UNIVERSITI MALAYA

WQD7009

Big Data Applications and Analytics

Group 10

Group Assignment

Name	**Matric Number**
Low Boon Kiat	17138399
Wen Hui See	S2176564
Shi Sun Lai	S2175592
Jia Wei Eng	S2172813
Then Tsze Yen	S2194020

Contents

1. Introduction

1.1. Introduction

International trade plays an important role in the world's economy, such as driving growth; in addition, providing career opportunities and facilitating the exchange of goods and services across borders (Yang et al., 2023). However, taking the latest instance, whereby Covid-19 has severely impacted the trade flows; hence, reshaping the patterns of globalization (Mena et al., 2022) which could require data-driven approach to understand these complex dynamics.

In the context of rapidly evolving global markets and trade policies, there is a need to analyse and understand the patterns and implications of trade volume as it will provide insights into overall economic health of the countries, enables business to allocate appropriate resources for countries with higher demands. Still, with the complexity of trade data covering various metrics such as tariffs, trade values and comparative advantages, this would pose a challenge for stakeholders who require insights into trade dynamics to make informed decisions.

Hence, in order to get useful and reliable insights, we have to ensure data is stored, managed and analyzed efficiently. In today's globalized and digital world, cloud computing has become more affordable due to "pay-as-you-go" pricing and using basic hardware which enable us to process large amount of data quickly and effectively (Gharpure & Ghodke, 2021). Thus, this study will be taking a more structured approach by using Google Cloud Platform (GCP) to facilitate meaningful analysis.

1.2. Objectives

- To perform data lifecycle processes using different cloud-based tools throughout the data lifecycle framework.
- To perform trade volume, tariff line and rate analysis.
- To compare the query performance among the analysis performed in BigQuery in terms of latency, memory and number of records for read and written.

2. Dataset Description

2.1. Dataset Description

The dataset "World Export & Import Dataset (1989-2023)" was retrieved from Kaggle. This dataset covers information on international trade and trade policies with 33 features and 8096 rows. The table below displays the features we used in the analytic phase along with their descriptions.

Features	Description
Partner Name	Country involved in export or import.
Year	Year in which export or import occurred.
Export (US$ Thousand)	Total value of goods exported.
Import (US$ Thousand)	Total value of goods imported.
World Growth (%)	Percentage growth in world trade during a specific year.
AHS Simple Average (%)	Simple average of applied tariffs across Applied Harmonized System (AHS) tariff lines.
AHS Weighted Average (%)	Weighted average of applied tariffs considering the value of trade for each AHS tariff line.
AHS Total Tariff Lines	Total number of tariff lines in AHS.
AHS Dutiable Tariff Lines Share (%)	Percentage of tariff lines in AHS subject to duties.
AHS Duty Free Tariff Lines Share (%)	Percentage of tariff lines in AHS that are duty-free.
AHS Specific Tariff Lines Share (%)	Percentage of tariff lines in AHS subject to specific duties.
AHS AVE Tariff Lines Share (%)	Percentage of tariff lines in AHS subject to ad valorem equivalent (AVE) duties.
AHS SpecificDuty Imports (US$ Thousand)	Total value of imports subject to specific duties in AHS.
AHS Dutiable Imports (US$ Thousand)	Total value of imports subject to duties in AHS.
AHS Duty Free Imports (US$ Thousand)	Total value of imports that are duty-free in AHS.

2.2 Dataset Selection Meeting Minute

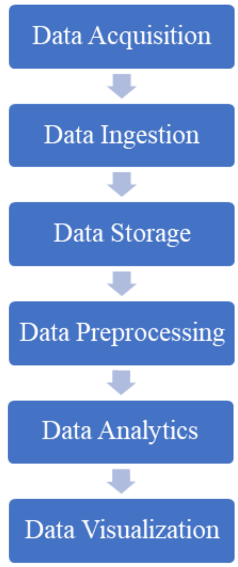
Date	6th Dec 2023 (Wed)
Time	9.00pm – 9.30pm
Venue	Microsoft Teams
Agenda	Review datasets and choose the best one which fulfils criteria.

Discussion:

The table below summarizes justification for accepting or rejecting each dataset. The chosen one is World Export & Import Dataset (1989 - 2023).

Dataset	Link	Justification
Crude Oil Price Dataset (from 1983 to present)	Link	• Rejected because it has very few attributes and observations (only 4 columns and 490 observations).
World Export & Import Dataset (1989 - 2023)	Link	• Accepted because it has sufficient attributes and observations (33 columns and 8097 observations). • Closely related to the theme of global economy.
Gold Price Prediction	Link	• Rejected because it has very few attributes (only 6 columns) despite having a significant number of observations.
Health & Income Outcomes of OECD & OPEC Countries	Link	• Contains sufficient attributes and observations (10 columns and 10546 observations). • Rejected because the data is not up to date (from 1960 until 2016), meanwhile two attributes i.e. infant mortality and GDP have many missing values.
CO_2 and Greenhouse Gas Emissions by Our World in Data	Link	• Contains 80 attributes and 50599 observations. • Rejected because most columns have a significant number of missing values, i.e. for those columns which have missing values, around 30-40% of the values are missing.

3. Data Lifecycle Framework



3.1. Data Acquisition

The chosen dataset for our analysis is the "World Export & Import Dataset (1989-2023)," obtained from Kaggle. This dataset offers a thorough compilation of information on international trade and trade policies, serving as a valuable resource for researchers, policymakers and analysts aiming to comprehend global trade dynamics. The rationale behind the selection of this dataset due to its inclusion of various trade aspects, such as trade values, tariff rates, and trade policy indicators, ensures a comprehensive view of international trade trends. In addition, it is notable that the dataset is originated from a variety of sources including official trade statistics, government reports and international organizations' databases which adds credibility to conduct analysis.

3.2. Data Ingestion

Data ingestion is the process of transporting data from one or more sources to a target site for further processing and analysis (Mohemmed, 2019). This data can originate from a range of sources, including data lakes, IoT devices or on-premises databases. In this project, the "World Export & Import Dataset (1989-2023)" was downloaded and saved from Kaggle to the local machine. Next, the data from local machine is uploaded to the cloud storage service like Google Cloud Storage, which is considered a form of batch data ingestion. This is because we are taking a dataset from our local environment and transferring it to the cloud storage in a discrete, typically one-time operation at a scheduled interval. We will be using Google Cloud Console to upload the dataset to Google Cloud Storage.

3.3. Data Storage

Next, data storage is tasked with ensuring the appropriate storage of entries and provide effective support for retrieving data efficiently (Mazumdar et al., 2019). In this project, we will be using BigQuery to store the data. BigQuery is able to organize table data in a columnar format, where each column is stored independently. This design is advantageous for analytic workloads that will be performed in the following phases of this framework that involve aggregating data across a large number of entries (Mucchetti, 2020).

3.4. Data Processing

The process of converting gathered raw data into information that is useful is known as data processing (Peng et al., 2022). Processing is essential to preparing data for analysis and interpretation after it has been gathered. It involves several processes, including organising, sanitising, and transforming unprocessed data (Wang et al., 2020). Data preprocessing is a specific step in data processing. It involves 5 tasks such as data cleaning, data optimization, data transformation, data integration and data conversion (Joshi & Patel, 2021). These various data preprocessing tasks are important for reducing noisy data and providing quality data for efficient data analysis results.

3.5. Data Analytics

Data analytics represents a crucial stage that involves extracting valuable insights and trends from the stored and processed data to aid in informed decision-making. This stage involves exploratory data analysis which aims to understand data distributions, patterns and correlations, as well as to identify potential areas of interest for further investigation (Sivarajah et al., 2017). In this project, we will be using BigQuery for exploratory data analysis. BigQuery allows us to run SQL queries on large datasets stored in Google Cloud Platform for fast and scalable data exploration. We can perform data sampling, descriptive statistics, aggregation, grouping, and basic data profiling using SQL queries in BigQuery (Lakshmanan, 2022).

3.6. Data Visualisation

Data is driven by the vast data generated from the diverse source like computers, social media and mobile devices (Mustafa et al., 2020). Data visualisation involves the systemic representation of data, incorporating various attributes and variables of information units (Khan & Sarwar, 2011). Furthermore, these methods can incorporate advanced analytics to develop interactive graphics, which are accessible on various devices including desktops, laptops, and mobile devices (Sucharitha et al., 2014). First, Looker is used for the data visualisation in Google Cloud Platform. The export and import of the countries are focused who is the highest and lowest performances by introducing map visualisation Looker throughout the years. Moreover, a multiple line graph visualisation on countries' growth is done with highest balance of trade, which balance of trade.

4. Justification of on-premises or cloud-based tools selection

4.1. Data Ingestion – Google Cloud Storage

Google Cloud Storage is selected for data ingestion due to its versatility to support various data formats. Given that the dataset we used in this project is structured in nature, Cloud Storage provides a flexible platform for uploading and managing data. In addition, Cloud Storage seamlessly integrates with various data sources and tools, making it easier to ingest data from different locations (Yu et al., 2021; Amiri-Zarandi et al., 2022). This is important when dealing with a historical dataset like ours spanning from 1989 to 2023.

4.2. Data Storage – BigQuery

BigQuery is selected for data storage purposes due to its serverless data warehouse designed for high-performance analysis of large datasets. Given the significant volume of our dataset with 33 attributes and 8096 entries, BigQuery's analytical capabilities are suitable for complex queries and data exploration. This is because it stores data in a columnar format which will enhance query performance with less execution time and reduce costs (L'Esteve, 2023). Moreover, it seamlessly integrates with various data analysis tools and visualization platforms available on Google Cloud. This will facilitate a smooth transition from data storage to analysis (Ali et al., 2021).

Possible alternative technology – Hadoop Distributed File System (HDFS):

HDFS can scale horizontally, making it well-suited for handling large volume of historical data from 1989 to 2023. This scalability enables the addition of machines to the cluster as the dataset grows, ensuring for unlimited storage capacity. However, selecting Google Cloud Storage for ingestion and BigQuery for storage over HDFS offers more advantages (Google Cloud, 2018):

- Cost saving: From past research, switching to Google Cloud Storage from continuous HDFS usage can reduce total ownership costs by up to 57%, this is because we only pay for the storage used and as well avoiding initial hardware expenses.
- Enhanced performance: Google Cloud Storage is designed to be compatible with HDFS, offering similar or even better performance level. In addition, it also supports traditional Hadoop or Spark jobs.

4.3. Data Processing – Vertex AI Colab Enterprise

Vertex AI's Colab Enterprise in Google Cloud is an ideal tool for preprocessing data. Colab Enterprise is derived from cloud based Jupyter notebook of Google which is named as Colab. Since the dataset used in this project contains a total of 11,147 null values, Colab Enterprise provides access to powerful libraries such as Pandas and NumPy, which offer efficient tools for handling missing data, including methods for imputation, dropping null values, and filling missing values based on various strategies. Its seamless integration with BigQuery allows for direct access to the dataset, enabling efficient querying and subsequent storage of the preprocessed data

(Pamma, 2023). Furthermore, its collaborative environment supports group project dynamics (Ghoshal, 2023), while its cloud-based infrastructure ensures the efficient execution of preprocessing tasks, providing a comprehensive solution for the data processing needs.

Apache Spark can be served as the alternative technology for data processing. It is a robust open-source distributed computing system that offers in-memory processing, scalability, and rich APIs in Java, Scala, Python, and R, making it an excellent alternative technology. Its in-memory processing capabilities can significantly accelerate data processing, while its scalability allows it to handle large datasets with ease. It also offers robust functionality for cleaning and preprocessing data, including dealing with missing or null values.

4.4. Data Analytics – BigQuery

BigQuery is selected for exploratory data analysis. This is because they offer several advantages in terms of integration, scalability and ease of use (Riahi et al., 2018). First, BigQuery seamlessly integrates with other Google Cloud Platform services for data ingestion, storage and processing, allowing smooth workflows within the GCP ecosystem. Second, BigQuery offers high-speed querying, enabling quick exploration and analysis on massive volume of data. Its serverless nature and scalability make it suitable for handling large datasets without requiring provisioning the infrastructure.

The alternative technology for data analysis is Apache Hive. Hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale, making it suitable for data analysis. However, Hive has some limitations, such as not supporting OLTP, subqueries, and having high latency (GeeksforGeeks, 2022). These limitation cause BigQuery is selected instead of Apache Hive.

4.5. Data Visualization - Looker Studio

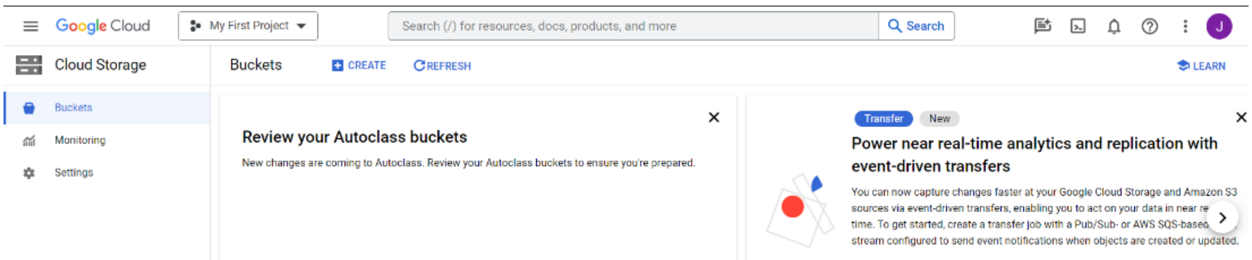
Looker Studio facilitates data visualization by leveraging the powerful data analysis and modeling capabilities of BigQuery and BigQueryML. This robust tool enables the crafting of interactive and engaging visualizations like chart and graphs, enhancing the analysis and presentation of data from the web and social media (Santana M. K. T. & Padmamma, 2023). With the integration between Looker Studio and Google Cloud Platform, users can easily ingest, store and process data, creating a smooth workflow. The high-speed querying features of BigQuery allow for swift exploratory data analysis of large datasets without the need for infrastructure management. This integration means the data can be visualized in Looker Studio in real-time, using SQL commands to reflect the insights derived from the machine learning models giving a comprehensive and interactive experience.

The alternative technology for data visualization is PowerBI. PowerBI is a popular tool for interactive data visualization developed by Microsoft. As this project is using the Google Cloud Platform, using its own visualization tools which is Looker Studio is way more compatible with. Overall, Looker Studio is a good solution for those who already used Google Cloud Platform whereas Microsoft Power BI will likely be a better fit for those Microsoft users.

