WORLD EXPORTS & GLOBAL TRADE PERFORMANCE ANALYTICS PLATFORM

CAPSTONE PROJECT

INTRODUCTION

NAME: Sai Pranavi Arra

COLLEGE: VNR Vignana Jyothi Institute of Engineering and Technology

TRACK: Data Engineering

PROBLEM STATEMENT

🌍 **Business Problem**

International trade organizations and policy bodies collect massive amounts of export data across:

- Countries
- Regions
- Product categories

However, they face major challenges:

- Raw export data scattered across multiple files
- Manual consolidation causing inconsistencies & errors
- No centralized system for analyzing global trade trends
- Difficulty identifying high-performing & declining markets
- Limited visibility for policy makers to make data-driven decisions

OBJECTIVE

The objective of this project is to build an End-to-End Global Trade Analytics Platform that can support the following:

- Automates ingestion and processing of global export datasets
- Cleans, standardizes, and validates trade data
- Performs advanced analytics using Pandas & PySpark
- Orchestrates ETL workflows using Apache Airflow
- Generates meaningful insights & visuals using Power BI Dashboards
- Supports policy makers, economists, and analysts in decision making

The project aims to develop a centralized analytics platform that automates the processing of global trade data, enabling efficient analysis of export performance and market trends. By transforming raw data into meaningful insights, the platform will support data-driven economic and trade decision-making while providing interactive dashboards that help policymakers, analysts, and stakeholders easily explore and understand global trade dynamics.

TECHNOLOGY STACK

Data Processing

Python- Used as the primary programming language for implementing data processing logic, handling datasets, and building ETL workflows due to its simplicity, flexibility, and strong ecosystem.

Pandas & NumPy- Pandas is used for data cleaning, transformation, and analysis, while NumPy provides high-performance numerical computing support. Together, they help in handling structured datasets efficiently.

Big Data Analytics

Databricks- A cloud-based unified analytics platform used to process large-scale datasets, build pipelines, and run distributed computing workloads. It enables collaboration and simplifies big data processing.

PySpark- The Python API for Apache Spark, used to perform large-scale data processing, transformations, and analytics on massive datasets in a distributed environment.

TECHNOLOGY STACK

Workflow Orchestration

Apache Airflow- Used to schedule, automate, and manage data pipelines. It ensures tasks run in sequence, handles retries, monitors workflows, and maintains pipeline reliability.

Data Visualization

Power BI- Used to build interactive dashboards and visual analytics that help stakeholders explore trade insights, monitor KPIs, and make data-driven decisions.

Others

Git & GitHub- Used for version control and collaboration, ensuring code management, tracking changes, and maintaining project history.

Databricks Notebooks- Interactive notebook environments used for data exploration, development, testing, and visualization of intermediate outputs during the data pipeline development process.

