UNIVERSITY OF MALAYA

WQD7009

BIG DATA APPLICATIONS AND ANALYTICS

IMPACT OF EDUCATION EXPENDITURE ON GDP GROWTH ANALYSIS

GROUP ASSIGNMENT REPORT GROUP 16

GROUP MEMBERS:

NAME	MATRIX
SHALINI A/P GUNALAN	S2147287
DHIVASHINI A/P LINGADARAN	S2127834
NG YAN TING	17204322
TARSVINI A/P RAVINTHER	17193844
ZHANCANG HEI	s2188090

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Project Background

Societal advancement and economic development are significantly influenced by education. To analyse how a country's economic trajectory is affected by the GDP share allotted to education, it is crucial to comprehend the relationship between GDP growth and education spending. This study investigates the connection between a nation's total economic growth and its GDP investment in education and how it affects other economic sectors such as industry & service sectors. To gain insight into how education influences economic environments, the research looks for trends and relationships using a World Bank dataset that contains a variety of economic, demographic, and developmental metrics for countries all over the world. Policymakers, economists, and educators should find value in this investigation as it provides useful data to inform strategic choices for promoting equitable and sustainable economic growth.

1.2 Project Objectives & Descriptions

This project's major goal is to discover whether an investment in education affects its ability for economic resilience and rapid recovery from downturns or crisis. Additionally, to promote more equitable economic development, this research aims to recognise and comprehend the challenges that each country faces as well as how investment in education can address these differences. The chosen dataset, World Bank Data on Countries is obtained from Kaggle.com https://www.kaggle.com/datasets/yusufglcan/country-data?select=Countries.csv . This dataset is a collection of global economic, demographic, and developmental indicators. The information was compiled from the World Bank, with many nation's data collected between 2000 and 2022. It offers a comprehensive picture of all the different facets of the social and economic landscape of every nation and covers a broad range of features. This dataset helps to compare nations according to indicators, identify patterns and trends in the social and economic development of various nations or look at how indicators have evolved over time for nations or regions.

2. MEETING MINUTES REPORT

Meeting Minute

4 Dec, 2023

Agenda	Attendees
• Brainstorm and identify suitable dataset for group assignment. • Select a research topic. • Identify research objectives.	• SHALINI (Team Leader) • DHIVASHINI (Member) • NG YAN TING (Member) • TARSVINI (Member) • ZHANCANG HEI (Member)

Updates

- Research title is IMPACT OF EDUCATION EXPENDITURE ON GDP GROWTH ANALYSIS
- Research objectives are:
  a. To discover whether an investment in education affects its ability for economic resilience and rapid recovery from downturns or crisis.
  b. To recognise and comprehend the challenges that each countries faces as well as how investment in education can address these differences.
- Ng Yan Ting's World Bank Data on Countries dataset was chosen for this research.
- The dataset was chosen because it is the largest dataset with many features compared to other members' dataset.
- The dataset contains variety of economic, demographic, and developmental metrics for countries all over the world.
- The dataset contains insightful features that are relevant to research objectives. Missing values in these features are less than 10%.
- Policymakers, economists, and educators can be benefitted from this research.

Action Items

Action	Owner
• Meeting Minute • Cloud-based data analytics framework proposal	Shalini
• Justify one potential tool relevant to each layer of the data lifecycle or data architecture process.	Dhivashini
• Produce practical output evidence by implementing at least 30% of the proposed data lifecycle process.	Tarsvini
• Analyse the implemented framework using any three performance evaluation metrics for query processing. • Produce three graphs to check their performance.	Ng Yan Ting & Zhanchang Hei

Timeline

Task List	Dec-23				Jan-24	
	W1	W2	W3	W4	W1	W2
Cloud-based data analytics framework proposal						
Justify one potential tool relevant to each layer						
Produce practical output evidence (cover 2 tools in framework development)						
Analyze three performance evaluation metrics for query processing & produce three graph						

EVIDENCE

N	Matric No	Name	Name of the Dataset	Status	Link for dataset
34	17204322	Ng Yan Ting	World Bank Data on Countries	Approved	https://www.kaggle.com/datasets/yusufglcan/country-data?select=Countries.csv

1.0 Introduction On Dataset

The dataset "World Bank Data on Countries" was chosen from Kaggle. This dataset contains an extensive variety of economic, demographic, and development data for countries all around the world. It is collected from the World Bank, an International organization dedicated to assisting developing countries with economic advancement through financing, advice and research. This data set comprises 25 attributes and 5106 entries. Table 1 in section 1.1 shows all the parameters in the dataset with explanation.

This is the link to the dataset:
https://www.kaggle.com/datasets/yusufglcan/country-data?select=Countries.csv

Figure: Part of Ng Yan Ting's Individual Assignment

3. PROJECT DESIGN

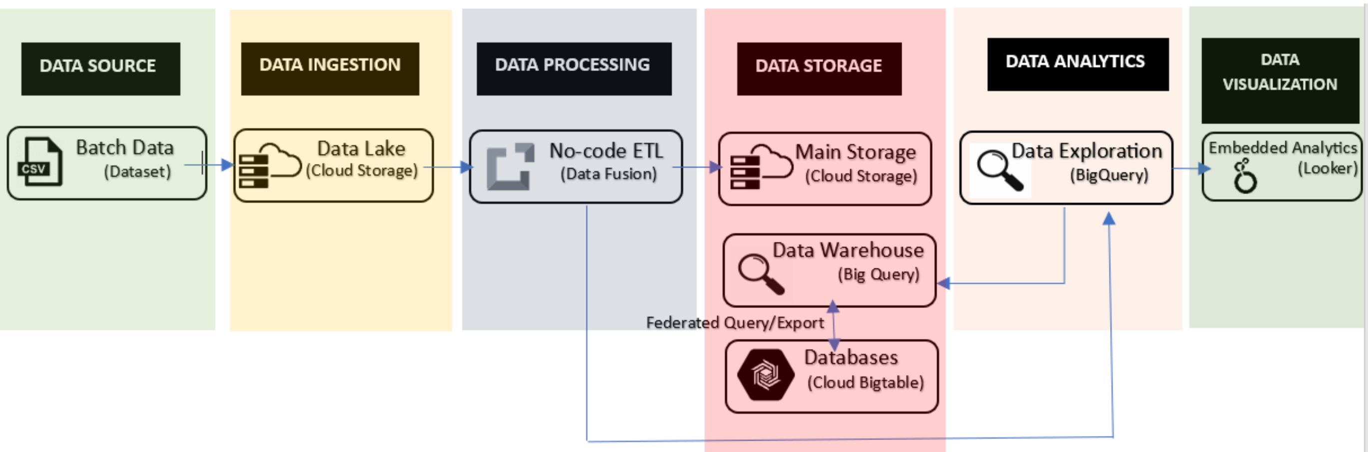


Figure 1: Overall system framework for cloud-based Impact of Education Expenditure on GDP Growth Analysis

3.1 Project Framework

In this project, batch data from World Bank dataset used to analyse relationships between relationship between GDP growth and education spending. The blue line represents the data flow of the batch data. In this section, five layers of data analytics cycle will be discussed, and overview of the framework will be discussed in the next subsection.

3.2 Data Life Cycle Processes

3.2.1 Data Ingestion

Data ingestion is the process of gathering and preparing data for analysis, involving the extraction of information from diverse sources like databases, logs, APIs, and file systems. This task employs various tools such as ETL (Extract, Transfer & Load) tools.In our project, we plan to upload the World Bank Data on Countries dataset from Kaggle to Cloud Storage, using it as the data ingestion tool for batch data. The data stored in Cloud Storage can then be accessed by other tools. Specifically, Data Fusion will extract the dataset for further processing in data pipelines and analysis.

3.2.2 Data Processing

Data processing is a crucial phase in the data journey, ensuring that information is accurate, consistent, and aligned with its intended use. This involves various activities like cleaning, filtering, aggregating, and transforming data to enhance its quality. The ultimate aim of data processing is to ready the data for subsequent analysis or visualization, extracting valuable insights. In our project, we're crafting a framework pipeline in Data Fusion. This tool takes batch data from Cloud Storage, previously stored, and processes it by cleaning and transforming the information. The result is refined data, neatly organized and directed to the specified output.

3.2.3 Data Storage

Storing data serves the purpose of creating a reliable, secure, and easily accessible repository for future use. Once we've gathered the necessary data, it's crucial to safeguard it with security measures and encryption keys, ensuring the protection of user identities and enabling backup and recovery. Our processed data stored in Cloud Storage and BigQuery. Cloud Storage acts as a secure backup, while BigQuery serves as the storage hub for analytics. Additionally, we leverage Cloud Bigtable for its efficient read and write throughput, low latency, and excellent performance in scanning massive amounts of data. This makes it particularly adept at quickly retrieving records, enhancing overall efficiency.

3.2.4 Data Analytics

Data analysis occurs after data processing, serving as a crucial step in interpreting and identifying patterns, trends, or relationships. Our particular focus is on evaluating the influence of education expenditure on GDP growth over time, guiding decisions through data-driven insights. Notably, this project involves analysing data directly within Big Query, allowing for seamless exploratory and descriptive analyses. The outcomes of these analyses will come to life through visualization in Looker, offering a clear and comprehensible representation of the findings.

3.2.5 Data Visualization

Presenting information visually is all about creating clear and understandable representations, like charts and graphs, to convey data analysis outcomes in a way that anyone can grasp. The selection of visualization techniques depends on the nature of the data and the audience it's intended for. By transforming results into visual formats, communication becomes more impactful. In our project, Looker Studio acts as the space for visualizing the results of exploration and descriptive analytics. Visualizations are utilized to illustrate relationships between relationship between GDP growth and education spending.

3.3 Project Implemented Tools & Justification

3.3.1 Cloud Storage

Google Cloud Storage (GCS) is a scalable and long-lasting object storage solution for storing and retrieving data. It offers a secure and dependable method of ingesting data into the data lifecycle. GCS allows users to upload data to the Google Cloud Platform, after which any application service in the Google Cloud Platform can directly access the supplied data, considerably improving the researcher's comprehension of the data. This justifies that GCS supports large data sets and can handle a variety of data types, making it suitable for storing the World Bank Data on Countries dataset. Its integration with other Google Cloud services simplifies the data transfer process to subsequent layers. Storing data in Google Cloud Storage speeds up processing and allows academics to do follow-up data exploration study. Aside from that, it makes data available at any time and in any location. It also offers flexible serverless development for multicloud setups.

3.3.2 Cloud Data Fusion

Another feature supplied by Google Cloud Platform is data fusion. It is a fully managed, cloud-native business data integration solution that allows for the rapid development and management of data pipelines. The Cloud Data Fusion enables us to build scalable data integration solutions to clean, prepare, blend, transmit, and convert data without the need for infrastructure maintenance. This utility is useful and popular due to its compatibility with practically all applications. Google's Big Data tools, such as Cloud Storage, Big Query, and other tools required for this project. Another reason Data Fusion is employed is that it can be used without the requirement for coding, which is quite useful in completing the project. Aside from that, data fusion enables real-time data integration, allowing us to provide change streams into BigQuery for continuous analytics. Thus, the Data Fusion tool provides a scalable and economical solution for the World Bank dataset, which may require cleaning, transformation, and enrichment. Because of its no-code characteristics, it is usable by users with varied levels of technical ability.

3.3.3 Big Query

Google BigQuery is a fully managed, serverless commercial data warehouse that provides for scalable analysis and management of petabytes of data in seconds and minutes. It also includes BigQuery ML, which models using SQL syntax. The reason choosing Big Query is to aid us in the processing and analysis of large amounts of data. We may leverage Google Cloud PlatBigQuery is a fantastic solution because of its ability to democratize insights through a scalable and secure platform that includes machine learning features and can adapt to data of any size, from bytes to petabytes, with no operational overhead making it an ideal choice for the World Bank dataset. Furthermore, its ability to integrate with other GCP products utilized in this project, such as Looker, Data Fusion, and so on, is a benefit in selecting this tool. This makes it easy to transfer data for free to Data Fusion and other data integration technologies, which can successfully capture data and transfer it automatically, and it works both ways form features in BigQuery to integrate, transform, analyze, and implement visual data reports. Besides, BigQuery enables ML models to be trained in SQL rather than languages such as Python and Java at high speed due to less complexity from avoiding transporting and formatting vast amounts of data. Last but not least, BigQuery serves a dual role by also being utilized for data exploration at the data analytics layer because it can handle massive datasets quickly, analysts and data scientists can perform ad-hoc queries, discover patterns, and derive insights from the World Bank dataset. Hence, its SQL-like syntax, it is understandable to users who are experienced with traditional database querying.

3.3.4 PowerBI & Looker Studio

Based on data visualization layer there are two tools can be used. First, PowerBI is a sophisticated corporate analytics application that smoothly connects with a variety of data sources, including cloud-based solutions. PowerBI may connect to Google Cloud Storage or BigQuery to build dynamic and meaningful visuals based on the World Bank dataset for the data visualization layer. Its easy-to-use interface and comprehensive collection of visualization choices make it ideal for building engaging dashboards and reports. Second, Looker Studio is another component of GCP that focuses on visualization. We can directly visualize the data in BigQuery using Looker Studio, which is faster and more comfortable, and the visualization impact of Looker Studio is also clean and simple. Furthermore, Looker studio includes a data connection that serves as a pipeline connecting the Looker tool to its underlying data. Aside from that, both PowerBI and Looker Studio has a highly user-friendly web interface, which helped our project because we had no prior expertise applying the program. This facilitates data integration from several data sources into a single system, which was quite useful in this project.

3.4 Implementation of the proposed framework (Practical)

Google Cloud storage

A new bucket named '**bucket_countries**' was created in google cloud storage.
Our dataset '**Countries.csv** ' is uploaded to this bucket.

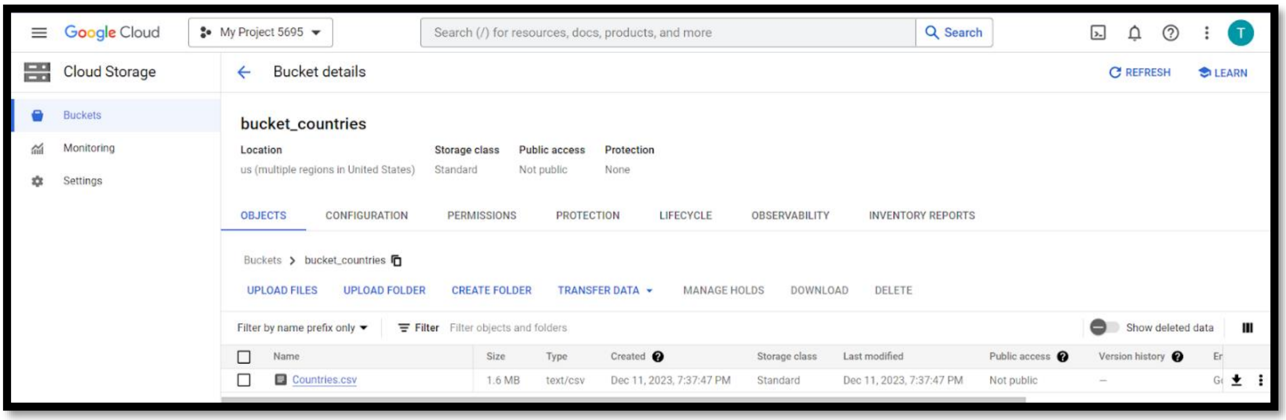


Figure 2: Uploading dataset in Google Cloud Storage

Data Fusion

Then, Data Fusion instance named '**countries**' was created as shown in below Figure.

