GHG and Methane Emissions Monitoring and Analysis Using AWS Glue and AWS Quicksight

Big Data Analytics Project Report

CIS 602: Special Topics in CIS, Summer 2024

Group 17

Members:

(3)Sindhuja Baikadi(02128756)

(38)Vaishnavi Paineni(02125170)

Project Title: GHG and Methane Emissions Monitoring and Analysis Using AWS Glue and AWS Quicksight

Abstract:

The "Dynamic GHG and Methane Monitoring and Analysis Using AWS Glue and AWS Quicksight" project aims to establish a comprehensive and scalable system for tracking real-time GHG and Methane emissions. By leveraging Amazon S3 for data ingestion, AWS Glue for data processing, and AWS Quicksight for Analysis, the project will enable continuous monitoring and in-depth analysis of GHG and Methane levels. This system is designed to provide immediate insights and critical information to support environmental management and policy-making efforts.

Data Set:

The dataset contains data from all the countries over the world. It has the following columns: Entity(Country), Code, Year, GHG, and Methane. The data is recorded over some time from 1850 to 2022.

- https://ourworldindata.org/co2-and-greenhouse-gas-emissions

Objectives:

1. Data Acquisition: Implement Amazon S3 Data Stream to consistently collect GHG and Methane emissions data from multiple sources such as industrial sensors, satellite data, and public APIs.

2. Data Preprocessing and Enrichment: Use AWS Glue to clean, normalize, and enhance incoming GHG and Methane emissions data with additional contextual information like geographic and temporal details.

3. Instantaneous Data Insights: Design real-time analytics to measure critical GHG and Methane metrics (e.g., concentration levels, emission trends) for specific regions, providing rapid insights into changes in GHG and Methane emissions.

4. Querying the data through Athena: The data stored in the s3 bucket i.e., raw data and processed data is queried using Athena.

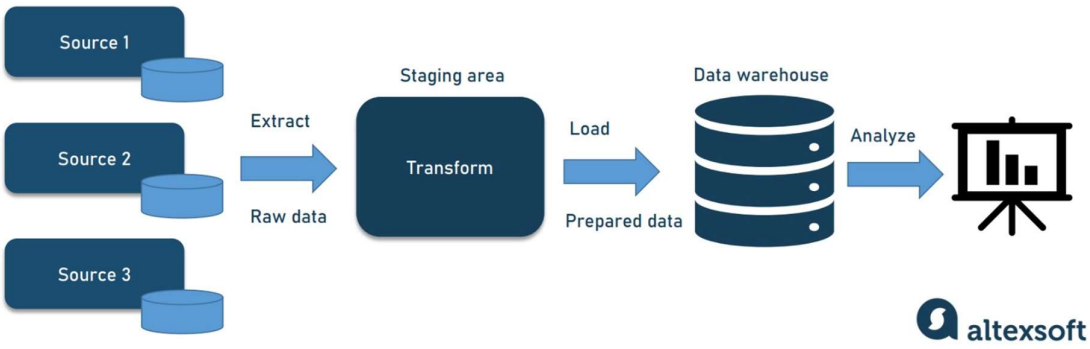
5. Structured Data Storage: Store processed GHG and Methane emissions data in Amazon S3, systematically organized by geographic location, and indexed using AWS Glue Data Catalog for easy retrieval and analysis.

6. Scheduled Data Analysis: Set up AWS Glue jobs to conduct regular batch processing, including historical GHG and Methane data analysis to detect long-term trends and anomalies in emissions.

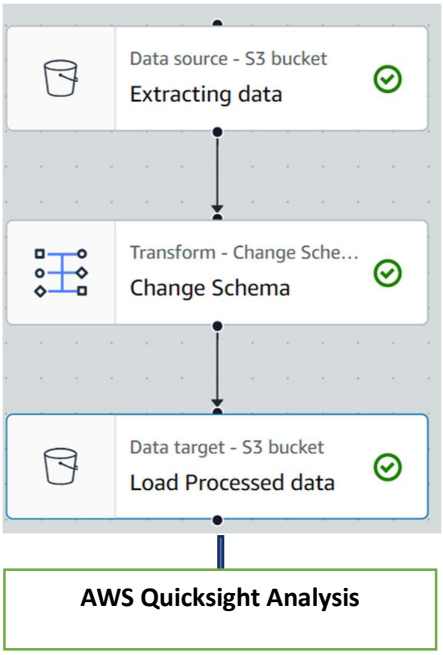
7. Interactive Data Representation: Utilize Amazon QuickSight or similar data visualization tools to create dynamic dashboards and reports that display real-time and historical GHG and Methane emissions data in an easily interpretable format.

8. Performance Monitoring: Amazon CloudWatch continuously monitors the system's performance and health, ensuring reliability and efficiency.

AWS Data Pipeline:



Our Data Pipeline:



1. Data Ingestion: Ingest data related to GHG and Methane into an AWS S3 bucket.

2. Data Transformation: Use AWS Glue to preprocess and transform the incoming data on GHG and Methane. This step includes data cleaning, renaming columns, converting data types, and converting the data into Parquet format. Additionally, catalog the data using AWS Glue Data Catalog for streamlined access and discovery.

3. Data Storage: Store the transformed data in Amazon S3.

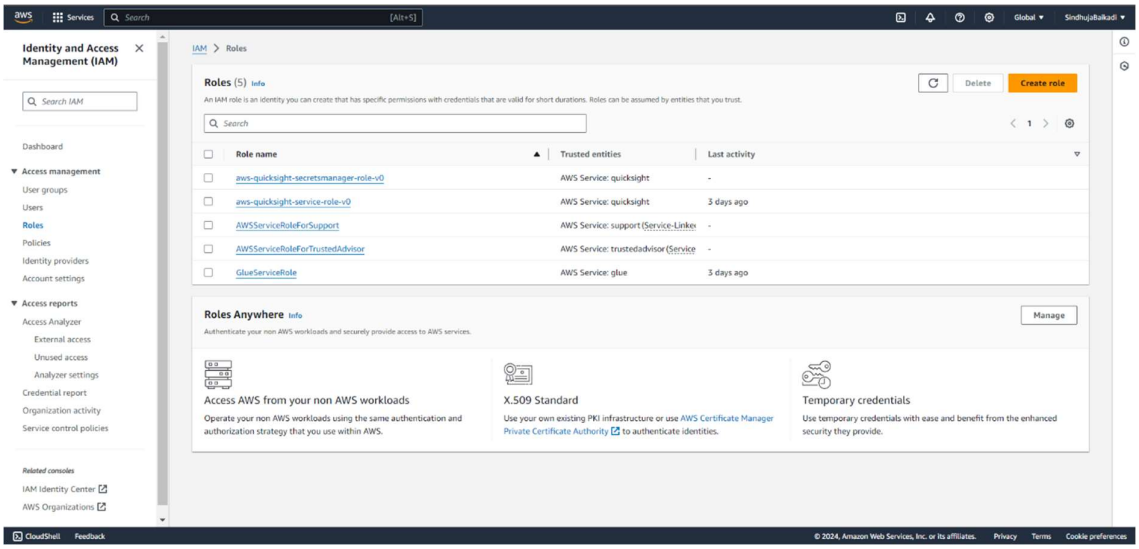
4. Batch Data Processing: Configure AWS Glue jobs to execute periodic batch processing tasks on the GHG and Methane data.