DATASETS EXPLANATION

📌 Overview

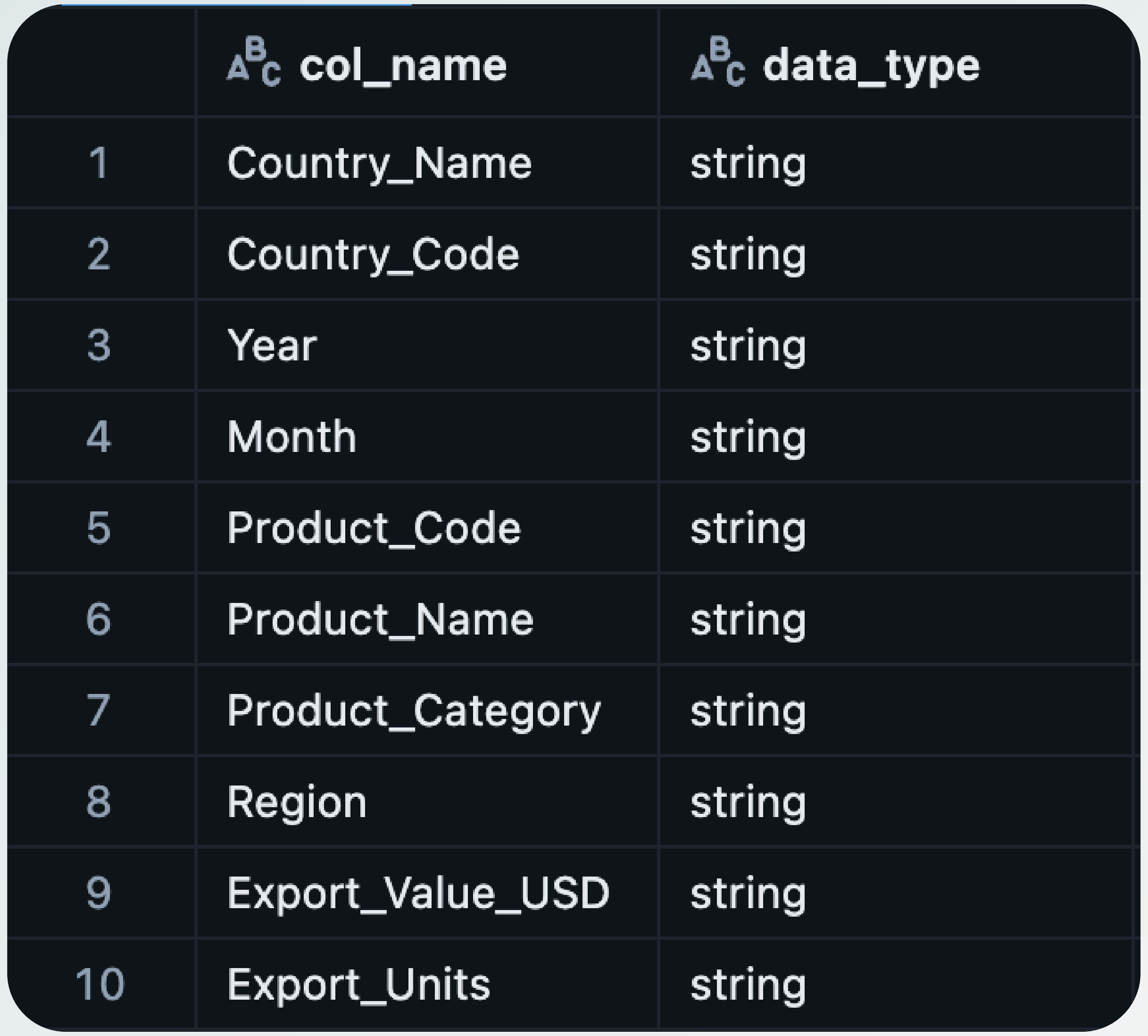
The project uses global export datasets containing country-wise and product-wise trade information. The data represents export volumes, trade values, product categories, and time-based performance trends across different countries. The datasets for this project follows a star model.

📁 Datasets Used

1.Global Export Fact Dataset

- Primary transactional dataset
- Contains country-level export records
- Includes metrics like export value, quantity, year, and product details
- ~7000 records (moderately large dataset)
- Contains ~5% null values and some intentional inconsistencies to simulate real-world data

Schema of Global Export Fact Dataset



	col_name	data_type
1	Country_Name	string
2	Country_Code	string
3	Year	string
4	Month	string
5	Product_Code	string
6	Product_Name	string
7	Product_Category	string
8	Region	string
9	Export_Value_USD	string
10	Export_Units	string

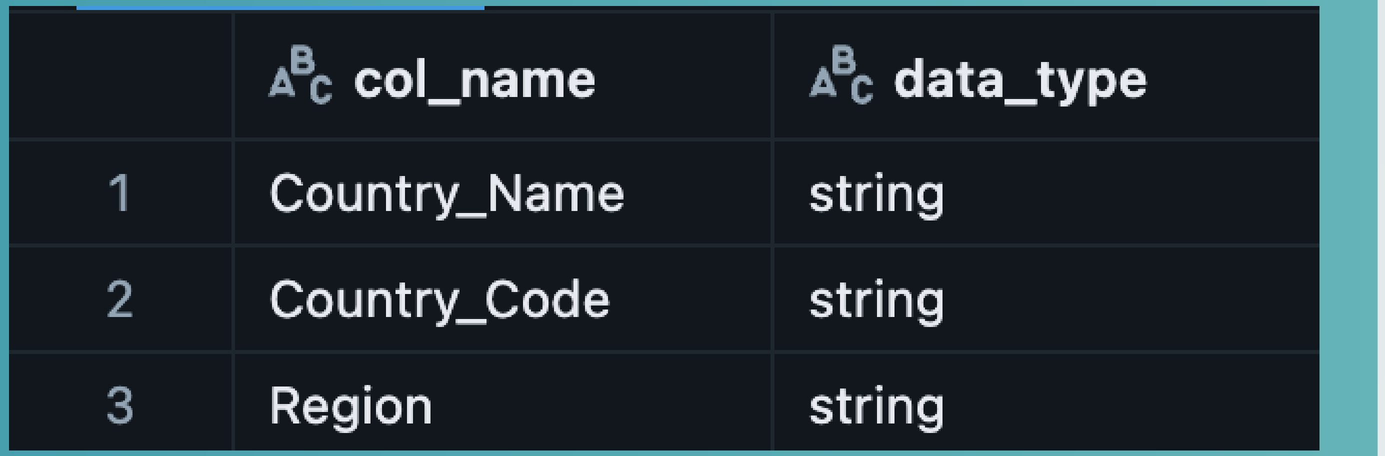
DATASETS EXPLANATION

❓ Why is Reference datasets important?