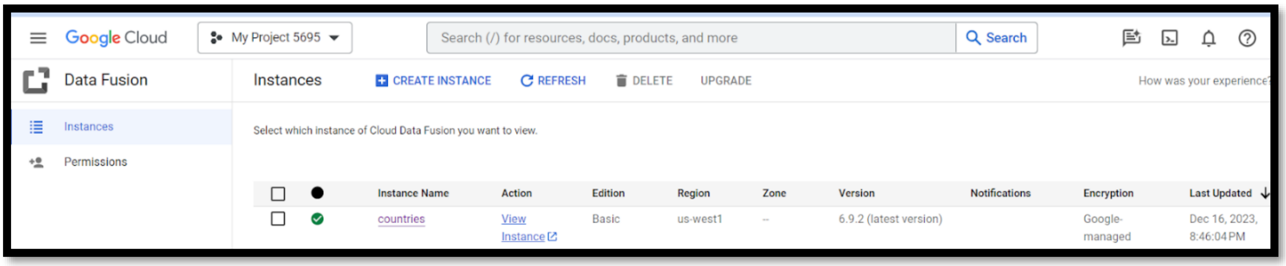


Figure 3: Creating instance in Data Fusion

Data Fusion- wrangler

Then, one of the cloud fusion functionality wrangler was used to preprocess the dataset.
We located the dataset from the connections (Google Cloud storage bucket) and parsed the file using the configurations as shown in figure below:

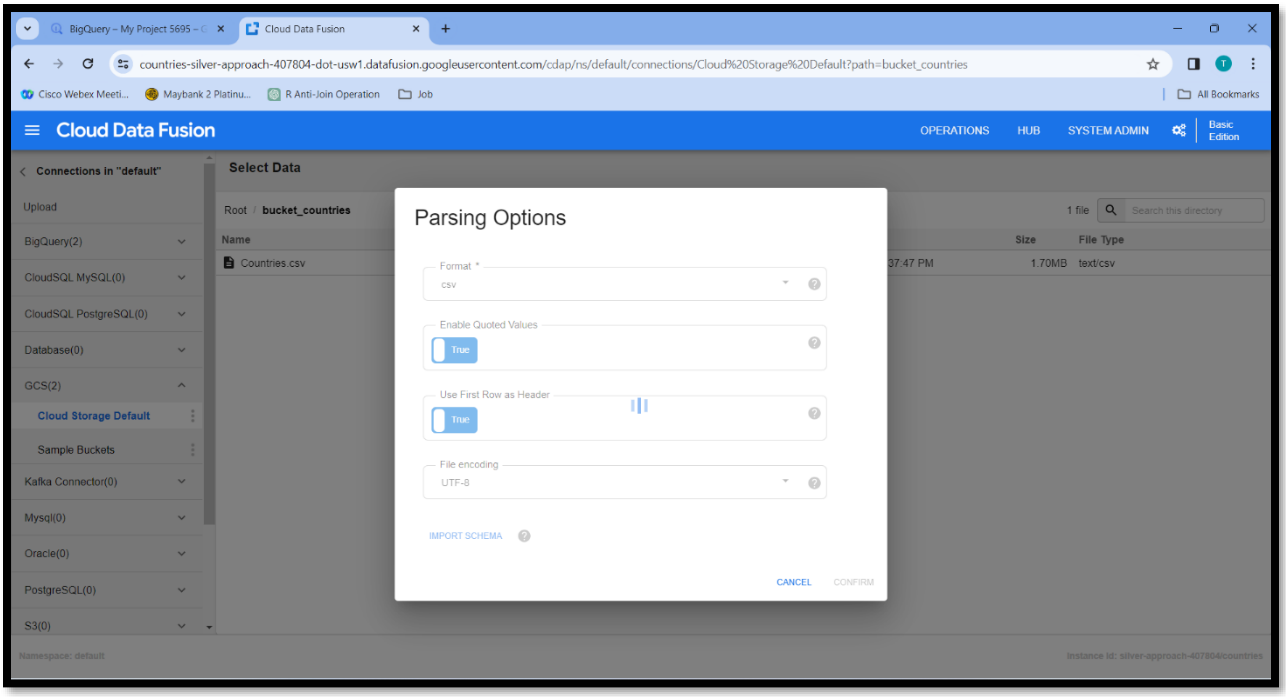


Figure 4: Parsing dataset in data fusion-wrangler

Our dataset has the first row as header and also has value which includes single quotes.
To cater the above conditions, we enabled these 2 configurations in wrangle:

1. 'Enable Quoted Values'
2. Use first row as header

Once the dataset is parsed in cloud data fusion, the interface showed preview of some rows of as below:

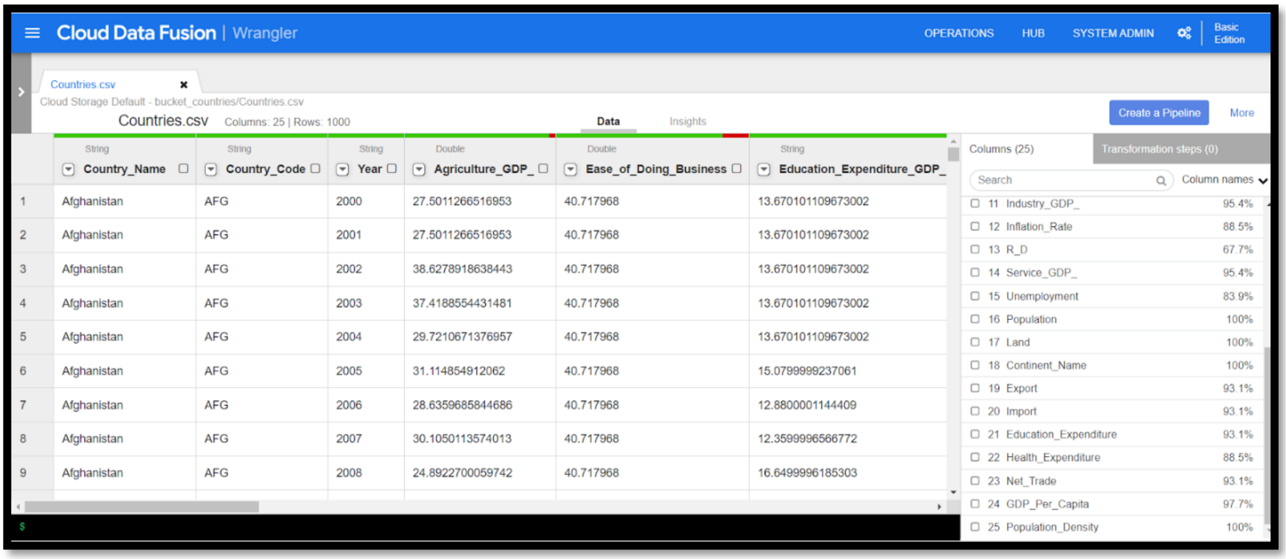


Figure 5: Preview of Dataset in Cloud data Fusion

1. Irrelevant attributes were removed from the dataset.
2. The data type of numeric columns were changed from string to decimal (4 d.p) to standardize.
3. The data type of population column was changed from string to integer.
4. The columns (numeric data type) having empty values were replaced with 0.
5. Some columns were renamed to make it more understandable.

