Building an ETL Pipeline and Visualization Dashboard Using GCP

1. Introduction

In today's data-driven world, businesses rely on effective pipelines to turn raw data into actionable insights. For this project, I designed an end-to-end ETL (Extract, Transform, Load) pipeline to process an e-commerce dataset. The goal was to integrate cloud-based tools for data transformation, storage, and visualization, making the dataset more usable and insightful for decision-making.

This report walks you through the process of applying concepts learned in this course, such as data transformation and distributed storage, and combining them with new tools like Google Data Fusion and Looker Studio. By integrating both familiar and advanced technologies, the project showcases a comprehensive pipeline to transform raw data into insightful visualizations.

2. Background

I observed how e-commerce has become a central part of our lives. From ordering groceries and medicines to clothing and electronics, people rely on online platforms for almost everything. This shift to online shopping generates vast amounts of data, offering a unique opportunity to analyze trends and understand consumer behavior. This sparked my interest in choosing an e-commerce dataset for this project.

The dataset provides insights into product details, sales performance, and customer preferences, making it a perfect candidate for building an ETL pipeline. Analyzing this type of data can reveal patterns and trends that are not immediately obvious, and designing a pipeline ensures the data is structured and ready for exploration.

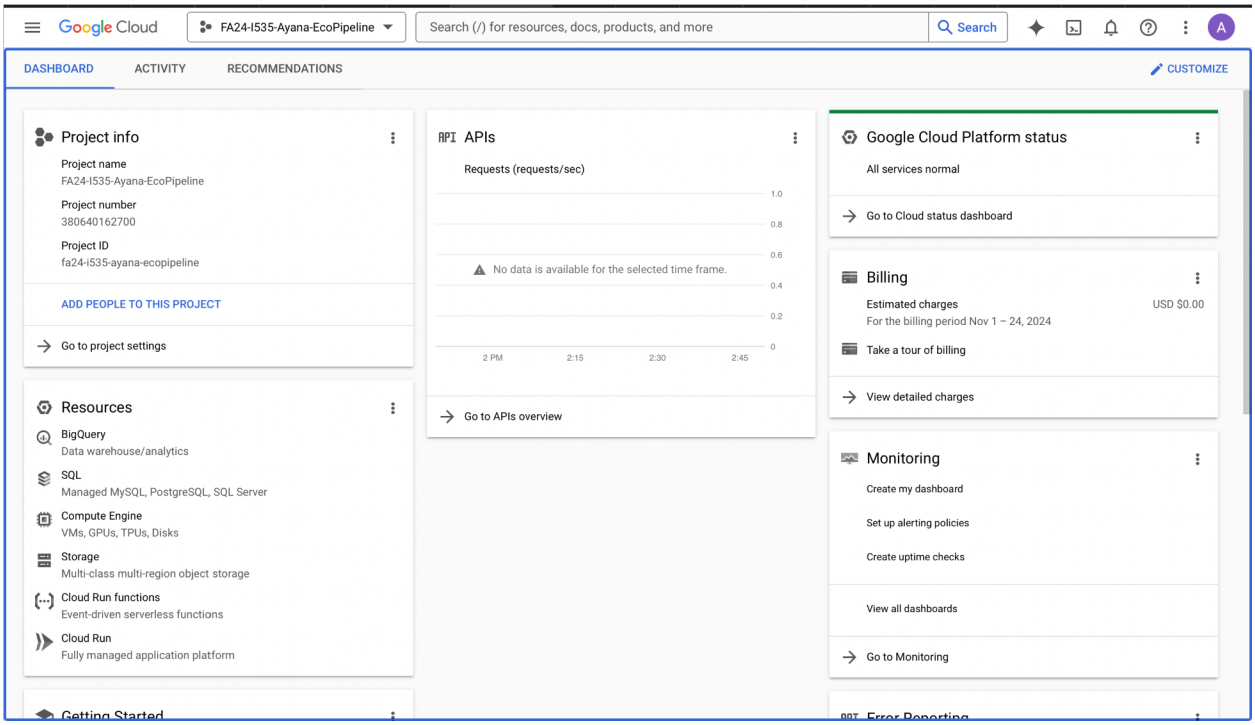
3. Methodology

The following steps outline the process I followed to build the project. These steps, along with the tools used, are described in detail to guide you in replicating this project:

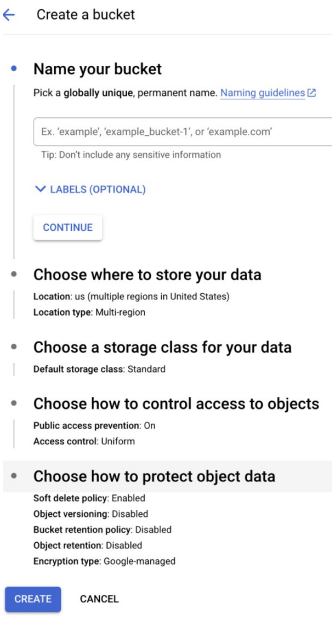
1. Data Extraction
2. Data Transformation (renaming columns, combining columns, calculating revenue)
3. Data loading in Big Query
4. Data Visualization (lookerStudio)

Steps to recreate the project: -

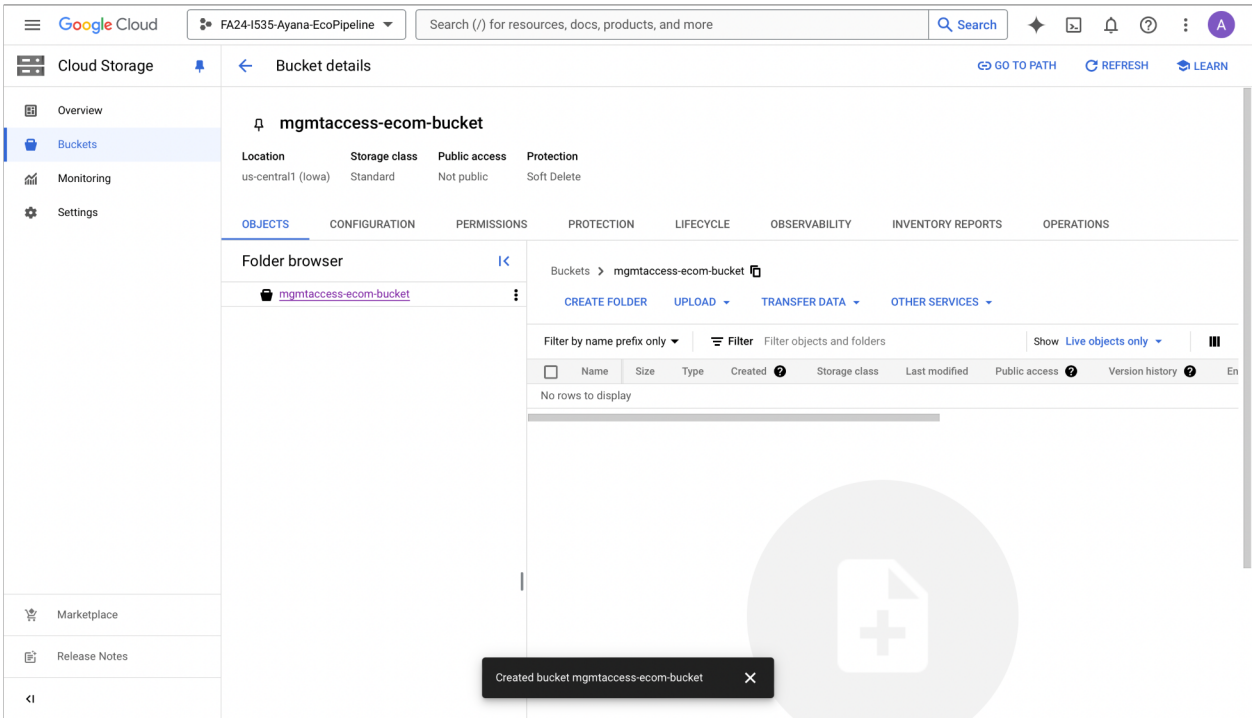
Start by creating a new project in GCP. Once that is done you will land on the project dashboard page as shown below.



Next step is to make a storage bucket. To do this navigate to google cloud storage->buckets> create a bucket as shown in the picture below name your bucket and let everything else be the default setting.