Reference datasets play a crucial role in ensuring accuracy, reliability, and consistency within the global trade analytics platform. They help standardize key attributes like eliminating spelling variations and formatting differences that may exist in the raw data. They also improve overall data quality by enabling validation, helping identify missing or incorrect records, and ensuring only trusted data progresses through the pipeline.

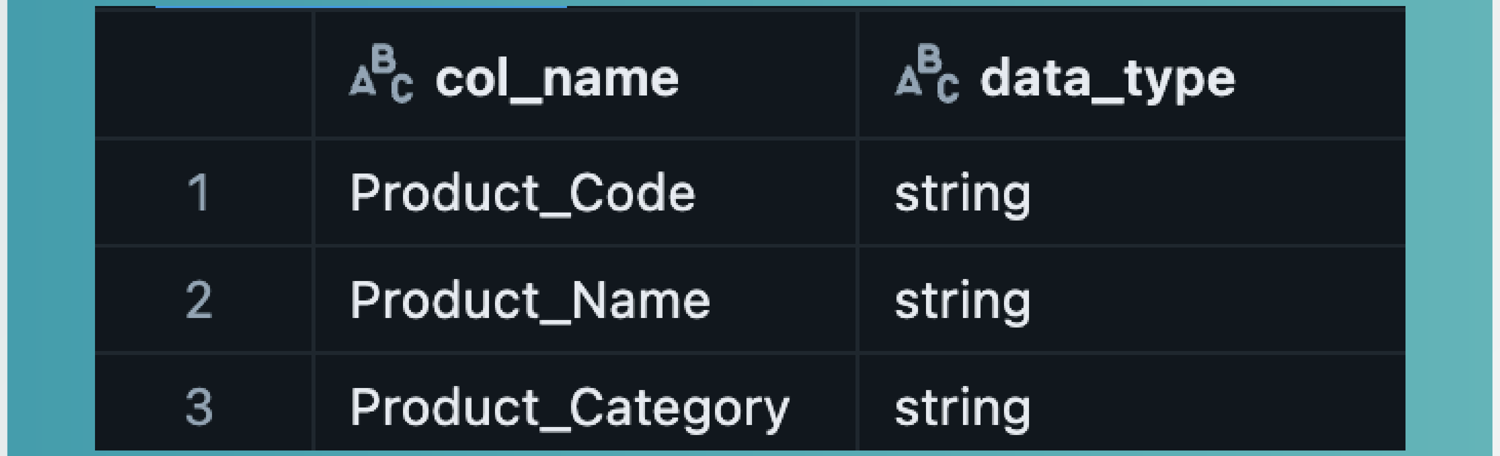
📂 Reference Datasets Used

Country Reference Dataset



	col_name	data_type
1	Country_Name	string
2	Country_Code	string
3	Region	string

Product Reference Dataset



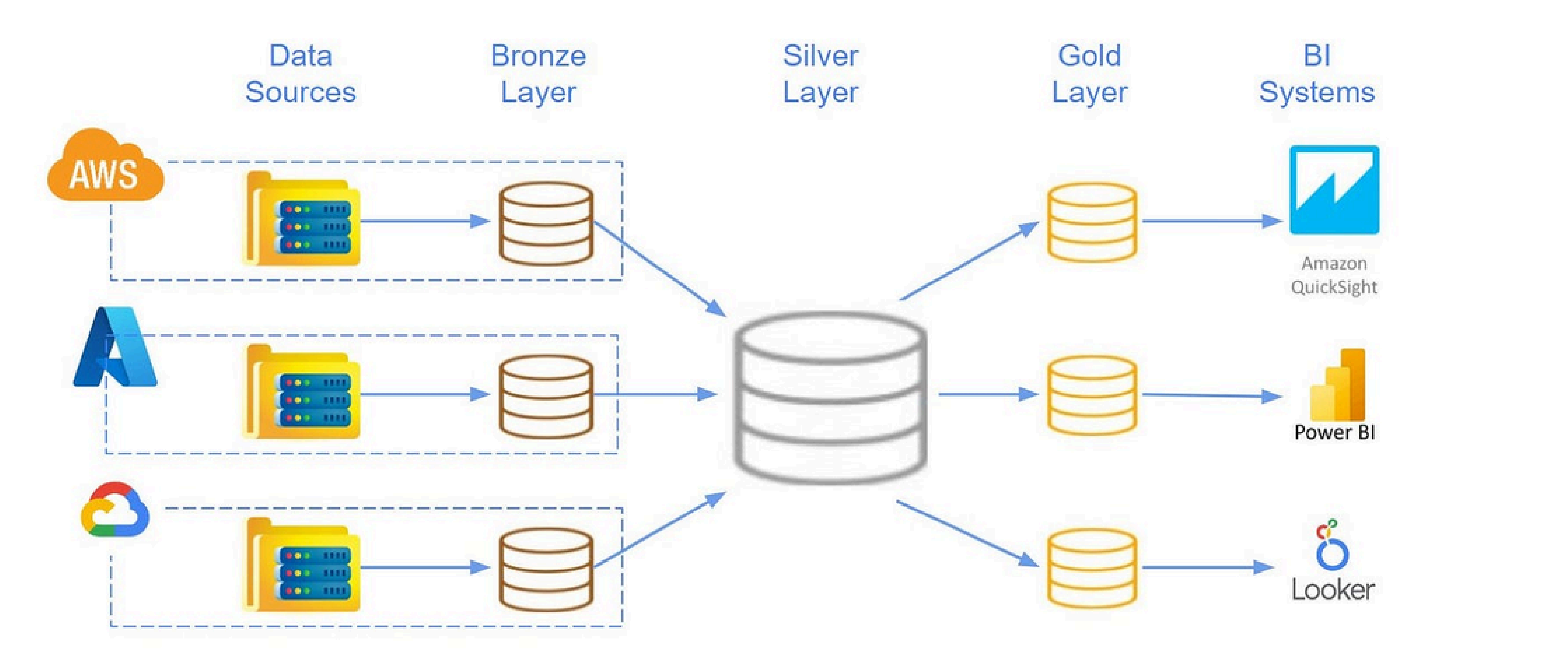
	col_name	data_type
1	Product_Code	string
2	Product_Name	string
3	Product_Category	string

METHODOLOGY

Overall Architecture

The project follows the Medallion Architecture, a layered data design pattern widely used in modern data engineering to progressively improve data quality and analytical value. It consists of three structured layers: Bronze, Silver, and Gold.

