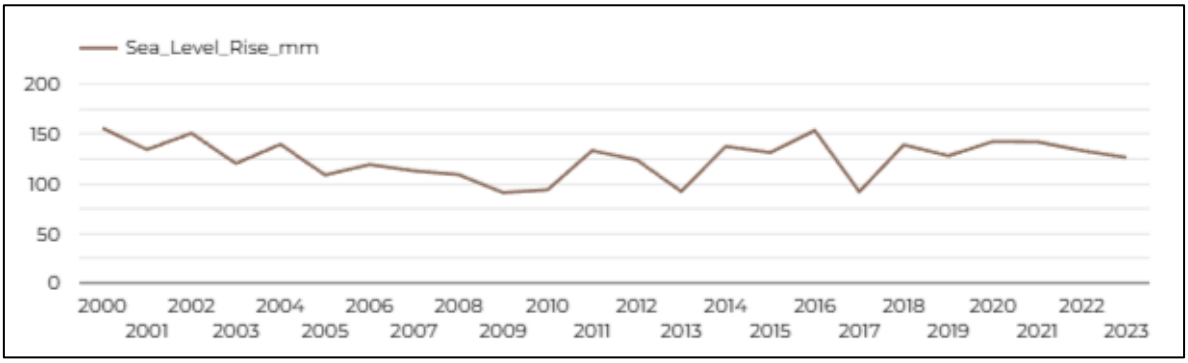


Figure 45: Line chart of Sea Level Rise over Time

Based on ***Figure 45***:

- The line chart shows a consistent rise in sea levels, with fluctuations that may correspond to regional climatic events.
- The overall upward trend aligns with increased ice melt and thermal expansion due to global warming.

5.4.3.5 Line Chart: CO2 Emissions Per Capita Over Time

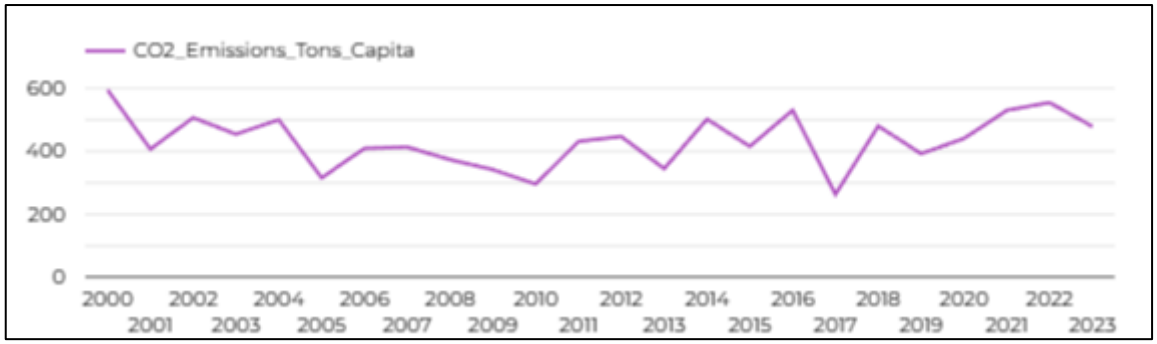


Figure 46: Line chart of CO2 Emissions Per Capita over Time

Based on ***Figure 46***:

- The CO2 emissions show fluctuations without a significant overall decrease, highlighting the challenges in transitioning to low-carbon economies.
- Certain periods reflect peaks, potentially due to industrial growth or policy changes.