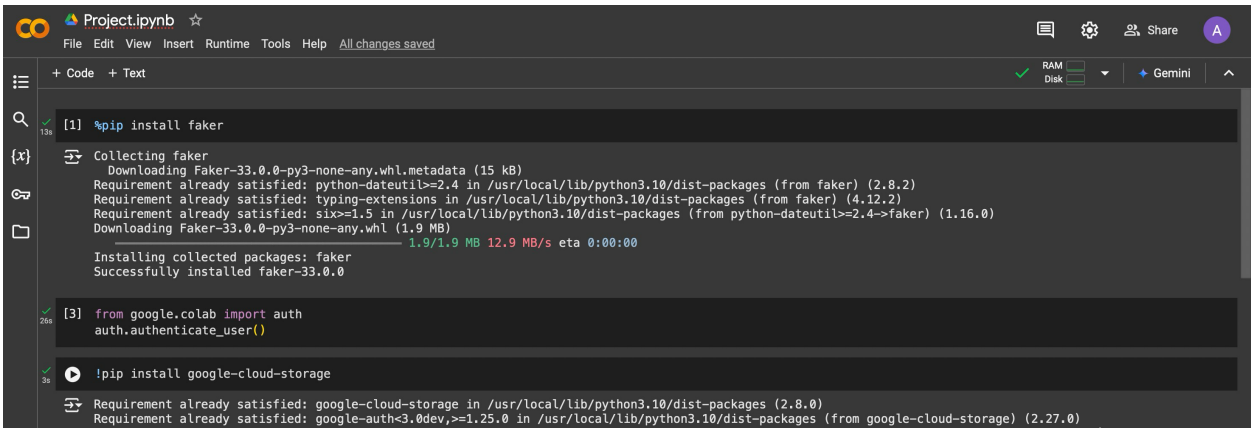


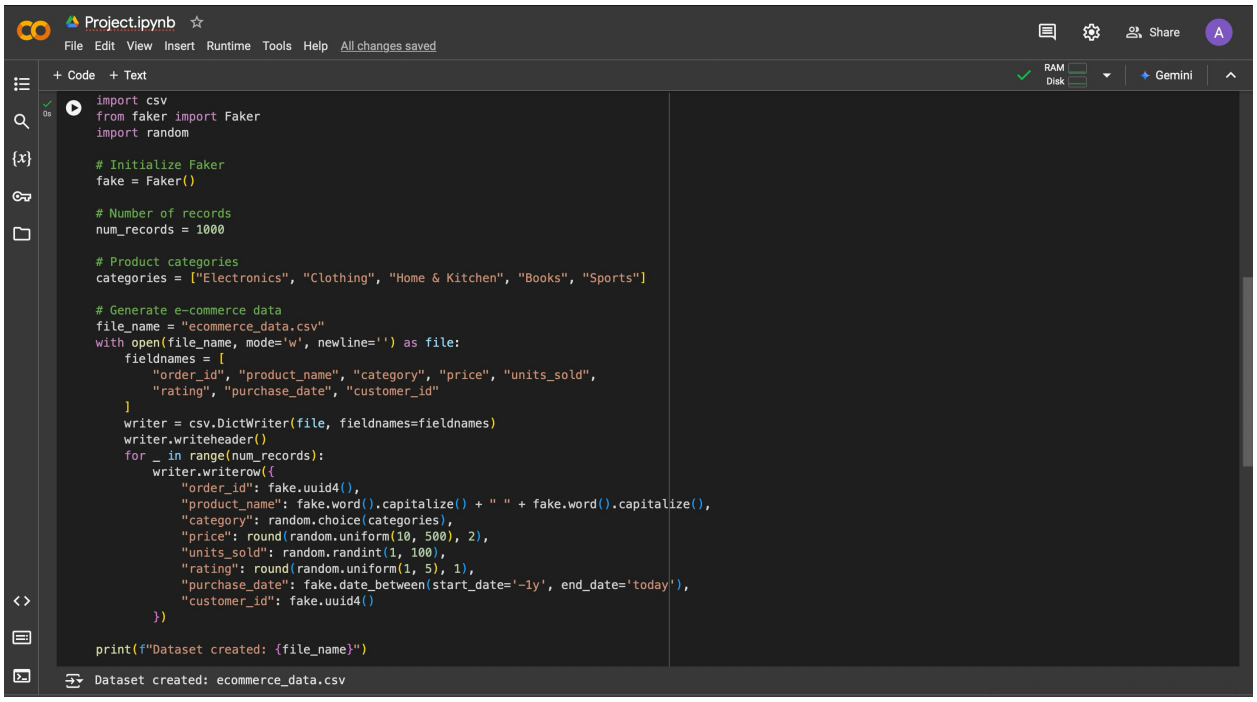
Initially the bucket will be empty as shown below.

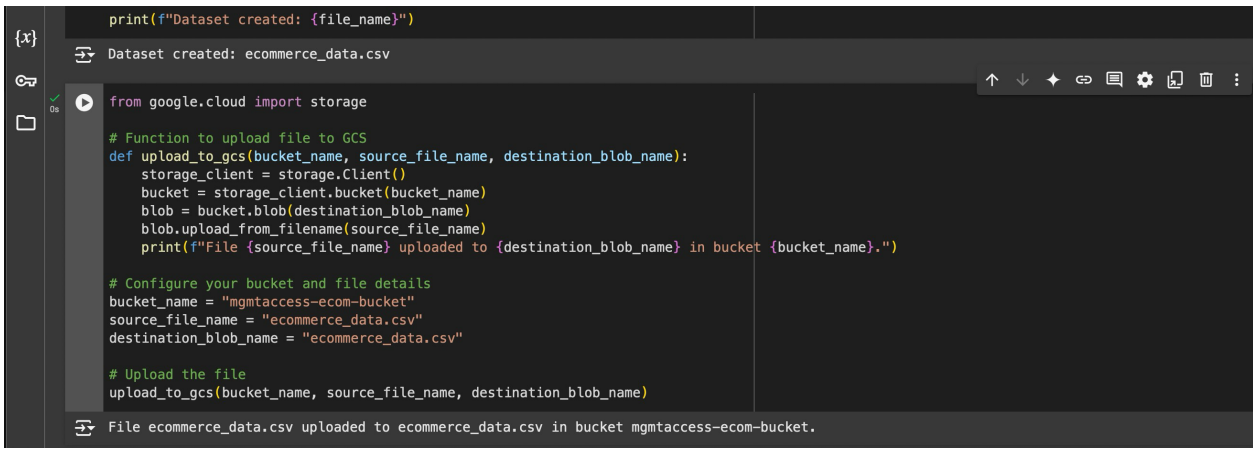


Now we will write the script for extracting the data in a google collab notebook and converting it to csv format you can take the code from below or you can access it from this link: https://colab.research.google.com/drive/1vxxP3VdCGRbtPHJaJnWrQQrbRUuu8Ywo?usp=sharing

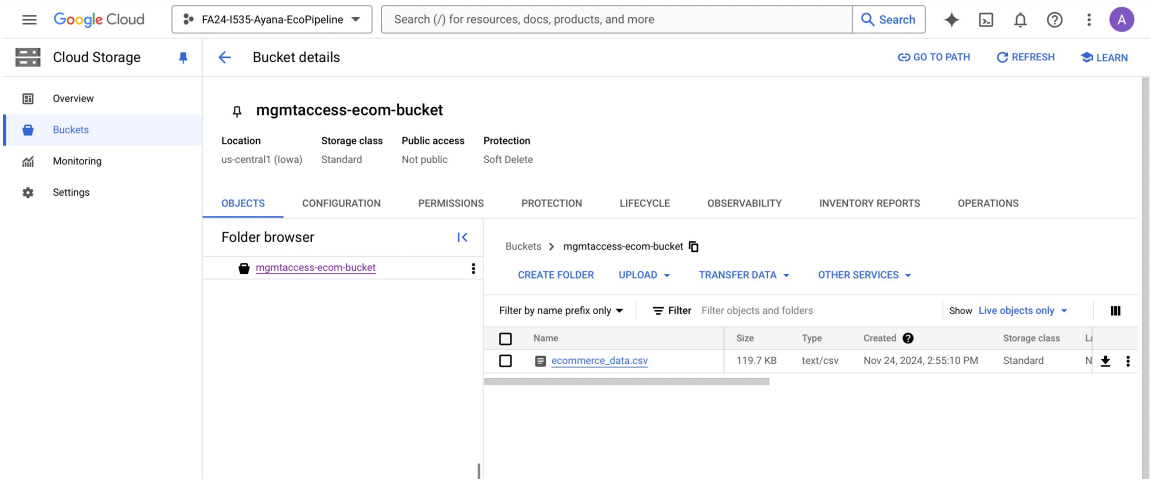
You now have your bucket set up, so simply update the bucket name in the provided code. I've also included the authentication commands in the Colab file, so you don't need to worry about that.





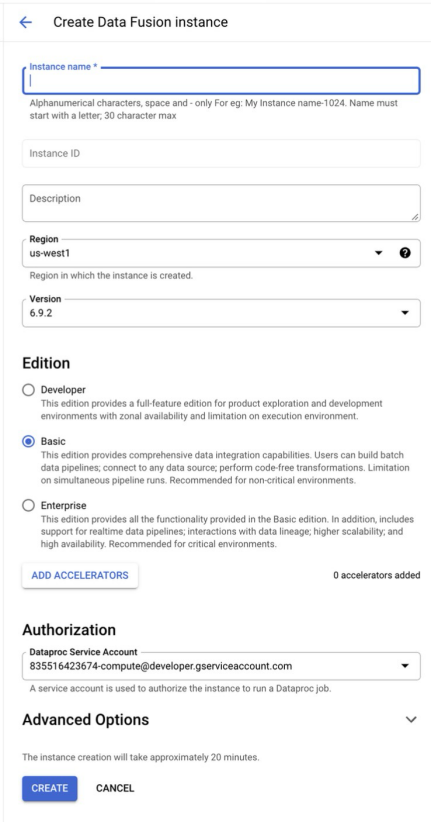


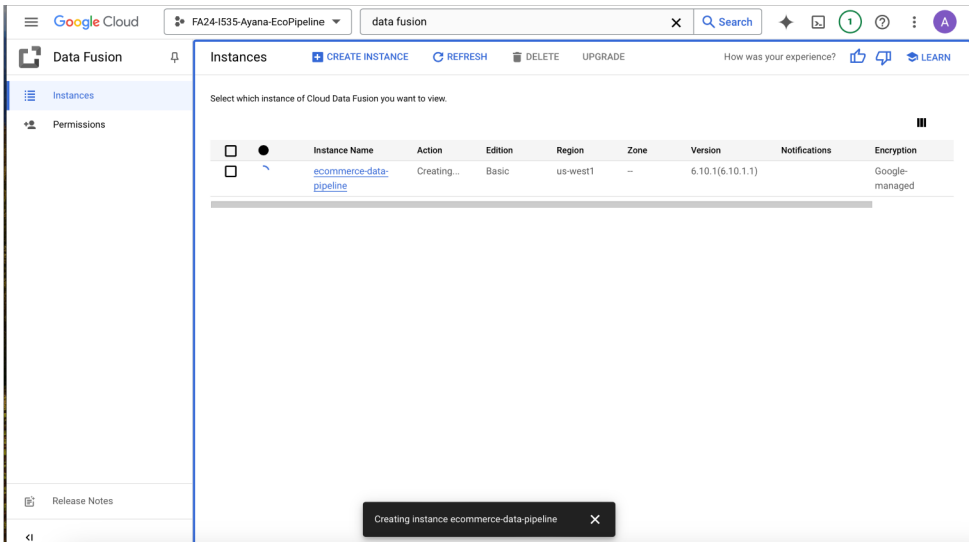
Once you run the code, you should see a file named ecommerce_data.csv in the bucket.