- **Bronze Layer:** Stores raw ingested data exactly as received from source systems, preserving original integrity.
- **Silver Layer:** Cleans, standardizes, and enriches data to make it structured, reliable, and analytics-ready.
- **Gold Layer:** Delivers highly curated, business-focused datasets optimized for reporting and advanced analytics.



METHODOLOGY

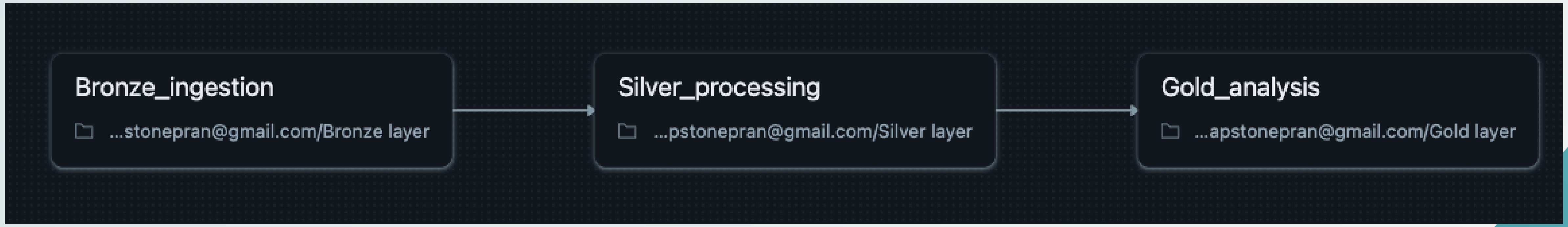
STAGE 1: Databricks platform

Databricks serves as the core data engineering and analytics platform for this project. Built on Apache Spark, it provides powerful distributed computing capabilities to handle large datasets efficiently. Databricks allows collaborative notebook development, scalable compute environments, and seamless data transformation workflows. In this project, Databricks was used to ingest raw global export data, perform large-scale transformations, enrich data using reference datasets, and structure it into layered storage zones.