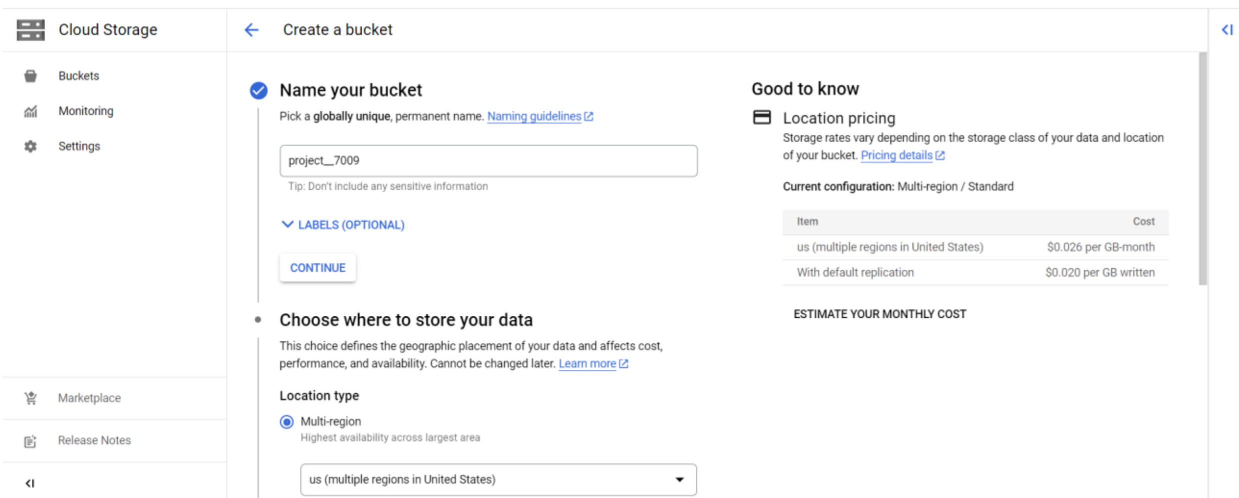
5. Proposed Framework Implementation

5.1. Data Ingestion – Google Cloud Storage

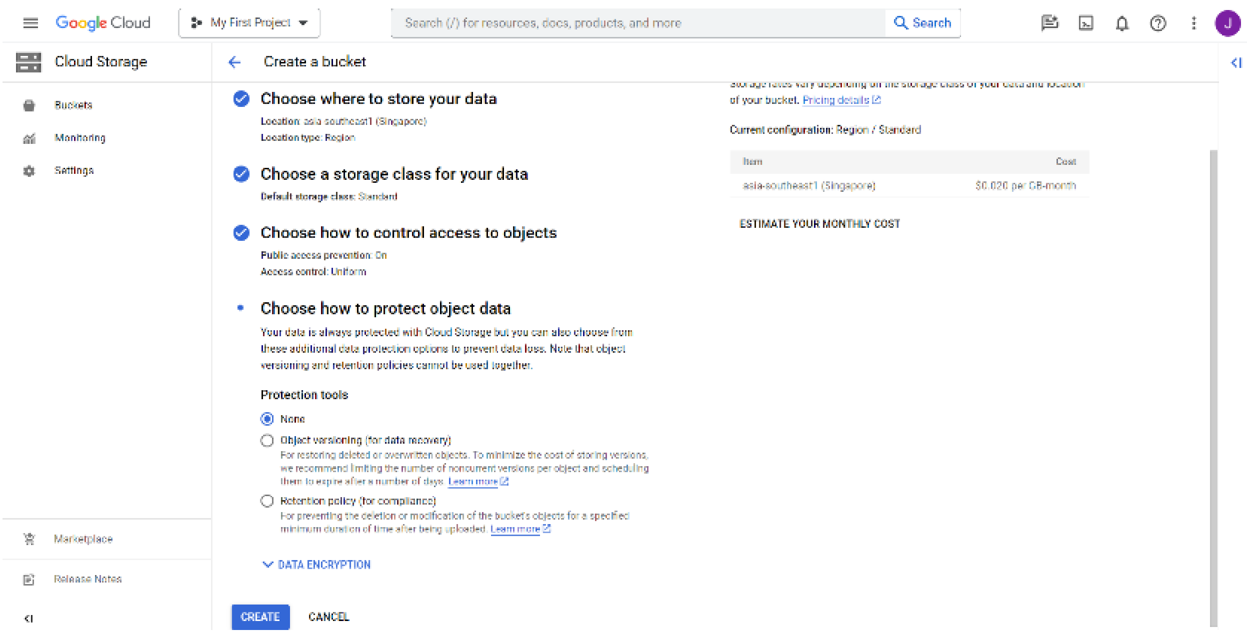
Step 1: Create a bucket by clicking "Create" button.



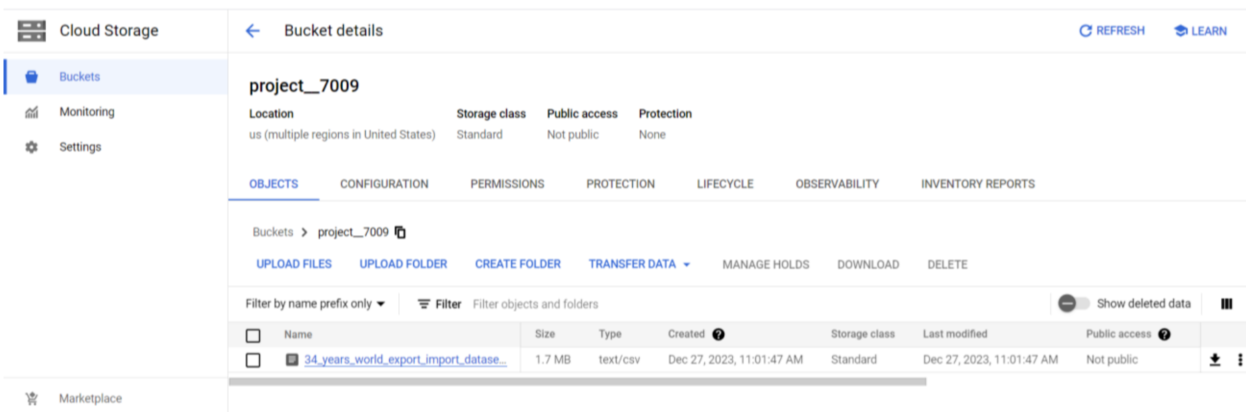
Step 2: Enter the name of the bucket – project__7009. Here, we need to ensure the name is globally unique.



Step 3: Once we finished setting up all the storage locations, storage class, control access and data protection, click "Create" button.

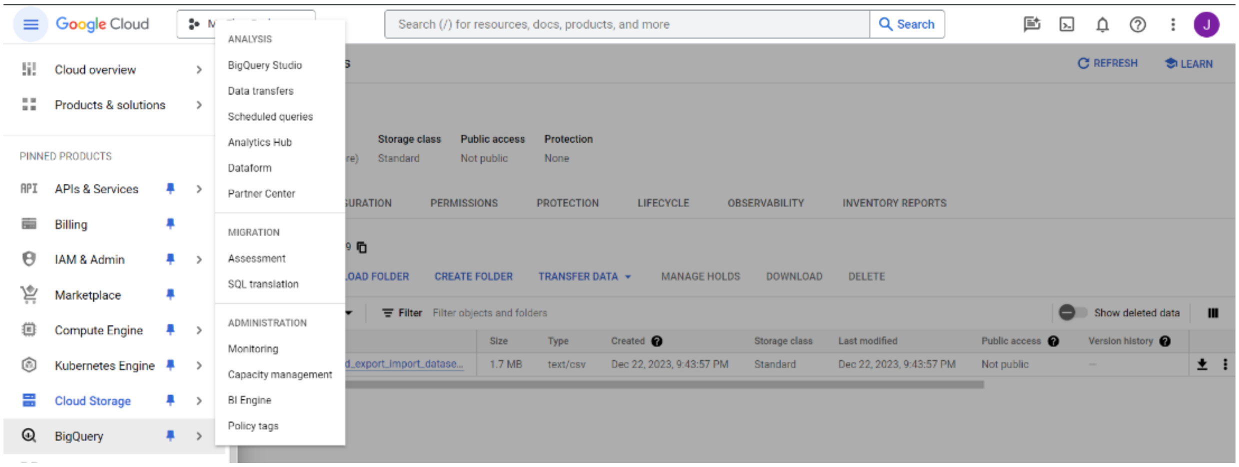


Step 4: After clicking the “Create” button, we will be taken to the bucket panel. Here, we will upload our dataset which is a csv file named “34_years_world_export_import_dataset”.



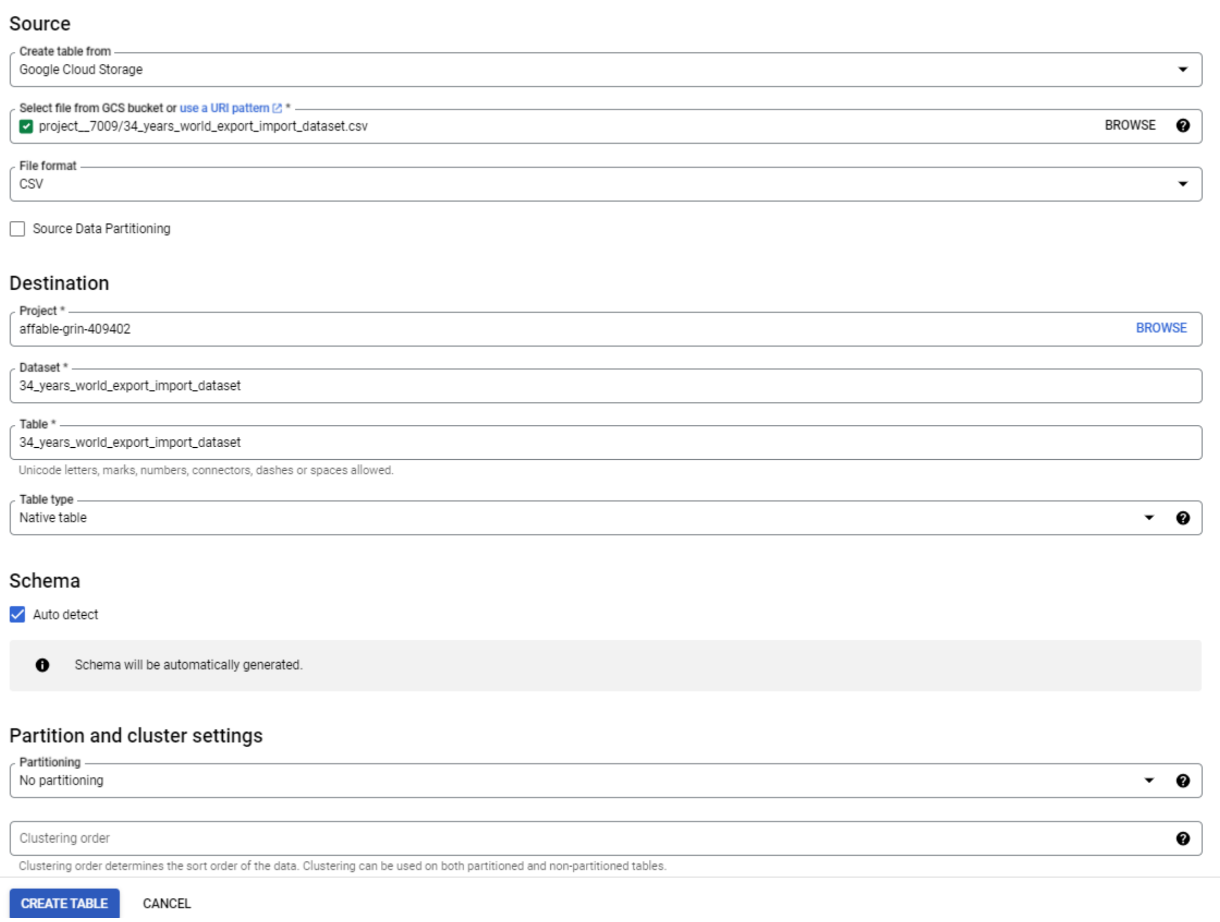
5.2. Data Storage – BigQuery

Step 1: We click on the “BigQuery” button.

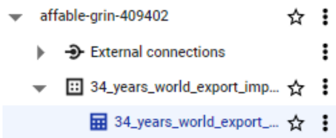


Step 2: Click on the "Create Table" button.

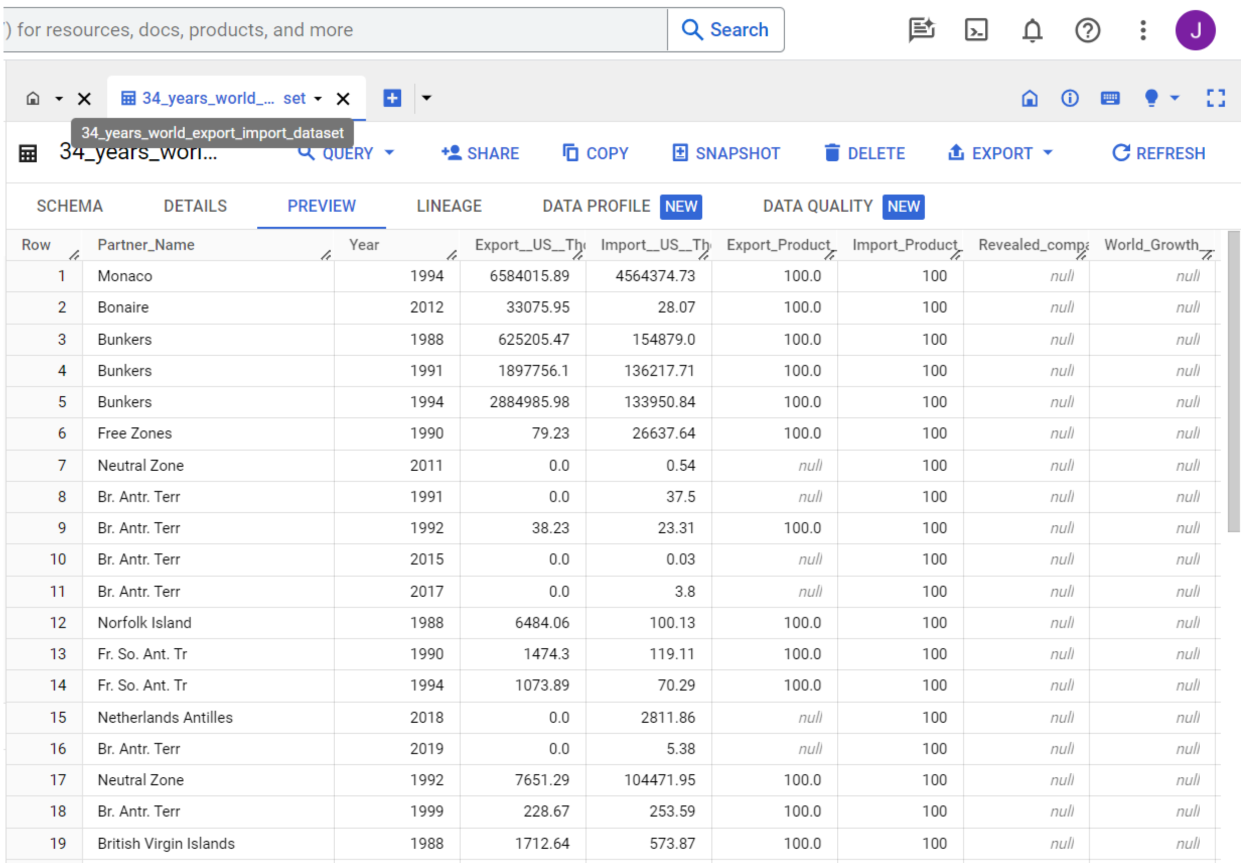
Step 3: We can observe the pop out window, and we will go for the "Google Cloud Storage" selection. Later, we click on the browse and select our "34_years_world_export_import_dataset" from GCS bucket and fill up the table name with "34_years_world_export_import_dataset" as well. Click "Create Table".