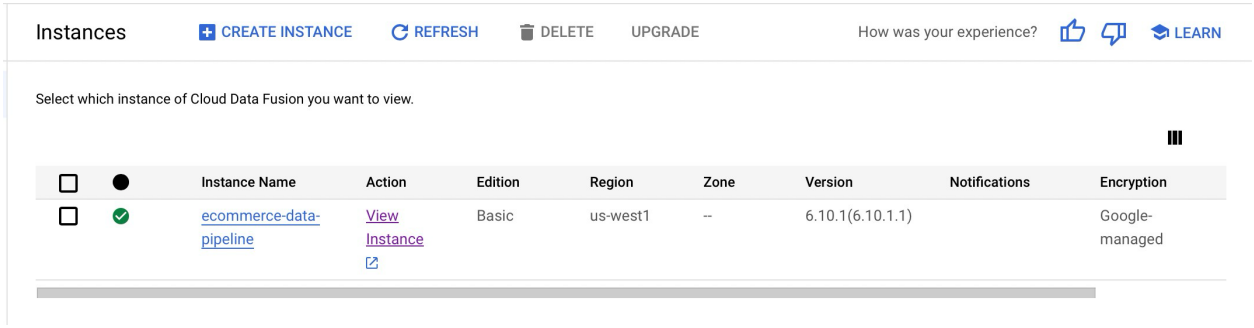
1. Data Extraction:

We will create a Data Fusion instance, you just must enter the name of the cluster and select the right region (in my case uswest1), and leave everything on default, it will take some time to run approximately 15 minutes.

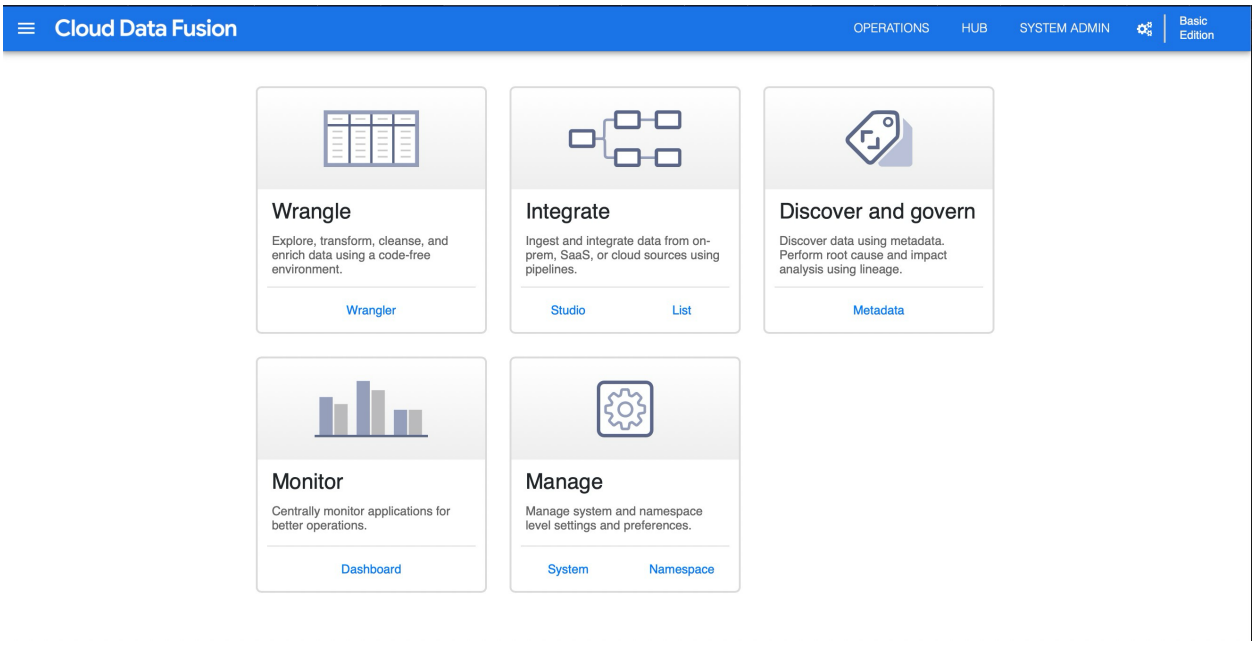




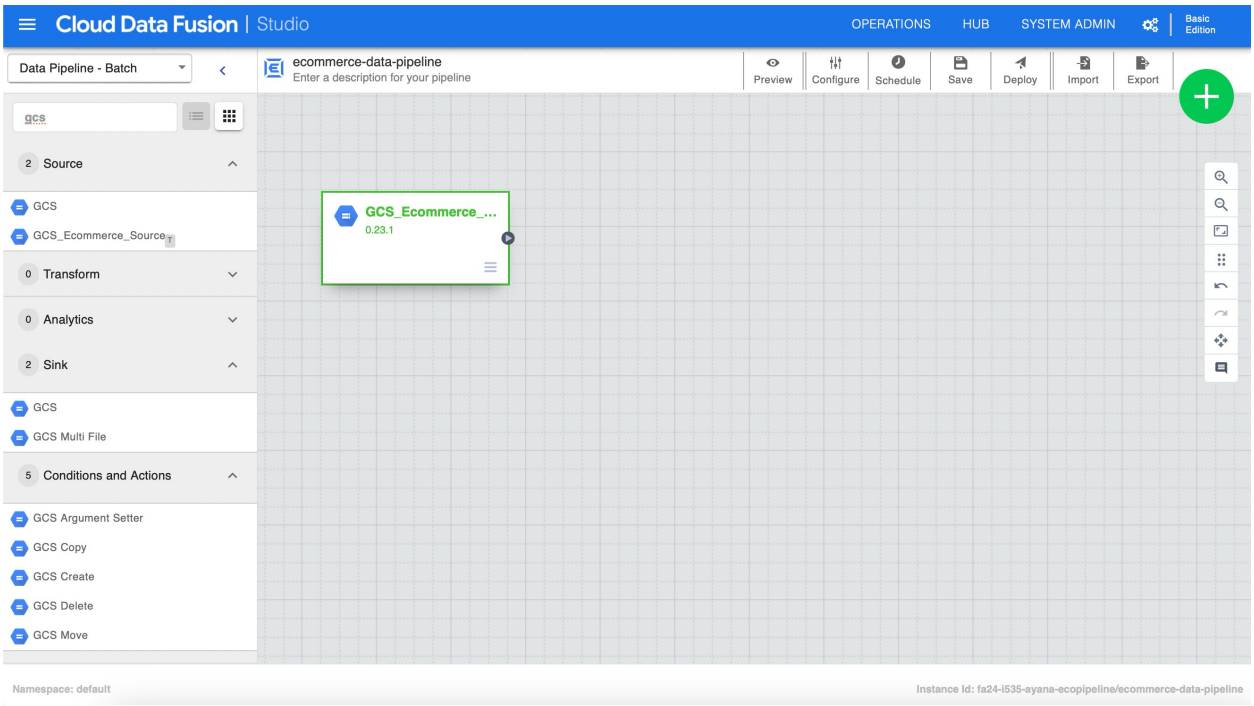
Your Data Fusion instance, which is a Dataproc instance, should now be ready. Click on the "View Instance" button, and you will be prompted to authenticate yourself.



Once authenticated, you will land on the following page. Select "Studio" to create the pipeline.



- In Cloud Data Fusion Studio, click on **"Create Pipeline"**.
- Select **"Batch"** as the pipeline type.
- Add the **"Cloud Storage"** source icon to the canvas and double click to open properties configure it to point to your e-commerce CSV file stored in Google Cloud Storage.
- Set default properties but make sure you name the path to your bucket and point to your project.
- Click validate to check for any errors.

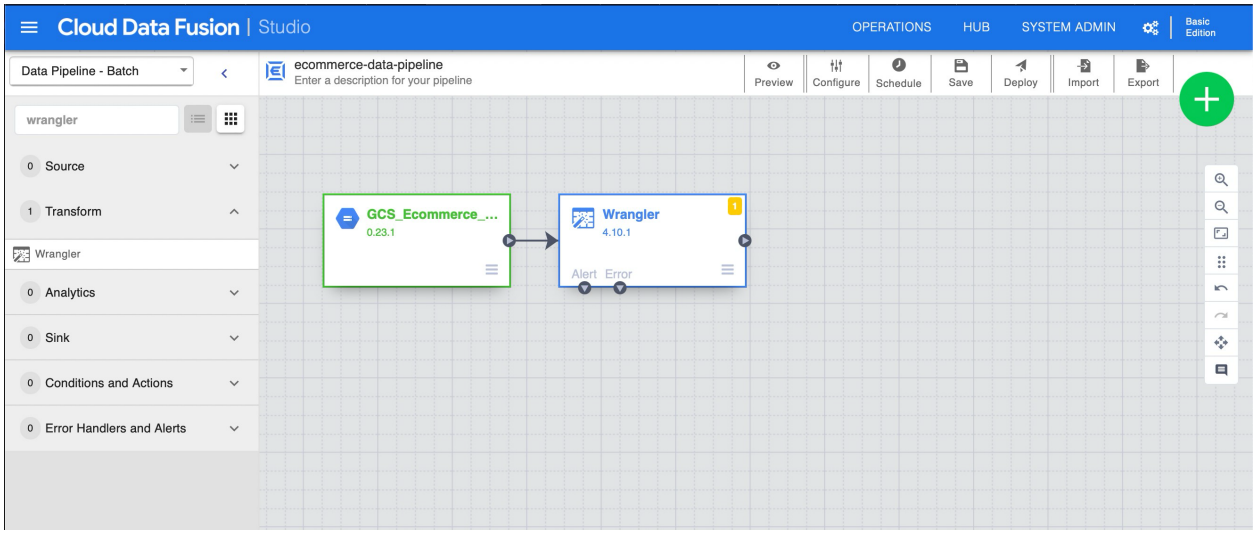


2. Data Transformation

Data Fusion's Wrangler tool was used to clean and enhance the dataset by applying several transformations.