5. Data Visualization: Employ a visualization tool like Amazon QuickSight to create interactive dashboards for analyzing and presenting the data.

Step-1: Setting IAM Role

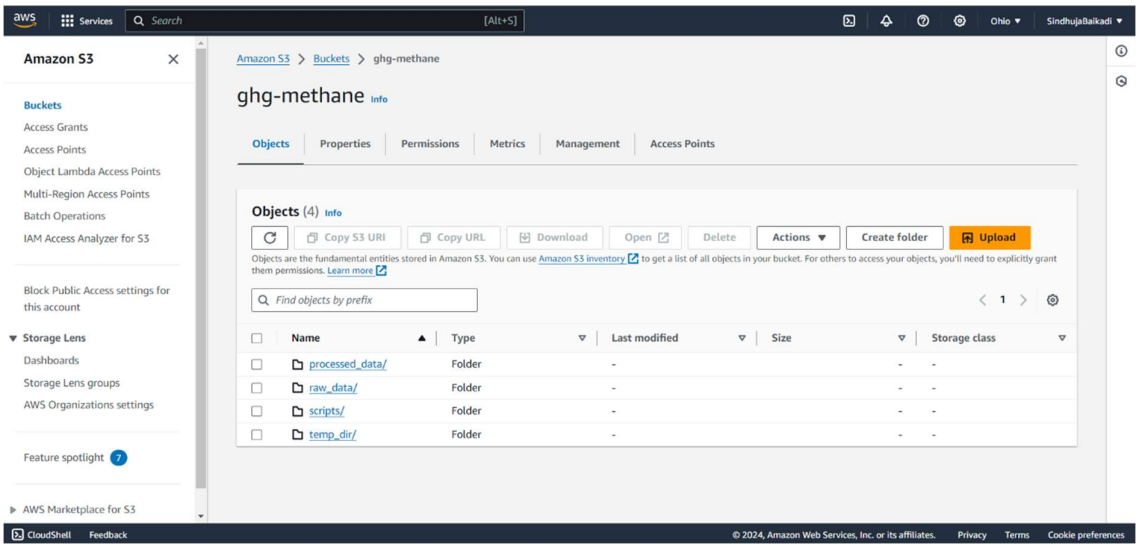
An IAM role in AWS is a set of permissions that allows users or services to perform actions on AWS resources. Unlike IAM users, roles are not tied to specific individuals and provide temporary access through assumed credentials. They are useful for granting access to resources without needing to share long-term credentials.

Created an IAM role with the name **"GlueServicerole"** with administrative access so that it can be used in all applications of AWS like AWS GLUE.



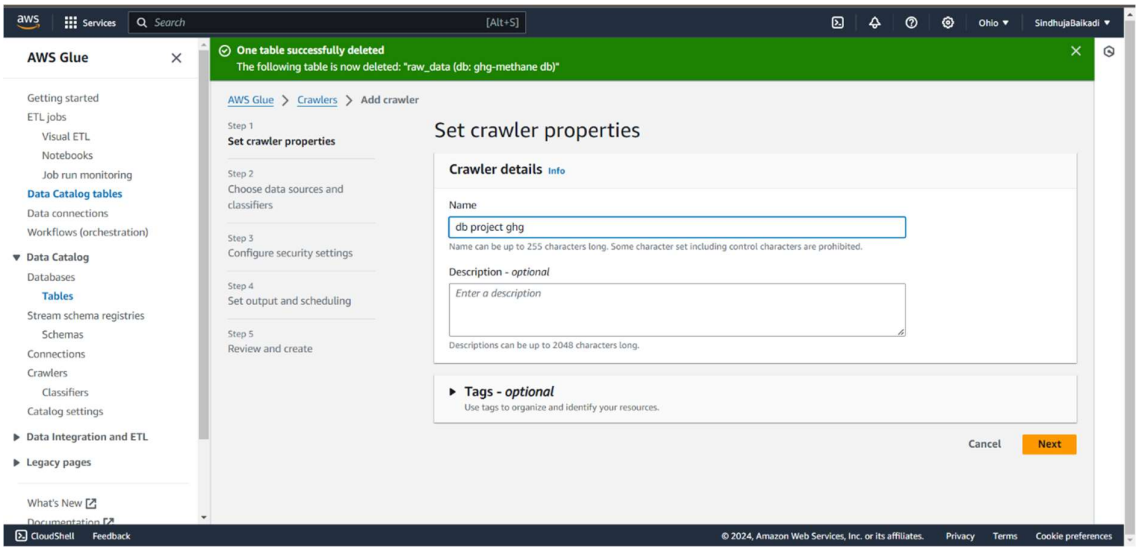
Step-2: Creating S3 buckets

An S3 bucket named **"ghg-methane"** was set up, containing subfolders for different purposes. The `**raw_data**` subfolder is used for uploading the initial raw data, and `**processed_data**` stores data after it has been transformed. A `**scripts**` folder holds the Py-Spark scripts necessary for these transformations, and a `**temp_dir**` folder is designated for temporary storage by AWS Glue during the querying process.