Step 4: When the table is successfully created, we can see it over here.



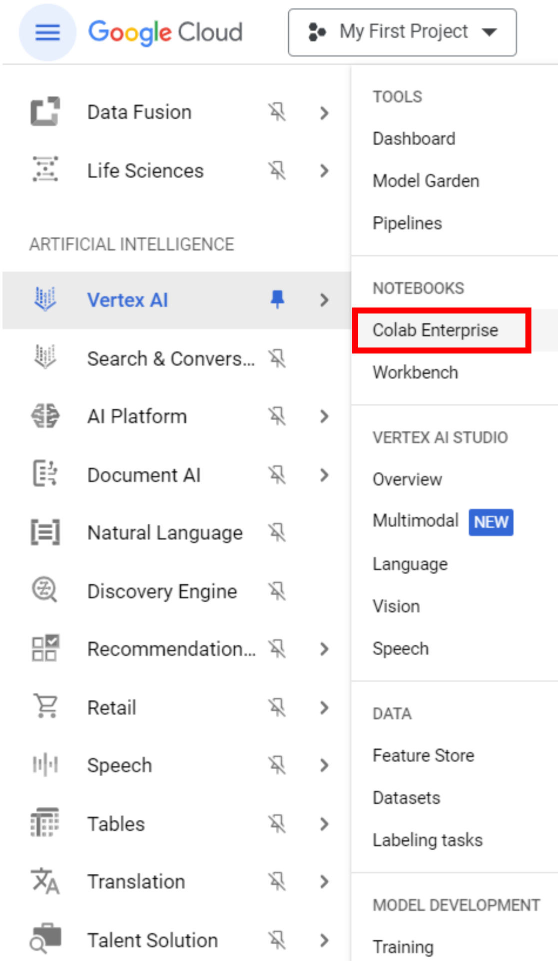
Step 5: A preview is shown below to claim that we have stored the data into BigQuery for further preprocessing and analysis.



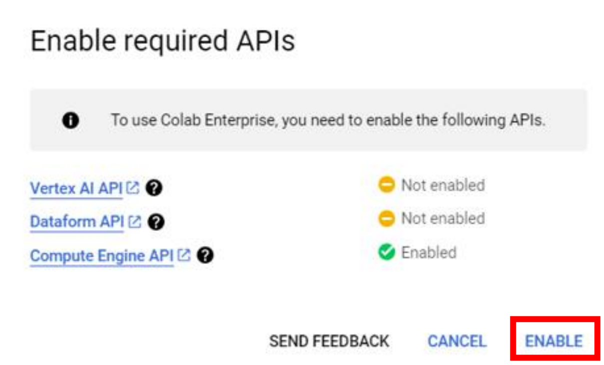
Row	Partner_Name	Year	Export_US_Th	Import_US_Th	Export_Product_	Import_Product_	Revealed_comp	World_Growth_
1	Monaco	1994	6584015.89	4564374.73	100.0	100	null	null
2	Bonaire	2012	33075.95	28.07	100.0	100	null	null
3	Bunkers	1988	625205.47	154879.0	100.0	100	null	null
4	Bunkers	1991	1897756.1	136217.71	100.0	100	null	null
5	Bunkers	1994	2884985.98	133950.84	100.0	100	null	null
6	Free Zones	1990	79.23	26637.64	100.0	100	null	null
7	Neutral Zone	2011	0.0	0.54	null	100	null	null
8	Br. Antr. Terr	1991	0.0	37.5	null	100	null	null
9	Br. Antr. Terr	1992	38.23	23.31	100.0	100	null	null
10	Br. Antr. Terr	2015	0.0	0.03	null	100	null	null
11	Br. Antr. Terr	2017	0.0	3.8	null	100	null	null
12	Norfolk Island	1988	6484.06	100.13	100.0	100	null	null
13	Fr. So. Ant. Tr	1990	1474.3	119.11	100.0	100	null	null
14	Fr. So. Ant. Tr	1994	1073.89	70.29	100.0	100	null	null
15	Netherlands Antilles	2018	0.0	2811.86	null	100	null	null
16	Br. Antr. Terr	2019	0.0	5.38	null	100	null	null
17	Neutral Zone	1992	7651.29	104471.95	100.0	100	null	null
18	Br. Antr. Terr	1999	228.67	253.59	100.0	100	null	null
19	British Virgin Islands	1988	1712.64	573.87	100.0	100	null	null

5.3. Data Processing – Vertex AI Colab Enterprise

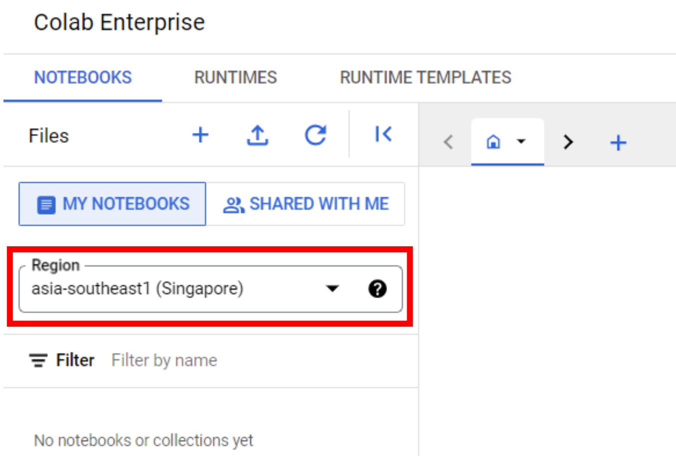
Step 1: Choose Colab Enterprise tool from Vertex AI under Artificial Intelligence Section from the Navigator.



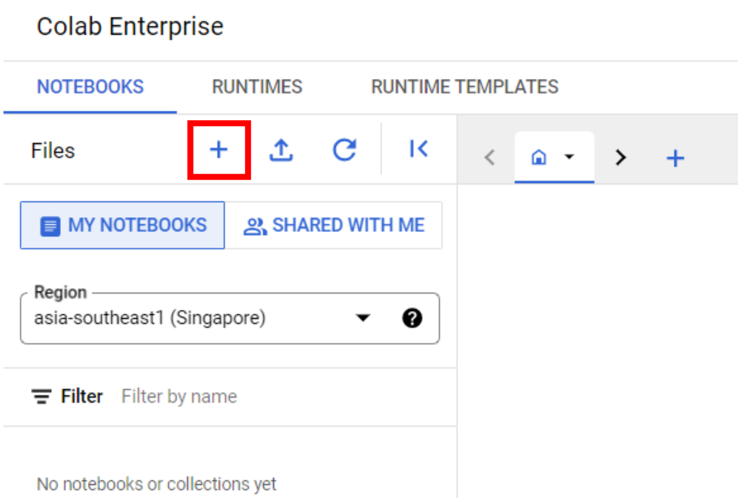
Step 2: Click “Enable” to enable the required APIs.



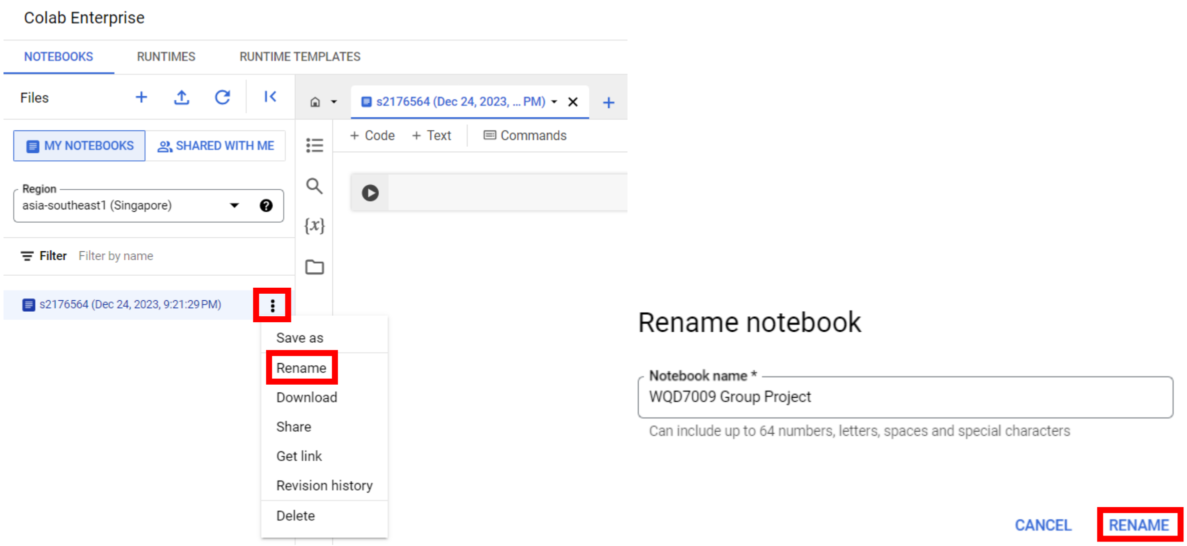
Step 3: Choose the preferred Region, we select asia-southeast1 (Singapore) as preferred region.



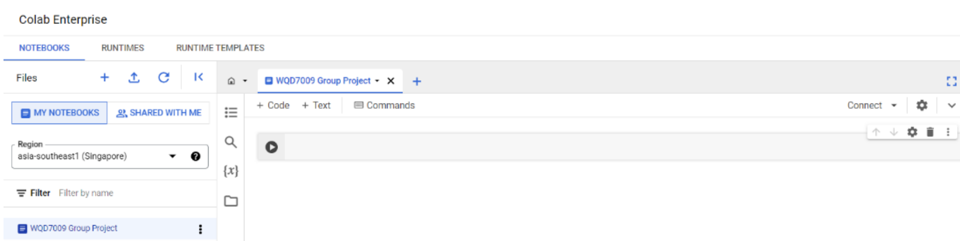
Step 4: Click "+" button to create a new notebook.



Step 5: Click the 3 dots and select "Rename" to rename the notebook, here we rename it as "WQD7009 Group Project". Once done inserting the notebook name, click "Rename".



Step 6: Now, we can start coding in the notebook.



Step 7: Load the data from the BigQuery table into a dataframe.

```
# Query the table in BigQuery and extract to a Pandas DataFrame named 'df'

%%bigquery df --project affable-grin-409402
SELECT * FROM `affable-grin-409402.34_years_world_export_import_dataset.34_years_world_export_import_dataset`
```

Job ID dd2ecf82-2bf8-4dca-90be-fff181ec2ddd successfully executed: 100%

Downloading: 100%

Step 8: Next, we view the first 5 rows of the dataframe. The dataframe contains 33 columns.

```
[2] # View the first 5 rows of the DataFrame
    df.head()
```

	Partner_Name	Year	Export__US__Thousand_	Import__US__Thousand_	Export_Product_Share___	Import_Product_Share___	Revealed_c
0	Monaco	1994	6584015.89	4564374.73	100.0	100	
1	Bonaire	2012	33075.95	28.07	100.0	100	
2	Bunkers	1988	625205.47	154879.00	100.0	100	
3	Bunkers	1991	1897756.10	136217.71	100.0	100	
4	Bunkers	1994	2884985.98	133950.84	100.0	100	

5 rows × 33 columns

Step 9: Get some basic info and statistics about the dataframe.

```python
# Get the basic info of the DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8096 entries, 0 to 8095
Data columns (total 33 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Partner_Name                          8096 non-null   object 
 1   Year                                  8096 non-null   Int64  
 2   Export__US__Thousand_                 8096 non-null   float64
 3   Import__US__Thousand_                 8096 non-null   float64
 4   Export_Product_Share____              8076 non-null   float64
 5   Import_Product_Share____              8096 non-null   Int64  
 6   Revealed_comparative_advantage        4712 non-null   float64
 7   World_Growth____                      4410 non-null   float64
 8   Country_Growth____                    4410 non-null   float64
 9   AHS_Simple_Average____                8080 non-null   float64
 10  AHS_Weighted_Average____              8080 non-null   float64
 11  AHS_Total_Tariff_Lines                8080 non-null   float64
 12  AHS_Dutiable_Tariff_Lines_Share____   8080 non-null   float64
 13  AHS_Duty_Free_Tariff_Lines_Share____  8080 non-null   float64
 14  AHS_Specific_Tariff_Lines_Share____   8080 non-null   float64
 15  AHS_AVE_Tariff_Lines_Share____        8080 non-null   float64
 16  AHS_MaxRate____                       8080 non-null   float64
 17  AHS_MinRate____                       8080 non-null   float64
 18  AHS_SpecificDuty_Imports__US__Thousand_  8081 non-null   float64
 19  AHS_Dutiable_Imports__US__Thousand_   8081 non-null   float64
 20  AHS_Duty_Free_Imports__US__Thousand_  8081 non-null   float64
 21  MFN_Simple_Average____                8081 non-null   float64
 22  MFN_Weighted_Average____              8081 non-null   float64
 23  MFN_Total_Tariff_Lines                8081 non-null   float64
 24  MFN_Dutiable_Tariff_Lines_Share____   8081 non-null   float64
 25  MFN_Duty_Free_Tariff_Lines_Share____  8081 non-null   float64
 26  MFN_Specific_Tariff_Lines_Share____   8080 non-null   float64
 27  MFN_AVE_Tariff_Lines_Share____        8080 non-null   float64
 28  MFN_MaxRate____                       8081 non-null   float64
 29  MFN_MinRate____                       8081 non-null   float64
 30  MFN_SpecificDuty_Imports__US__Thousand_  8081 non-null   float64
 31  MFN_Dutiable_Imports__US__Thousand_   8081 non-null   float64
 32  MFN_Duty_Free_Imports__US__Thousand_  8081 non-null   float64
dtypes: Int64(2), float64(30), object(1)
memory usage: 2.1+ MB
```

```python
# Get the basic statistics about the DataFrame
df.describe()
```

	Year	Export__US__Thousand_	Import__US__Thousand_	Export_Product_Share____	Import_Product_Share____	Revealed_comparative_advantage
count	8096.0	8.096000e+03	8.096000e+03	8076.0	8096.0	4712.0
mean	2004.908226	1.421192e+08	1.305216e+08	100.0	100.0	1.0
std	9.707831	9.928417e+08	9.073802e+08	0.0	0.0	0.0
min	1988.0	0.000000e+00	3.000000e-02	100.0	100.0	1.0
25%	1997.0	4.274264e+05	1.601335e+05	100.0	100.0	1.0
50%	2005.0	3.719683e+06	2.053967e+06	100.0	100.0	1.0
75%	2013.0	2.585514e+07	2.102937e+07	100.0	100.0	1.0
max	2021.0	2.422743e+10	2.193121e+10	100.0	100.0	1.0

8 rows × 32 columns

Step 10: Since the World Growth & Country Growth are similar, it may introduce redundancy and potentially lead to confusion or errors in data analysis. Therefore, we decide to drop the Country Growth. The total columns become 32.

```python
# Drop the Country_Growth____ column
df = df.drop('Country_Growth____', axis=1)
df.head()
```