- From the component's menu, drag and drop the **"Wrangle"** transformation icon onto the canvas.
- Connect it to the Cloud Storage source.
- Within the Wrangle transformation (double click to open), you can apply your transformations, like renaming columns, creating new calculated fields (e.g., product_details and revenue).

An image of when both icons are added to the pipeline.

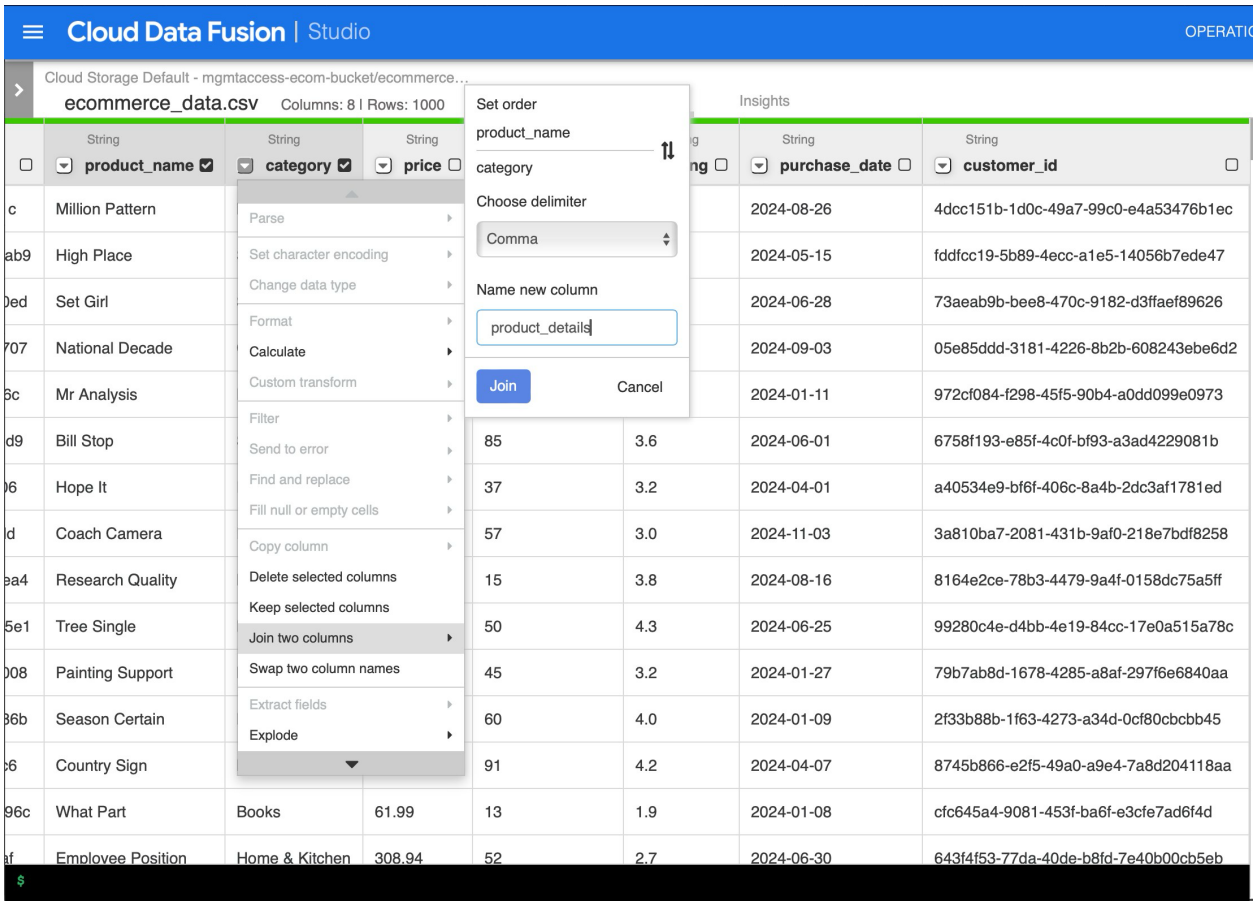


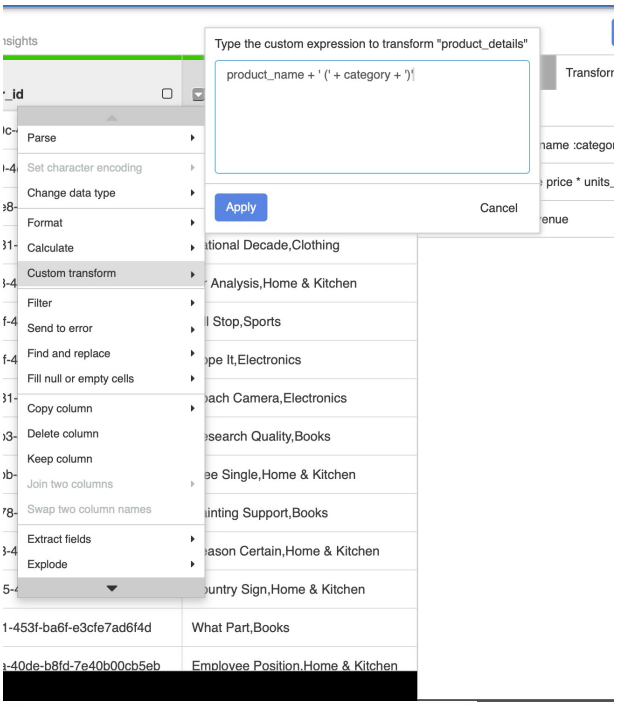
For the all the transformations I have mentioned below you just have to select the column you want to apply the transformation and click on the dropdown menu icon at the top of the column of any column (if you have selected more than one) and you will get the list of options from which you can choose.

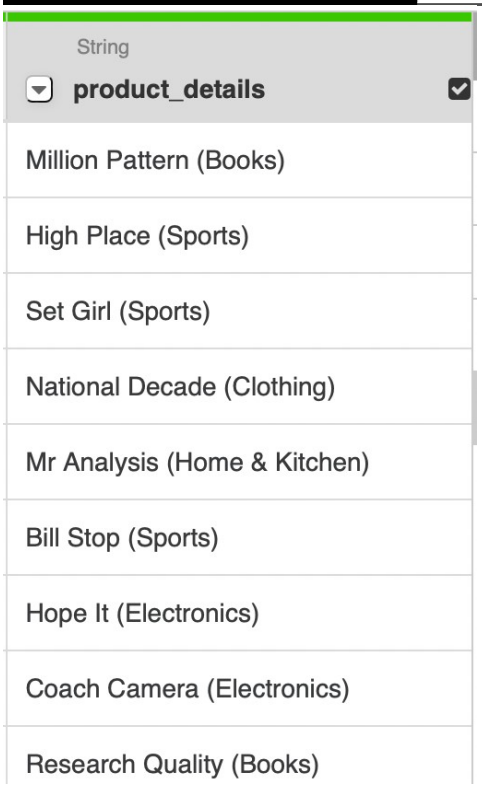
Transformations I did:

Combining Columns:

- A new column, product_details, was created by merging the product_name and category columns. This provided a descriptive label for the products.
- **Formula Used:** product_name + ' (' + category + ')'
- *Purpose:* To make product analysis easier by creating a detailed product description.