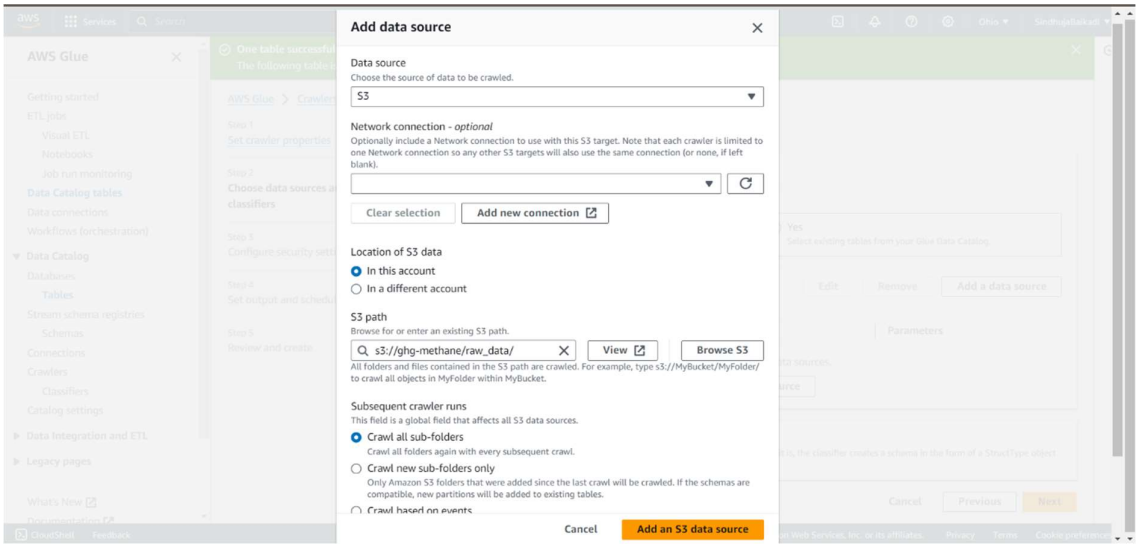


Step-3: AWS GLUE – ETL Setup and Process

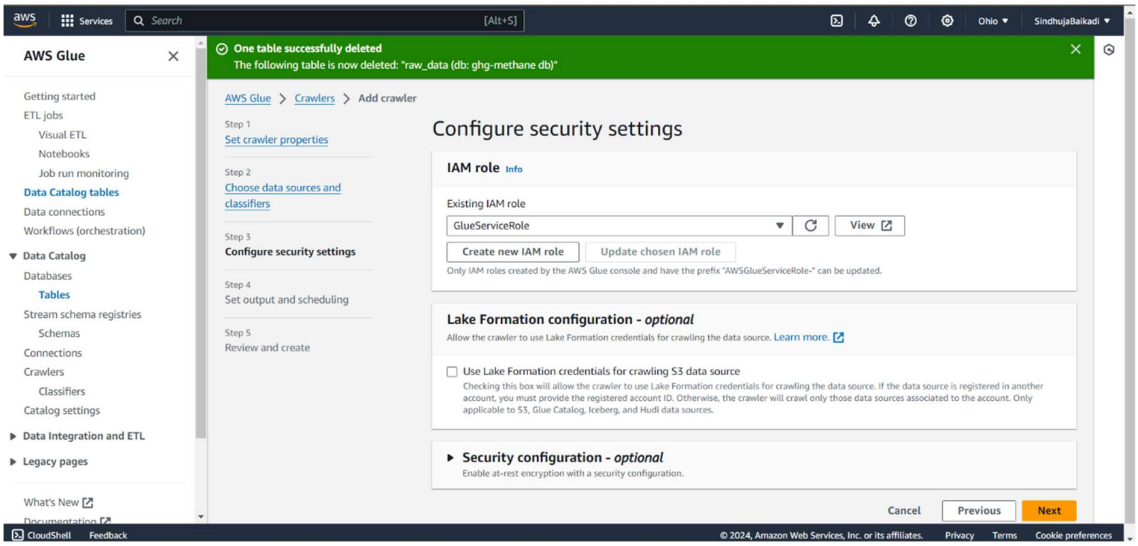
Open the AWS Management Console, search for "AWS Glue," and select it to access the AWS Glue console. In the navigation pane, choose Tables under Databases, then click "Add Tables using a crawler" and enter "db project ghg" as the name. Click "Next."



To add a data source, select S3 and choose "In this account" for the data location. Enter the S3 bucket path for raw_data, set the crawler to "Crawl all sub-folders," and add the S3 data source.

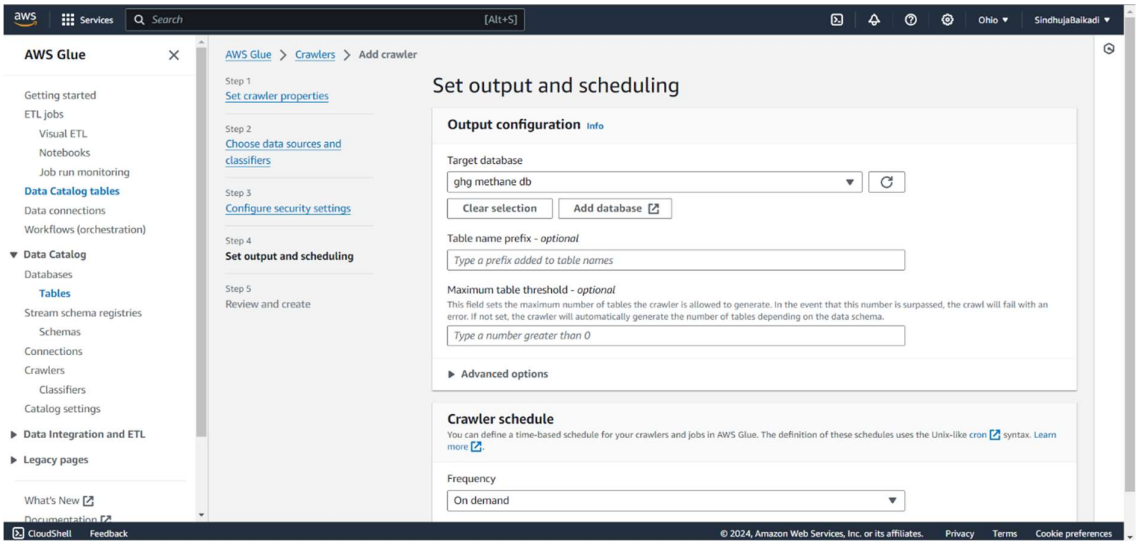


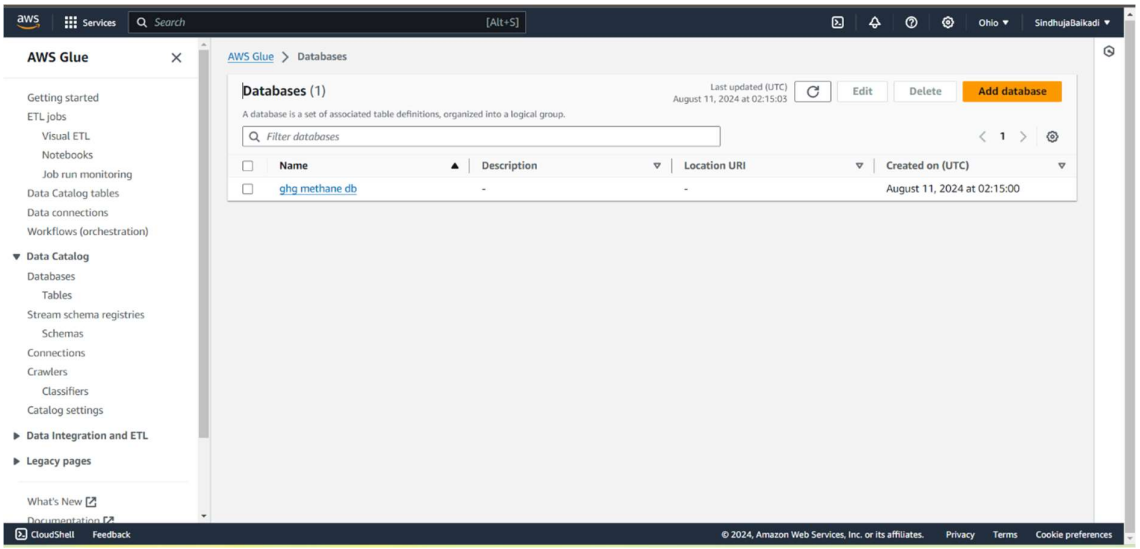
Click "Next," select "GlueServiceRole" for the IAM role, and proceed by clicking "Next" again.