5.4.3.6 Line Chart: Rainfall Over Time

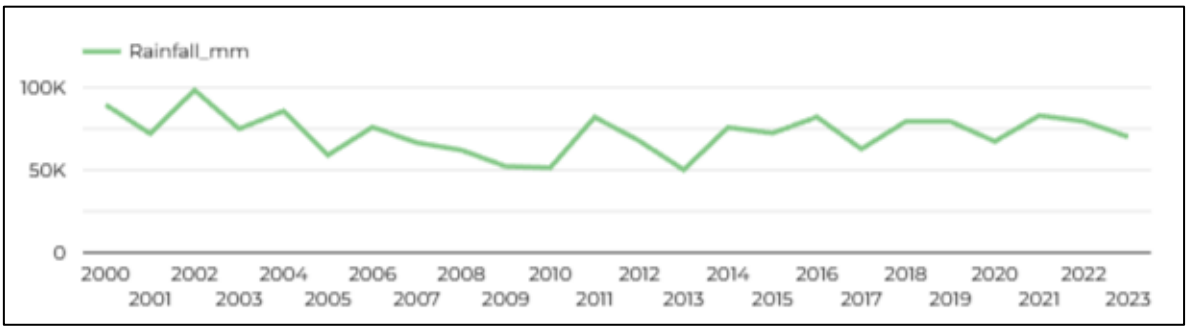


Figure 47: Line chart of Rainfall over Time

Based on ***Figure 47***,

- The chart shows variations in global rainfall over time.
- The fluctuating pattern indicates regional and seasonal shifts in precipitation, likely linked to climate change.

CHAPTER 6: EVALUATION METRICS WITH GRAPHS

This section presents the evaluation metrics employed to assess the performance of various tools, with a primary focus on BigQuery.

6.1 BigQuery Execution Time Insights

6.1.1 Elapsed Time for Each Query:

The query **"Yearly_Change_in_CO2_Emission"** exhibits the longest elapsed time, ***Figure 48***, making it the most time-intensive to execute, while other queries, such as **"Most_Extreme_Weather_Events_Countries"** demonstrate comparatively shorter execution times.

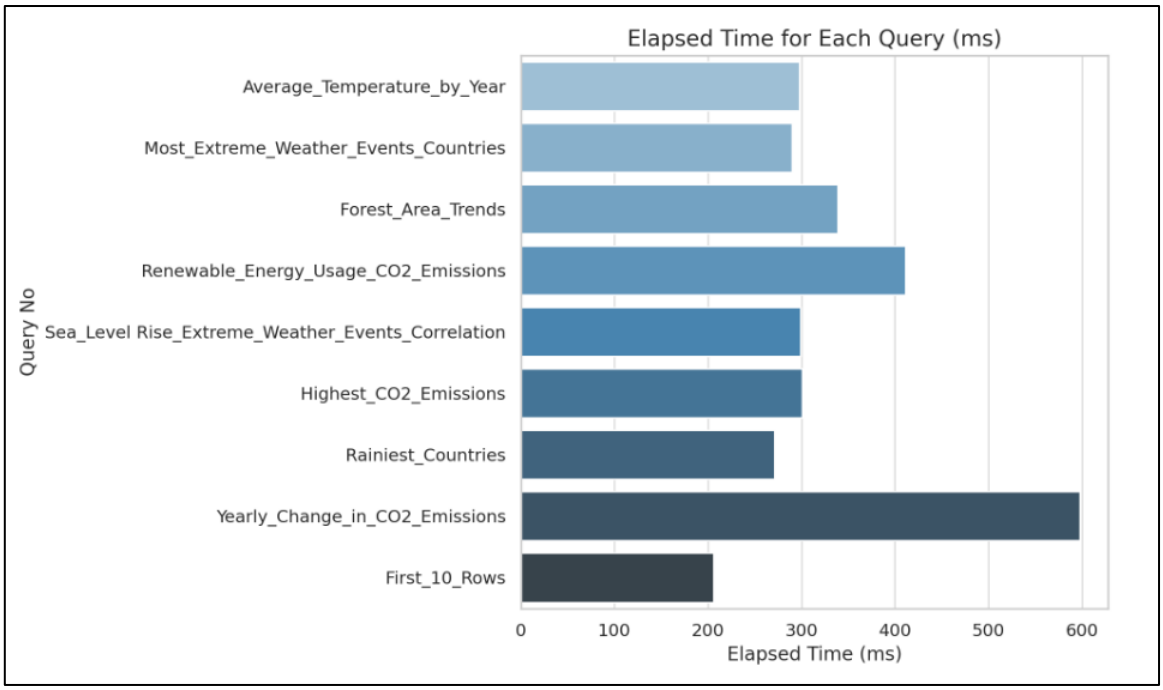


Figure 48: Horizontal Bar Chart of Elapsed Time for Each Query (ms)

6.1.2 Slot Time Consumed for Each Query:

The query "**Yearly_Change_in_CO2_Emissions**" records the highest slot time consumption, ***Figure 49***, whereas queries like "**Average_Temperature_by_Year**" consume significantly less slot time.