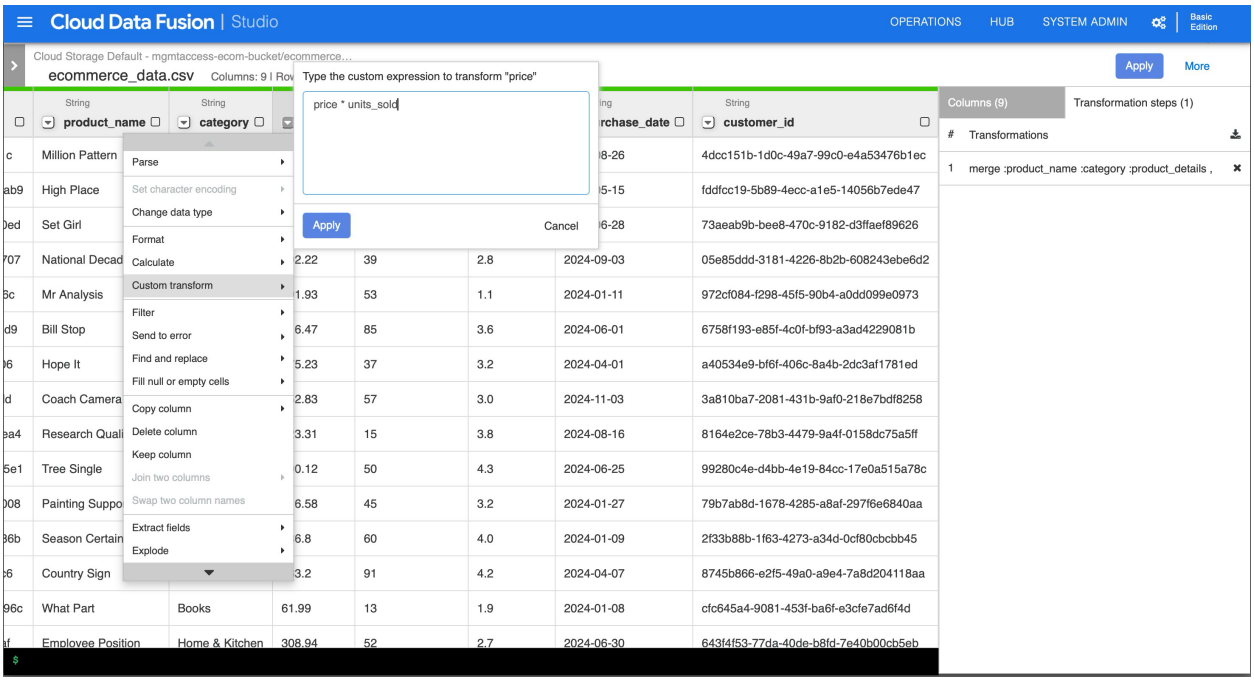

Cloud Data Fusion | Studio
OPERATIO
Cloud Storage Default - mgmtaccess-ecom-bucket/ecommerce...
ecommerce_data.csv
Columns: 8 | Rows: 1000
Insights
Parse
Set character encoding
Change data type
Format
Calculate
Custom transform
Filter
Send to error
Find and replace
Fill null or empty cells
Copy column
Delete selected columns
Keep selected columns
Join two columns
Swap two column names
Extract fields
Explode
Set order
product_name
category
Choose delimiter
Comma
Name new column
product_details
Join
Cancel
String product_name
String category
String price
String purchase_date
String customer_id
c Million Pattern 2024-08-26 4dcc151b-1d0c-49a7-99c0-e4a53476b1ec
ab9 High Place 2024-05-15 fddfcc19-5b89-4ecc-a1e5-14056b7ede47
ed Set Girl 2024-06-28 73aeab9b-bee8-470c-9182-d3ffaef89626
707 National Decade 2024-09-03 05e85ddd-3181-4226-8b2b-608243ebe6d2
6c Mr Analysis 2024-01-11 972cf084-f298-45f5-90b4-a0dd099e0973
d9 Bill Stop 85 3.6 2024-06-01 6758f193-e85f-4c0f-bf93-a3ad4229081b
6 Hope It 37 3.2 2024-04-01 a40534e9-bf6f-406c-8a4b-2dc3af1781ed
d Coach Camera 57 3.0 2024-11-03 3a810ba7-2081-431b-9af0-218e7bdf8258
ea4 Research Quality 15 3.8 2024-08-16 8164e2ce-78b3-4479-9a4f-0158dc75a5ff
5e1 Tree Single 50 4.3 2024-06-25 99280c4e-d4bb-4e19-84cc-17e0a515a78c
008 Painting Support 45 3.2 2024-01-27 79b7ab8d-1678-4285-a8af-297f6e6840aa
36b Season Certain 60 4.0 2024-01-09 2f33b88b-1f63-4273-a34d-0cf80cbcbb45
c6 Country Sign 91 4.2 2024-04-07 8745b866-e2f5-49a0-a9e4-7a8d204118aa
96c What Part Books 61.99 13 1.9 2024-01-08 cfc645a4-9081-453f-ba6f-e3cfe7ad6f4d
af Emplovee Position Home & Kitchen 308.94 52 2.7 2024-06-30 643f4f53-77da-40de-b8fd-7e40b00cb5eb
$





Calculating Revenue:

- To compute the total revenue for each transaction, a new column, revenue, was created by multiplying the price and units_sold columns.
- **Formula Used:** price * units_sold
- *Purpose:* To evaluate the monetary performance of products and categories.



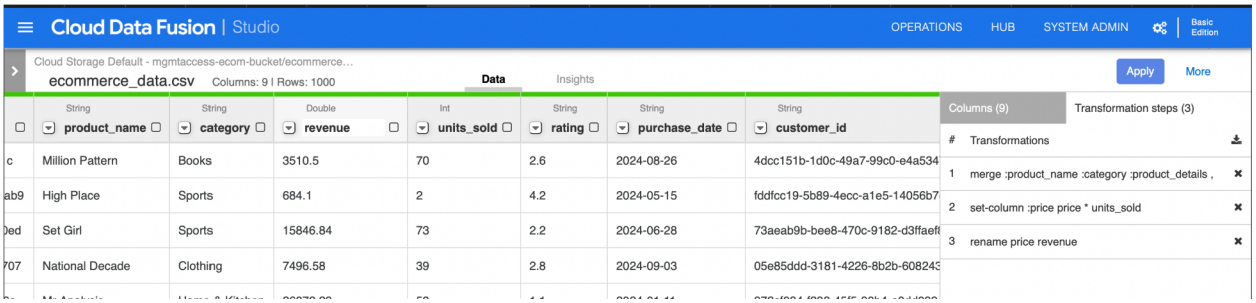
Renaming Columns:

- Some columns were renamed for better readability and alignment with the analysis goals.Example: The column price was renamed to revenue after applying calculations.

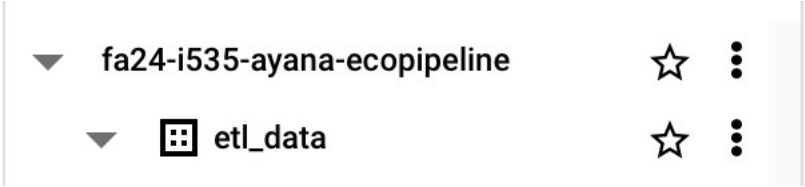
Validating Data:

- Checked for null or inconsistent values across columns and ensured transformations were applied correctly.

This image shows the transformations that I have applied on this data:

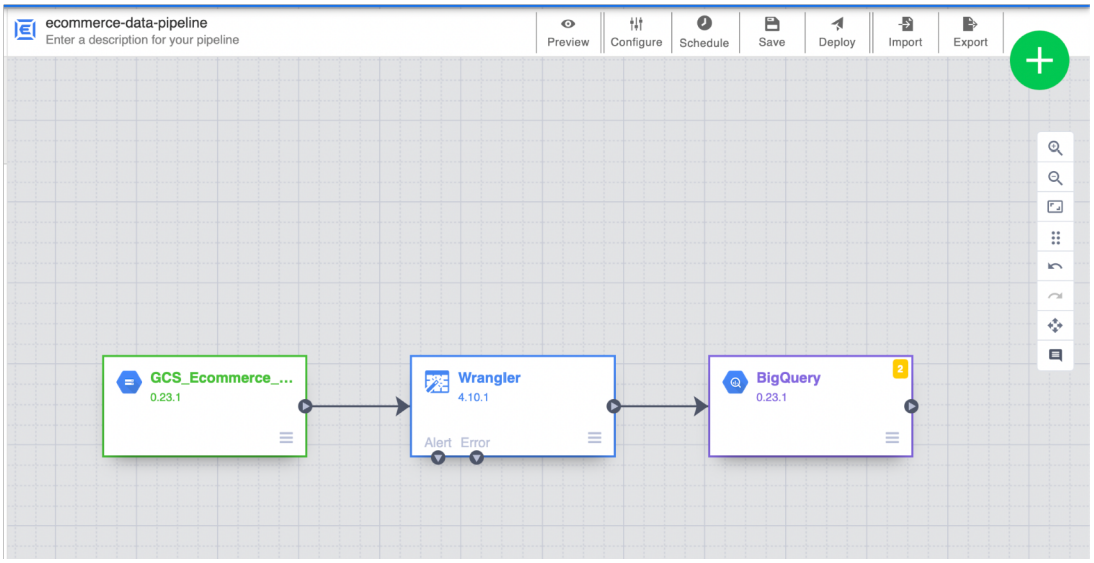


Create a dataset in BigQuery like I have and use the same table name in the BigQuery proeprties below in the pipeline