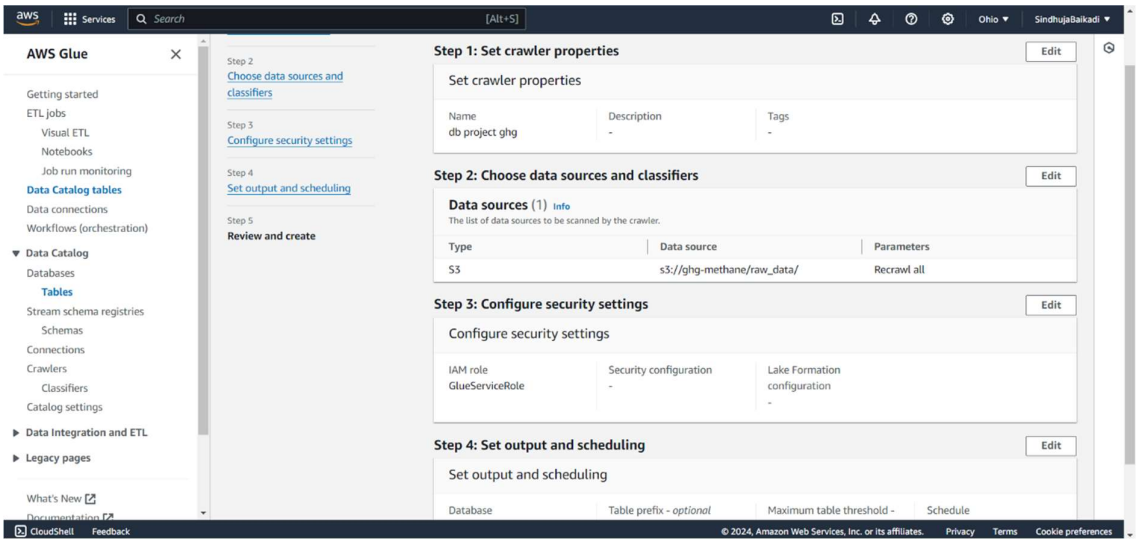
In the Output configuration section, select "Add database" to open a new tab, name the database "ghg methane db," and click "Create database."

Return to the Set output and scheduling page, and for the Target database, select "ghg methane db." In the Crawler schedule section, keep the Frequency set to "On demand" and click "Next."

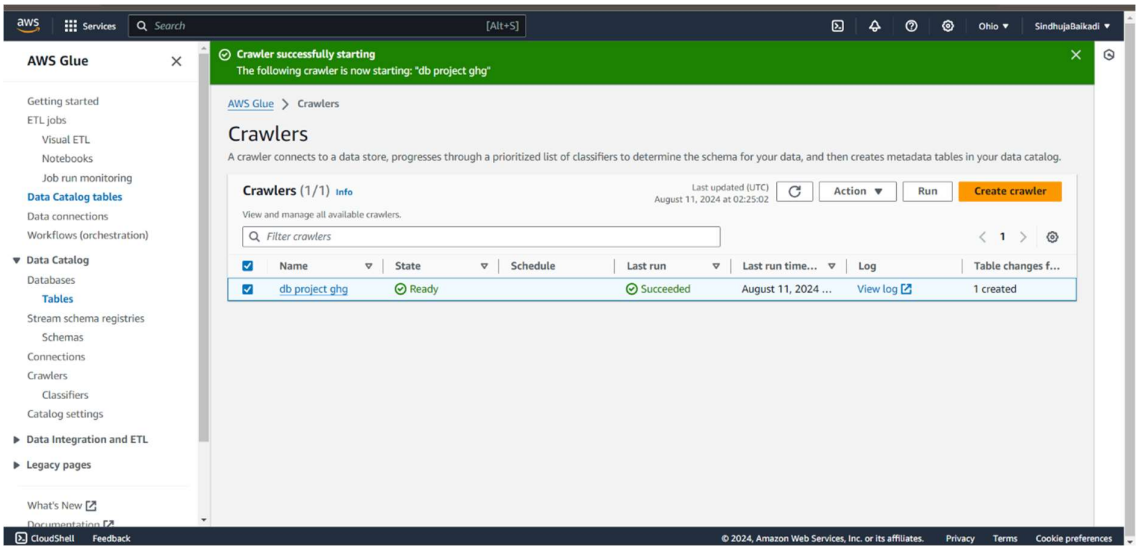




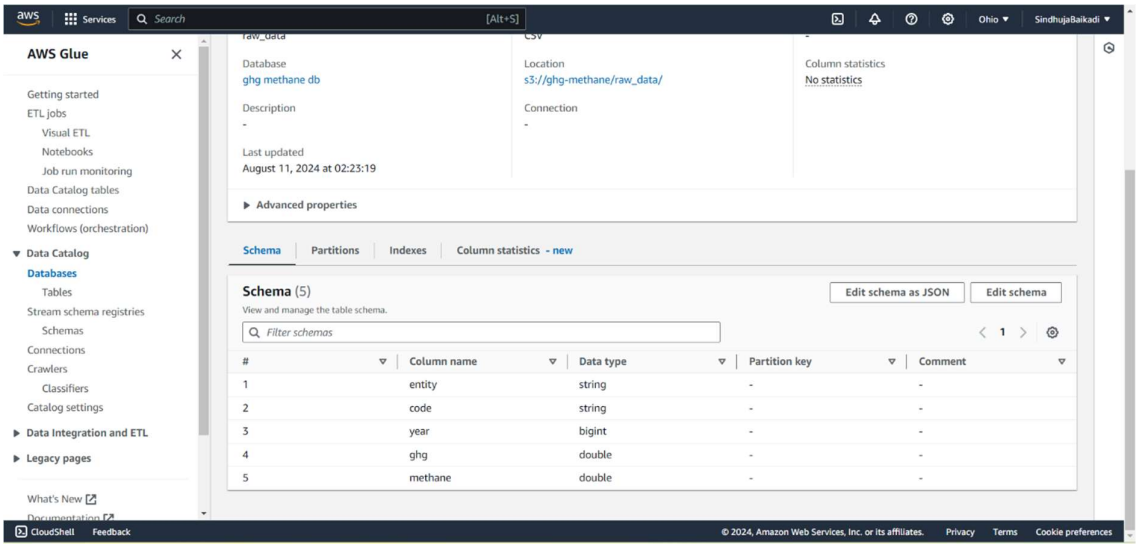
Review the configuration, then click "Create crawler."



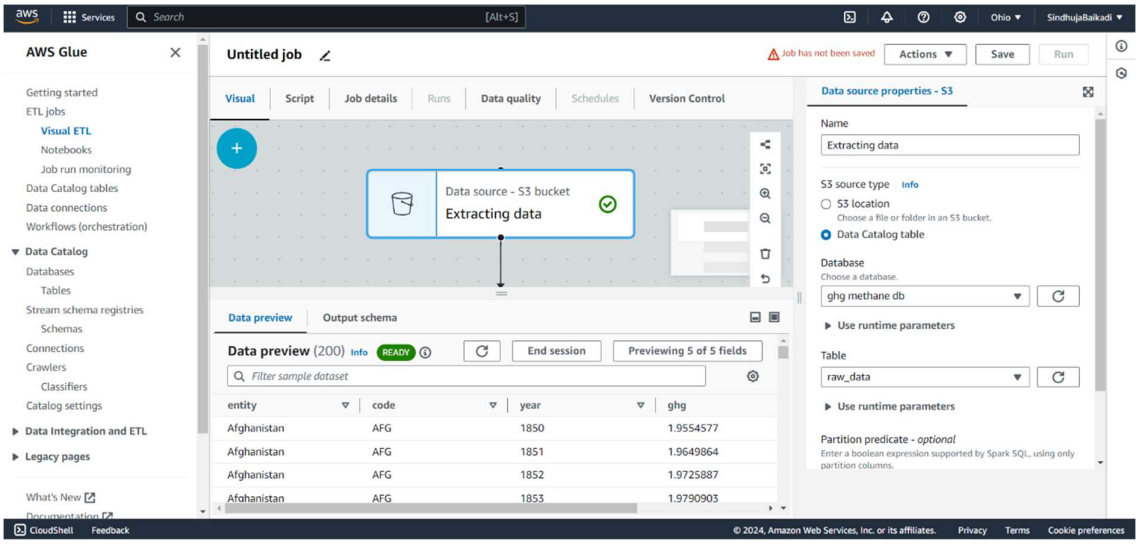
Start the AWS Glue crawler you set up to execute the extract and load steps of your ETL process. Go to the Crawlers page, select the "db project ghg" crawler, and click Run. The status will update to Running as it creates the database and builds the metadata.