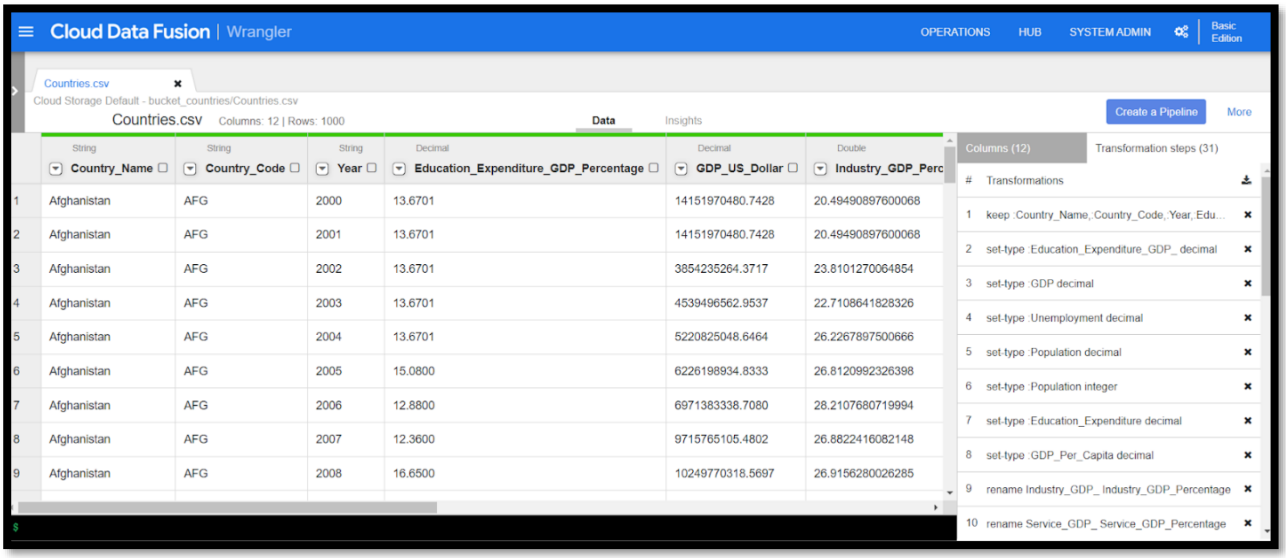


Figure 6: Data Fusion Transformation Steps 1 -10

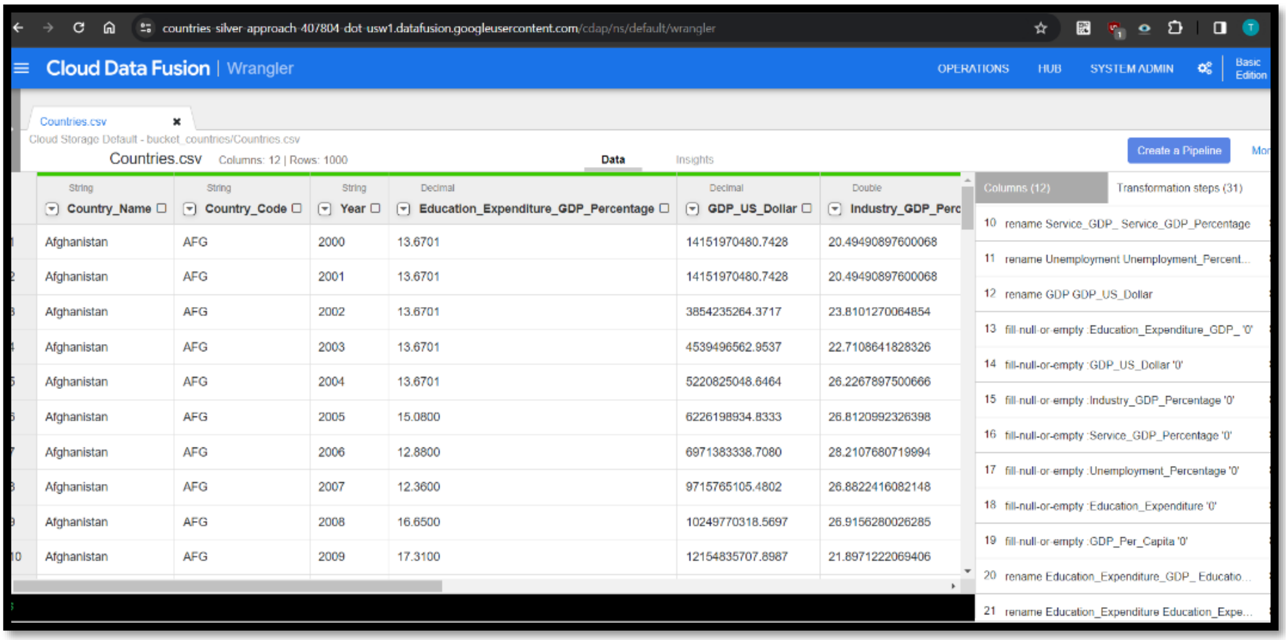


Figure 6.1: Data Fusion Transformation Steps 11 -20

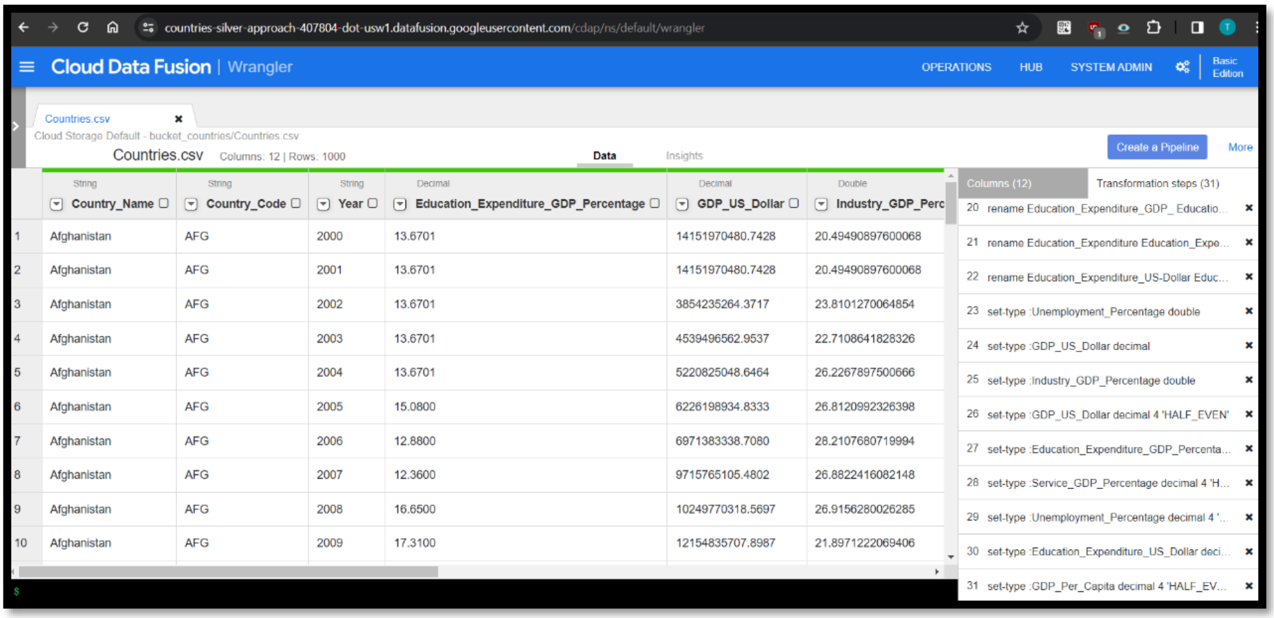


Figure 6.2: Data Fusion Transformation Steps 21 -31

Data Fusion- Pipeline creation

1. Pipeline 'country pipeline' was created and deployed.

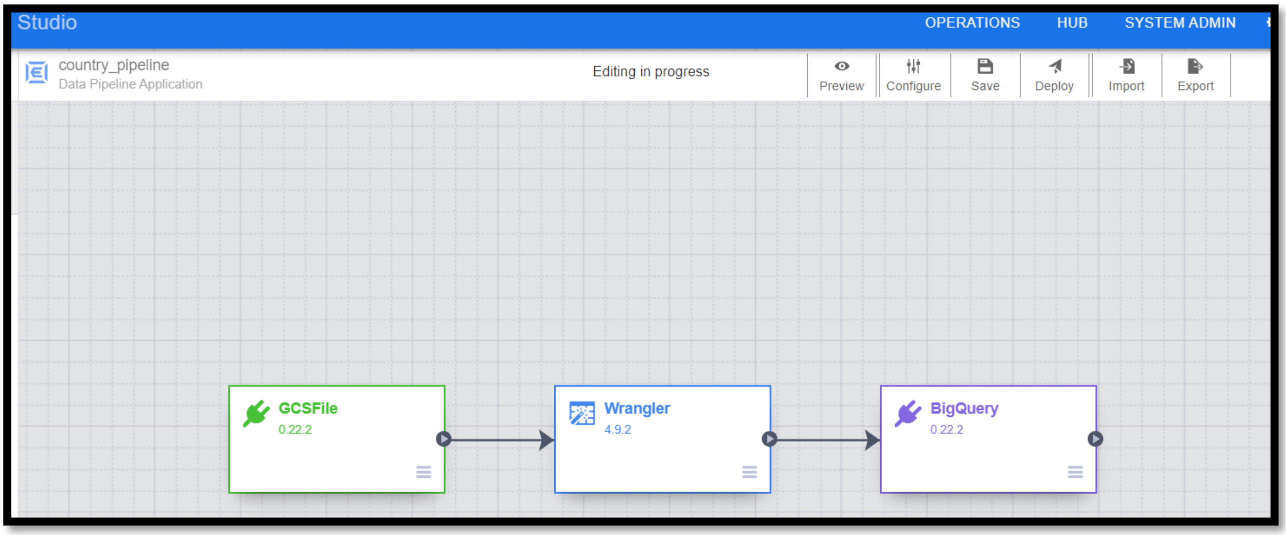


Figure 7: Pipeline in data fusion studio

2. Then, the pipeline was executed successfully

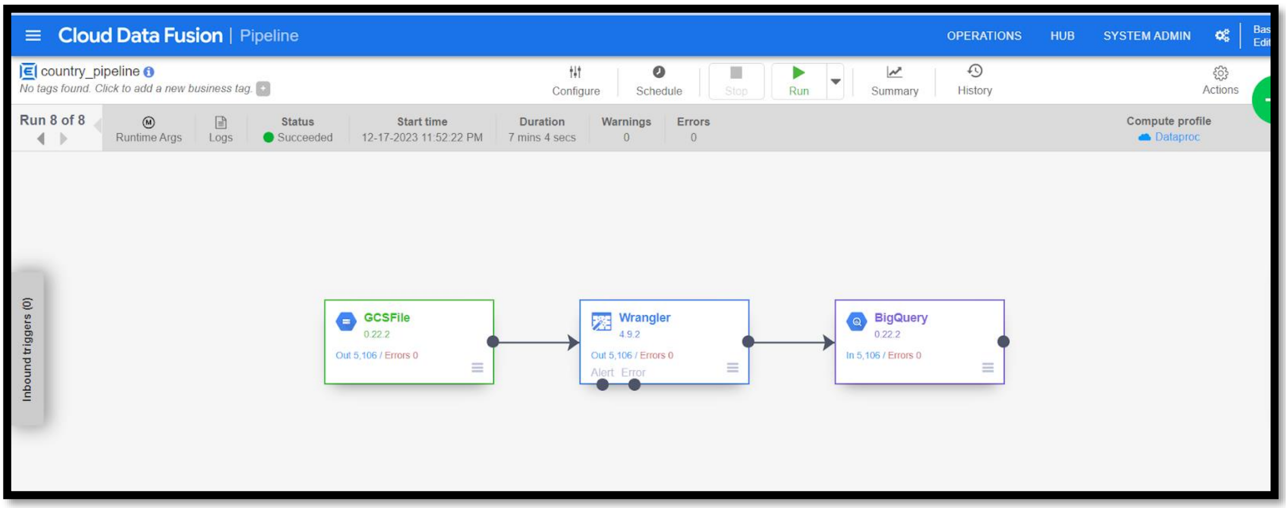
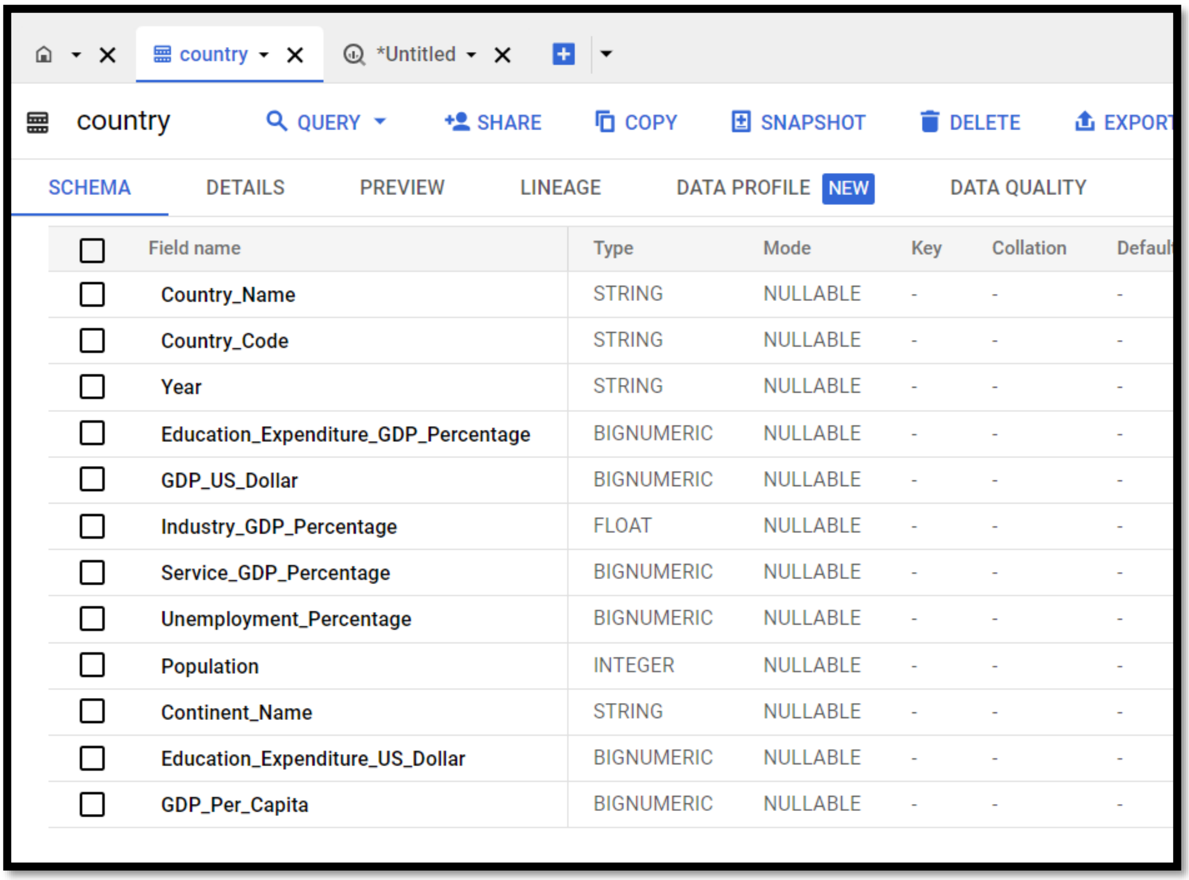


Figure 8: Successful Execution of Data Fusion Pipeline

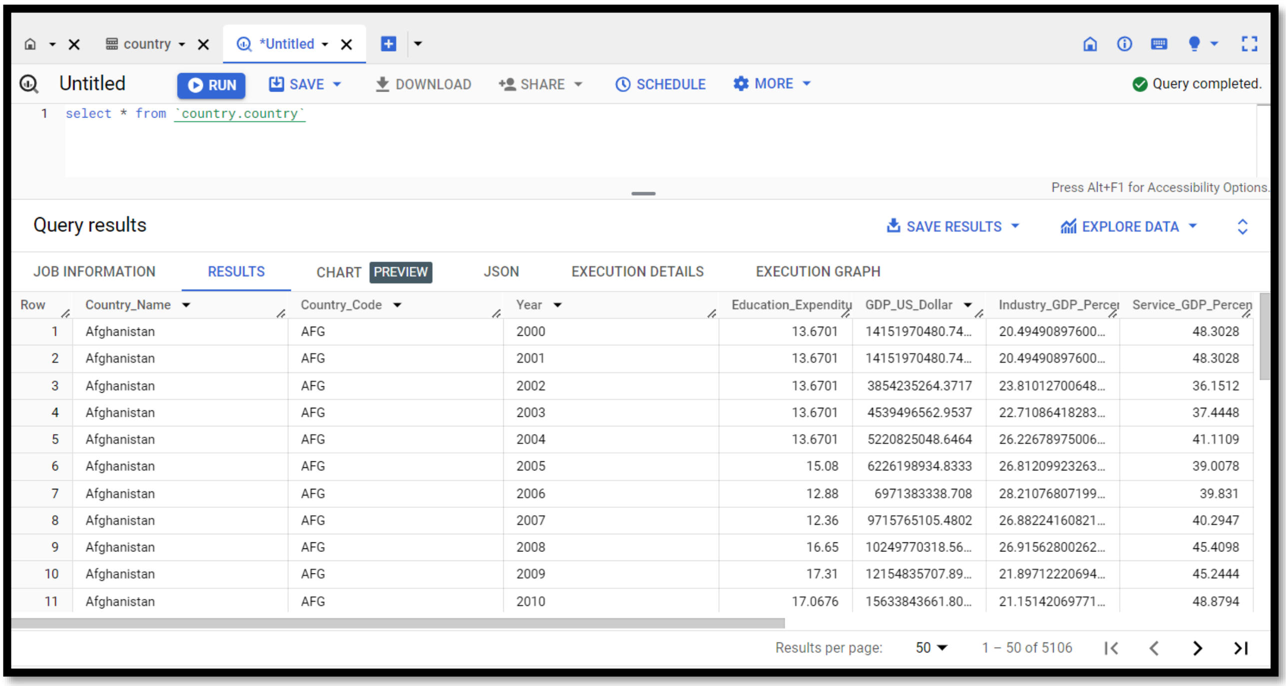
3. The cleaned data was processed successfully in Big Query as shown below



Field name	Type	Mode	Key	Collation	Default
Country_Name	STRING	NULLABLE	-	-	-
Country_Code	STRING	NULLABLE	-	-	-
Year	STRING	NULLABLE	-	-	-
Education_Expenditure_GDP_Percentage	BIGNUMERIC	NULLABLE	-	-	-
GDP_US_Dollar	BIGNUMERIC	NULLABLE	-	-	-
Industry_GDP_Percentage	FLOAT	NULLABLE	-	-	-
Service_GDP_Percentage	BIGNUMERIC	NULLABLE	-	-	-
Unemployment_Percentage	BIGNUMERIC	NULLABLE	-	-	-
Population	INTEGER	NULLABLE	-	-	-
Continent_Name	STRING	NULLABLE	-	-	-
Education_Expenditure_US_Dollar	BIGNUMERIC	NULLABLE	-	-	-
GDP_Per_Capita	BIGNUMERIC	NULLABLE	-	-	-

Figure 9: Country table schema in BigQuery

The select query was run to check the output from the table as below:



```
select * from `country.country`
```

Row	Country_Name	Country_Code	Year	Education_Expenditu	GDP_US_Dollar	Industry_GDP_Percen	Service_GDP_Percen
1	Afghanistan	AFG	2000	13.6701	14151970480.74...	20.49490897600...	48.3028
2	Afghanistan	AFG	2001	13.6701	14151970480.74...	20.49490897600...	48.3028
3	Afghanistan	AFG	2002	13.6701	3854235264.3717	23.81012700648...	36.1512
4	Afghanistan	AFG	2003	13.6701	4539496562.9537	22.71086418283...	37.4448
5	Afghanistan	AFG	2004	13.6701	5220825048.6464	26.22678975006...	41.1109
6	Afghanistan	AFG	2005	15.08	6226198934.8333	26.81209923263...	39.0078
7	Afghanistan	AFG	2006	12.88	6971383338.708	28.21076807199...	39.831
8	Afghanistan	AFG	2007	12.36	9715765105.4802	26.88224160821...	40.2947
9	Afghanistan	AFG	2008	16.65	10249770318.56...	26.91562800262...	45.4098
10	Afghanistan	AFG	2009	17.31	12154835707.89...	21.89712220694...	45.2444
11	Afghanistan	AFG	2010	17.0676	15633843661.80...	21.15142069771...	48.8794

Results per page: 50 1 – 50 of 5106

Figure 10: Select query results

The table in BigQuery was visualized as below in Looker Studio:

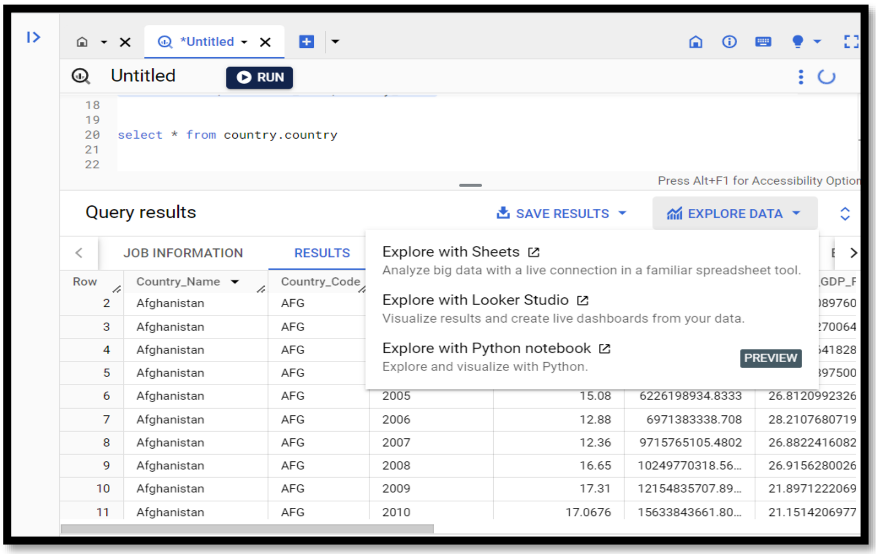


Figure 11: Explore Looker Studio

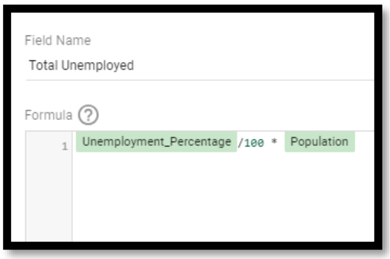
These results were further analysed in Looker Studio.