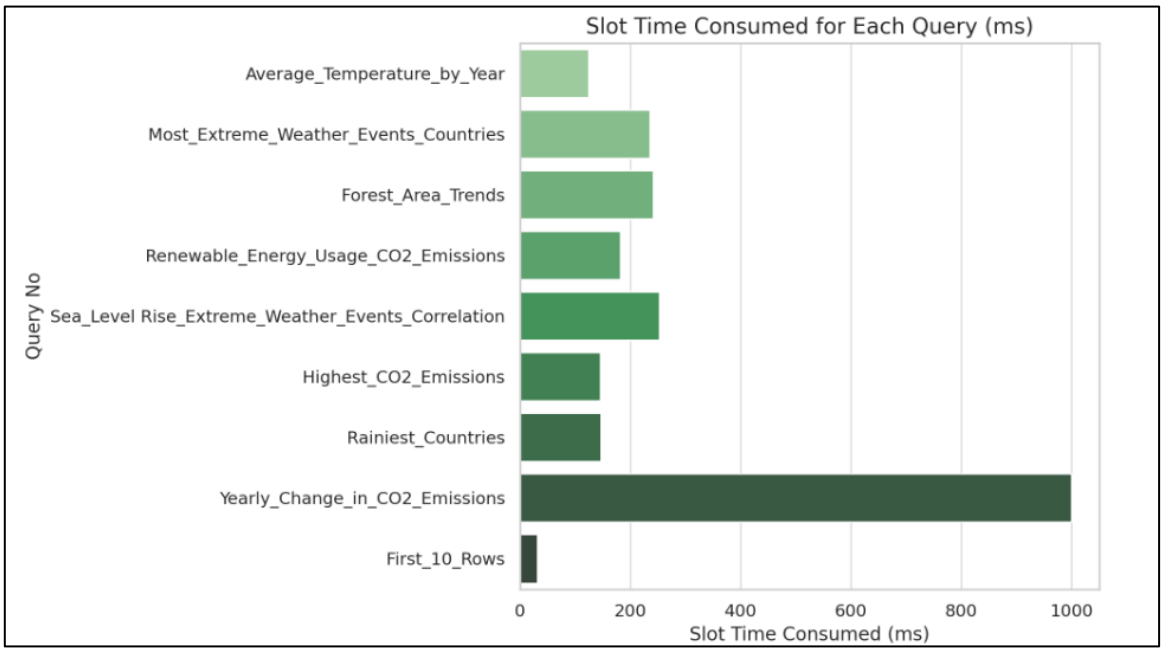


Figure 49: Slot Time Consumed for Each Query

6.1.3 Bytes Shuffled for Each Query:

The query "**Yearly_Change_in_CO2_Emission**" involves the highest data shuffling, ***Figure 50***, indicating intensive data processing, while the query "**Highest_CO2_Emissions**" entails significantly less data shuffling.