	Partner_Name	Year	Export__US__Thousand_	Import__US__Thousand_	Export_Product_Share____	Import_Product_Share____	Revealed_comparative_advantage
0	Monaco	1994	6584015.89	4564374.73	100.0	100	NaN
1	Bonaire	2012	33075.95	28.07	100.0	100	NaN
2	Bunkers	1988	625205.47	154879.00	100.0	100	NaN
3	Bunkers	1991	1897756.10	136217.71	100.0	100	NaN
4	Bunkers	1994	2884985.98	133950.84	100.0	100	NaN

5 rows × 32 columns

Step 11: Check for duplication and there is no duplication in the dataframe.

```
# Check if there is duplication in the DataFrame
df.duplicated().sum()
```

```
0
```

Step 12: Then, check for null values. The results show that there are null values in some attributes.

```
# Check if there is null value in the DataFrame
df.isnull().sum()
```

```
Partner_Name                                   0
Year                                           0
Export__US__Thousand_                          0
Import__US__Thousand_                          0
Export_Product_Share____                      20
Import_Product_Share____                       0
Revealed_comparative_advantage              3384
World_Growth____                            3686
AHS_Simple_Average____                        16
AHS_Weighted_Average____                      16
AHS_Total_Tariff_Lines                        16
AHS_Dutiable_Tariff_Lines_Share____           16
AHS_Duty_Free_Tariff_Lines_Share____          16
AHS_Specific_Tariff_Lines_Share____           16
AHS_AVE_Tariff_Lines_Share____                16
AHS_MaxRate____                               16
AHS_MinRate____                               16
AHS_SpecificDuty_Imports__US__Thousand_       15
AHS_Dutiable_Imports__US__Thousand_           15
AHS_Duty_Free_Imports__US__Thousand_          15
MFN_Simple_Average____                        15
MFN_Weighted_Average____                      15
MFN_Total_Tariff_Lines                        15
MFN_Dutiable_Tariff_Lines_Share____           15
MFN_Duty_Free_Tariff_Lines_Share____          15
MFN_Specific_Tariff_Lines_Share____           16
MFN_AVE_Tariff_Lines_Share____                16
MFN_MaxRate____                               15
MFN_MinRate____                               15
MFN_SpecificDuty_Imports__US__Thousand_       15
MFN_Dutiable_Imports__US__Thousand_           15
MFN_Duty_Free_Imports__US__Thousand_          15
dtype: int64
```

Step 13: As noticed from the data description earlier, the mean, mode, median of Export Product Share are 100 and that of Revealed comparative advantage are 1. Therefore, we decide to perform mean imputation for the null values in both attributes.

```
# Impute mean for null values in Export_Product_Share
df['Export_Product_Share____'].fillna(value=df['Export_Product_Share____'].mean(), inplace=True)
df['Revealed_comparative_advantage'].fillna(value=df['Revealed_comparative_advantage'].mean(), inplace=True)

# Check if the null values in Export_Product_Share and Revealed_comparative_advantage have been imputed successfully
df.isnull().sum()
```

```
Partner_Name                                 0
Year                                         0
Export__US__Thousand_                        0
Import__US__Thousand_                        0
Export_Product_Share____                     0
Import_Product_Share____                     0
Revealed_comparative_advantage               0
World_Growth____                          3686
AHS_Simple_Average____                      16
AHS_Weighted_Average____                    16
AHS_Total_Tariff_Lines                      16
AHS_Dutiable_Tariff_Lines_Share____         16
AHS_Duty_Free_Tariff_Lines_Share____        16
AHS_Specific_Tariff_Lines_Share____         16
AHS_AVE_Tariff_Lines_Share____              16
AHS_MaxRate____                             16
AHS_MinRate____                             16
AHS_SpecificDuty_Imports__US__Thousand_     15
AHS_Dutiable_Imports__US__Thousand_         15
AHS_Duty_Free_Imports__US__Thousand_        15
MFN_Simple_Average____                      15
MFN_Weighted_Average____                    15
MFN_Total_Tariff_Lines                      15
MFN_Dutiable_Tariff_Lines_Share____         15
MFN_Duty_Free_Tariff_Lines_Share____        15
MFN_Specific_Tariff_Lines_Share____         16
MFN_AVE_Tariff_Lines_Share____              16
MFN_MaxRate____                             15
MFN_MinRate____                             15
MFN_SpecificDuty_Imports__US__Thousand_     15
MFN_Dutiable_Imports__US__Thousand_         15
MFN_Duty_Free_Imports__US__Thousand_        15
dtype: int64
```

Step 14: For all the AHS & MFN related attributes, we decide to perform mean imputation as well for the null values. The null values will be imputed based on the mean value of attribute of each country.

```
# Specify the AHS and MFN related columns with missing values
AHS_MFN_columns_with_missing_values = [
    'AHS_Simple_Average____',
    'AHS_Weighted_Average____',
    'AHS_Total_Tariff_Lines',
    'AHS_Dutiable_Tariff_Lines_Share____',
    'AHS_Duty_Free_Tariff_Lines_Share____',
    'AHS_Specific_Tariff_Lines_Share____',
    'AHS_AVE_Tariff_Lines_Share____',
    'AHS_MaxRate____',
    'AHS_MinRate____',
    'AHS_SpecificDuty_Imports__US__Thousand_',
    'AHS_Dutiable_Imports__US__Thousand_',
    'AHS_Duty_Free_Imports__US__Thousand_',
    'MFN_Simple_Average____',
    'MFN_Weighted_Average____',
    'MFN_Total_Tariff_Lines',
    'MFN_Dutiable_Tariff_Lines_Share____',
    'MFN_Duty_Free_Tariff_Lines_Share____',
    'MFN_Specific_Tariff_Lines_Share____',
    'MFN_AVE_Tariff_Lines_Share____',
    'MFN_MaxRate____',
    'MFN_MinRate____',
    'MFN_SpecificDuty_Imports__US__Thousand_',
    'MFN_Dutiable_Imports__US__Thousand_',
    'MFN_Duty_Free_Imports__US__Thousand_',
]

# Specify the column containing country names
country_column = 'Partner_Name'

# Impute missing values based on the mean of each country for all specified columns
for column in AHS_MFN_columns_with_missing_values:
    df[column] = df.groupby(country_column)[column].transform(lambda x: x.fillna(x.mean()))

# Check if the null values have been imputed successfully
df.isnull().sum()
```

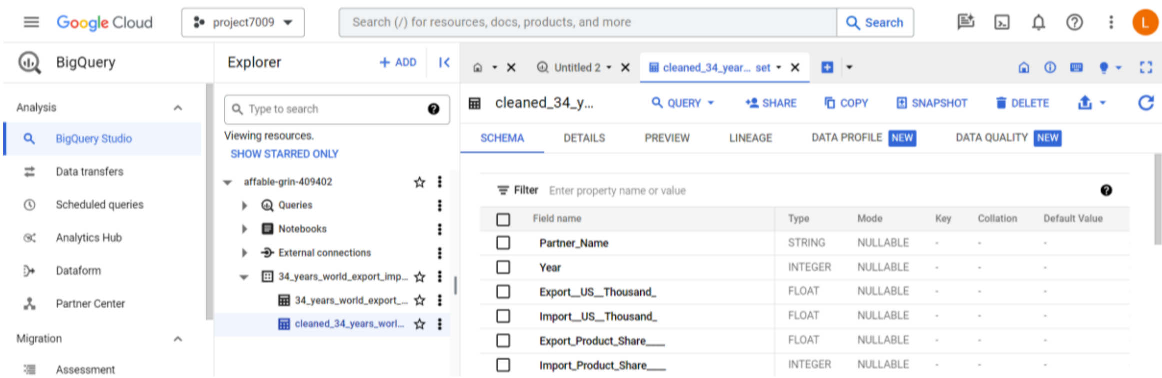
```
Partner_Name                                0
Year                                        0
Export__US__Thousand_                       0
Import__US__Thousand_                       0
Export_Product_Share____                    0
Import_Product_Share____                    0
Revealed_comparative_advantage              0
World_Growth____                         3686
AHS_Simple_Average____                      0
AHS_Weighted_Average____                    0
AHS_Total_Tariff_Lines                      0
AHS_Dutiable_Tariff_Lines_Share____         0
AHS_Duty_Free_Tariff_Lines_Share____        0
AHS_Specific_Tariff_Lines_Share____         0
AHS_AVE_Tariff_Lines_Share____              0
AHS_MaxRate____                             0
AHS_MinRate____                             0
AHS_SpecificDuty_Imports__US__Thousand_     0
AHS_Dutiable_Imports__US__Thousand_         0
AHS_Duty_Free_Imports__US__Thousand_        0
MFN_Simple_Average____                      0
MFN_Weighted_Average____                    0
MFN_Total_Tariff_Lines                      0
MFN_Dutiable_Tariff_Lines_Share____         0
MFN_Duty_Free_Tariff_Lines_Share____        0
MFN_Specific_Tariff_Lines_Share____         0
MFN_AVE_Tariff_Lines_Share____              0
MFN_MaxRate____                             0
MFN_MinRate____                             0
MFN_SpecificDuty_Imports__US__Thousand_     0
MFN_Dutiable_Imports__US__Thousand_         0
MFN_Duty_Free_Imports__US__Thousand_        0
dtype: int64
```

However, there is still 1 column, World Growth left with 3686 nulls values. Some countries like Yemen Democratic only contain about 3 years of information, and we realised that all the 3 rows remain blank for the World Growth attribute. Other than the World Growth attribute, all information of Yemen Democratic are completed. Therefore, we decide to keep the null values in the column.

Step 15: Once completed processing the data, save the dataframe into BigQuery table named “cleaned_34_years_world_export_import_dataset”.

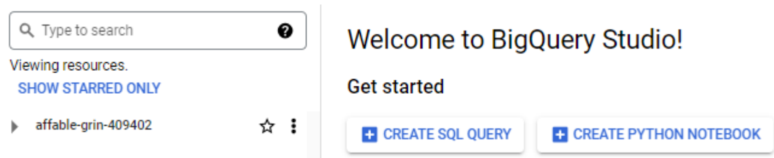
```
# Indicate the Google Cloud project ID, BigQuery Dataset ID and Bigquery Table ID
project_id = 'affable-grin-409402'
dataset_id = '34_years_world_export_import_dataset'
table_id = 'cleaned_34_years_world_export_import_dataset'

# Save DataFrame to BigQuery table
df.to_gbq(destination_table=f"{project_id}.{dataset_id}.{table_id}", project_id=project_id, if_exists='replace')

100%|██████████| 1/1 [00:00<00:00, 6223.00it/s]
```

Step 16: Go to BigQuery, there are 2 tables, 1 contains the original dataset and another 1 table contains the cleaned dataset.

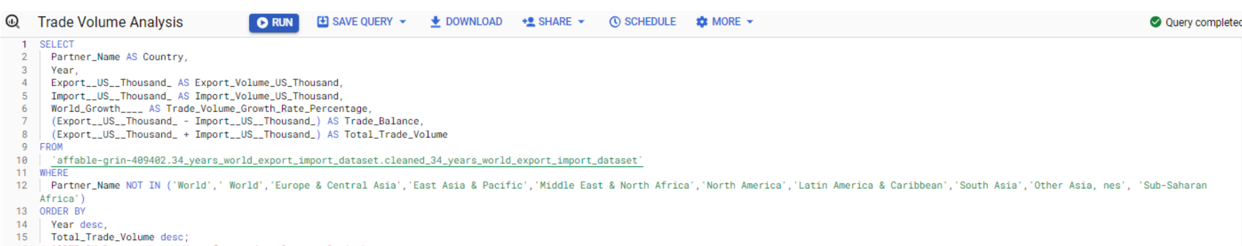


5.4. Data Analytics – BigQuery

Step 1: Go to BigQuery Studio. Click on the "Create SQL Query" button.

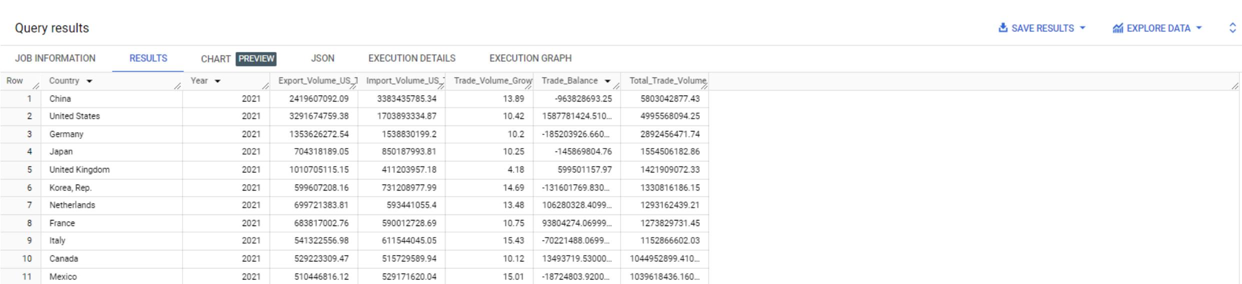


Step 2: Enter a valid SQL query in the query editor. For example, query the cleaned dataset to determine the countries with highest trade volumes in 2021.



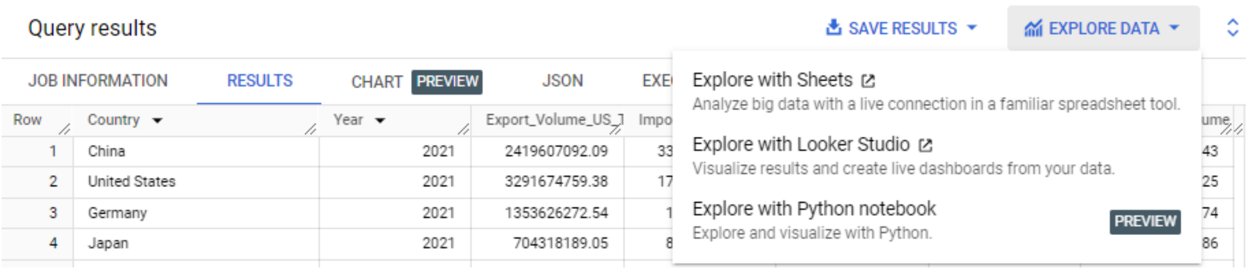

```sql
SELECT
  Partner_Name AS Country,
  Year,
  Export__US__Thousand_ AS Export_Volume_US_Thousand,
  Import__US__Thousand_ AS Import_Volume_US_Thousand,
  World_Growth____ AS Trade_Volume_Growth_Rate_Percentage,
  (Export__US__Thousand_ - Import__US__Thousand_) AS Trade_Balance,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name NOT IN ('World',' World','Europe & Central Asia','East Asia & Pacific','Middle East & North Africa','North America','Latin America & Caribbean','South Asia','Other Asia, nes', 'Sub-Saharan Africa')
ORDER BY
  Year desc,
  Total_Trade_Volume desc;
# ORDER BY Partner_Name, Year for trade volume analysis by country;
```

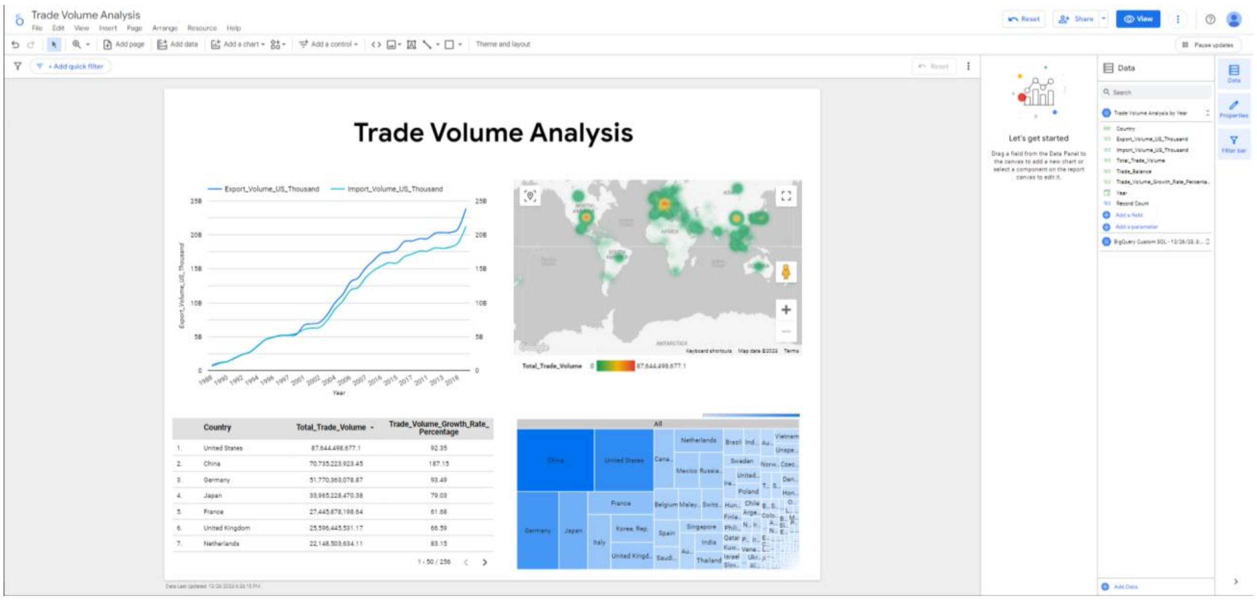
Step 3: Click "Run" to obtain the query results. It showed that the countries with the highest trade volume in 2021 is China, followed by United States, Germany, and Japan.