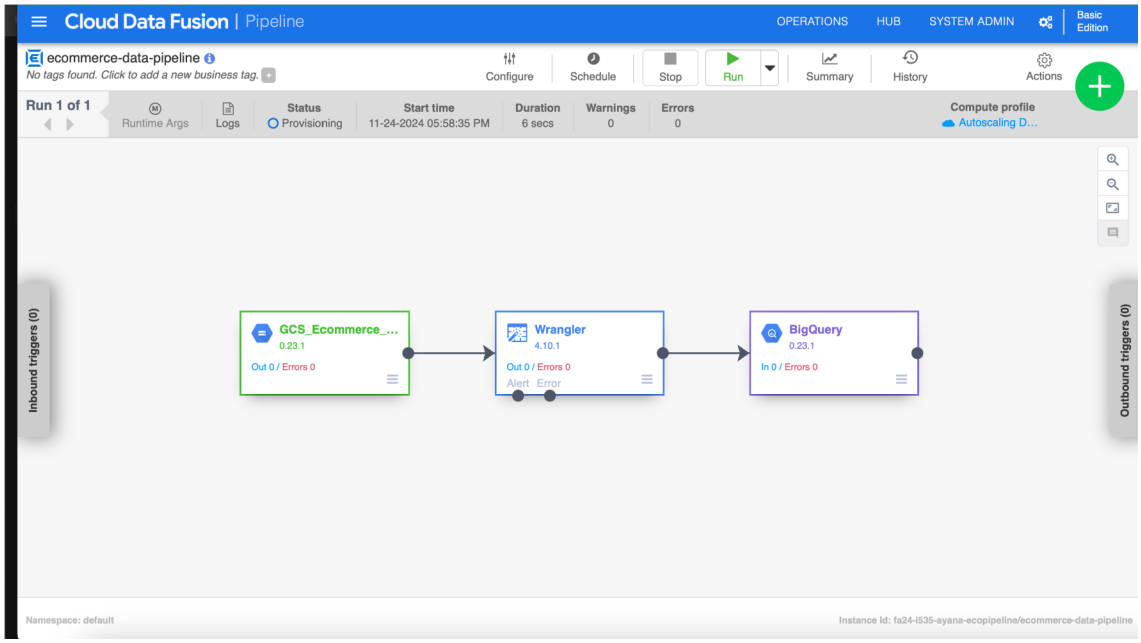


3. Load Data into BigQuery:

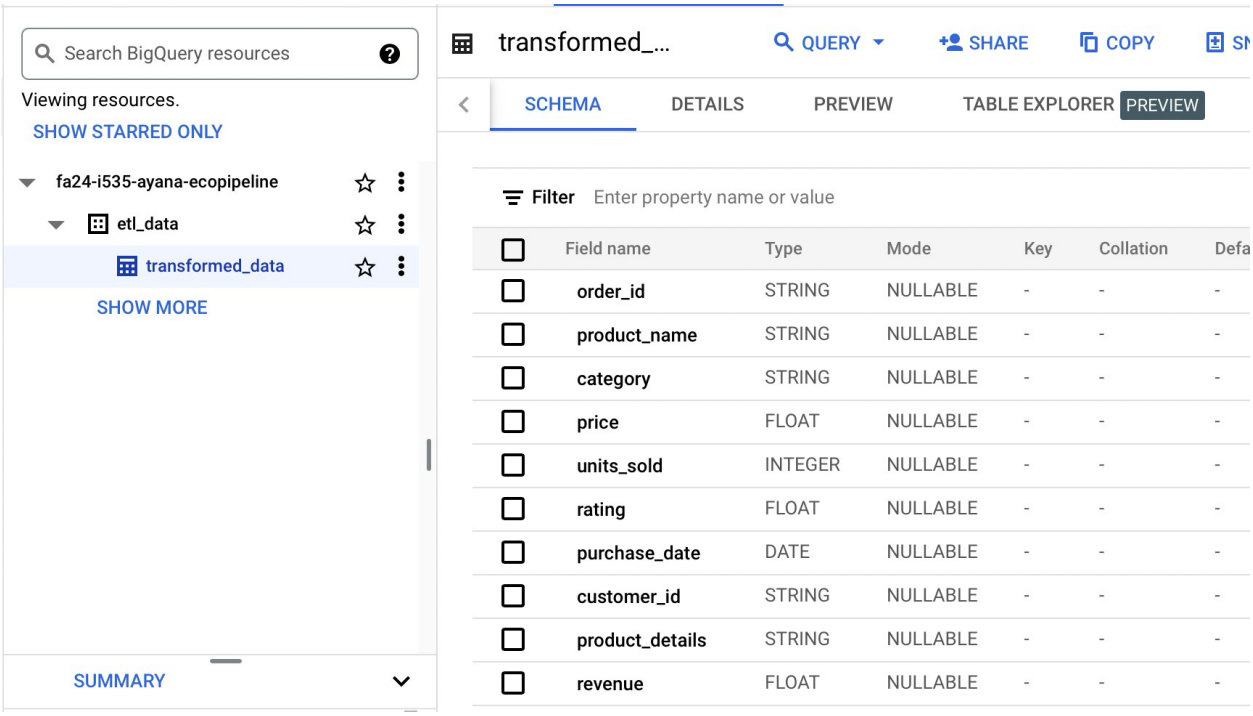
- Drag and drop the **"BigQuery"** sink (for structured and queryable format) icon to the canvas.
- Connect it to the "Wrangle" transformation.
- Configure the BigQuery sink to load the transformed data into a BigQuery table.
- Specify the name of the new transformed data, in my case I names it as transformed_data.
- Validate for errors



Hit the save button and then click on the Deploy button on the Cloud Data Fusion page. It will take you to the below page where you should click “Run” to execute the ETL process. It will take some time to run, and you can see your transformed_data in the dataset of BigQuery.



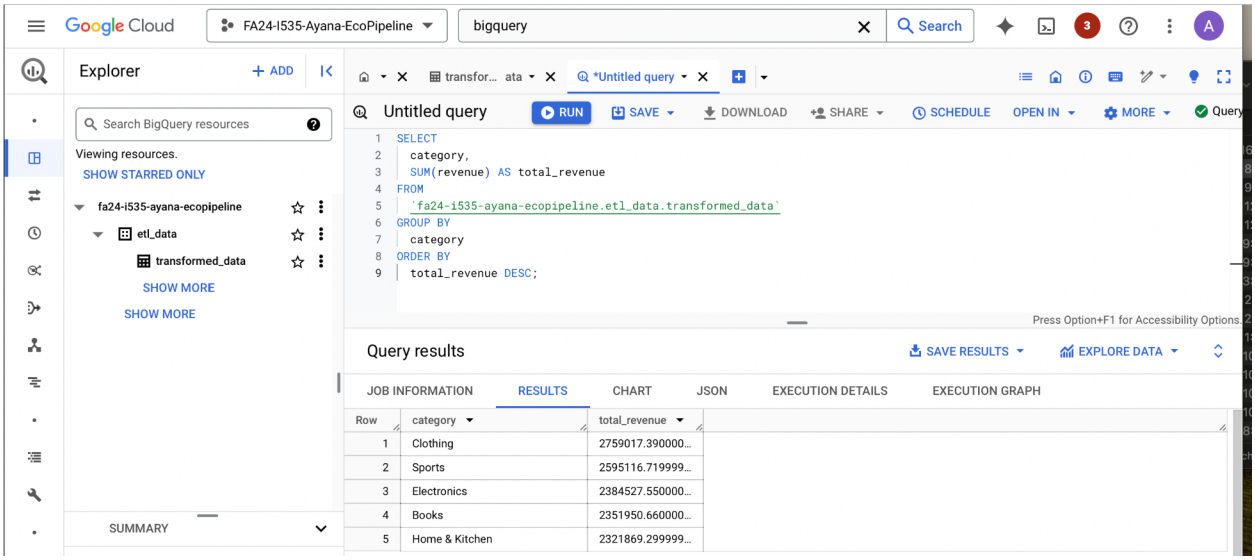
Once the data is loaded into BigQuery you can see it would look alike the image below.



After creating the dataset in BigQuery, I executed the following aggregation queries to analyze the data and extract meaningful insights. These queries provided valuable insights and served as the basis for creating visualizations that highlighted key trends and patterns in the dataset.

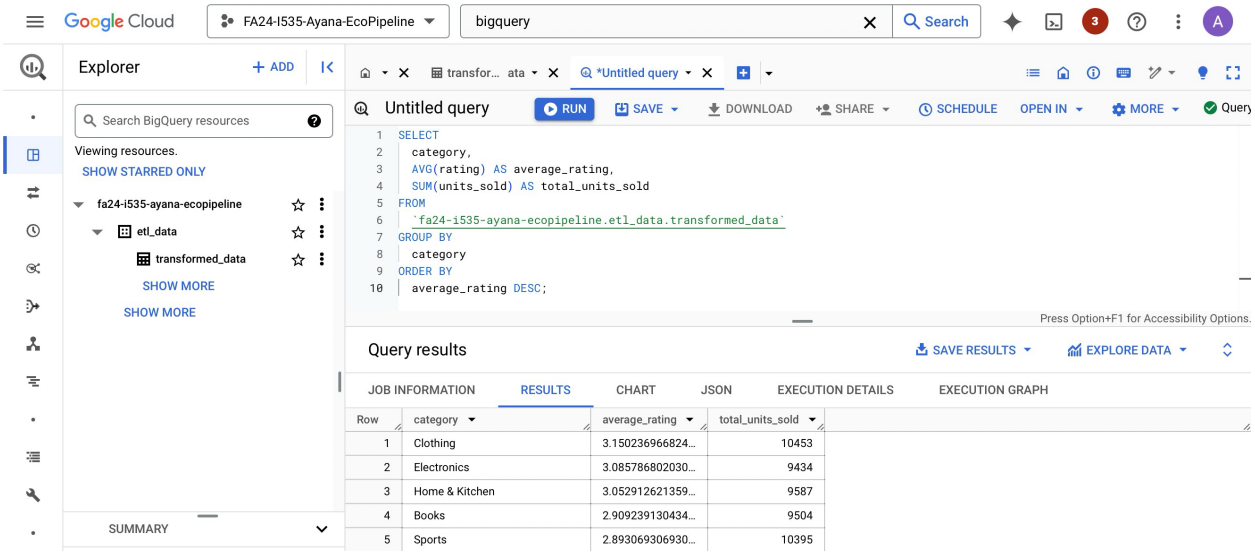
Query 1: Total Revenue per Category

Calculates the total revenue for each product category.



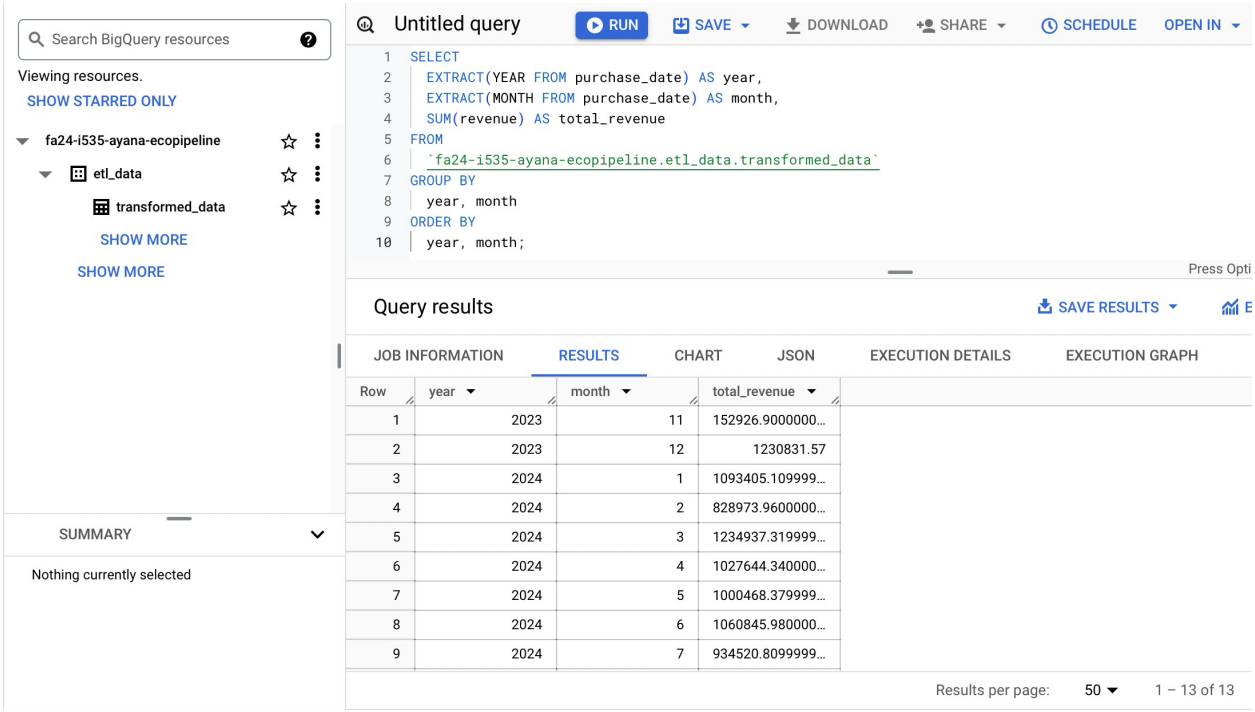
Query 2: Average Rating and Units Sold per Product Category

Calculates the average rating and total units sold for each product category.