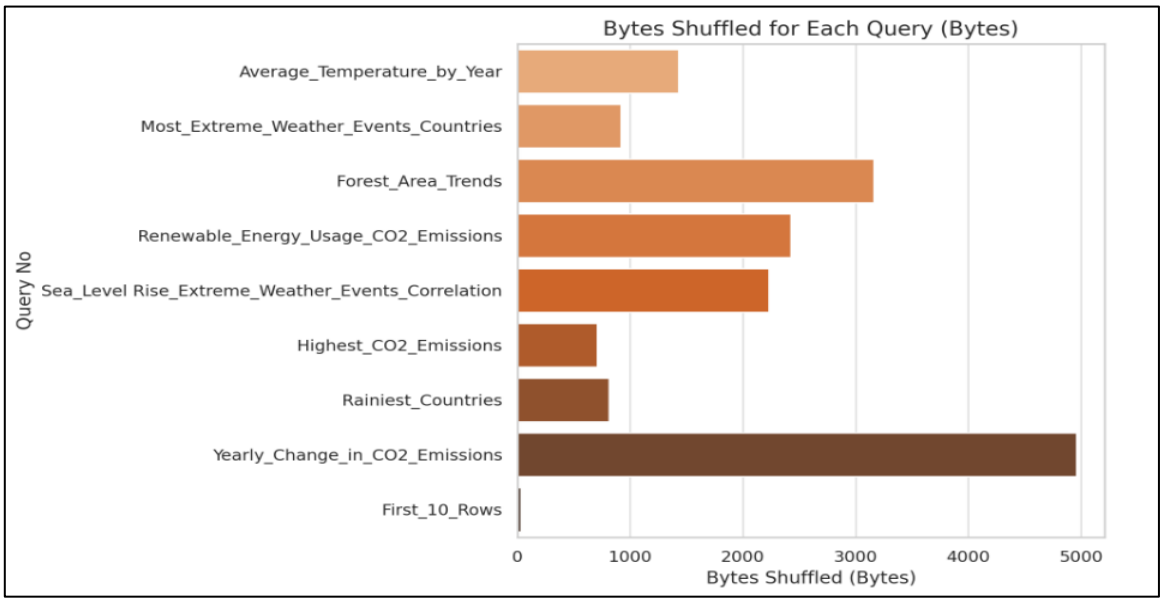


Figure 50: Bytes Shuffled for Each Query

6.1.4 Correlation Between Metrics:

There is a strong positive correlation (0.87) between **elapsed time** and **slot time consumed**, ***Figure 51***, indicating that longer execution times are associated with higher slot time usage. Additionally, **bytes shuffled** exhibit a moderate correlation with both elapsed time and slot time, suggesting their role in increasing execution complexity.

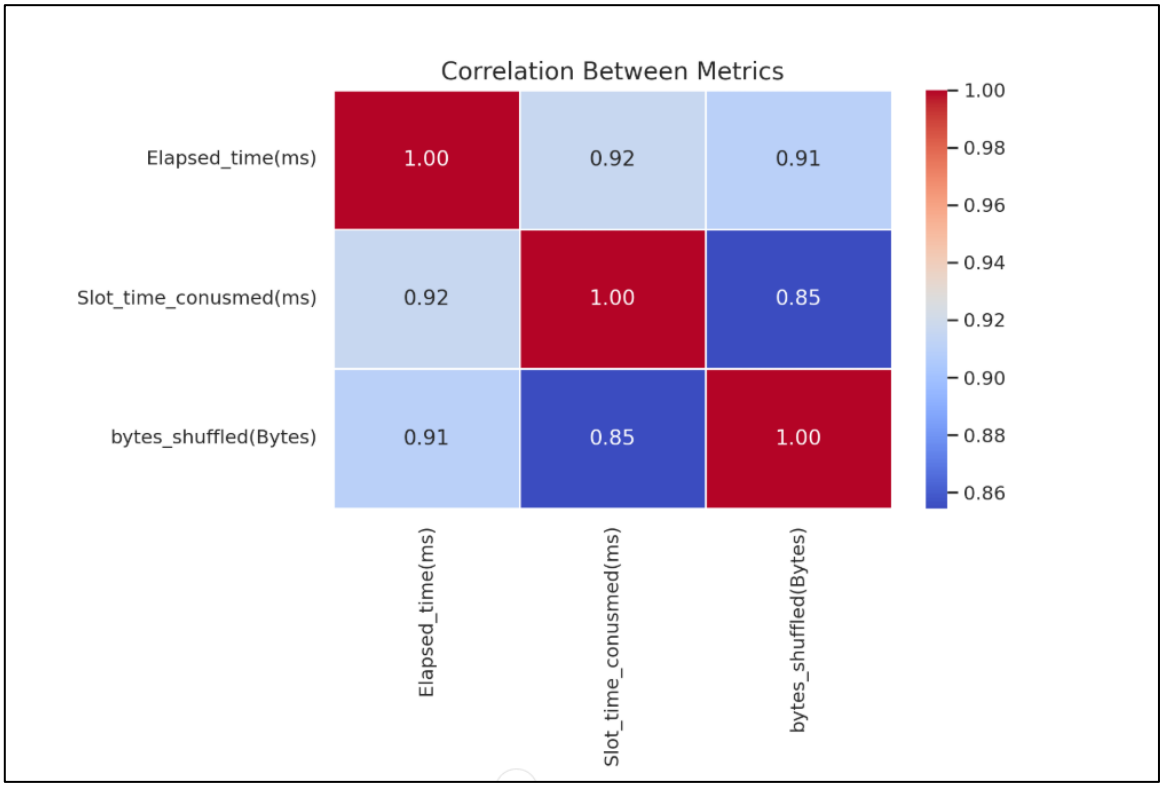


Figure 51: Correlation Between Metrics