The columns Unemployment Percentage, Service GDP Percentage and Industry GDP Percentage showed the percentage.

But, for report visualization, numeric value was needed. So, to cater this, 3 new calculated fields were added in looker.

New Calculated Fields:

1. Total Unemployed



2. Industry_sector_contribution_US_Dollar

3. Service_sector_contribution_US_Dollar

Looker Report

The looker report consists of 1 table and 3 bar graphs.

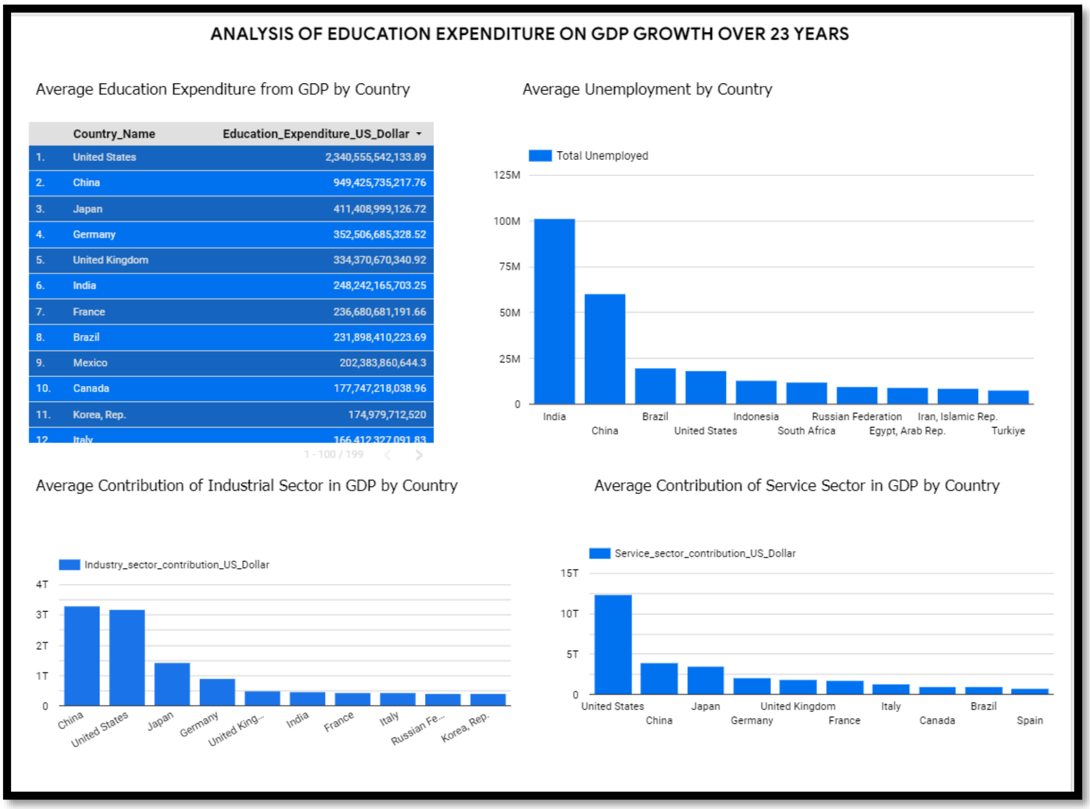
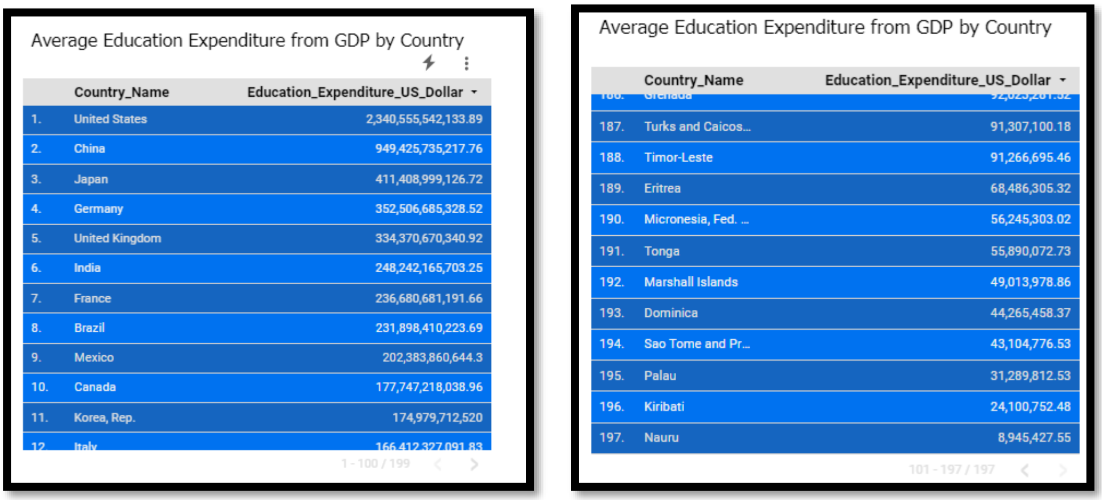


Figure 12: Report in Looker Studio

Each visual will be analysed.


Average Education Expenditure from GDP by Country

	Country_Name	Education_Expenditure_US_Dollar
1.	United States	2,340,555,542,133.89
2.	China	949,425,735,217.76
3.	Japan	411,408,999,126.72
4.	Germany	352,506,685,328.52
5.	United Kingdom	334,370,670,340.92
6.	India	248,242,165,703.25
7.	France	236,680,681,191.66
8.	Brazil	231,898,410,223.69
9.	Mexico	202,383,860,644.3
10.	Canada	177,747,218,038.96
11.	Korea, Rep.	174,979,712,520
12.	Italy	166,412,327,091.83

1 - 100 / 199

Average Education Expenditure from GDP by Country

	Country_Name	Education_Expenditure_US_Dollar
187.	Turks and Caicos...	91,307,100.18
188.	Timor-Leste	91,266,695.46
189.	Eritrea	68,486,305.32
190.	Micronesia, Fed. ...	56,245,303.02
191.	Tonga	55,890,072.73
192.	Marshall Islands	49,013,978.86
193.	Dominica	44,265,458.37
194.	Sao Tome and Pr...	43,104,776.53
195.	Palau	31,289,812.53
196.	Kiribati	24,100,752.48
197.	Nauru	8,945,427.55

101 - 197 / 197

Figure 13: Average Education Expenses by Country

Based on the graph above, United States has the highest education expenditure followed by China and Japan.
Whereas, Nauru (a country in Oceania) has the least average education expenditure among all the countries.

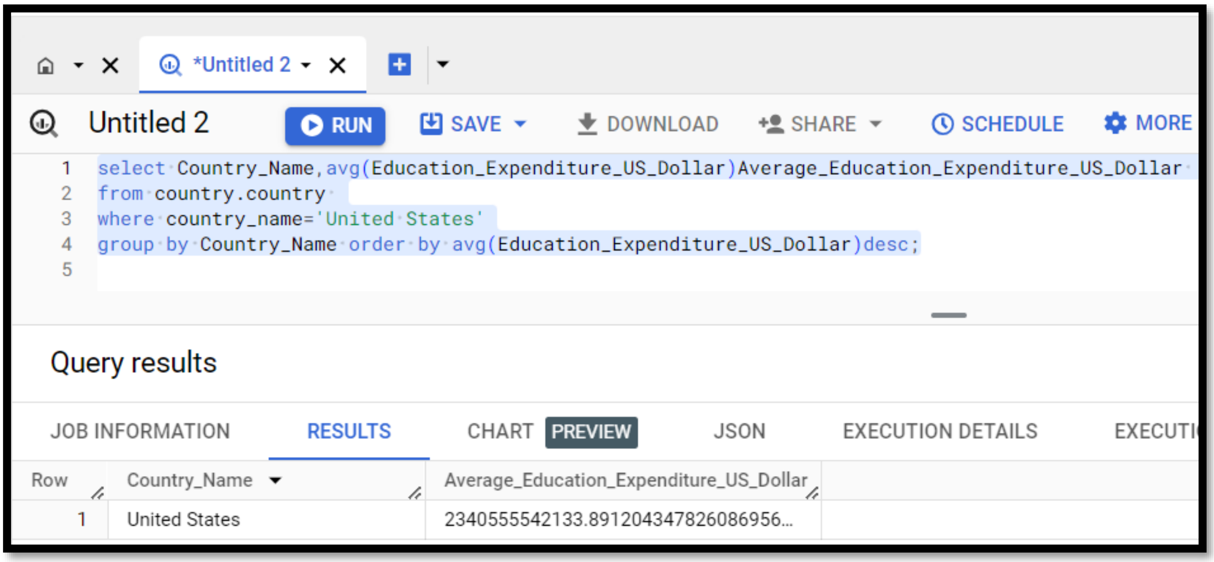


Figure 14: Average education expenses of United States in BigQuery

To confirm the accuracy of the average education expenditure displayed in the visual, the United States' results were queried from Big Query as shown above.

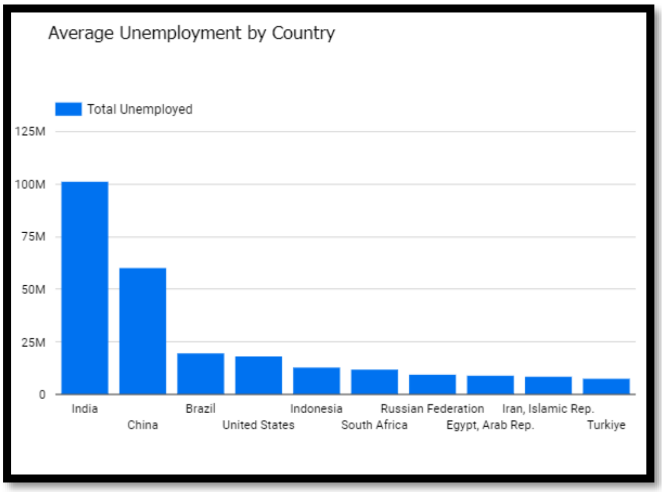


Figure 15: Average unemployment by Country

India, China,Brazil and United States seem to have the largest average unemployment among all the countries despite being in top 10 countries in education expenditure. This can be due to these countries having larger population compared to other countries and affected by other factors.

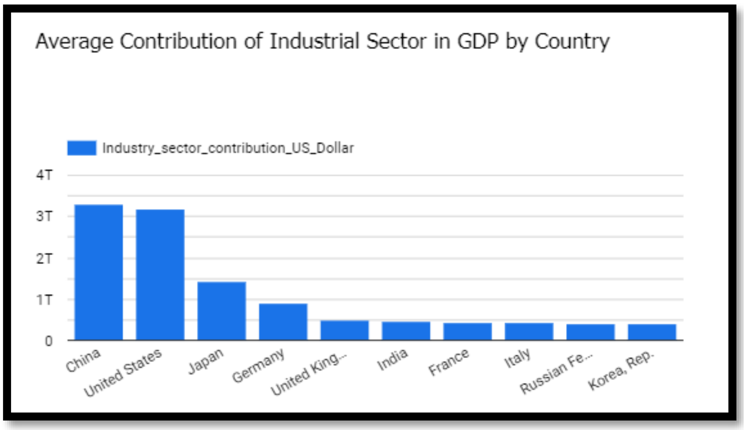


Figure 16: Average Contribution of Industrial Sector by Country

China tops in industry sector contribution followed by United States and Japan.
China and United States contributed about 3 trillion US dollar respectively over the 23 years from industry sector alone.

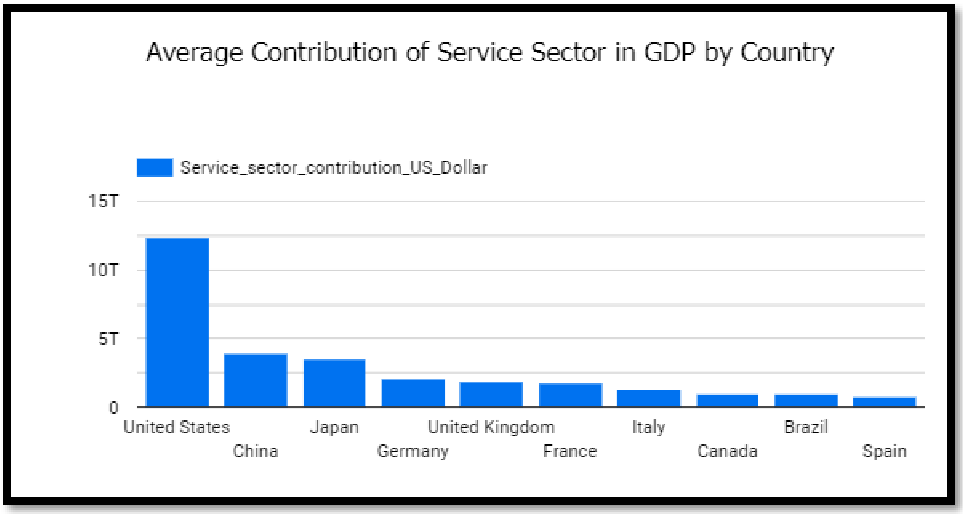


Figure 17: Average Contribution of Service Sector by Country

In terms of service sector, United States contributed around 10 trillion for its GDP. China on the other hand has contributed around 2.5 trillion from service sector. Interestingly, India is not in the top 10 countries for average contribution of service sector. This could be due to India's high unemployment rate.

Analysis highlights stark contrasts in education, employment, and economic contributions. The U.S., China, and Japan lead in education spending, while Nauru spends the least. Despite top education expenditures, India, China, Brazil, and the U.S. face significant unemployment, potentially influenced by large populations. China dominates industry contributions, matched by the U.S., both contributing $3 trillion over 23 years. In the service sector, the U.S. excels, contributing $10 trillion, whereas China contributes $2.5 trillion. Intriguingly, India, marked by a high unemployment rate, falls outside the top 10 contributors in the service sector.

3.5 Framework Performance

The implemented Data Fusion and Big Query will be assessed in this section.