Step 1: Storing Raw Data in Catalog (Volumes)

We first ingested all global trade datasets into Databricks Unity Catalog Volumes, which serve as secure and scalable storage locations within Databricks.

- A Volume is a storage container inside Unity Catalog used to store raw files such as CSV, JSON, or Parquet.
- It ensures centralized governance, access control, and easy integration with Delta pipelines.


Bronze_ingestion
...stonepran@gmail.com/Bronze layer
Silver_processing
...pstonepran@gmail.com/Silver layer
Gold_analysis
...apstonepran@gmail.com/Gold layer

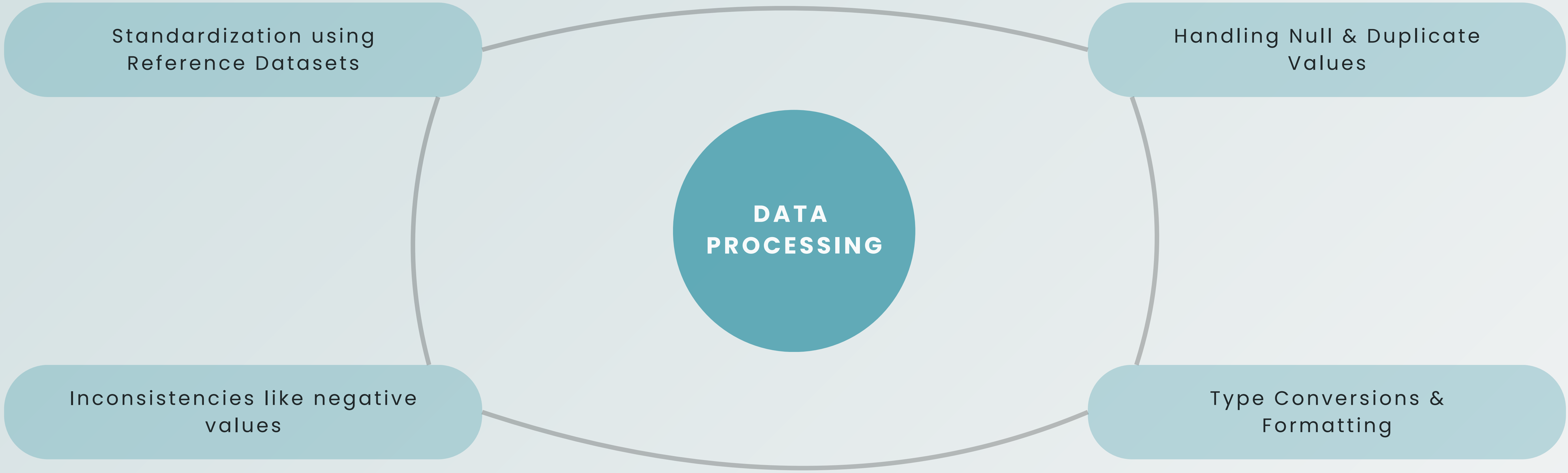
Step 2: Modular Design Using Three Notebooks

To maintain a clean and maintainable workflow, we followed a modular notebook approach:

- Bronze Notebook – Reads raw data from Volumes
- Silver Notebook – Cleans, standardizes, and transforms the data
- Gold Notebook – Creates final analytics-ready datasets

This modular separation improves readability, debugging, scalability, and reusability.