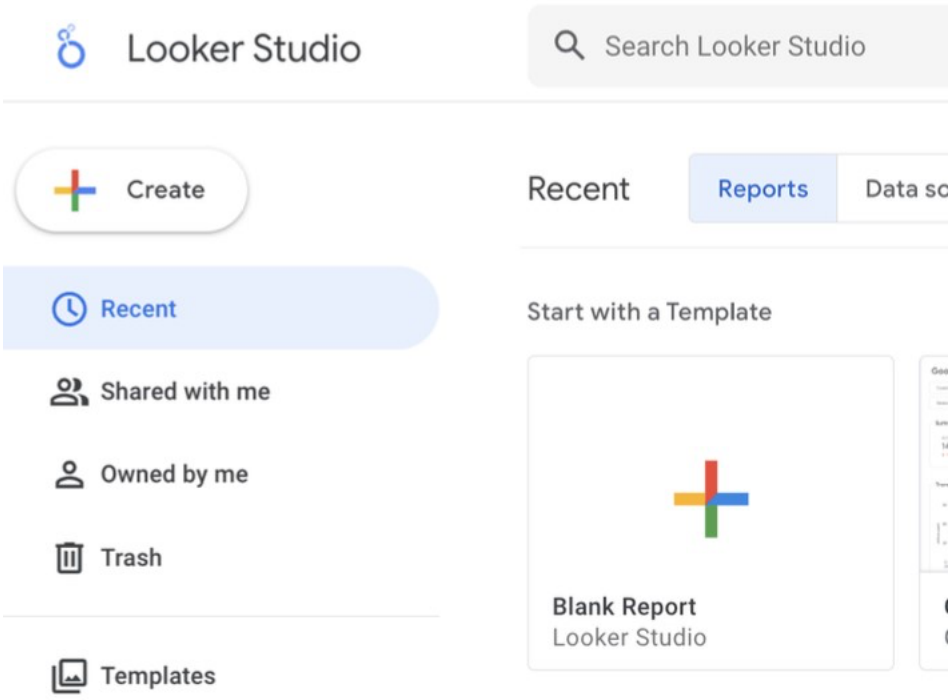
Query 3: Monthly Revenue

Calculates the total revenue generated each month based on the purchase_date.

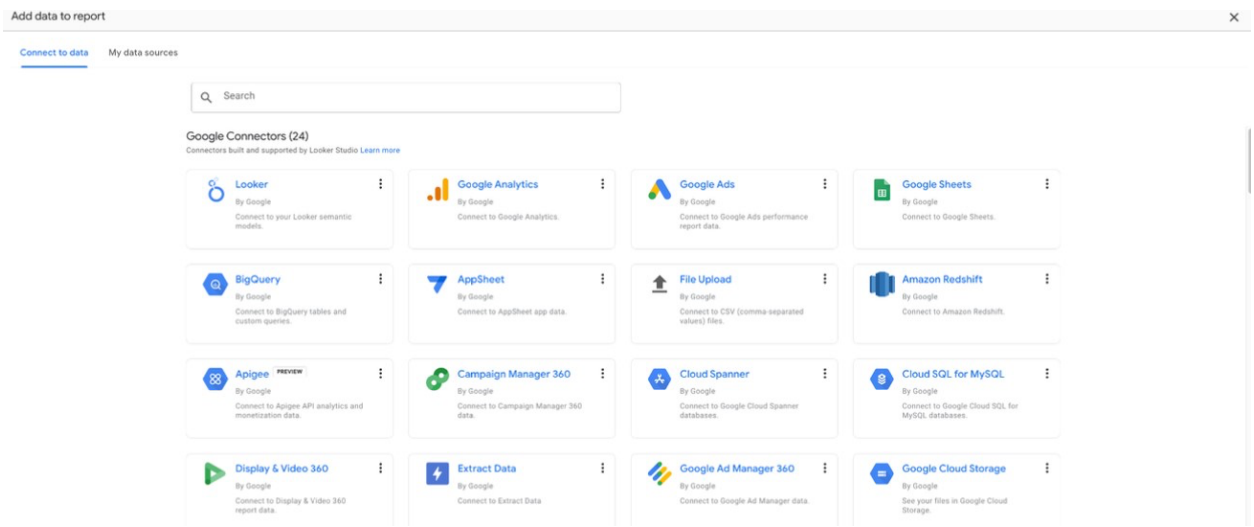


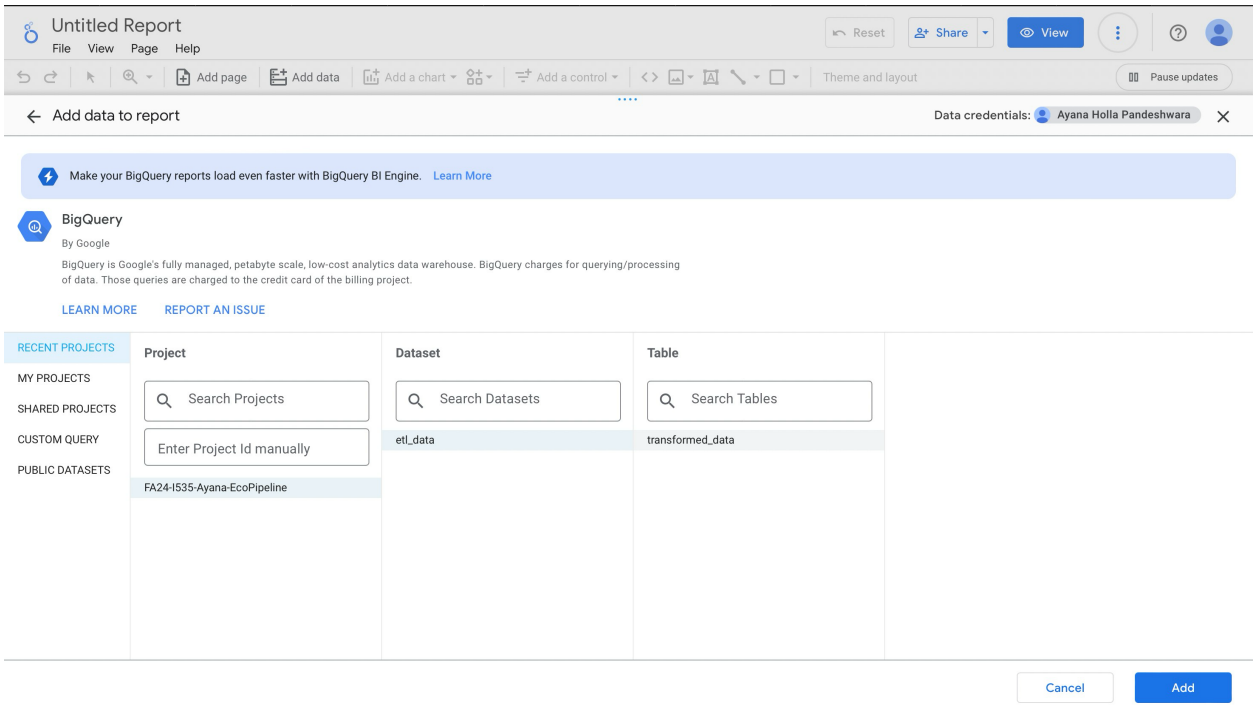
4. Data Visualization

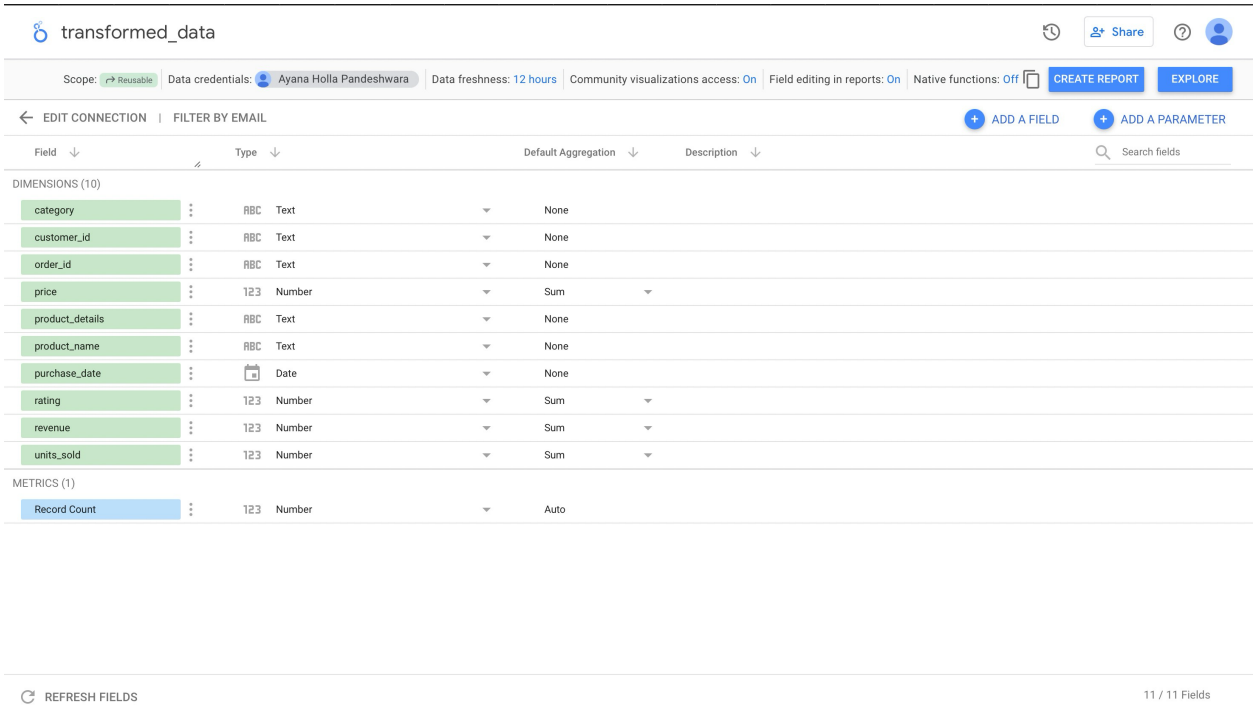
Visualizations were created using Looker Studio to derive insights from the dataset. Go to lookerStudio and click on blank report and now you will see multiple options to use the data of your choice.