3.5.1 Framework Performance for Data Fusion

Table 2: Dataset size is 561.9 KB

Test Case	Component	Record out per second	Min process time (one record)	Max process time (one record)	Standard deviation	Average processing time
Country	GCS properties	429689.472 ms	0 secs	1.455 ms	0.002 ms	0.002327 ms
Country	Wrangle	4343.757 ms	0 secs	360.679 ms	0.23 ms	0.230215 ms
Country	BigQuery	375330.785 ms	0 secs	2.774 ms	0.002 ms	0.002664 ms

Table 3: Dataset size is 1.6 MB

Test Case	Component	Record out per second	Min process time (one record)	Max process time (one record)	Standard deviation	Average processing time
Telco	GCS properties	435219.911 ms	0 secs	1.489 ms	0.002 ms	0.002298 ms
Telco	Wrangle	7493.506 ms	0 secs	159.169 ms	0.133 ms	0.133449 ms
Telco	BigQuery	307238.703 ms	0 secs	6.745 ms	0.003 ms	0.003255 ms

We tested the performance of data fusion by selecting two data sets of different sizes.
Figure 1 - Average processing time:
In the medium dataset (Country), Wrangle's average processing time is higher, but in the larger dataset (Telco), this time decreases, indicating that Wrangle is more efficient when processing larger datasets. GCS Properties and BigQuery both maintained low average processing times in both dataset sizes, indicating that these two components are relatively stable in processing efficiency.
Figure 2 - Record out per second:
The record output rates for both GCS properties and BigQuery are high for medium and larger datasets, indicating that they perform well in terms of data throughput. Although Wrangle's output rate improves on larger datasets, its output rate is still lower compared to GCS attributes and BigQuery.
Figure 3 - Max process time:
Wrangle's maximum processing time is higher on medium datasets, but improves significantly on larger datasets, which may mean that Wrangle's ability to process individual records increases as the dataset grows. Both GCS Properties and BigQuery have relatively low maximum processing

times for both dataset sizes, indicating that they are more stable in worst-case processing of individual records.

Figure 4 - Standard deviation:

Wrangle's standard deviation is higher on medium data sets but decreases on larger data sets, indicating improved consistency when dealing with larger data sets. Both GCS properties and BigQuery have lower standard deviations on medium and larger datasets, indicating that these two components have more consistent processing times and more predictable performance.

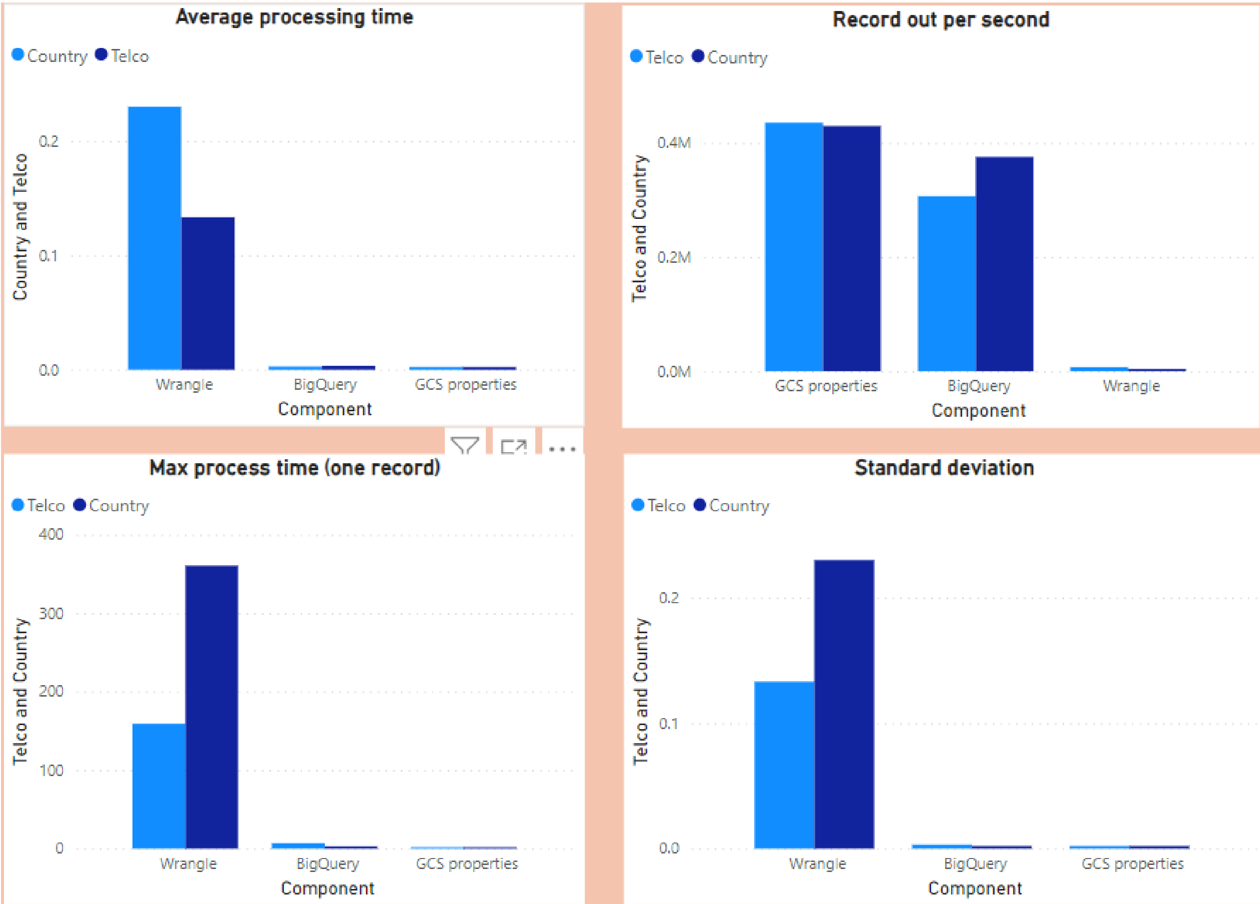
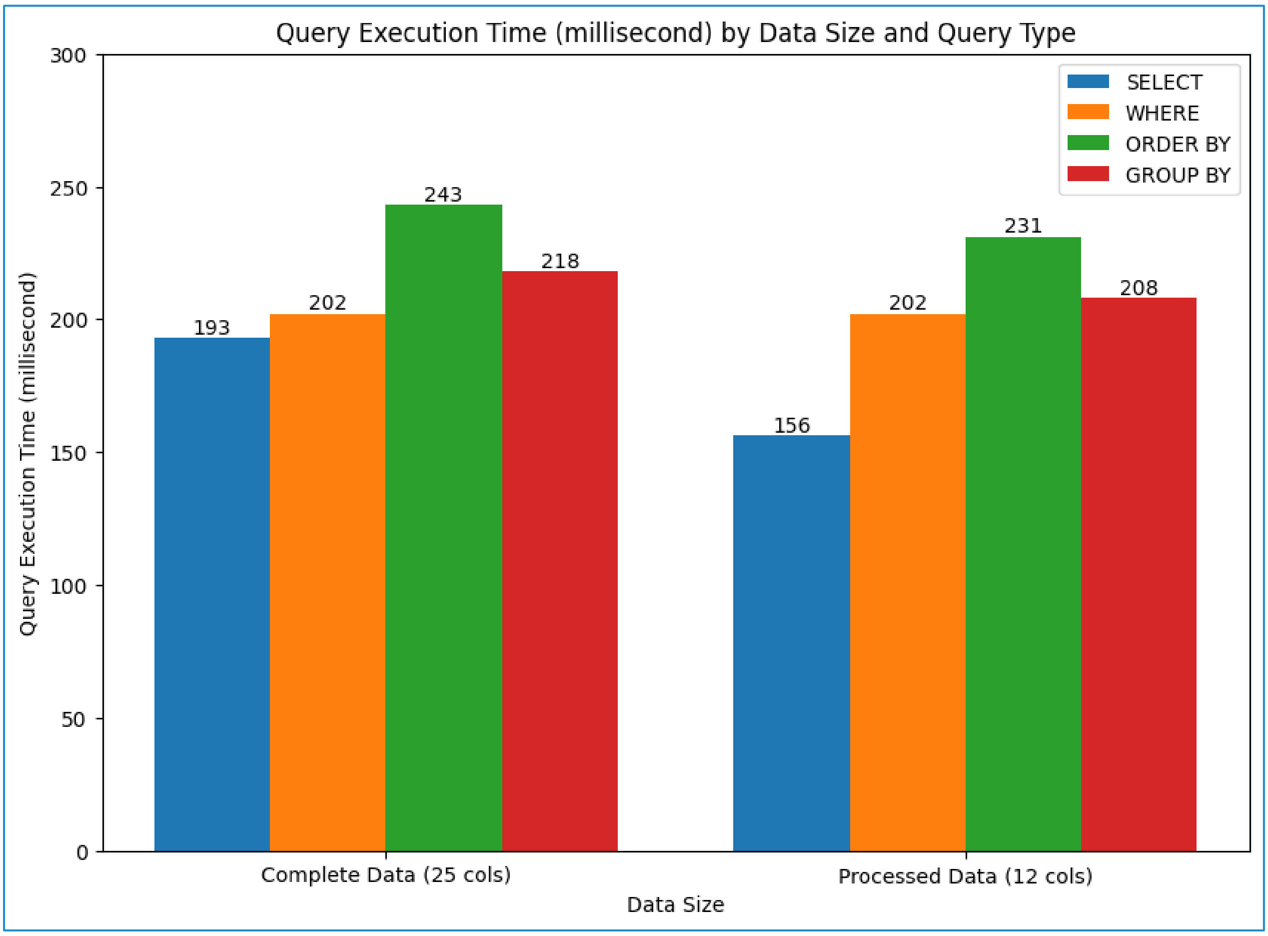


Figure: Performance Metrics of Data Fusion Components

In summary, GCS properties and BigQuery show high consistency for the processing of medium and larger data sets, indicating that they are suitable for processing data sets of various sizes. Dataset size has a significant impact on Wrangle's performance but has a smaller impact on GCS properties and BigQuery. This may be because Wrangle requires more computing resources and time when performing data cleaning and transformation operations, especially for complex or large data sets. Therefore, simplify the data model or preprocess the data to reduce the complexity of wrangle component transformation at runtime. This can effectively improve the performance of wrangle.

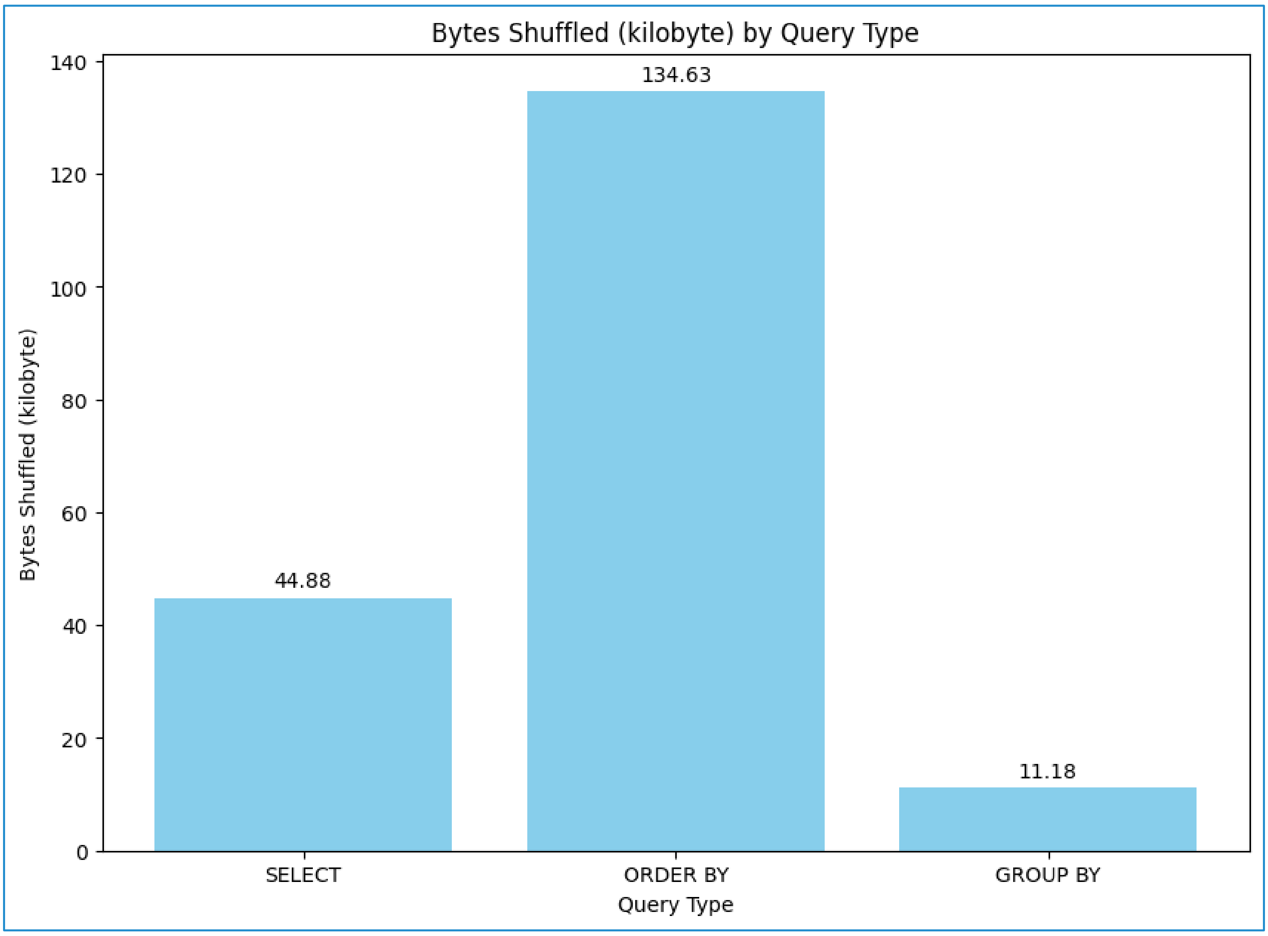
3.5.2 Framework Performance for Big Query

Graph 1: Query Execution Time (millisecond) by Data Size and Query Type



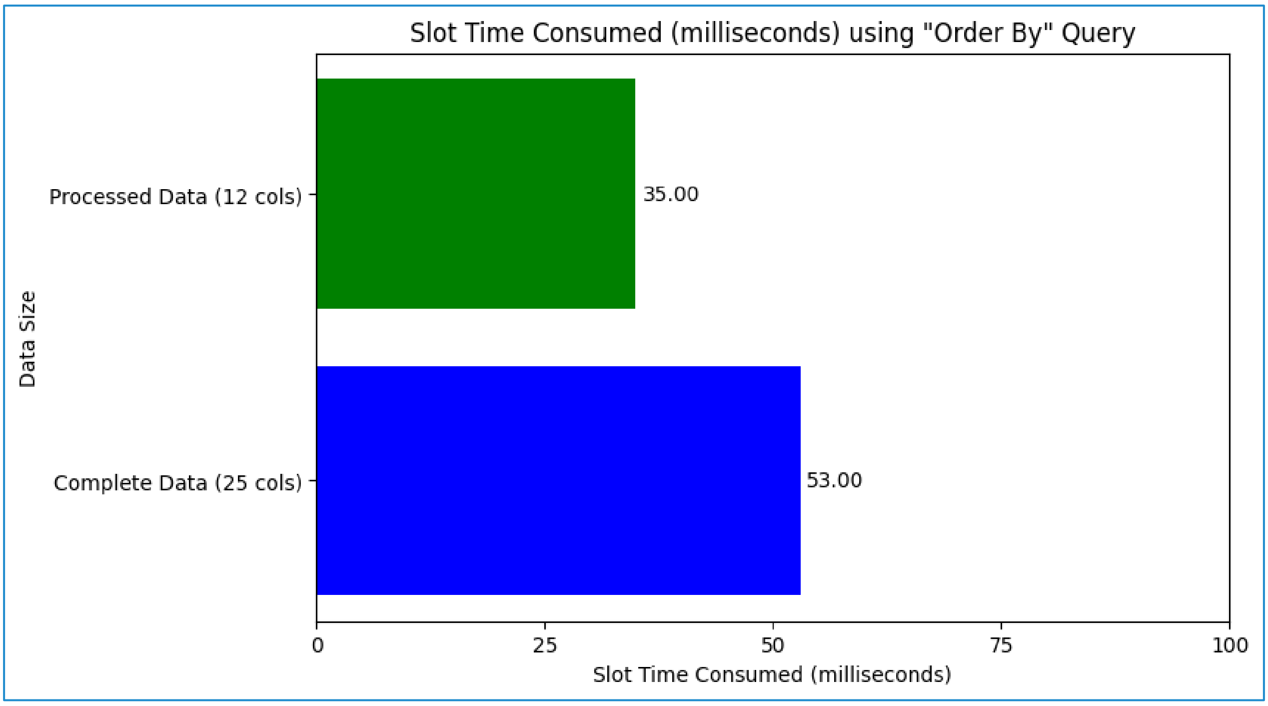
Query execution time measures the total time taken for a query to complete its execution. This clustered bar chart compares the execution time of different types of SQL queries on two datasets of different sizes. One of the datasets is the complete data obtained from Kaggle with 25 columns and the other is the processed dataset with 12 columns. The query types compared are SELECT, WHERE, ORDER BY, and GROUP BY. For both datasets, GROUP BY query took the longest time while SELECT query took the shortest time to execute. Similar trends are observed in both graphs, but the execution times are shorter in processed data, indicating that reducing the number of columns can improve performance.