Row	Country	Year	Export_Volume_US_T	Import_Volume_US_T	Trade_Volume_Grow	Trade_Balance	Total_Trade_Volume
1	China	2021	2419607092.09	3383435785.34	13.89	-963828693.25	5803042877.43
2	United States	2021	3291674759.38	1703893334.87	10.42	1587781424.510...	4995568094.25
3	Germany	2021	1353626272.54	1538830199.2	10.2	-185203926.660...	2892456471.74
4	Japan	2021	704318189.05	850187993.81	10.25	-145869804.76	1554506182.86
5	United Kingdom	2021	1010705115.15	411203957.18	4.18	599501157.97	1421909072.33
6	Korea, Rep.	2021	599607208.16	731208977.99	14.69	-131601769.830...	1330816186.15
7	Netherlands	2021	699721383.81	593441055.4	13.48	106280328.4099...	1293162439.21
8	France	2021	683817002.76	590012728.69	10.75	93804274.06999...	1273829731.45
9	Italy	2021	541322556.98	611544045.05	15.43	-70221488.0699...	1152866602.03
10	Canada	2021	529223309.47	515729589.94	10.12	13493719.53000...	1044952899.410...
11	Mexico	2021	510446816.12	529171620.04	15.01	-18724803.9200...	1039618436.160...

Step 4: In the Query Results section, click "Explore Data", then click "Explore with Looker Studio". This allows us to visualize results and create dashboard from the data.





Below summarizes different queries along with their corresponding query results. The results from the queries were visualized in Looker Studio.

(i) Trade volume analysis to determine countries with highest trade volumes in 2021.

<u>Query:</u>

Analysis (i) RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE Query completed

```
SELECT
  Partner_Name AS Country,
  Year,
  Export__US__Thousand_ AS Export_Volume_US_Thousand,
  Import__US__Thousand_ AS Import_Volume_US_Thousand,
  World_Growth____ AS Trade_Volume_Growth_Rate_Percentage,
  (Export__US__Thousand_ - Import__US__Thousand_) AS Trade_Balance,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name NOT IN ('World',' World','Europe & Central Asia','East Asia & Pacific','Middle East & North Africa','North America','Latin America & Caribbean','South Asia','Other Asia, nes', 'Sub-Saharan Africa')
  AND Year = 2021
ORDER BY
    Total_Trade_Volume desc;
```

<u>Query Results:</u>

Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Country	Year	Export_Volume_US_1	Import_Volume_US_	Trade_Volume_Grow	Trade_Balance	Total_Trade_Volume
1	China	2021	2419607092.09	3383435785.34	13.89	-963828693.25	5803042877.43
2	United States	2021	3291674759.38	1703893334.87	10.42	1587781424.510...	4995568094.25
3	Germany	2021	1353626272.54	1538830199.2	10.2	-185203926.660...	2892456471.74
4	Japan	2021	704318189.05	850187993.81	10.25	-145869804.76	1554506182.86
5	United Kingdom	2021	1010705115.15	411203957.18	4.18	599501157.97	1421909072.33
6	Korea, Rep.	2021	599607208.16	731208977.99	14.69	-131601769.830...	1330816186.15
7	Netherlands	2021	699721383.81	593441055.4	13.48	106280328.4099...	1293162439.21
8	France	2021	683817002.76	590012728.69	10.75	93804274.06999...	1273829731.45
9	Italy	2021	541322556.98	611544045.05	15.43	-70221488.0699...	1152866602.03
10	Canada	2021	529223309.47	515729589.94	10.12	13493719.53000...	1044952899.410...
11	Mexico	2021	510446816.12	529171620.04	15.01	-18724803.9200...	1039618436.160...
12	Belgium	2021	464825537.13	431548547.96	15.67	33276989.17000...	896374085.0899...
13	Switzerland	2021	506450948.04	388342418.48	5.52	118108529.56	894793366.52
14	India	2021	501730633.41	372312850.46	24.5	129417782.9500...	874043483.87

Insights:

- In 2021, the country with highest trade volume is China, followed by United States, Germany, and Japan.

(ii) Trade volume analysis for a specific country.

Query:

Analysis (ii) RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE Query completed.

```sql
SELECT
  Partner_Name AS Country,
  Year,
  Export__US__Thousand_ AS Export_Volume_US_Thousand,
  Import__US__Thousand_ AS Import_Volume_US_Thousand,
  World_Growth____ AS Trade_Volume_Growth_Rate_Percentage,
  (Export__US__Thousand_ - Import__US__Thousand_) AS Trade_Balance,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name = 'China'
ORDER BY
  Year desc,
  Total_Trade_Volume desc;
```

Query Results:

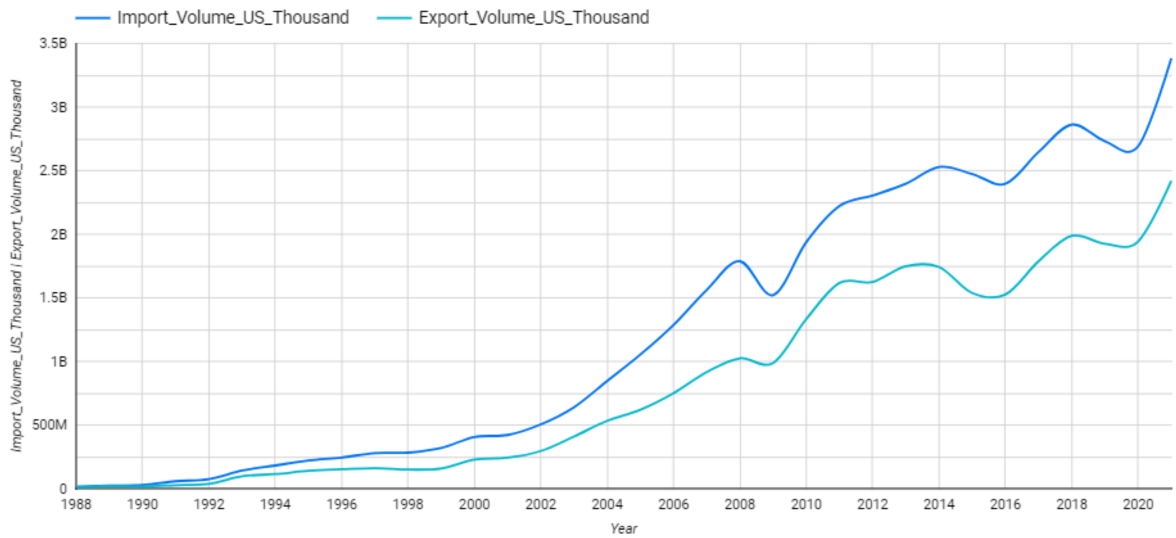
Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Country	Year	Export_Volume_US_1	Import_Volume_US_	Trade_Volume_Grow	Trade_Balance	Total_Trade_Volume
1	China	2021	2419607092.09	3383435785.34	13.89	-963828693.25	5803042877.43
2	China	2020	1943216340.33	2691026163.58	-0.23	-747809823.25	4634242503.91
3	China	2019	1924374506.81	2730865804.26	-1.28	-806491297.450...	4655240311.07
4	China	2018	1987292151.09	2862195428.63	7.57	-874903277.540...	4849487579.72
5	China	2017	1787239787.13	2647117289.59	7.76	-859877502.46	4434357076.72
6	China	2016	1525797542.12	2396469750.69	-2.77	-870672208.570...	3922267292.81
7	China	2015	1537067593.05	2473581490.39	-7.41	-936513897.339...	4010649083.439...
8	China	2014	1741960455.31	2529204919.96	0.24	-787244464.650...	4271165375.27
9	China	2013	1748337327.91	2397123775.55	3.56	-648786447.640...	4145461103.46
10	China	2012	1624172572.65	2304480905.73	2.12	-680308333.079...	3928653478.38
11	China	2011	1616784547.02	2221071164.59	11.75	-604286617.570...	3837855711.61
12	China	2010	1333059090.43	1937911681.5	17.83	-604852591.069...	3270970771.930...
13	China	2009	987126741.1	1520803383.21	-5.77	-533676642.11	2507930124.31
14	China	2008	1024298423.78	1787474945.0	8.84	-763176521.22	2811773368.779...
15	China	2007	915749711.81	1562106356.92	9.91	-646356645.110...	2477856068.73
16	China	2006	749094189.93	1287042055.06	9.51	-537947865.13	2036136244.989...

Explore Results in Looker Studio:



Insights:

- From 1988 to 2000, minimal increase in trade volume aligns with China's early stages of economic reforms and opening up to international trade, which started in the late 1970s.
- From 2000 onwards, significant surge in trade volume corresponds with China's substantial economic transformation. China became known as the "world's factory", leveraging its vast labor force, and competitive production costs to become a global manufacturing hub.

(iii) Tariff line analysis for China, United States, Germany, and Japan

Query:

Analysis (iii) RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE Query completed.

```sql
SELECT
  Partner_Name AS Country,
  Year,
  Export__US__Thousand_ AS Export_Volume_US_Thousand,
  Import__US__Thousand_ AS Import_Volume_US_Thousand,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume,
  AHS_Total_Tariff_Lines,
  AHS_Dutiable_Tariff_Lines_Share____ AS AHS_Dutiable_Tariff_Lines_Share,
  AHS_Duty_Free_Tariff_Lines_Share____ AS AHS_Duty_Free_Tariff_Lines_Share,
  AHS_Specific_Tariff_Lines_Share____ AS AHS_Specific_Tariff_Lines_Share,
  AHS_AVE_Tariff_Lines_Share____ AS AHS_AVE_Tariff_Lines_Share
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name IN ('China','United States','Germany','Japan')
ORDER BY
  Year desc;
```

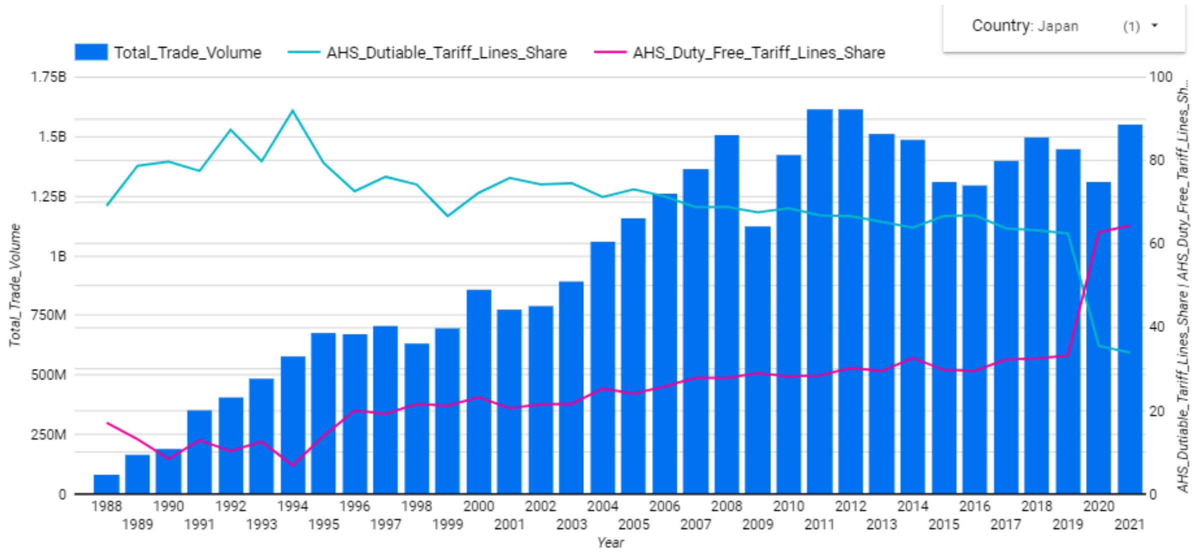
Query Results:

Query results SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS CHART PREVIEW JSON EXECUTION DETAILS EXECUTION GRAPH

Row	Country	Year	Export_Volume_US_T	Import_Volume_US_T	Total_Trade_Volume	AHS_Total_Tariff_Lin	AHS_Dutiable_Tariff_	AHS_Duty_Free_Tarif	AHS_Specific_Tariff_	AHS_AVE_Tariff_Line
1	China	2021	2419607092.09	3383435785.34	5803042877.43	905368.0	63.18	33.13	0.45	3.24
2	Japan	2021	704318189.05	850187993.81	1554506182.86	488442.0	33.91	64.42	0.32	1.35
3	Germany	2021	1353626272.54	1538830199.2	2892456471.74	440657.0	45.19	52.35	0.26	2.21
4	United States	2021	3291674759.38	1703893334.87	4995568094.25	770717.0	60.82	33.95	0.52	4.71
5	China	2020	1943216340.33	2691026163.58	4634242503.91	865922.0	65.6	30.88	2.56	0.95
6	Japan	2020	593173166.59	721069676.57	1314242843.16	453817.0	35.57	62.78	0.82	0.84
7	Germany	2020	1092044662.42	1291639475.47	2383684137.890...	418152.0	49.44	48.12	0.4	2.04
8	United States	2020	2702692879.03	1389318967.86	4092011846.890...	730205.0	61.35	33.51	3.74	1.4
9	China	2019	1924374506.81	2730865804.26	4655240311.07	720418.0	63.74	32.11	0.53	3.63
10	Japan	2019	654804747.44	795029187.31	1449833934.75	420262.0	62.53	33.21	0.69	3.57

Explore Results in Looker Studio:



Insights:

- The plateauing in Japan's trade volume since 2010s is caused by demographic challenges, including an aging population and declining birth rates, impacting labor force participation and economic growth potential.
- In response, the percentage of AHS duty free tariff lines in Japan was raised significantly in 2019. Higher percentages of duty-free tariff lines can stimulate trade by reducing barriers to imports and exports for specific goods and encouraging international trade relationships.