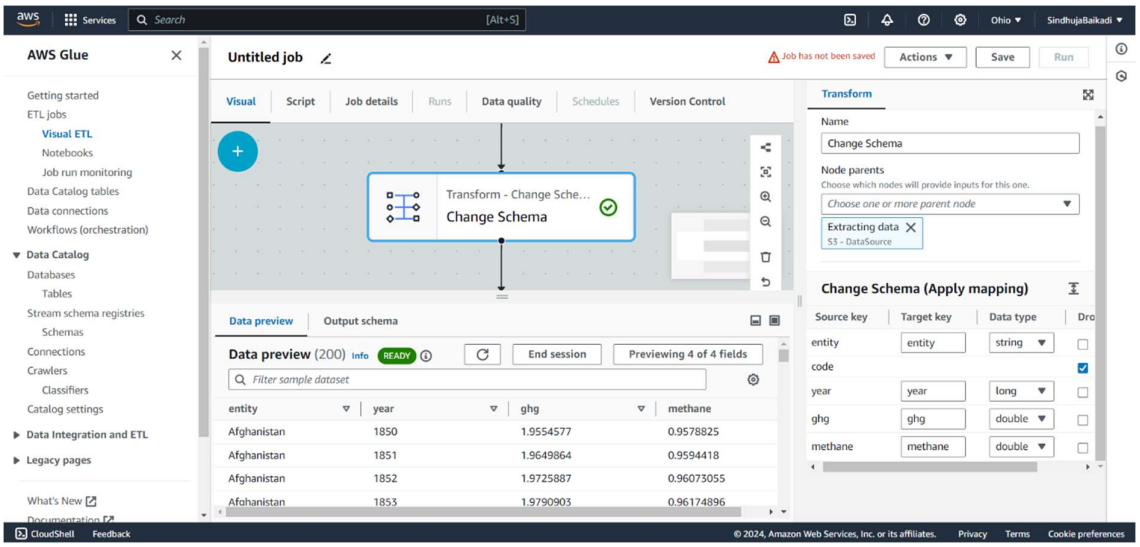
To check the AWS Glue metadata, go to the Databases section in the console, and select the "ghg methane db" database. This will show the schema with the columns identified by the crawler.



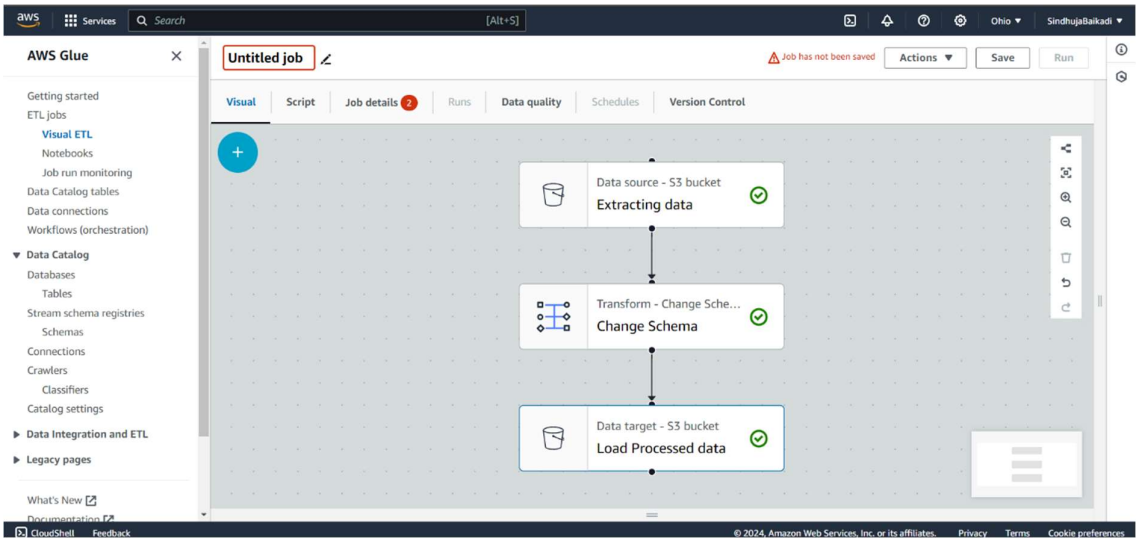
Under ETL Jobs, click on Visual ETL, **Extract** – The data was imported from the CSV file and uploaded to the designated location within the S3 bucket.



Transform – Various transformations were applied, such as checking for null values, renaming columns, and removing unnecessary columns that won't be needed in the subsequent stages of the ETL process. Dropped Code column and renamed entity column name as the country.

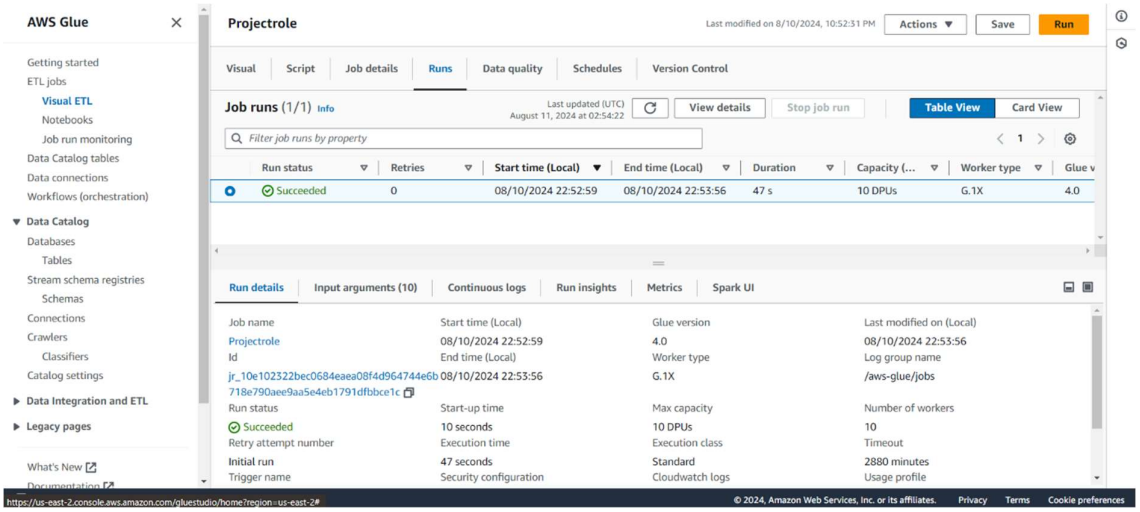


Load – After completing the transformations, the processed data was uploaded to a specific S3 location for storing processed data. Additionally, a data catalog table was created to store this processed data. The code column was removed and the entity column renamed as country.