Select bigQuery and you will see your dataset, select it and now select the table that we used till now as shown in the following screenshots:



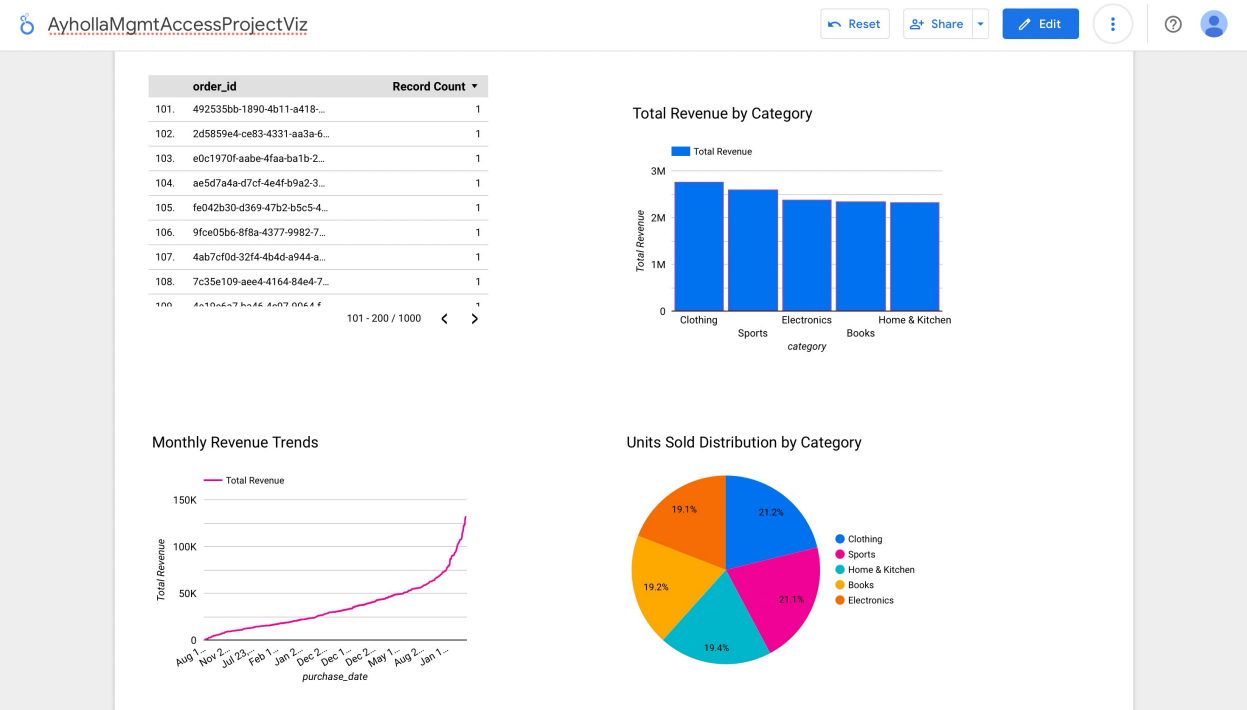




Now you have your table loaded in the lookerStudio you can use the menu for multiple options for visualisations. I have used tables, pie charts, line graph and bar chart for presentation of my data as shown below.

Three key visualizations were developed:

1. **Bar Chart - Revenue by Category:** Displayed the total revenue generated by each product category, highlighting the top-performing segments.
2. **Line Chart - Monthly Revenue Trends:** Captured the revenue trends over time, grouped by month, to provide an overview of seasonal patterns and growth.
3. **Pie Chart - Units Sold by Category:** Illustrated the proportion of units sold across di;erent categories, o;ering a snapshot of sales distribution.



4. Results

The project successfully implemented an ETL data pipeline using Google Cloud Technologies, transforming and analyzing an e-commerce dataset to derive actionable insights. Each stage of the pipeline, from data generation to visualization, contributed to uncovering meaningful patterns and trends.

Data Generation

The dataset was generated using the Faker library, which created a synthetic e-commerce dataset containing fields such as product names, categories, prices, units sold, and purchase dates. This dataset simulated real-world e-commerce scenarios, making it an excellent foundation for the ETL pipeline. The generated data was saved as a CSV file and uploaded to Google Cloud Storage, ensuring it was accessible for further processing.

Data Extraction

The dataset stored in Google Cloud Storage was accessed using Cloud Data Fusion, which seamlessly integrated the data into the ETL pipeline for transformations.

Data Transformation

The transformations were conducted in Cloud Data Fusion using the Wrangler tool. Two key transformations were applied to enrich the dataset:

- A new column, product_details, was created by combining product_name and category, providing a descriptive label for each product.
- A revenue column was calculated by multiplying price and units_sold, enabling insights into product and category financial performance.

These transformations structured the data for analysis and ensured it was ready for loading into BigQuery.

Data Loading into BigQuery

The transformed dataset was loaded into BigQuery under the dataset ecommerce_dataset and table transformed_data. BigQuery's scalable infrastructure allowed for e;icient querying and analysis of the enhanced dataset.

Visualizations and Insights

The final dataset was visualized using Looker Studio, producing the following charts:

- Bar Chart - Revenue by Category: This chart highlighted the total revenue generated by each category. Categories like Clothing and Sports were the top performers, driving the majority of sales revenue.
- Line Chart - Monthly Revenue Trends: Revenue trends over time showed steady growth, with occasional spikes suggesting seasonal demand or promotional impacts.
- Pie Chart - Units Sold Distribution by Category: This chart illustrated the proportion of units sold for each category, with Clothing leading slightly, reflecting its popularity.

Key insights from these visualizations include the identification of high-performing product categories, revenue seasonality, and balanced sales distribution across categories. These findings demonstrate how a well-designed pipeline can provide actionable insights for ecommerce businesses, helping them make informed decisions based on data.

5. Discussion

How the Project Aligns with Course Topics

This project incorporates several concepts and tools covered throughout the course. Key alignments include:

- **Big Data Lifecycle and Pipelines (Module 6):** The project exemplifies the lifecycle of data, starting from extraction, transformation, and loading (ETL) to final visualization. It also highlights the use of Cloud Data Fusion for pipeline automation, showcasing sequencing and scaling capabilities.

- **Cloud Computing (Module 3):** By leveraging Google Cloud Platform (GCP), the project demonstrates the practical use of cloud services for big data processing and storage.
- **Ingest and Storage (Module 7):** BigQuery, a data warehouse solution, was used to store the transformed data. This aligns with the module's focus on organizing and retrieving big data.
- **Processing and Analytics (Module 9):** Data transformations applied in Cloud Data Fusion ensured the quality and usability of the dataset, making it ready for further analysis.

These connections reinforce how theoretical concepts from the course translate into realworld applications, bridging the gap between learning and execution.

Challenges Faced

Throughout the project, I encountered several challenges:

1. **Quota Limitations in GCP:** The initial deployment of the pipeline failed due to insu;icient disk space quotas. This required modifying configurations to reduce the size of the dataset and adjusting resource allocations in GCP.
2. **Complexity of Data Fusion Interface:** Configuring the ETL pipeline in Cloud Data Fusion involved a learning curve. Issues such as misconfigured transformations and validation errors required additional troubleshooting.
3. **Visualization Customization in Looker Studio:** Creating interactive and accurate visualizations, especially grouping dates for monthly trends, required multiple iterations and adjustments to ensure clarity.

How Challenges Were Addressed

- To resolve the **quota limitations**, I optimized the pipeline configuration and used a smaller dataset to fit within the GCP limits.
- For the **Cloud Data Fusion interface**, I referred to tutorials and documentation to understand the configuration steps and successfully implement the pipeline.
- In Looker Studio, I experimented with data fields and visualization settings to overcome grouping issues, resulting in cleaner and more meaningful charts.

Skills Gained

This project deepened my understanding of:

- Building scalable ETL pipelines with cloud-based tools.
- Managing and transforming big data for analysis.
- Creating insightful visualizations to communicate findings e;ectively. These skills are essential for handling real-world big data scenarios and align with my learning objectives for this course.

6. Conclusion

The project highlights the potential of cloud-based tools like Google Cloud Data Fusion and BigQuery in managing and analyzing large datasets. By processing e-commerce data, I was able to uncover valuable insights, such as identifying high-revenue categories, analyzing monthly trends, and understanding customer purchasing patterns.

For e-commerce businesses, the findings demonstrate how data pipelines can streamline operations and support data-driven decision-making. For example, insights into revenue trends and product performance can help optimize inventory, improve marketing strategies, and enhance customer satisfaction.

Overall, this project reinforced the importance of combining scalable infrastructure with robust data processing techniques to derive actionable insights. While the journey presented challenges, it was a rewarding experience that bridged the gap between theoretical concepts and practical applications.

You can also access all the code templates on my github-
https://github.iu.edu/ayholla/I-535MgmtAccessProject

References:

- GCP SDK installation: https://cloud.google.com/sdk/docs/install-sdk • Login on Colab with gcloud without service account • Google Cloud Platform: YouTube. Available online: https://www.youtube.com/user/googlecloudplatform
- Towards Data Science: Medium. Available online: https://towardsdatascience.com/
- Google Cloud Blog: Available online: https://cloud.google.com/blog
- VPC networks: https://cloud.google.com/vpc/docs/vpc
- Use uniform bucket-level access: https://cloud.google.com/storage/docs/usinguniform-bucket-level-access
- GCP SDK setup: https://www.codingforentrepreneurs.com/blog/google-cloud-cliand-sdk-setup/
- Questions tagged 'google-cloud-platform': Stack Overflow. Available online: https://stackoverflow.com/questions/tagged/google-cloud-platform
- ChatGPT: https://chat.openai.com/c/051a2191-31a1-47b6-b1da-931f10581ce2