(iv) Tariff rate analysis for China, United States, Germany, Japan, and Malaysia

Query:

Analysis (iv)

```sql
SELECT
  Partner_Name AS Country,
  Year,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume,
  AHS_Simple_Average____ AS AHS_Simple_Average_Tariff_Rate,
  AHS_Weighted_Average____ AS AHS_Weighted_Average_Tariff_Rate,
  AHS_SpecificDuty_Imports__US__Thousand_ AS AHS_Specific_Duty_Imports_US_Thousand,
  AHS_Dutiable_Imports__US__Thousand_ AS AHS_Dutiable_Imports_US_Thousand,
  AHS_Duty_Free_Imports__US__Thousand_ AS AHS_Duty_Free_Imports_US_Thousand,
 CASE
    WHEN LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year) <> 0 THEN
      ((AHS_Simple_Average____ - LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year)) / NULLIF(LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year), 0)) * 100
    ELSE NULL
  END AS Tariff_Rate_Change_Percentage
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name IN ('China','United States','Germany','Japan','Malaysia')
ORDER BY
  Year desc,
  AHS_Duty_Free_Imports_US_Thousand desc;
```

Query Results:

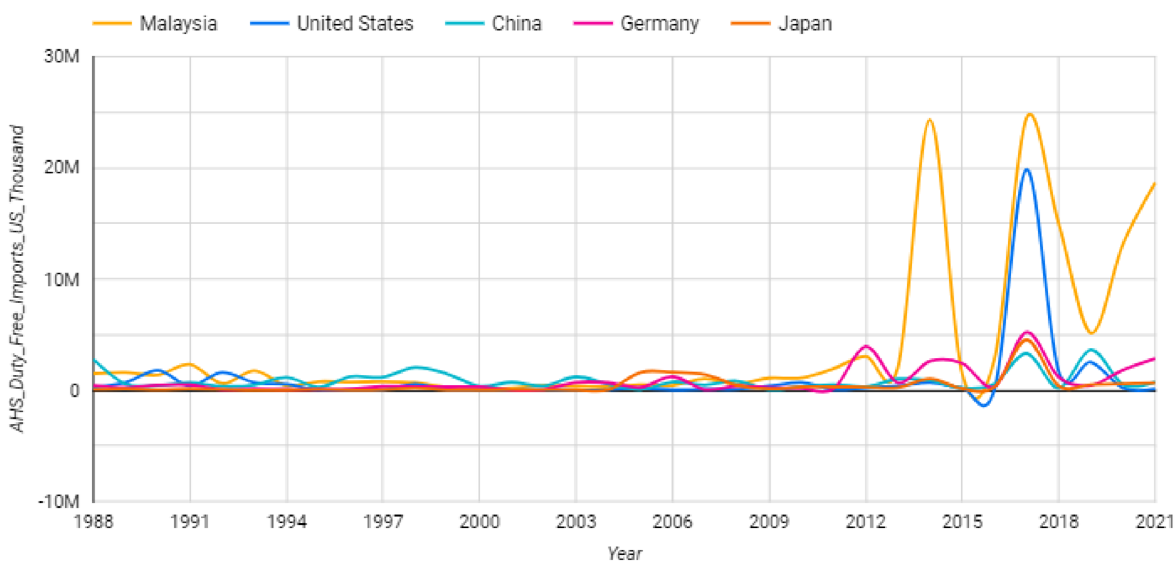
Query results

SAVE RESULTS   EXPLORE DATA

JOB INFORMATION   RESULTS   CHART PREVIEW   JSON   EXECUTION DETAILS   EXECUTION GRAPH

Row	Country	Year	Total_Trade_Volume	AHS_Simple_Averag	AHS_Weighted_Aver	AHS_Specific_Duty_I	AHS_Dutiable_Impor	AHS_Duty_Free_Imp	Tariff_Rate_Change
1	Malaysia	2021	686983709.61	3.7	2.71	10249354.49	394877911.76	18658728.51	2.493074792243...
2	Germany	2021	2892456471.74	5.76	5.55	109898884.04	601338793.91	2854060.22	-5.88235294117...
3	China	2021	5803042877.43	7.17	5.61	446975489.39	3600655113.5	718658.0	5.441176470588...
4	Japan	2021	1554506182.86	3.66	3.69	146131473.33	819719442.85	635740.45	-2.13903743315...
5	United States	2021	4995568094.25	7.07	5.06	244392878.4	1573467646.32	133209.79	10.64162754303...
6	Malaysia	2020	539644761.15	3.61	2.81	9279150.18	322290705.56	13172896.33	-8.37563451776...
7	Germany	2020	2383684137.890...	6.12	5.86	92765389.85	475591428.97	1829357.12	-13.3144475920...
8	Japan	2020	1314242843.16	3.74	4.09	137526582.59	712257675.57	582365.97	-33.4519572953...
9	China	2020	4634242503.91	6.8	5.98	326540960.89	3016693491.2	501207.44	-5.42420027816...
10	United States	2020	4092011846.890...	6.39	4.96	208281017.7	1404232629.96	196486.0	-11.7403314917...

<u>Explore Results in Looker Studio:</u>



<u>Insights:</u>

- Since mid-2010s, Malaysia's duty-free imports volume is higher than that of the China, United States, Germany, and Japan. This implied that Malaysia has specific policies that facilitate duty-free imports aimed at promoting trade and fostering economic development.
- Malaysia established several Free Trade Zones and Free Industrial Zones across the country. These zones offer various incentives to businesses, including exemptions on import duties for raw materials and machinery used in manufacturing purposes.

(v) Tariff Rate Analysis for China, United States, Germany, Japan, and India

<u>Query:</u>

Analysis (v) RUN SAVE QUERY DOWNLOAD SHARE SCHEDULE MORE Query completed.

```sql
SELECT
  Partner_Name AS Country,
  Year,
  (Export__US__Thousand_ + Import__US__Thousand_) AS Total_Trade_Volume,
  AHS_Simple_Average____ AS AHS_Simple_Average_Tariff_Rate,
  AHS_Weighted_Average____ AS AHS_Weighted_Average_Tariff_Rate,
  AHS_SpecificDuty_Imports__US__Thousand_ AS AHS_Specific_Duty_Imports_US_Thousand,
  AHS_Dutiable_Imports__US__Thousand_ AS AHS_Dutiable_Imports_US_Thousand,
  AHS_Duty_Free_Imports__US__Thousand_ AS AHS_Duty_Free_Imports_US_Thousand,
 CASE
    WHEN LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year) <> 0 THEN
      ((AHS_Simple_Average____ - LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year)) / NULLIF(LAG(AHS_Simple_Average____, 1) OVER (PARTITION BY Partner_Name ORDER BY Year), 0)) * 100
    ELSE NULL
  END AS Tariff_Rate_Change_Percentage
FROM
  `affable-grin-409402.34_years_world_export_import_dataset.cleaned_34_years_world_export_import_dataset`
WHERE
  Partner_Name IN ('China','United States','Germany','Japan','India')
ORDER BY
  Year desc,
  AHS_Specific_Duty_Imports_US_Thousand desc;
```

Query Results:

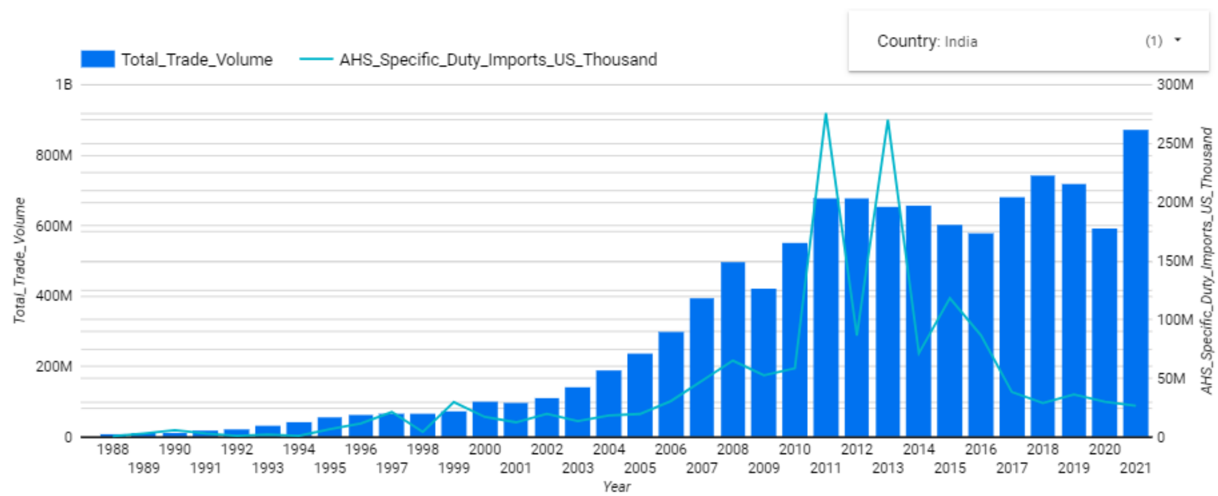
Query results

SAVE RESULTS EXPLORE DATA

JOB INFORMATION RESULTS CHART PREVIEW JSON EXECUTION DETAILS EXECUTION GRAPH

Row	Country	Year	Total_Trade_Volume	AHS_Simple_Averag	AHS_Weighted_Aver	AHS_Specific_Duty_I	AHS_Dutiable_Impor	AHS_Duty_Free_Imp	Tariff_Rate_Change_
1	China	2021	5803042877.43	7.17	5.61	446975489.39	3600655113.5	718658.0	5.441176470588...
2	United States	2021	4995568094.25	7.07	5.06	244392878.4	1573467646.32	133209.79	10.64162754303...
3	Japan	2021	1554506182.86	3.66	3.69	146131473.33	819719442.85	635740.45	-2.13903743315...
4	Germany	2021	2892456471.74	5.76	5.55	109898884.04	601338793.91	2854060.22	-5.88235294117...
5	India	2021	874043483.87	5.83	5.88	26715654.01	386114402.29	903859.81	1.567944250871...
6	China	2020	4634242503.91	6.8	5.98	326540960.89	3016693491.2	501207.44	-5.42420027816...
7	United States	2020	4092011846.890...	6.39	4.96	208281017.7	1404232629.96	196486.0	-11.7403314917...
8	Japan	2020	1314242843.16	3.74	4.09	137526582.59	712257675.57	582365.97	-33.4519572953...
9	Germany	2020	2383684137.890...	6.12	5.86	92765389.85	475591428.97	1829357.12	-13.3144475920...
10	India	2020	594792586.99	5.74	6.12	30275953.08	306501967.0	1094199.25	-20.0557103064...

Explore Results in Looker Studio:



Insights:

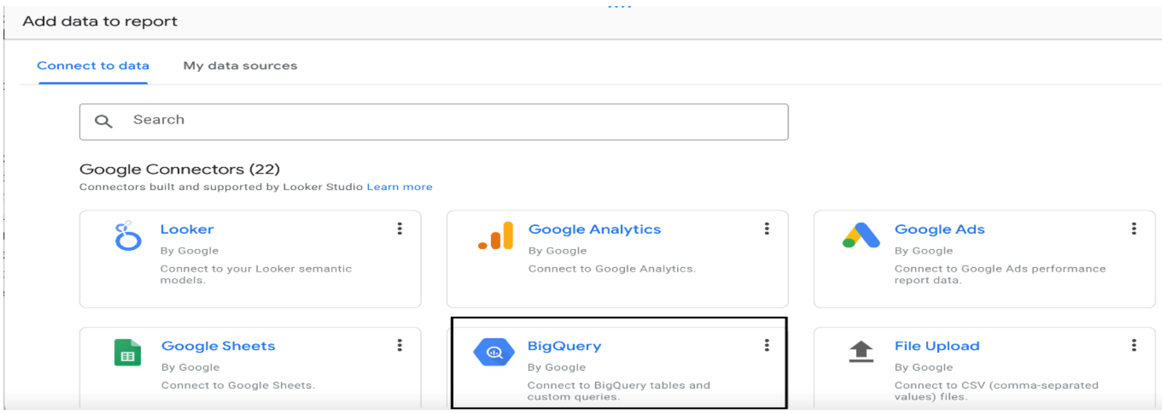
- In the early 1990s, India implemented economic reforms. The country’s commitment to economic liberalization gained momentum in the mid-2000s and led to significant growth in trade volumes.
- India’s trade volume growth decelerated in mid-2010s due to influence by global economic slowdown. In response, India implemented export promotion schemes i.e. Duty Entitlement Passbook (DEPB) scheme to incentivize exports of Indian goods.

5.5. Data Visualization - Looker Studio

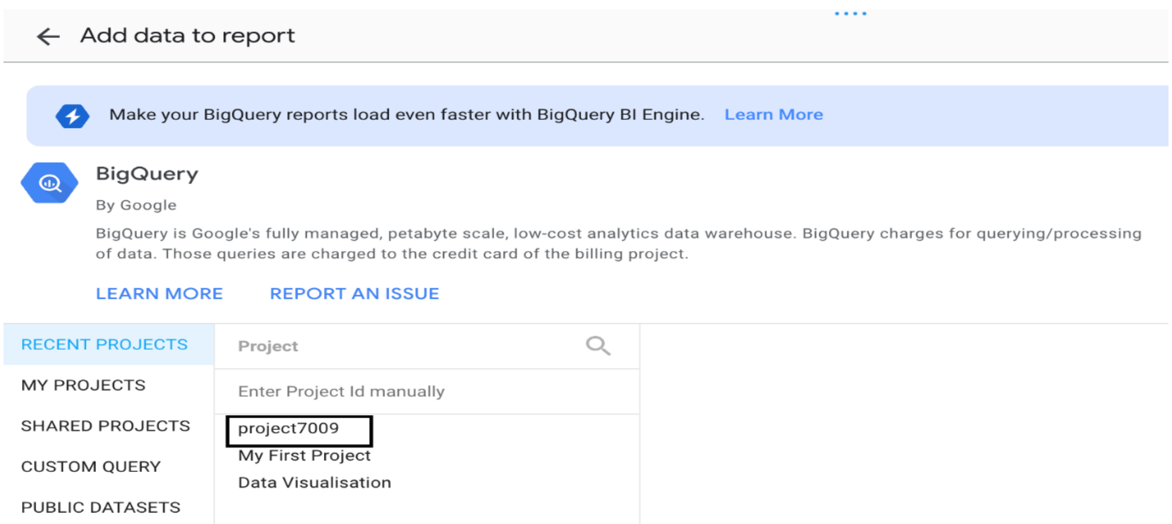
This is the link of dashboard of the Looker: https://lookerstudio.google.com/s/gIZmkCP62dc

Below is the step to do the visualisation.