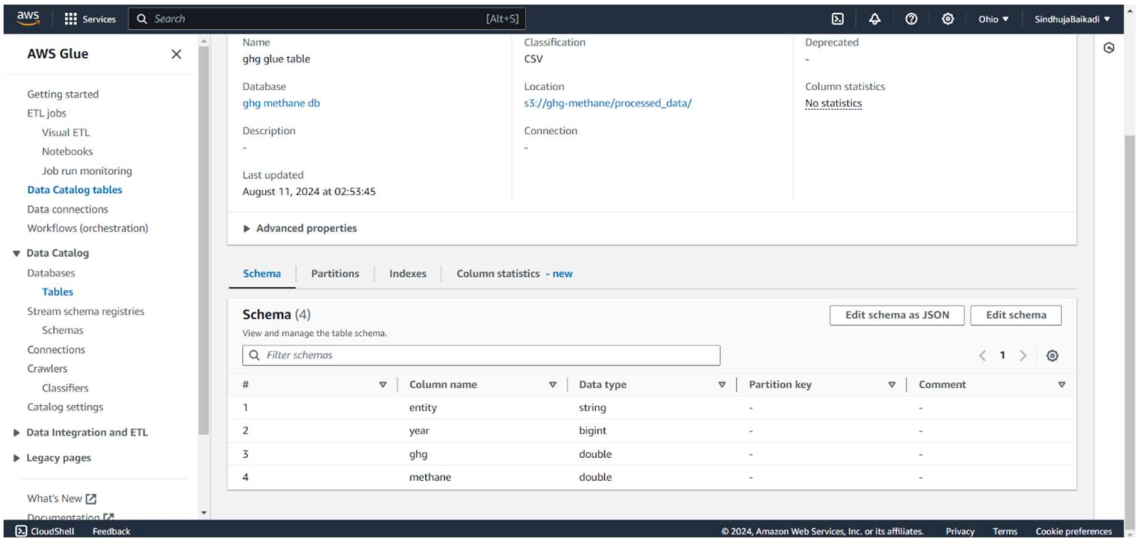
Output of ETL Jobs:

Jobs that ran successfully after ETL.



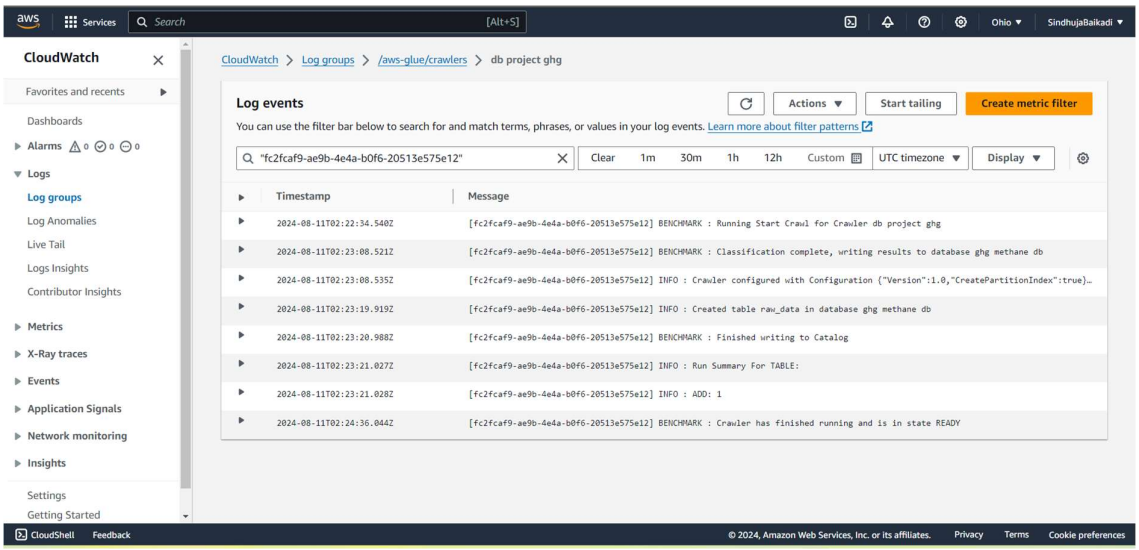
Processed data in data schema:

We can see that the code column is removed.



Step-4: Cloud Watch logs

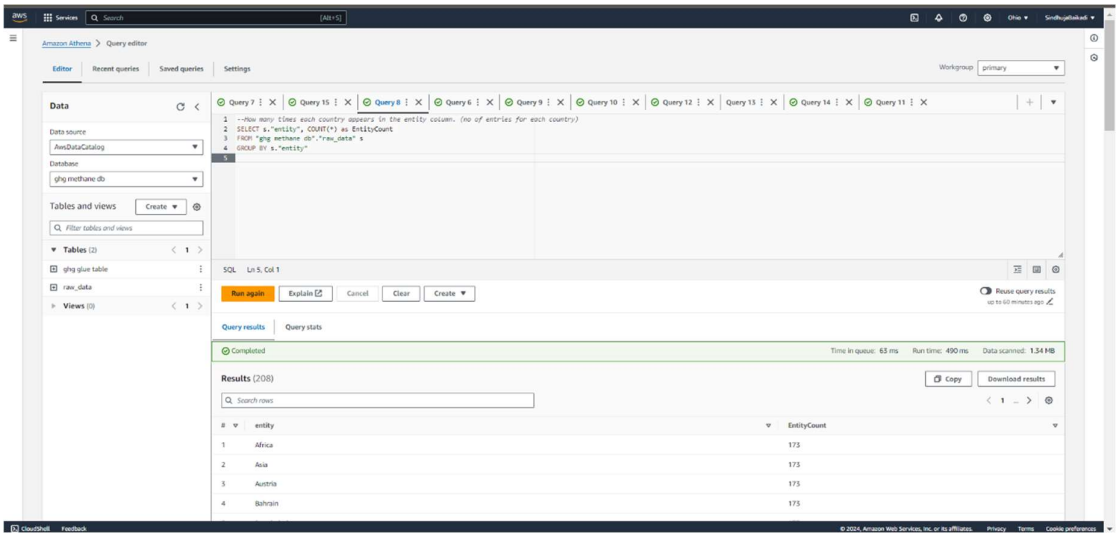
CloudWatch Logs is an AWS service that enables you to monitor, store, and access log files from your applications, systems, and AWS services. It allows you to collect, monitor, and analyze log data in real time, set alarms based on log metrics, and automatically trigger actions when specific conditions are met. CloudWatch Logs also provides long-term storage and retention of log data, making it easier to troubleshoot and audit activities in your AWS environment.