Graph 2: Bytes Shuffled by Query Type



Bytes shuffled refers to the amount of data that needs to be moved between different worker nodes during the execution of the query. This bar chart illustrates the amount of data (in kilobytes) that is shuffled during the execution of different query types using processed dataset. Shuffling can be a resource-intensive operation in distributed databases, as it involves transferring data between different nodes. SELECT query retrieves specific rows which leads to data shuffling of 44.88kB. The ORDER BY query shuffled 134.63kB of data, as significant data movement is required to achieve the desired order. The GROUP BY query aggregates data based on specified criteria, resulting in a moderate amount of data shuffle (11.18kB).

Graph 3: Slot Time Consumed (milliseconds) using "Order By" Query

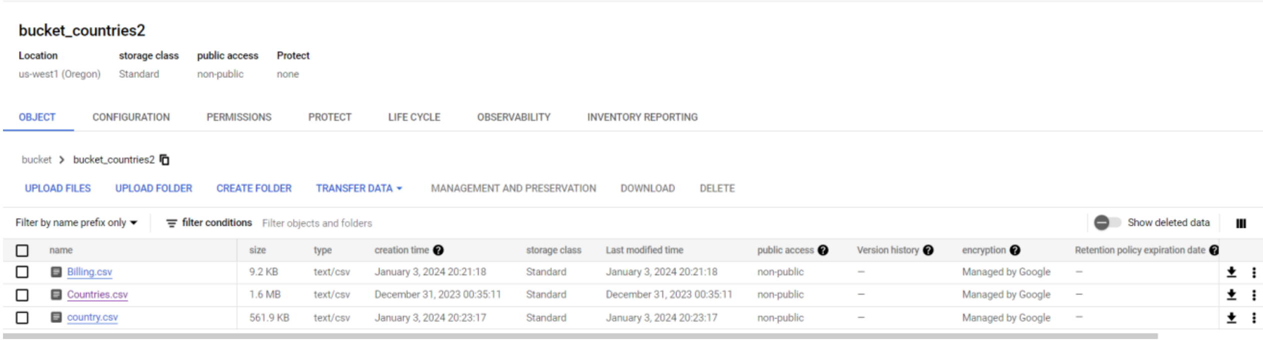


Slot time represents the duration for which a query occupies a slot, whether it is actively using the slot or waiting for resources to become available. It is useful for understanding the resource utilization in Big Query. This bar chart represents the slot time consumed when executing an ORDER BY query on two different sizes of datasets. From the diagram, the smaller dataset consumes significantly less slot time (35 milliseconds) than the larger dataset which consumes 53 milliseconds, implying higher efficiency or fewer computational resources needed.

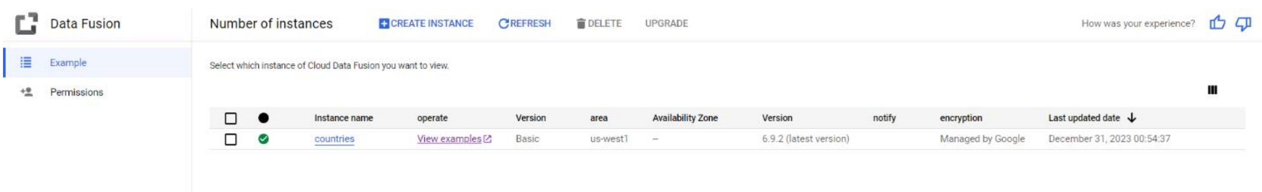
In summary, these three graphs suggest that data size and the type of SQL query significantly affect the performance of a database. Reducing data size and optimizing queries can lead to faster execution times, less data shuffle, and more efficient use of computational resources. These findings are useful for optimizing operations in distributed databases like Google Big Query, where resource utilization directly impacts cost and performance.

Appendix

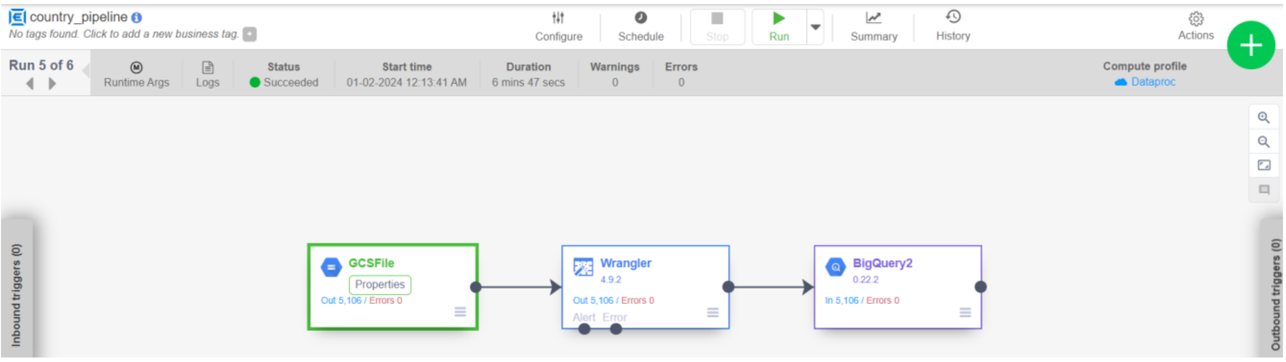
Create 2 Case bucket Test

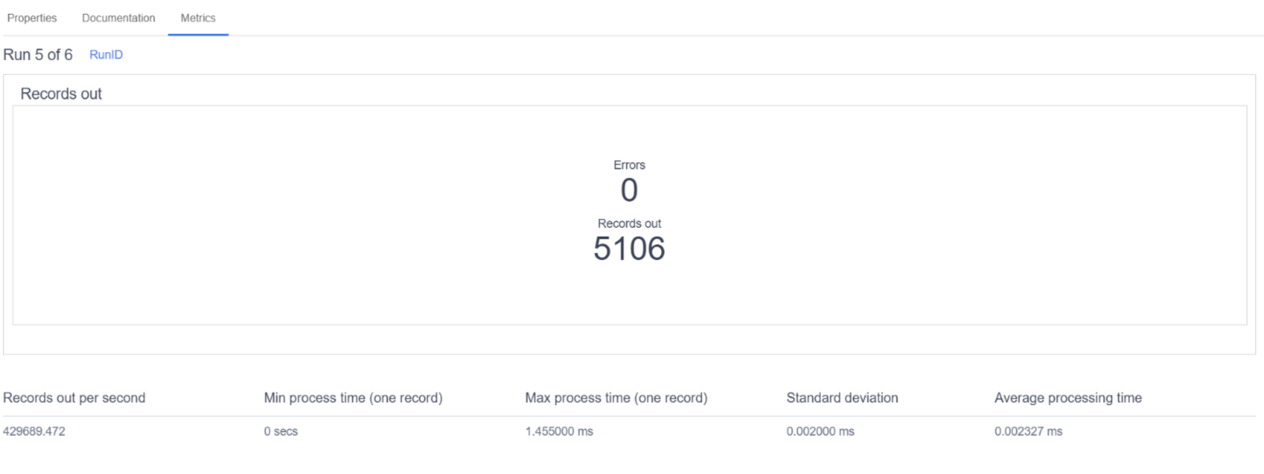


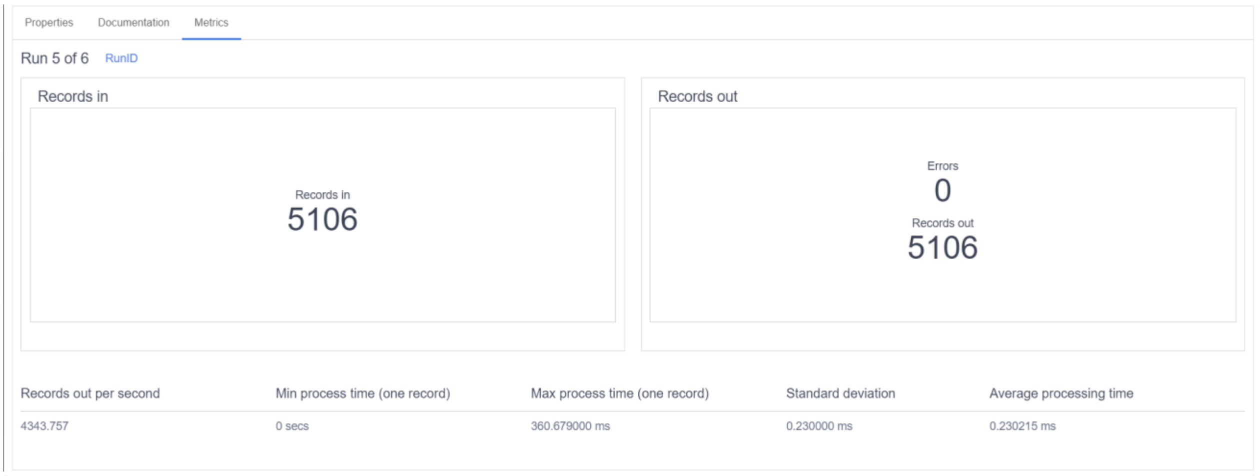
Create data fusion test instance

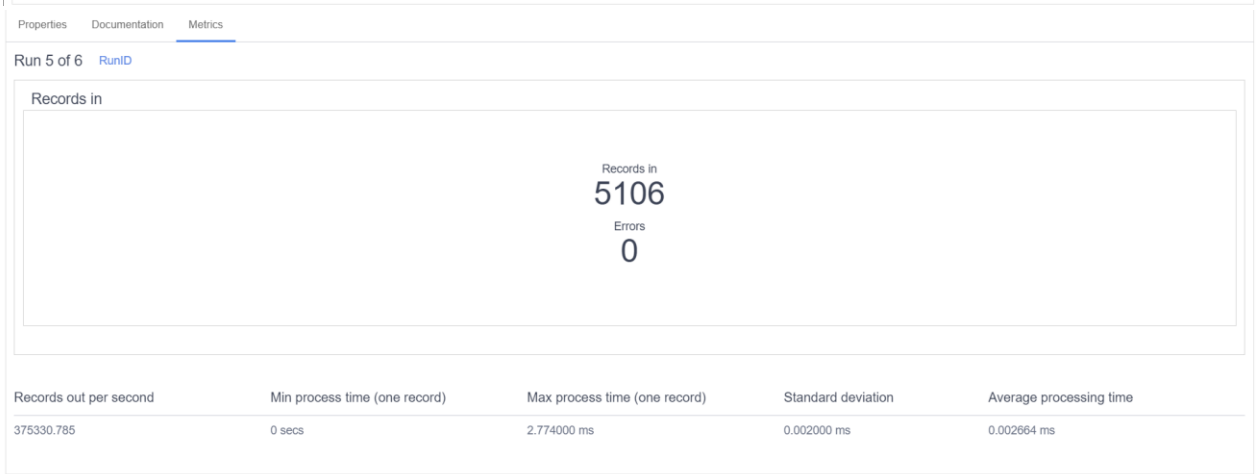


The figure below shows that the first test was successful.