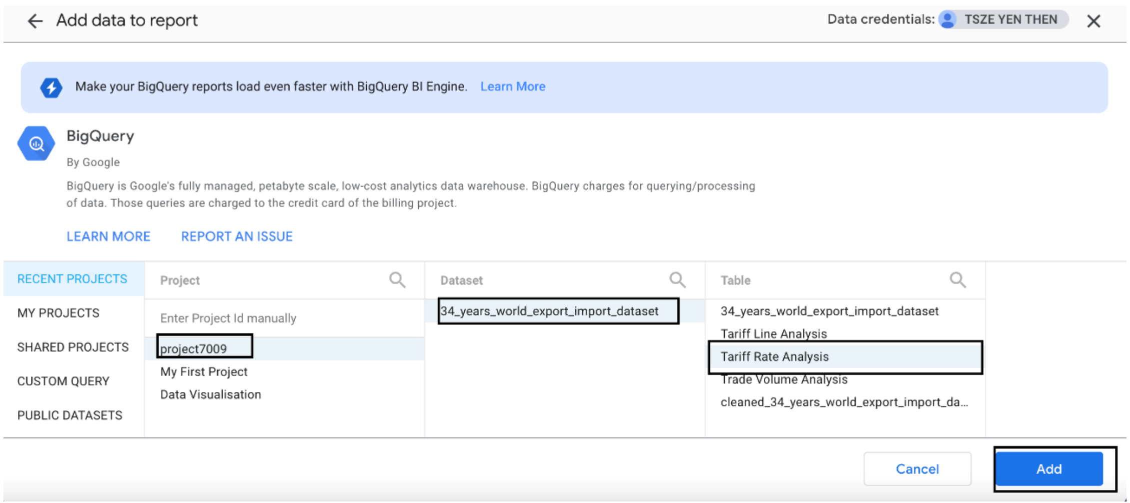
Step 1: Import the data from Big Query.



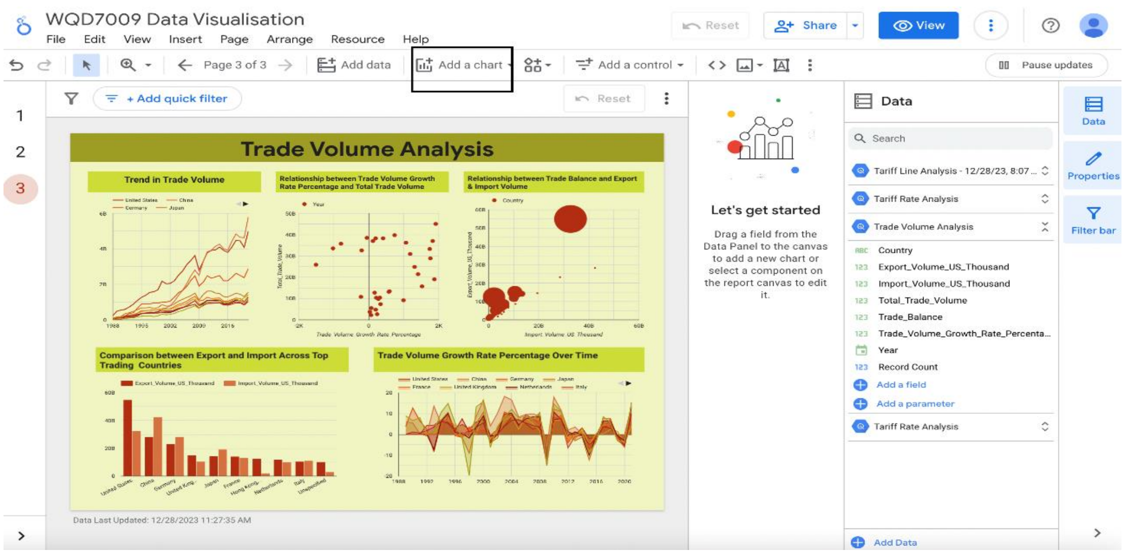
Step 2: Choose the project.



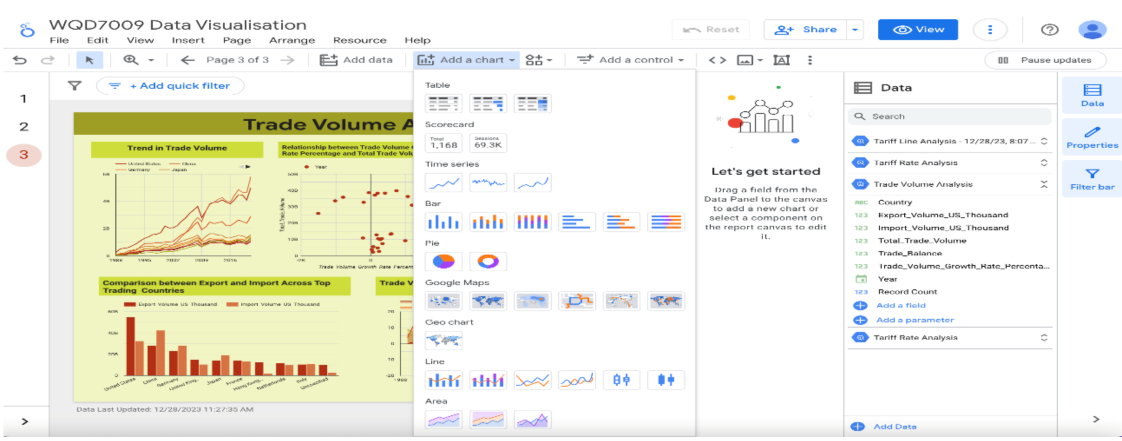
Step 3: Choose Dataset and Table.



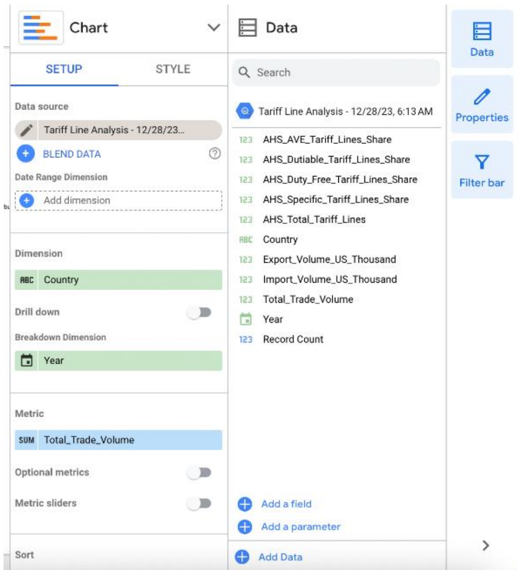
Step 4: Add a chart.



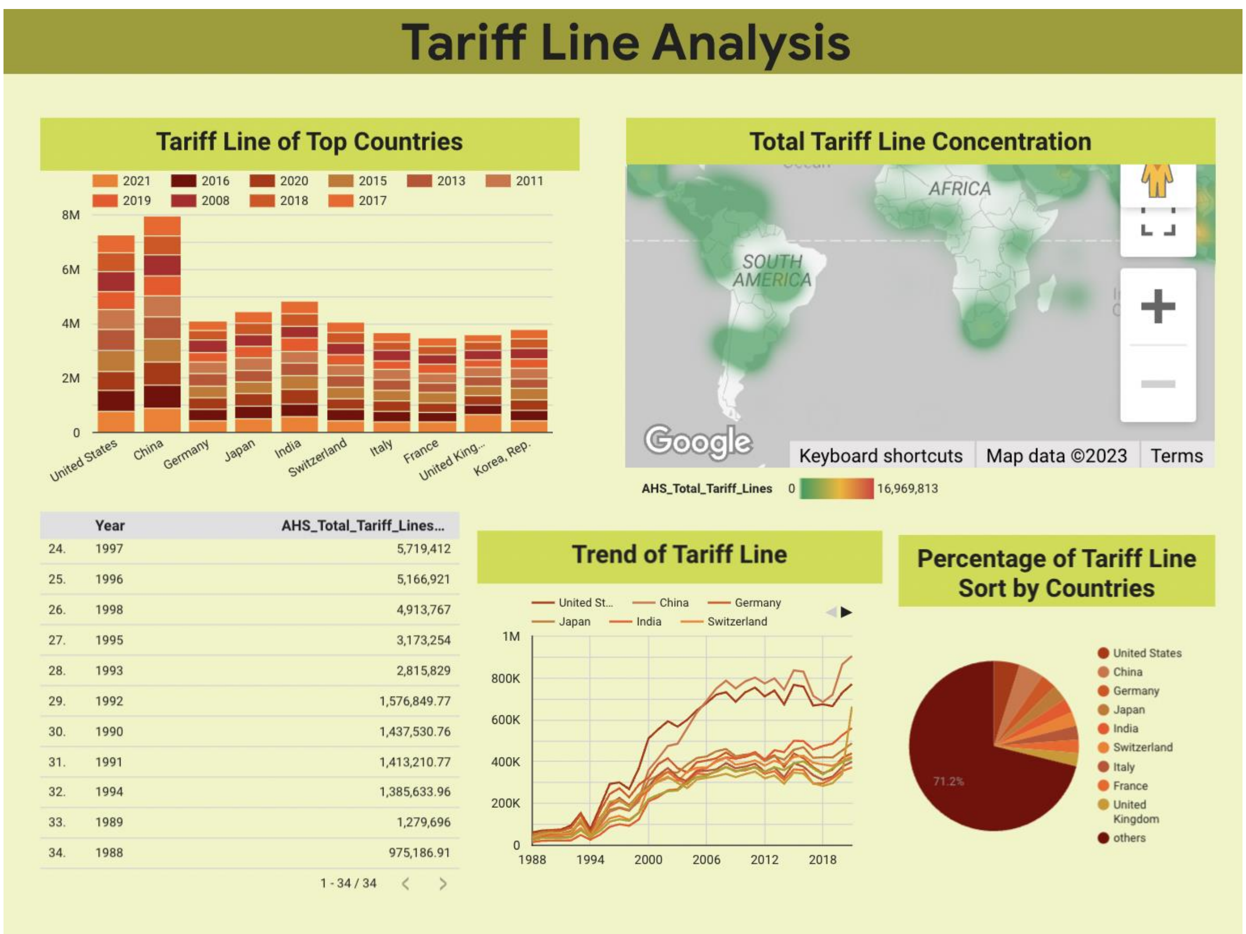
Step 5: Choose a chart.



Step 6: Choose the data source and dimensions.



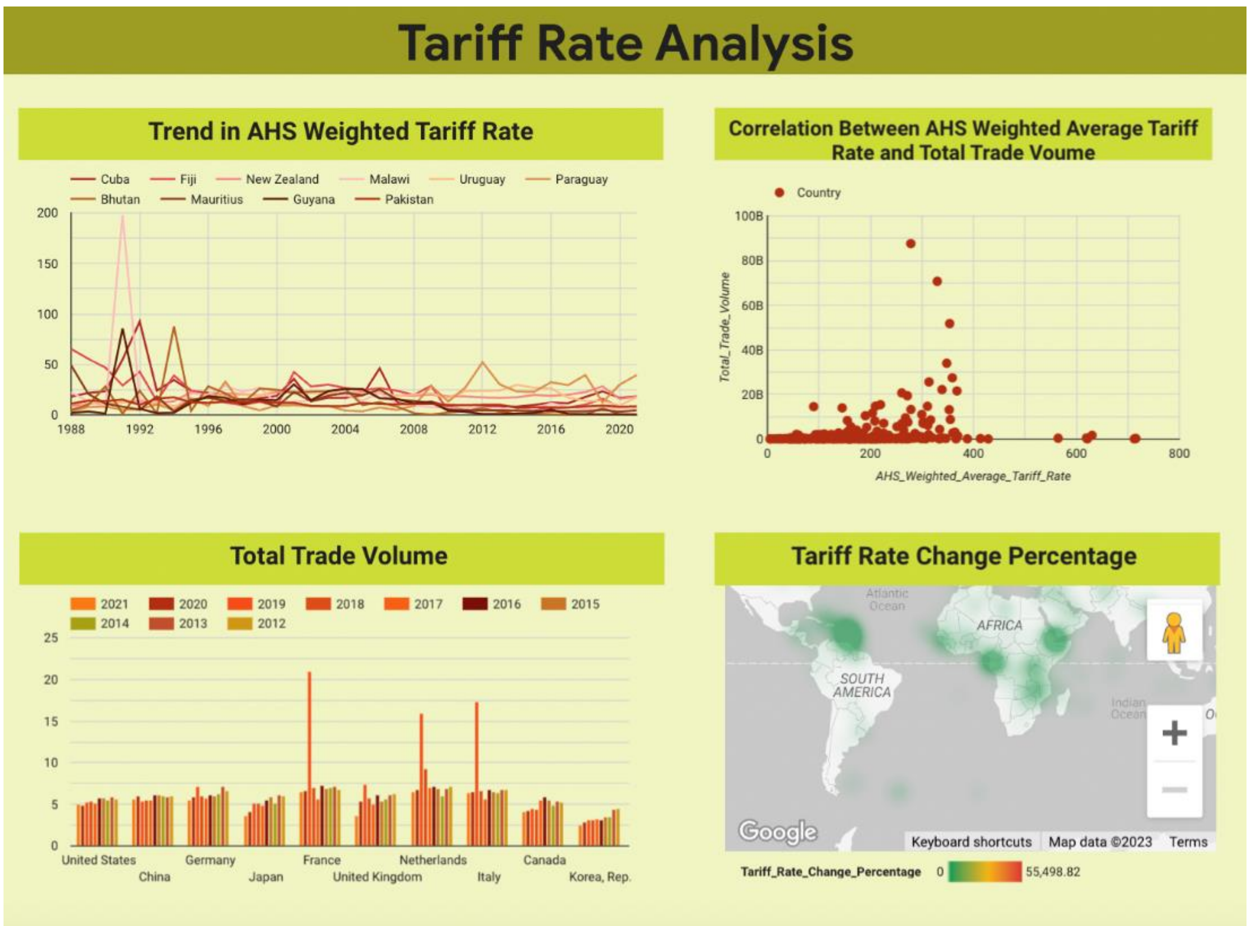
Data Visualization of Tariff Line Analysis:



	Year	AHS_Total_Tariff_Lines...
24.	1997	5,719,412
25.	1996	5,166,921
26.	1998	4,913,767
27.	1995	3,173,254
28.	1993	2,815,829
29.	1992	1,576,849.77
30.	1990	1,437,530.76
31.	1991	1,413,210.77
32.	1994	1,385,633.96
33.	1989	1,279,696
34.	1988	975,186.91