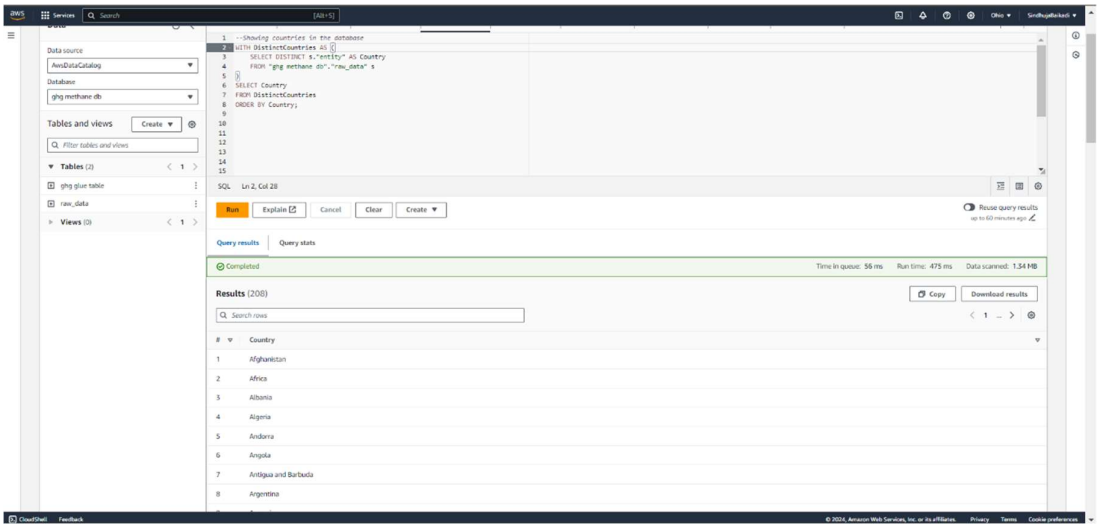
Step-5: Querying the Data Through Athena

Query the raw and processed data stored in the S3 bucket using Amazon Athena for ad-hoc analysis and reporting.

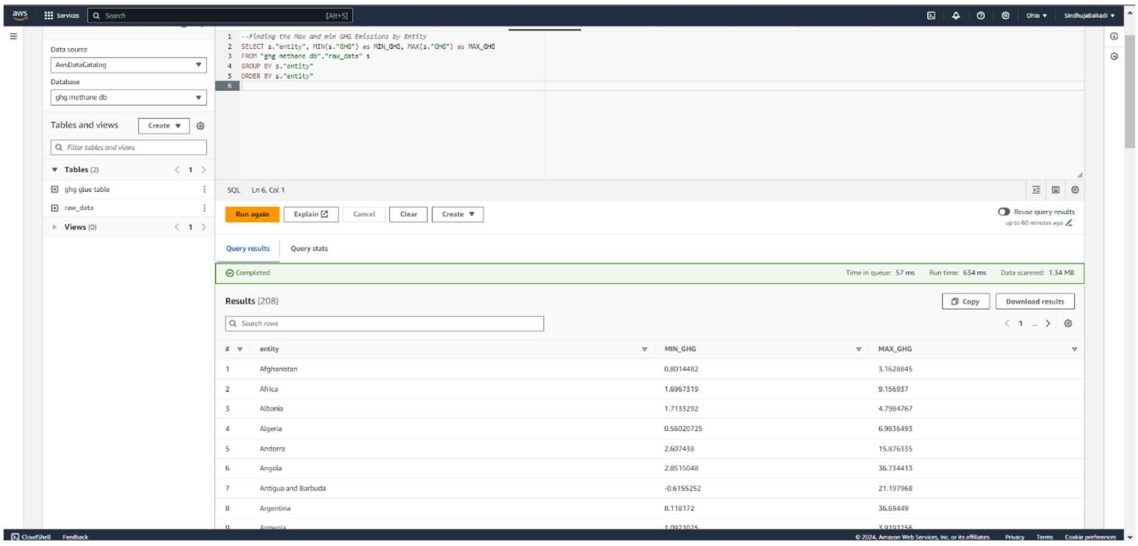
1. **How many times does each country appear in the entity column? (no of entries for each country)**



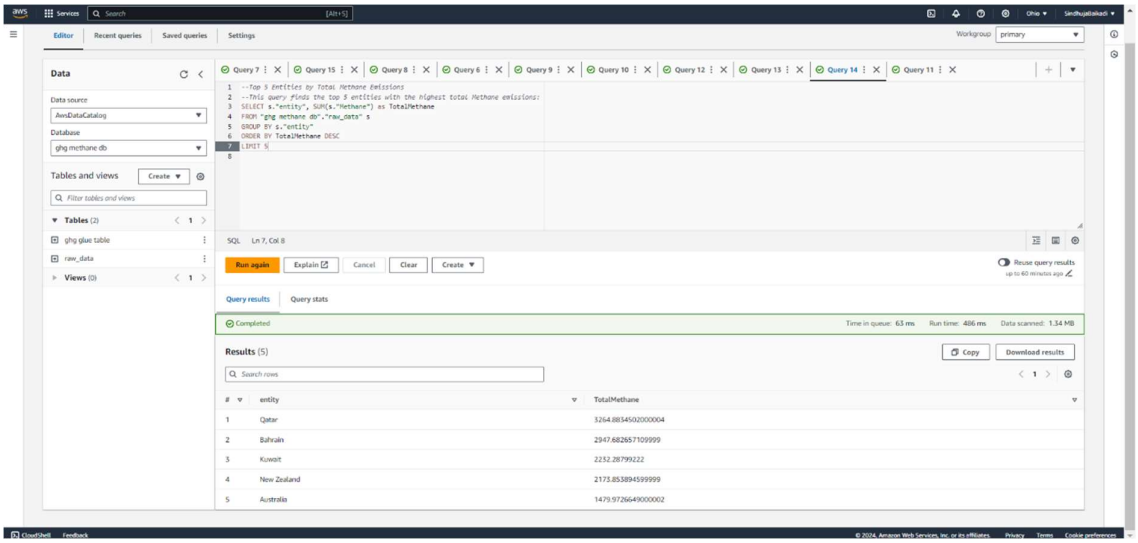
2. Showing countries in the database



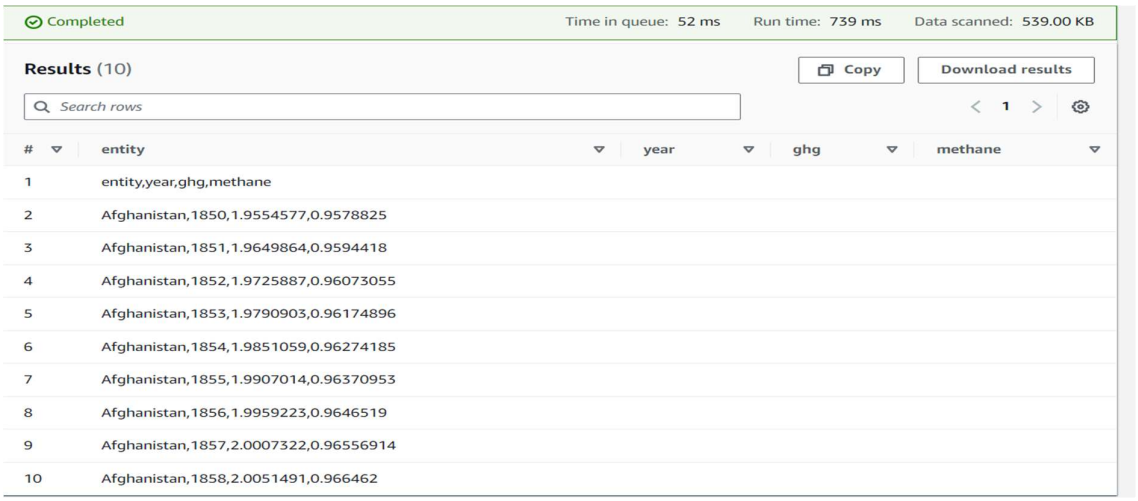
3. Finding the Max and minimum GHG Emissions by Entity.