Step 3: Data Processing



DATA PROCESSING

Standardization using Reference Datasets

Handling Null & Duplicate Values

Inconsistencies like negative values

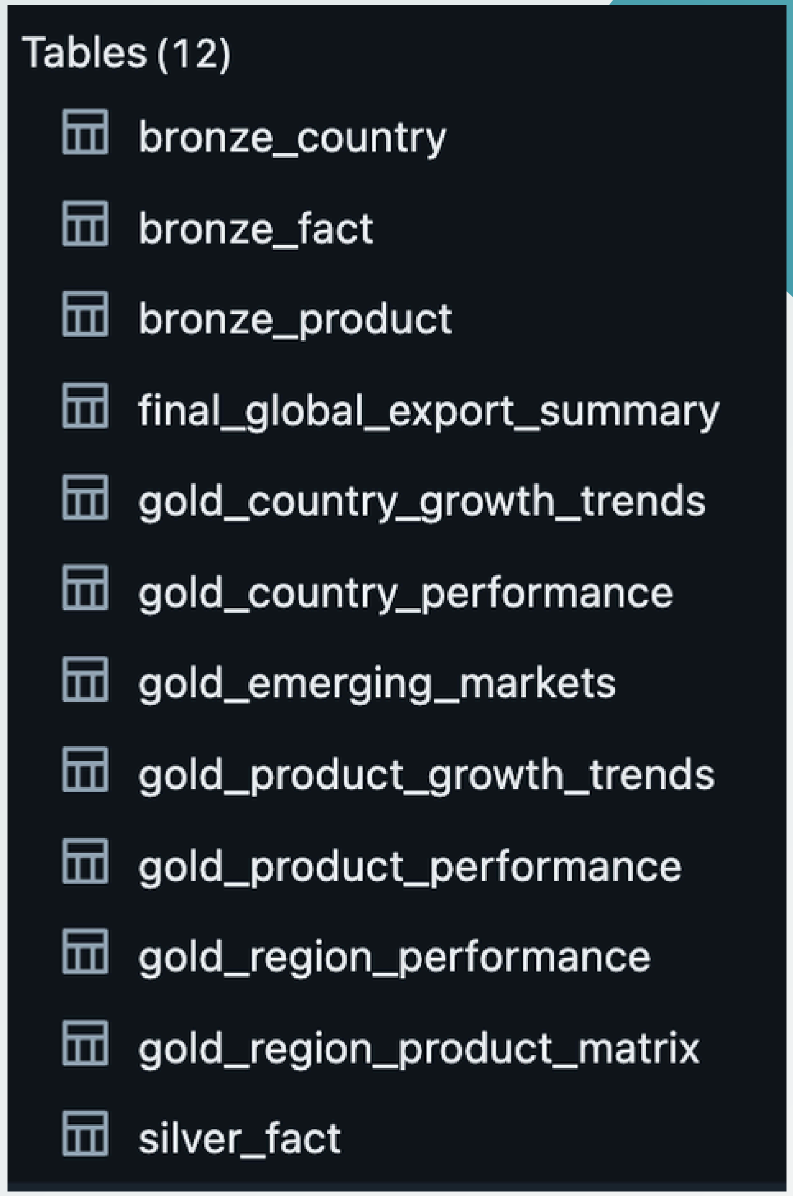
Type Conversions & Formatting

Step 4: Storage of Silver & Gold Data (Catalog Tables)

After processing, we stored refined datasets as Tables inside Unity Catalog.

What is a Table in Databricks?

- A Table is a structured, queryable dataset managed by Unity Catalog.
- It supports versioning, governance, ACID transactions, and optimized querying for analytics.



Tables (12)

bronze_country
bronze_fact
bronze_product
final_global_export_summary
gold_country_growth_trends
gold_country_performance
gold_emerging_markets
gold_product_growth_trends
gold_product_performance
gold_region_performance
gold_region_product_matrix
silver_fact

Step 5: Data Analysis & Insights

After processing the data, analytical models were built to understand trade performance trends. The project enabled trend analysis across multiple countries, time periods, and product segments.

Insights such as

- top exporting countries,
- high-performing product categories,
- growth behaviors,
- economic performance indicators were derived.

STAGE 2: Apache Airflow (Workflow Orchestration)

Apache Airflow acts as the orchestration engine for automating and managing the data pipeline. It enables scheduled execution of Databricks jobs, tracks pipeline success or failure, and ensures reliable workflow execution. Both manual and scheduled executions were supported, ensuring reusability and flexibility. Airflow helped establish a production-like automated ETL pipeline environment instead of running notebooks manually.

Step 1: Establish connection between Databricks and Airflow

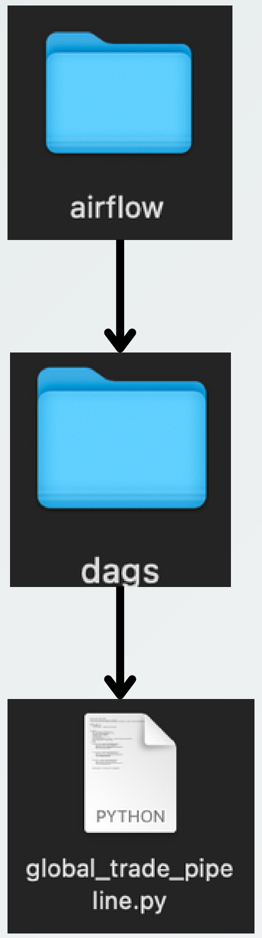
- Create a job for the databricks notebook and note the Job ID
- Generate a Personal Access Token in Databricks that acts as a connector in Airflow
- Create new connection in Airflow and configure details

Step 2: Create DAG

A DAG (Directed Acyclic Graph) in Airflow represents the data pipeline workflow.