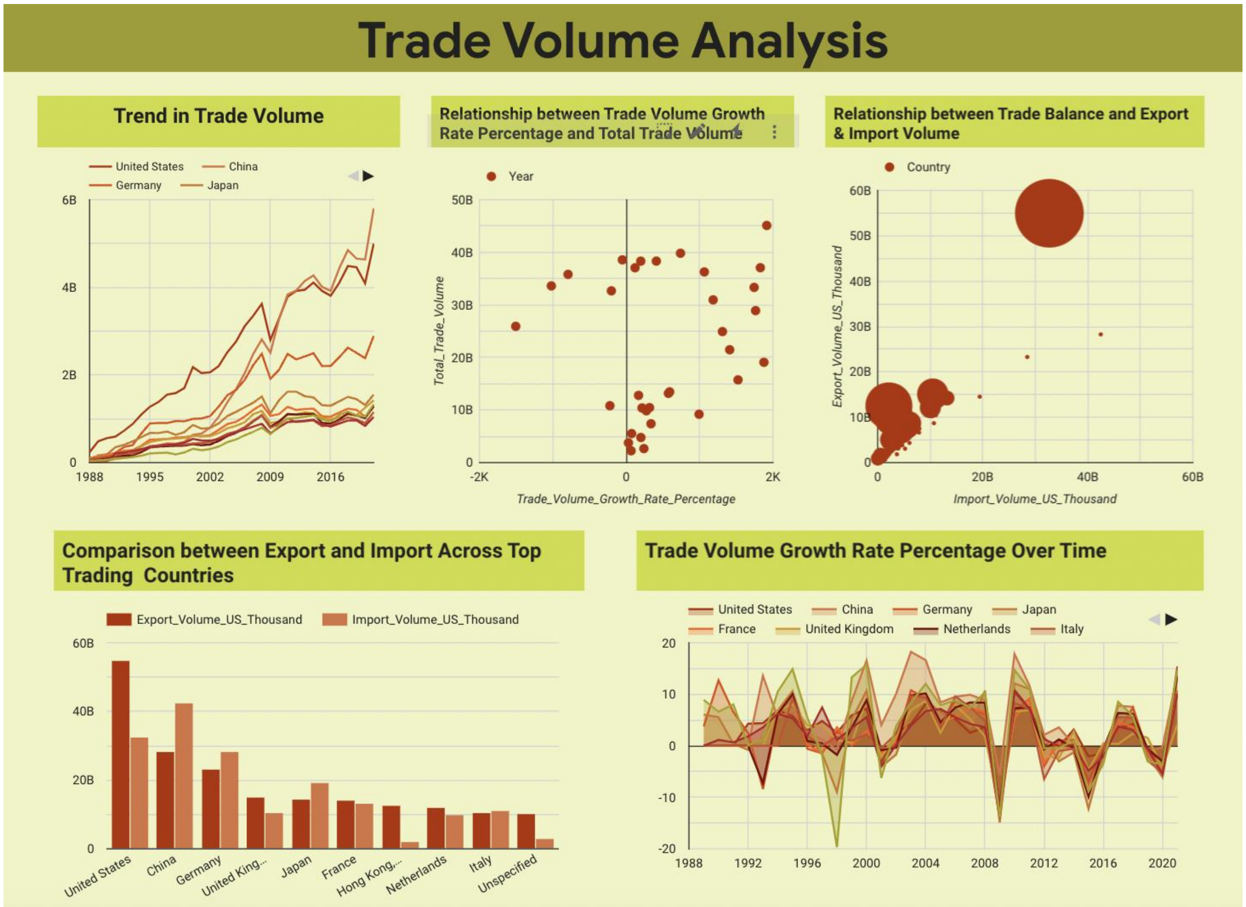
From the Tariff Line of Top Countries, China has the highest value of AHS Total Tariff Line accumulated from year 2017 to year 2021. The Heatmap shows the total tariff line concentration given the metric label where the least total tariff line towards to zero is green and the higher value of total tariff lines with dark red concentration. There is more total tariff line concentrated in Europe region. The table shows the AHS Total Tariff Line from the year 1988 to year 2021, the value roughly increases from year to year. Then, for the trend of tariff line, from the year 1988 to year 2005, United States scores the highest, however, begins from year 2005 to 2021, China has surpassed United States. Pie chart shows the percentage of the top AHS Total Tariff Line respectively and the other countries combined which obtain value of 71.2%.

Data Visualization of Tariff Rate Analysis:



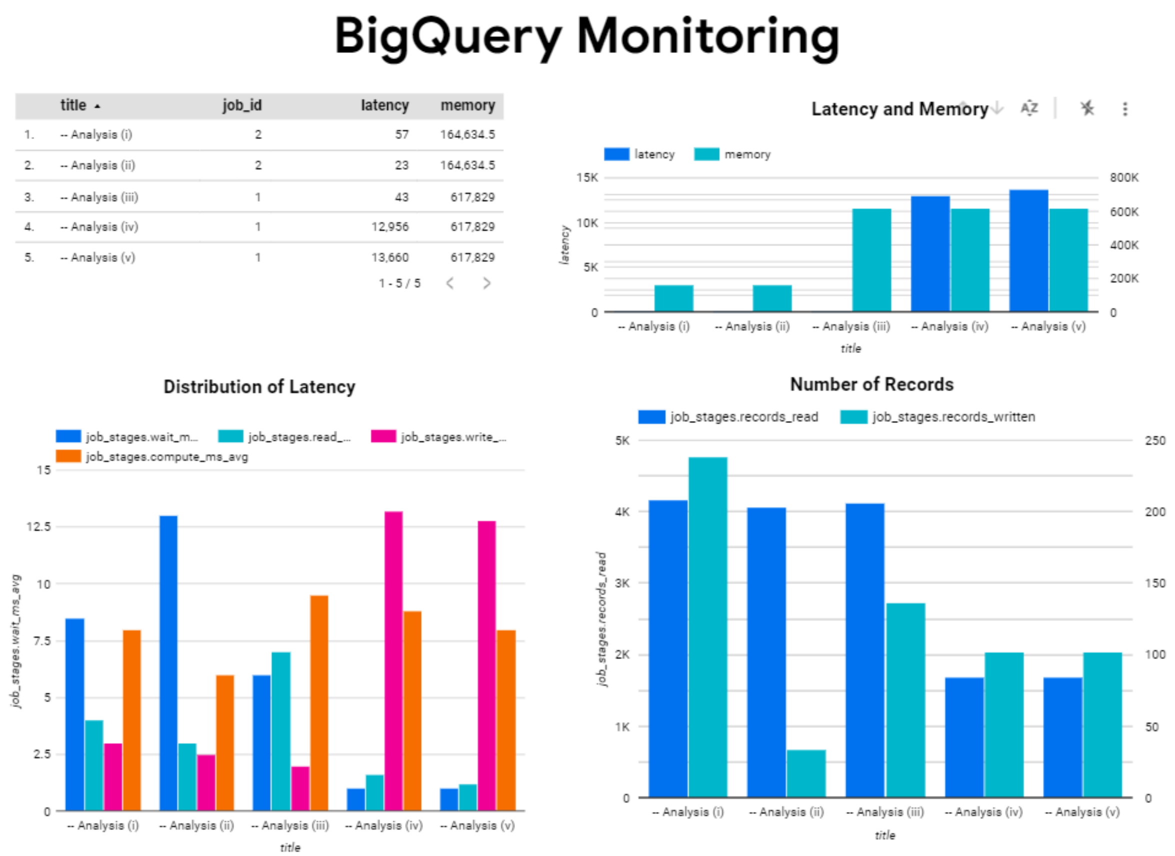
The line graph illustrates a generally stable weighted tariff rate among countries, with notable spikes in Bhutan and Guyana in 1990, Cuba in 1992, and Paraguay in 1994, possibly reflecting policy changes or trade imbalances. A bar chart shows trade values from 2012 to 2021, with France peaking in 2019 and spikes in Germany, the UK, the Netherlands, and Italy. The heatmap suggests little change in tariff rates overall, while the scatter plot indicates most countries maintain low tariff rates, with a few outliers representing high tariff rates and trade volumes.

Data Visualization of Trade Volume Analysis:



The line graph reveals rising trade volumes with the U.S. leading until China surpassed it post-2010. The scatter plot indicates no strong correlation between trade volumes and growth rates, with data clustered around moderate volumes and low growth rates. The bubble plot relates trade balance to export and import volumes, with bubble size denoting trade balance magnitude; the U.S. shows a surplus, while China, Germany, Japan, and Italy depict deficits. A 2009 area plot highlights a significant dip during the Great Recession, with China's and the U.S.'s trade growth rates dropping to –5.77% and –13.98%, respectively.

6. Evaluation Metrics with Graph



The metrics selected for our revaluation are latency, memory and number of records for read and written. The latency will have a further breakdown into four stages which is wait, read, compute and write to observe the distribution of the time taken for query.

7. Conclusion

In short, all our objectives have been met.

First, we have effectively integrated Google Cloud tools across the data lifecycle, using Google Cloud Storage for ingestion, BigQuery for storage, Vertex AI Colab for preprocessing, BigQuery for analysis and Looker Studio for visualization.

Second, insightful analysis of trade volume, tariff line and tariff rate has yielded several key findings:

- China’s Trade Growth (1988-2000): Corresponds with early economic reform stages, minimal increase has been observed in trade volume.
- China’s Economic Transformation (Post-200): Rapid import/export growth, establishing China as the “world’s factory”.

- Japan’s Trade Plateau (2010s): Indicates demographic challenges; responses include increased AHS duty-free tariff lines in 2019 to stimulate trade.

Finally, we have also conducted query performance comparisons in BigQuery among the analysis:

- Analysis 1 and 2 exhibited minimal memory use and latency.
- Starting from Analysis 3, there’s a marked increase in memory consumption and latency.
- A decreasing trend in number of records from Analysis 1 to 5.

8. Reference

Ali, M. H., Hosain, M. S., & Hossain, M. A. (2021). Big Data analysis using BigQuery on cloud computing platform. *Australian JofEng Inno Tech, 3*(1), 1-9. https://universepg.com/public/storage/journal-pdf/Big%20data%20analysis%20using%20bigquery%20on%20cloud%20computing%20platform.pdf

Amiri-Zarandi, M., Hazrati Fard, M., Yousefinaghani, S., Kaviani, M., & Dara, R. (2022). A platform approach to smart farm information processing. *Agriculture, 12*(6), 838. https://doi.org/10.3390/agriculture12060838

GeeksforGeeks. (2022, September 29). *Apache HIVE features and Limitations*. https://www.geeksforgeeks.org/apache-hive-features-and-limitations/

Gharpure, R., & Ghodke, M. (2021). Effect of cloud computing technology adoption on reduction in costs: A critical review from the perspective of business. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(10), 4391-4399. https://www.proquest.com/openview/ab10da4307c658b8608b0e2f3735bb89/1?pq-origsite=gscholar&cbl=2045096

Ghoshal, A. (2023, August 29). *Google Cloud's Colab Enterprise environment to help tune LLMs*. Google Cloud's Colab Enterprise environment to help tune LLMs | InfoWorld. https://www.infoworld.com/article/3705496/google-cloud-s-colab-enterprise-environment-to-help-tune-llms.html

Google Cloud. (2018). *HDFS vs. Cloud Storage: Pros, cons and migration tips.* https://cloud.google.com/blog/products/storage-data-transfer/hdfs-vs-cloud-storage-pros-cons-and-migration-tips

Joshi, A. P., & Patel, B. V. (2021). Data Preprocessing: the Techniques for Preparing Clean and Quality Data for Data Analytics Process. *Oriental Journal of Computer Science and Technology*, *13*(0203), 78–81. https://doi.org/10.13005/ojcst13.0203.03

Khan, M., & Sarwar, S. K. (2011). Data and information visualization Methods, and Interactive Mechanisms: a survey. *International Journal of Computer Applications*, *34*(1), 1–14. https://research.ijcaonline.org/volume34/number1/pxc3875722.pdf

L'Esteve, R. C. (2023). *Decentralizing Data and Democratizing Analytics. In The Cloud Leader's Handbook: Strategically Innovate, Transform, and Scale Organizations* (pp. 79-104). Berkeley, CA: Apress.

Lakshmanan, V. (2022). *Data Science on the Google Cloud Platform Implementing End-to-End Real-time Data Pipelines: From Ingest to Machine Learning*. O'Reilly.

Mazumdar, S., Seybold, D., Kritikos, K., & Verginadis, Y. (2019). A survey on data storage and placement methodologies for Cloud-Big Data ecosystem. *Journal of Big Data 6,* 15. https://doi.org/10.1186/s40537-019-0178-3

Mena, C., Karatzas, A., & Hansen, C. (2022). International trade resilience and the Covid-19 pandemic. *Journal of Business Research, 138,* 77-91. https://doi.org/10.1016/j.jbusres.2021.08.064

Md. Anwar, H. (2021). Big Data Analysis using BigQuery on Cloud Computing Platform. *Australian Journal of Engineering and Innovative Technology*, 1–9. https://doi.org/10.34104/ajeit.021.0109

Mohemmed, S. M. (2020). Cloud-based healthcare data management framework. *KSII Transactions on Internet and Information Systems, 14*(3), 1014-1025. https://doi.org/10.3837/tiis.2020.03.006

Mucchetti, M. (2020). *BigQuery for Data Warehousing*. https://link.springer.com/book/10.1007/978-1-4842-6186-6

Pamma, A. (2023, August 29). *Google upgrades Vertex AI, foundation models, adds Meta's Llama 2, Anthropic to portfolio and more*. Google upgrades Vertex AI, foundation models, adds Meta's Llama 2, Anthropic to portfolio and more | IT World Canada. https://www.itworldcanada.com/article/google-upgrades-vertex-ai-foundation-models-adds-metas-llama-2-anthropic-to-portfolio-and-more/545592

Peng, J., Wu, W., Yan, J. N., Qi, D., Rzeszotarski, J. M., & Wang, J. (2022). *User Interfaces for Exploratory Data Analysis: A Survey of Open-Source and Commercial Tools*.

Riahi, Y., & Riahi, S. (2018). Big Data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524–528. https://doi.org/10.21276/ijre.2018.5.9.5

Santana M. K. T., G., & Padmamma, S. (Eds.). (2023). Data Visualization of E-journal Collection of Kuvempu University Library Using Google's Looker Studio. *Proceedings of the 5th National Conference on Management of Modern Libraries (NACML) - 2023*.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data Challenges and analytical methods. *Journal of Business Research*, *70*, 263–286. https://doi.org/10.1016/j.jbusres.2016.08.001

Sucharitha, V., Sasidharan, S., & Prakash, P. G. O. (2014). Visualization of Big Data: Its tools and challenges. *International Journal of Applied Engineering Research*, *9*(18). https://www.amrita.edu/publication/visualization-big-data-its-tools-and-challenges

Wang, J., Yang, Y., Wang, T., Sherratt, R. S., & Zhang, J. (2020). Big Data Service Architecture: A Survey. *Journal of Internet Technology*, *21*(2), 393–405. https://doi.org/10.3966/160792642020032102008

Yang, C. H., Lee, C. F., & Chang, P. Y. (2023). Export-and import-based economic models for predicting global trade using deep learning. *Expert Systems with Applications, 218,* 119590. https://doi.org/10.1016/j.eswa.2023.119590

Yu, W., Liu, Y., Dillon, T., Rahayu, W., & Mostafa, F. (2021). An integrated framework for health state monitoring in a smart factory employing IoT and big data techniques. *IEEE Internet of Things Journal*, *9*(3), 2443-2454. https://doi.org/10.1109/JIOT.2021.3096637.