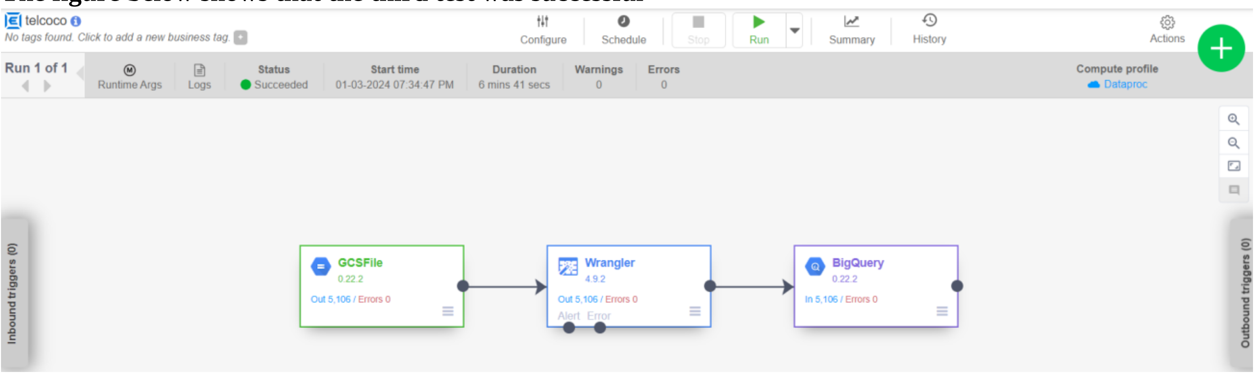


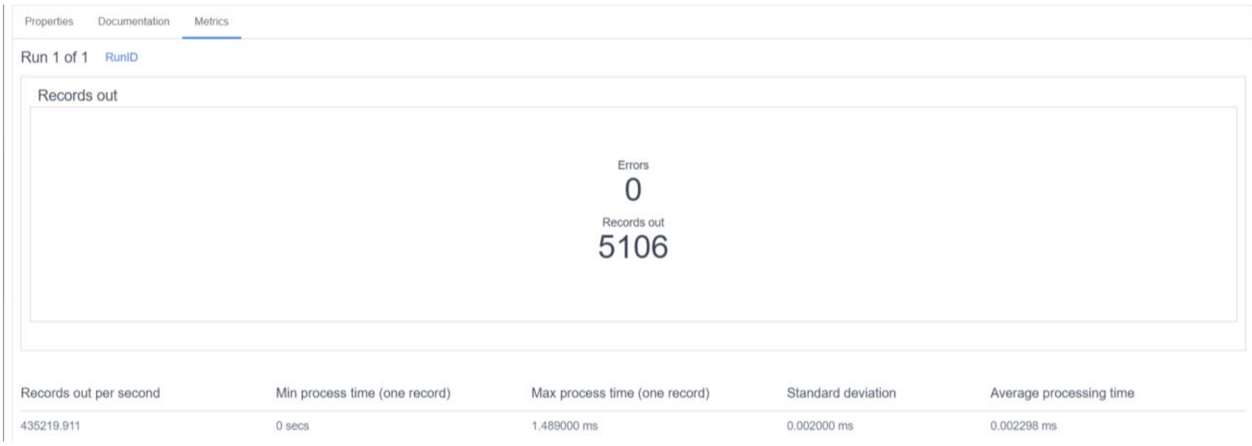


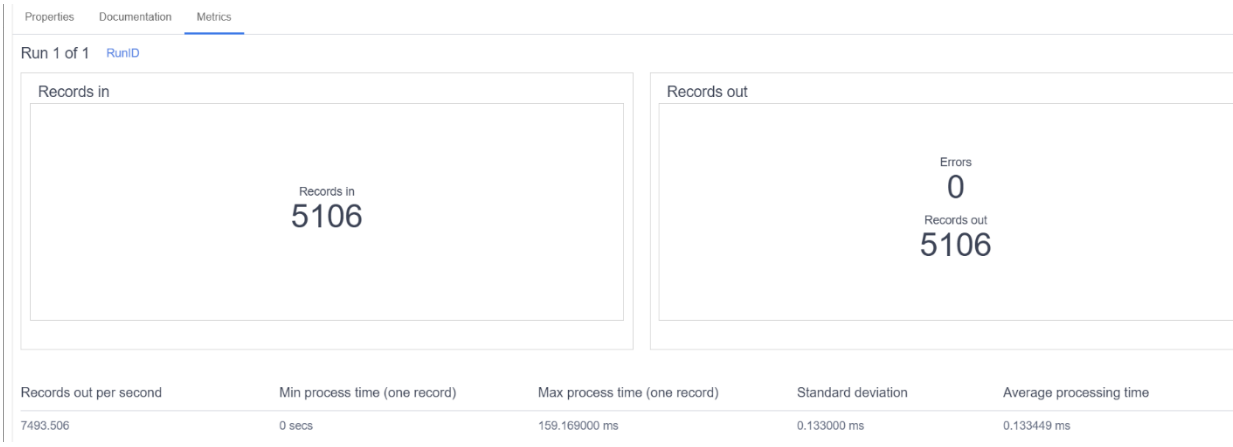




The figure below shows that the third test was successful




Properties
Documentation
Metrics
Run 1 of 1
RunID
Records out
Errors
0
Records out
5106
Records out per second
Min process time (one record)
Max process time (one record)
Standard deviation
Average processing time
435219.911
0 secs
1.489000 ms
0.002000 ms
0.002298 ms


Properties
Documentation
Metrics
Run 1 of 1
RunID
Records in
Records in
5106
Records out
Errors
0
Records out
5106
Records out per second
Min process time (one record)
Max process time (one record)
Standard deviation
Average processing time
7493.506
0 secs
159.169000 ms
0.133000 ms
0.133449 ms

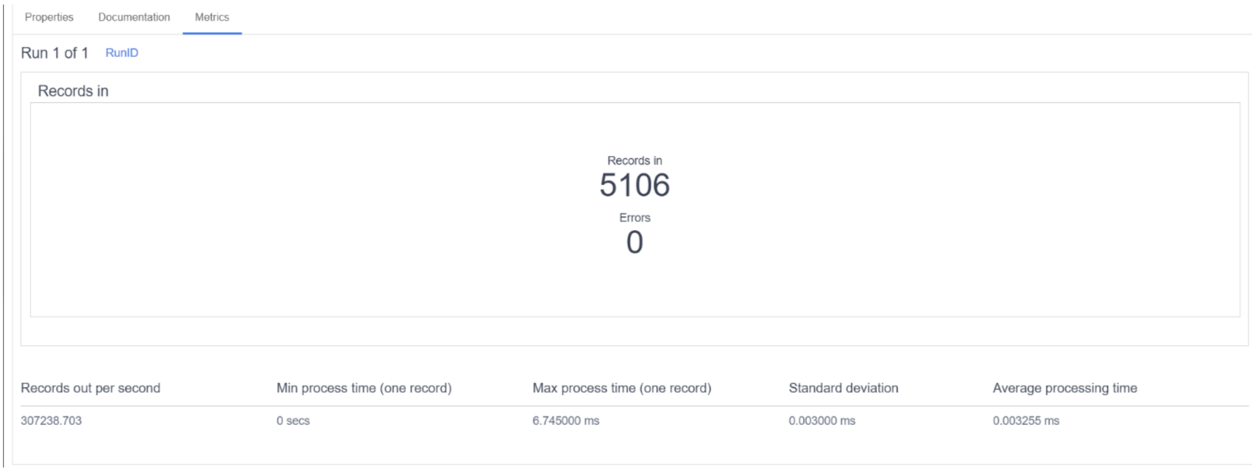

Properties
Documentation
Metrics
Run 1 of 1
RunID
Records in
Records in
5106
Errors
0
Records out per second
Min process time (one record)
Max process time (one record)
Standard deviation
Average processing time
307238.703
0 secs
6.745000 ms
0.003000 ms
0.003255 ms

Figure: Costs incurred during the use of the project

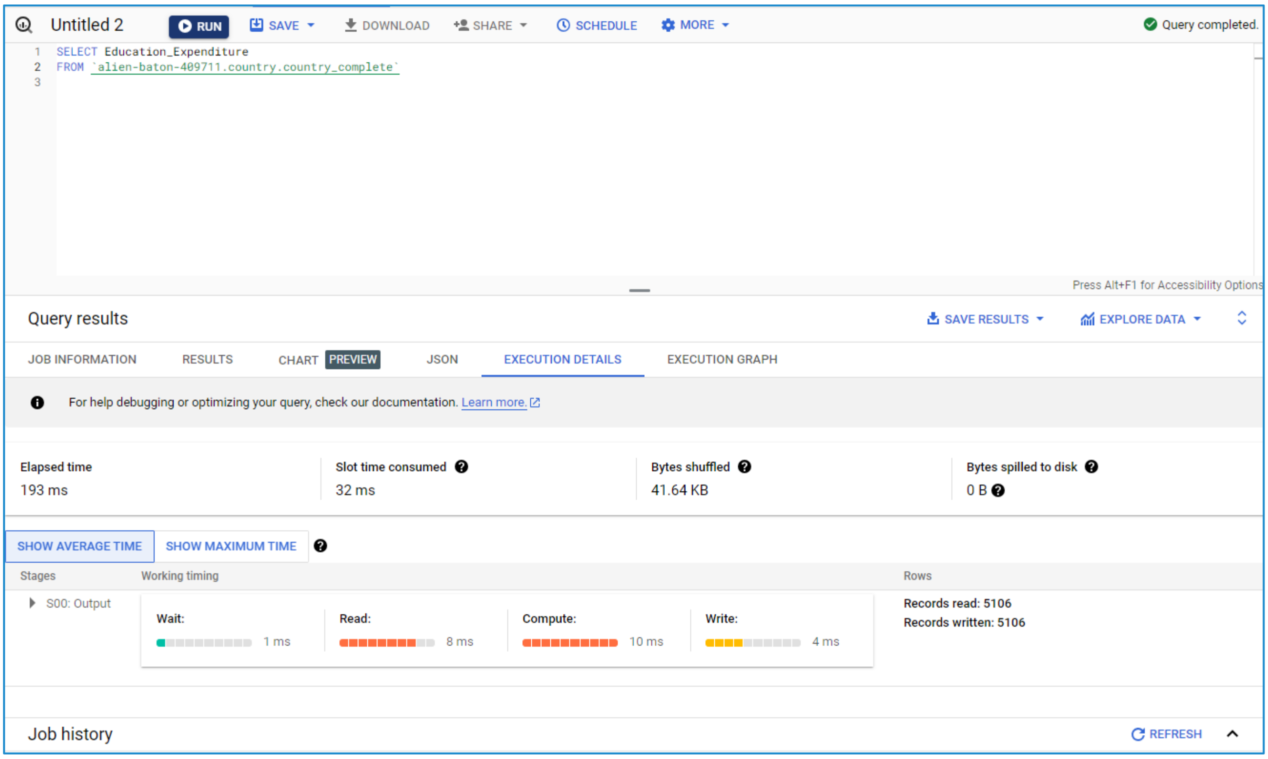


Figure: SELECT query with complete dataset (25 columns)

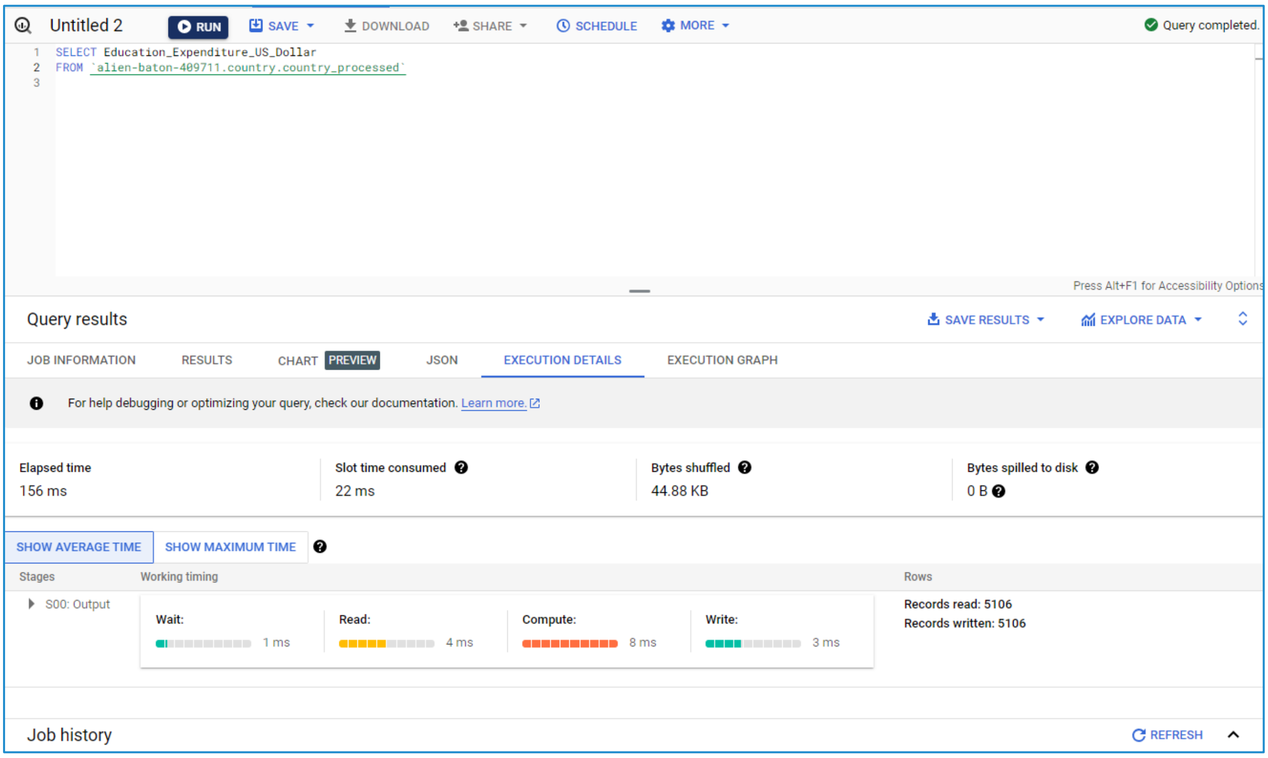


Figure: SELECT query with processed dataset (12 columns)

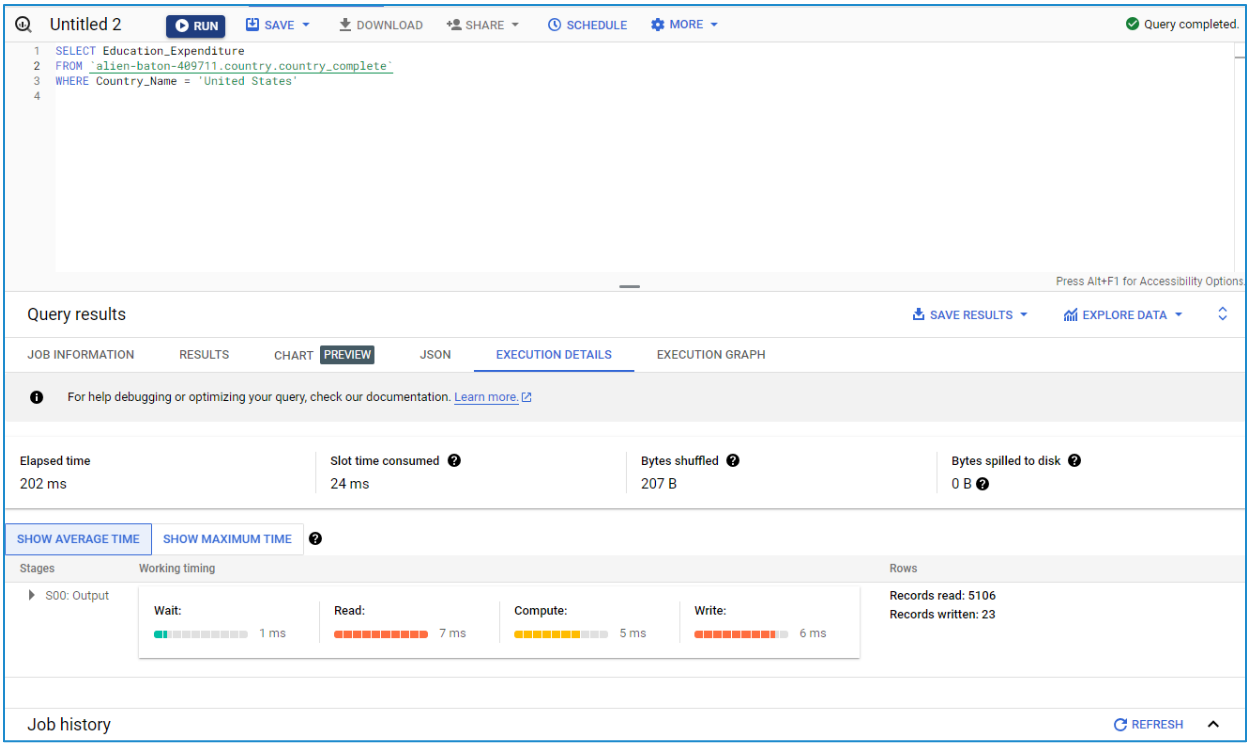


Figure: WHERE query with complete dataset (25 columns)