It defines:

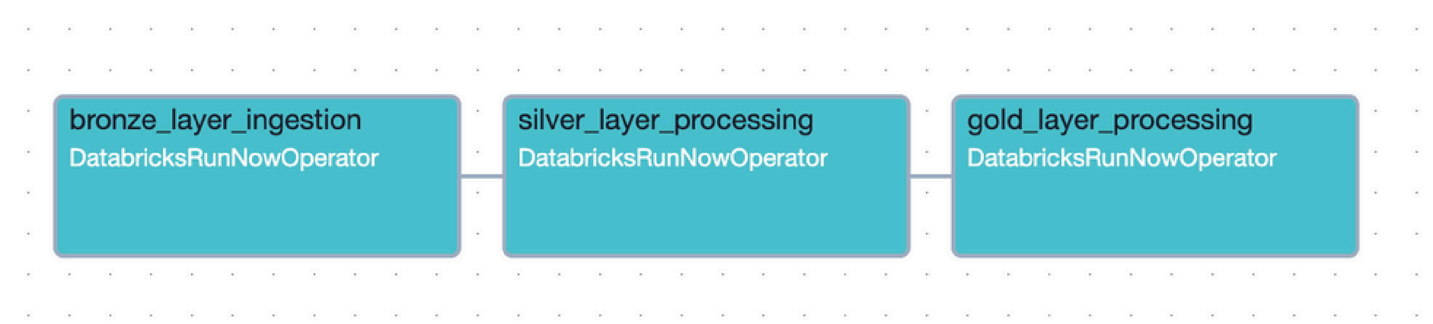
- What tasks need to run
- In what order
- How often
- Which system they connect to



Purpose of Our DAG in the project?

In our project, the DAG was used to:

- Trigger the Databricks ETL Job
- Automate the Bronze → Silver → Gold pipeline
- Allow scheduling (manual / weekly / periodic execution)
- Monitor and re-run failures easily



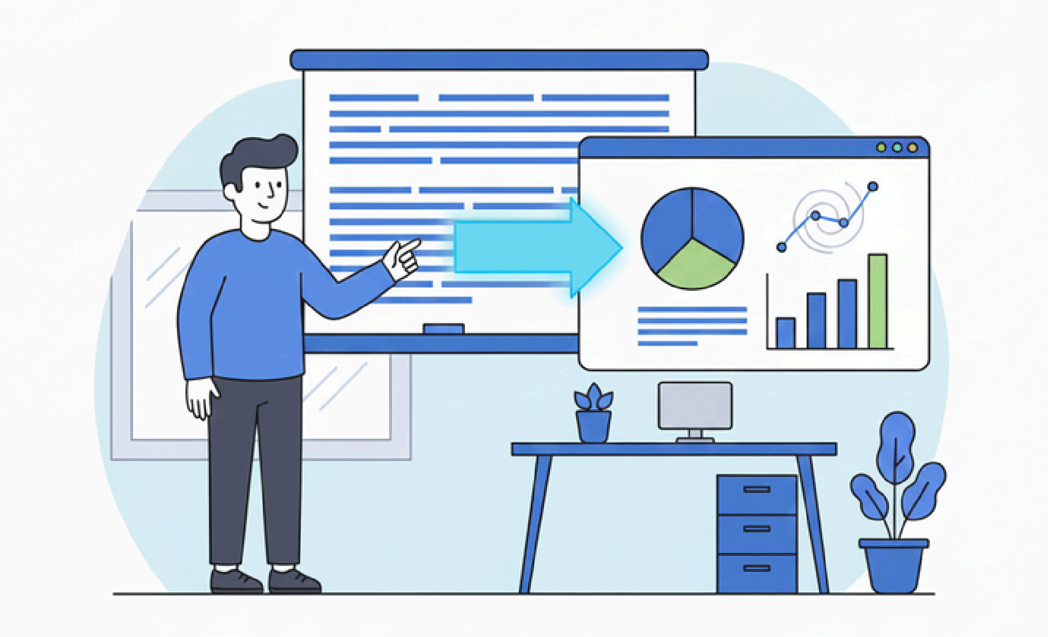
bronze_layer_ingestion
DatabricksRunNowOperator