4. **Top 5 Entities by Total Methane Emissions**



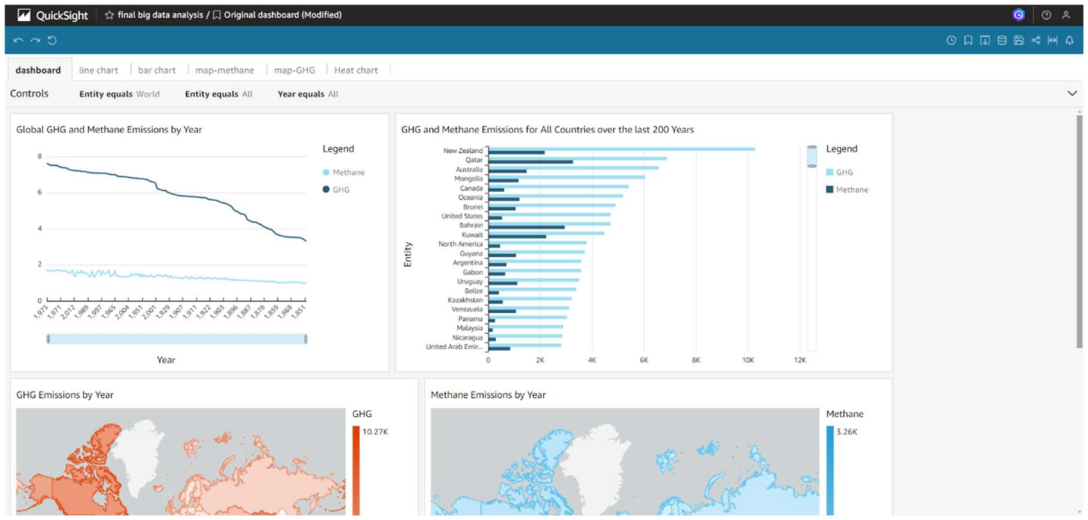
Results for the processed data:



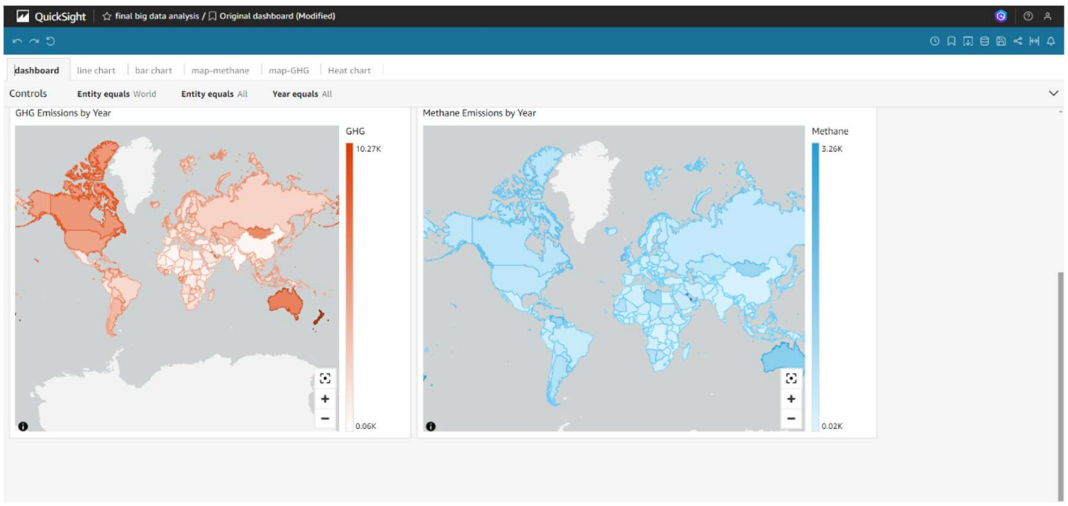
Step-6: AWS Quicksight Visualizations

- The dashboard provides a comprehensive view of both GHG and methane emissions globally, tracking their trends over time and across different countries.
- The comparative analysis highlights the countries that have historically been the largest contributors to these emissions. The heat maps offer a geographical perspective, making it easy to identify regions with the highest and lowest emissions.
- This dashboard can serve as a powerful tool for analyzing historical emission patterns and could be used to inform future environmental policies and strategies.
- **Global GHG and Methane Emissions by Year (Top Left)**
  Line Chart: This chart shows the trend of global GHG and methane emissions over time. It indicates how both types of emissions have evolved, with the GHG emissions showing a more pronounced decline compared to methane emissions. In this, we have incorporated the entity filter.

- **GHG and Methane Emissions for All Countries over the Last 200 Years (Top Right)**
  Bar Chart: This chart compares the cumulative GHG and methane emissions of various countries over the past 200 years. The countries are listed on the y-axis, and the x-axis represents the emission quantities. We have implemented year and entity filters in this.



- **GHG Emissions by Year (Bottom Left)**
  Heat Map: This map visually represents GHG emissions by country, with darker shades of red indicating higher levels of emissions. The color gradient on the right shows the range of emissions from 0.06K (lowest) to 10.27K (highest). The year filter is incorporated to change the year.
- **Methane Emissions by Year (Bottom Right)**
  Heat Map: This map is similar to the GHG heat map but focuses specifically on methane emissions. The color gradient here ranges from 0.02K to 3.26K. Again, darker shades of blue represent higher emissions. The year filter is incorporated to change the year.



Conclusion:

The "GHG and Methane Emissions Monitoring and Analysis Using AWS Glue and AWS Quicksight" project aims to provide actionable insights and real-time GHG and Methane emissions data for various applications. Leveraging AWS services, this project will enhance decision-making and safety in scenarios influenced by GHG and Methane levels.

References:

- https://ourworldindata.org/co2-and-greenhouse-gas-emissions
- [Towards Data Science: Free and Reliable Weather Data Sources](https://towardsdatascience.com/five-free-and-reliable-weather-data-sources-20b9ea6afac9)
- [OpenWeatherMap API](https://openweathermap.org/api)
- [NOAA Real-time Data](https://www.noaa.gov/education/resource-collections/data/real-time#:~:text=NOAA%20collects%20real%2Dtime%20data,%2C%20citizen%20scientists%2C%20and%20more)
- [Weather.gov API](https://www.weather.gov/documentation/services-web-api)
- [OpenWeatherMap API](https://openweathermap.org/api)