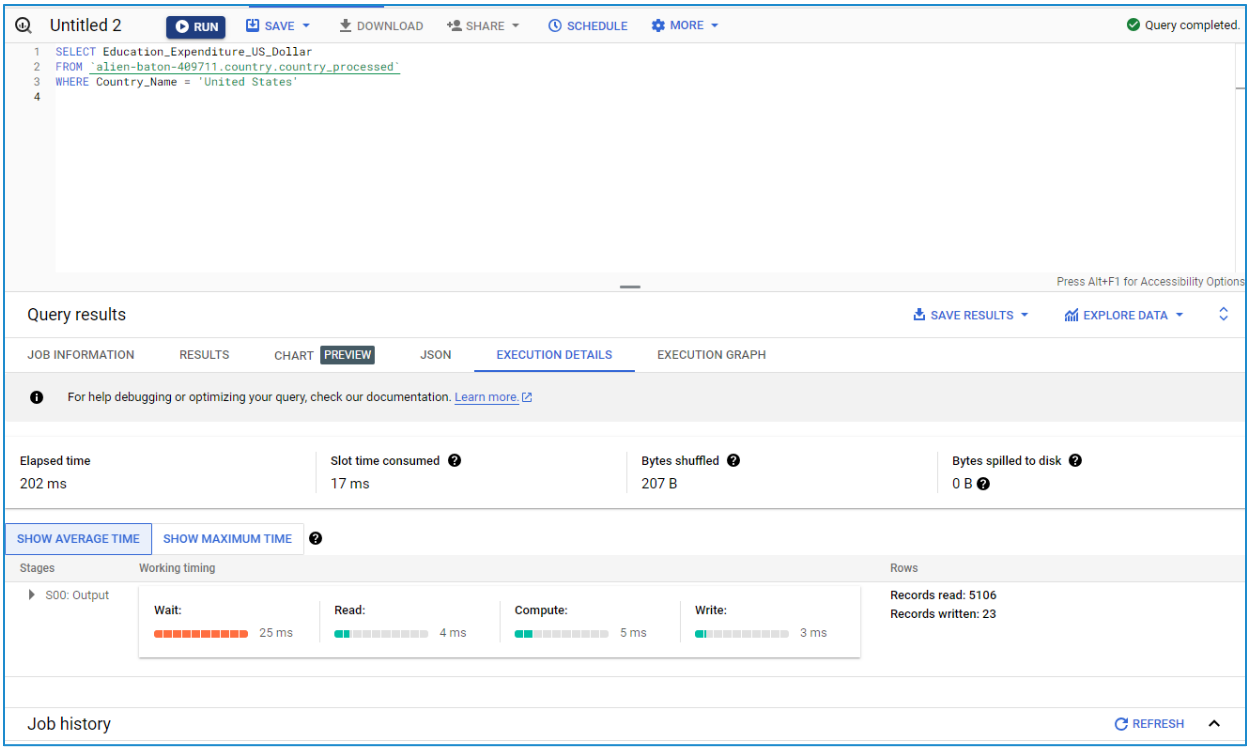


Figure: WHERE query with processed dataset (12 columns)

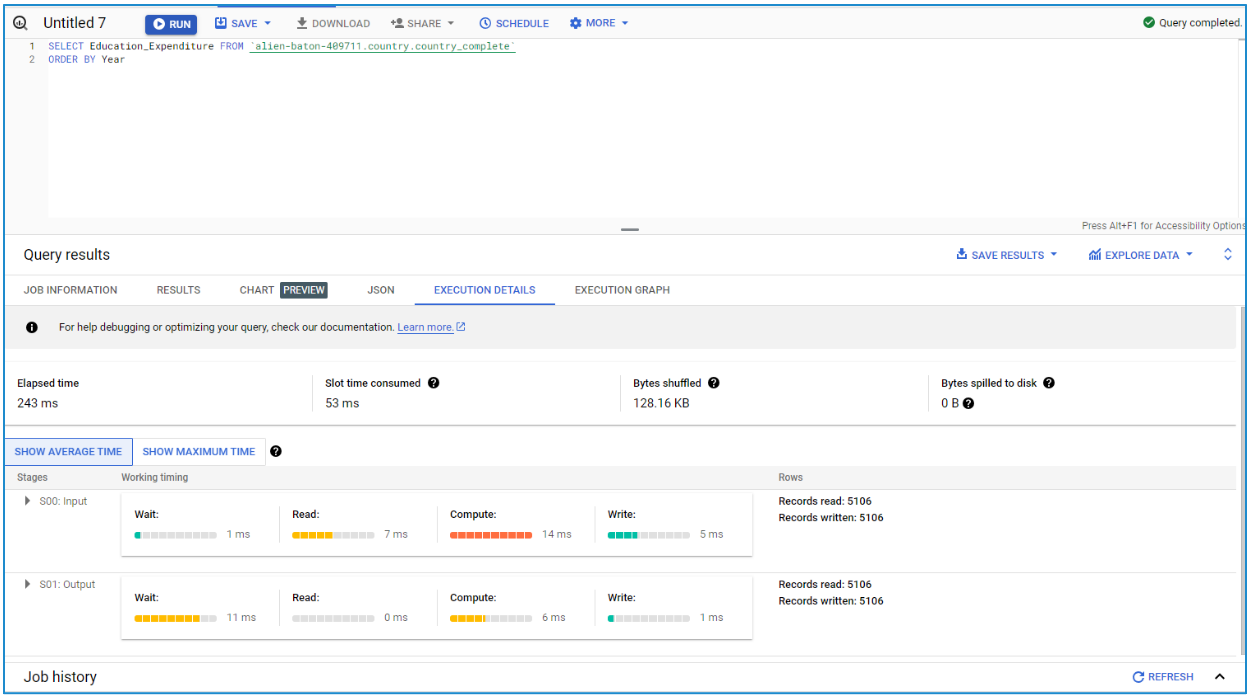


Figure: ORDER BY query with complete dataset (25 columns)

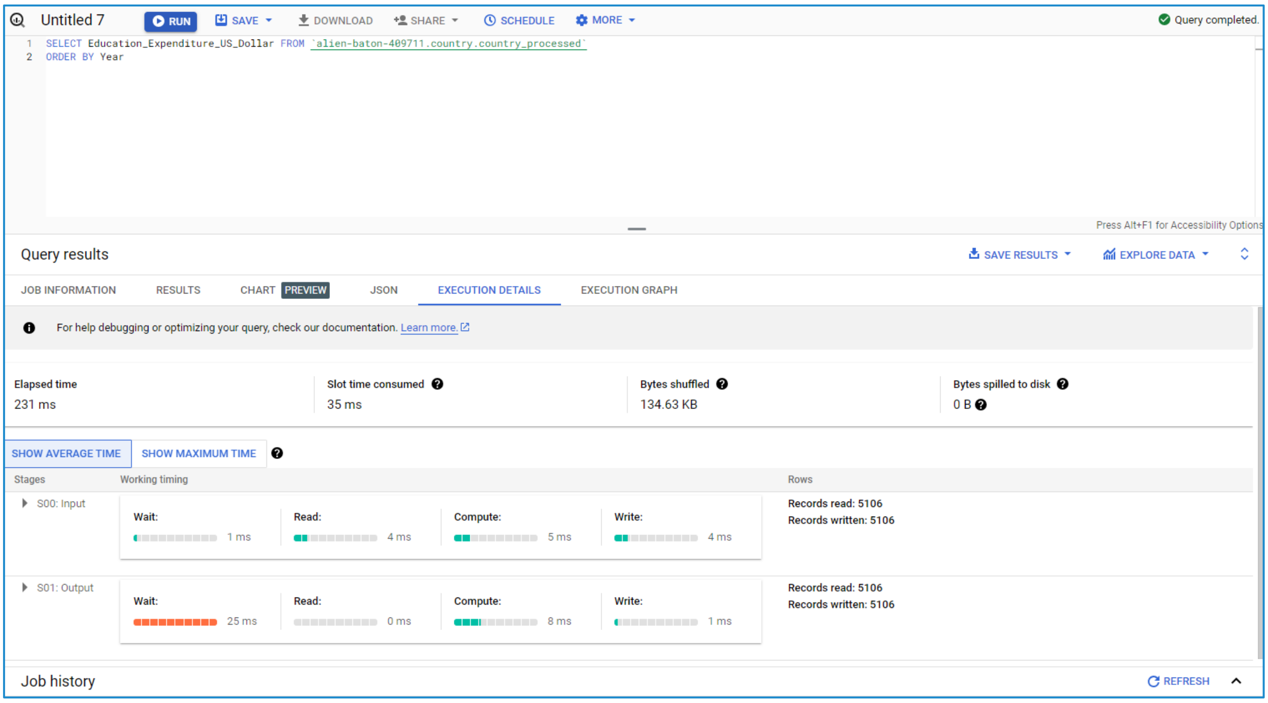


Figure: ORDER BY query with processed dataset (12 columns)

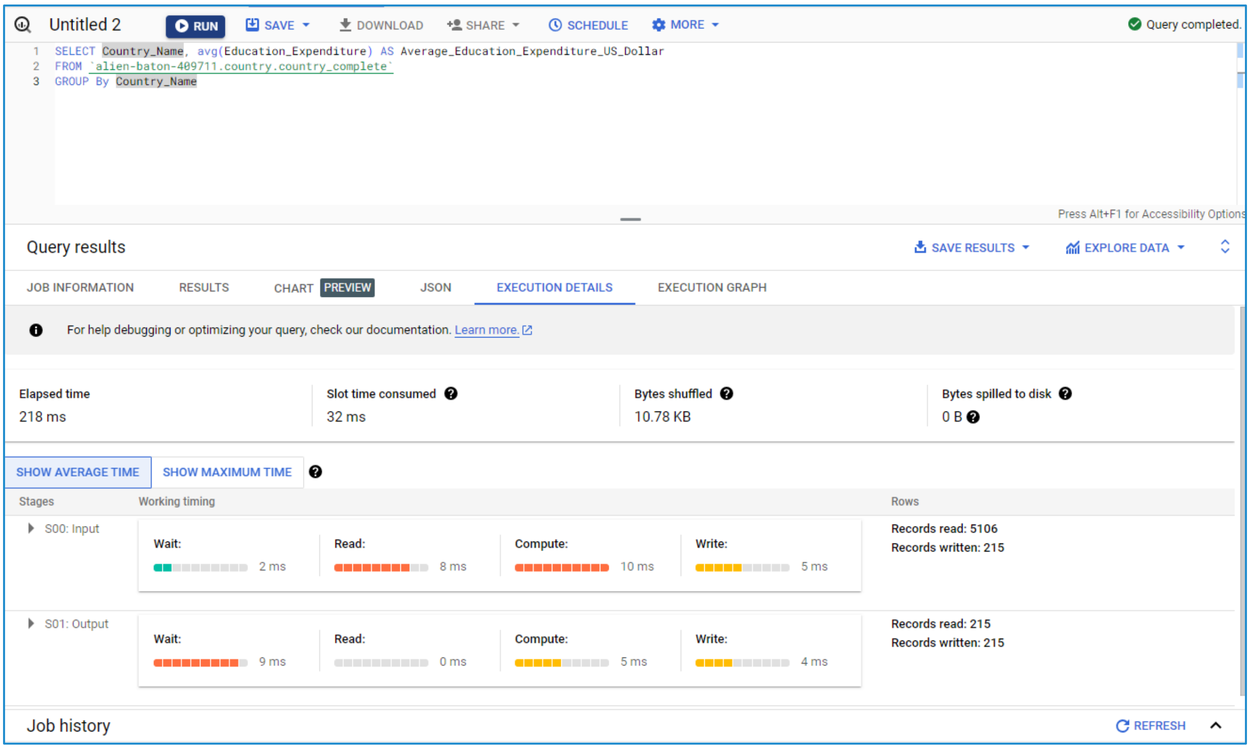


Figure: GROUP BY query with complete dataset (25 columns)

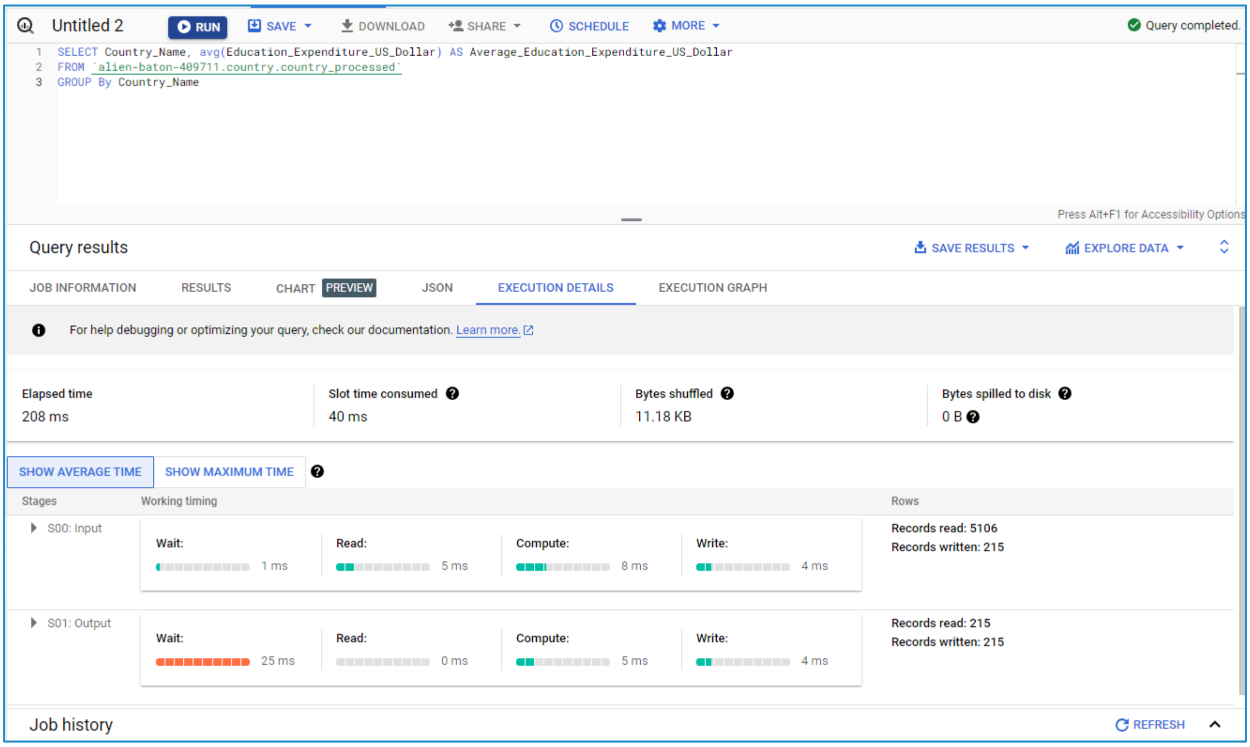


Figure: GROUP BY query with processed dataset (12 columns)