silver_layer_processing
DatabricksRunNowOperator

gold_layer_processing
DatabricksRunNowOperator

STAGE 3: Power BI Visualization

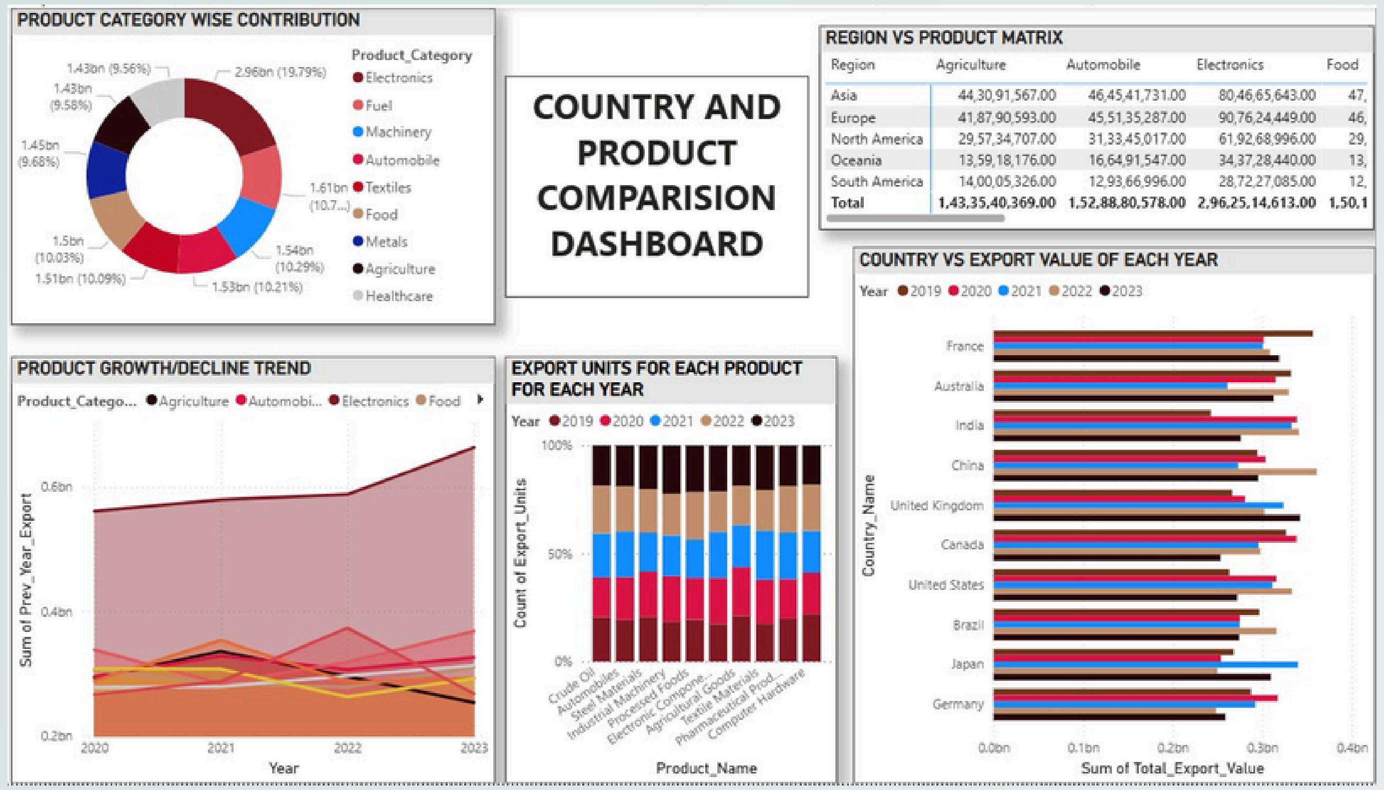
Power BI was used to transform processed Gold-layer data into meaningful visual dashboards. Interactive charts, time-series visualizations, country performance comparisons, and product category distributions were created to support strategic decision-making. Power BI enables users to write custom formulae, filter insights, and analyze trade performance dynamically.



Step 1: Establish connection between Databricks and Power BI

- Generate a Personal Access Token in Databricks that acts as a connector in Power BI
- Note down the Host Name, and HTTP Path from SQL Warehouses → Conection Details
- Select “Get Data” from the Power BI UI and search Azure Databricks
- Paste the configuration details and data from Databricks will be imported

Step 2: Create Dashboards



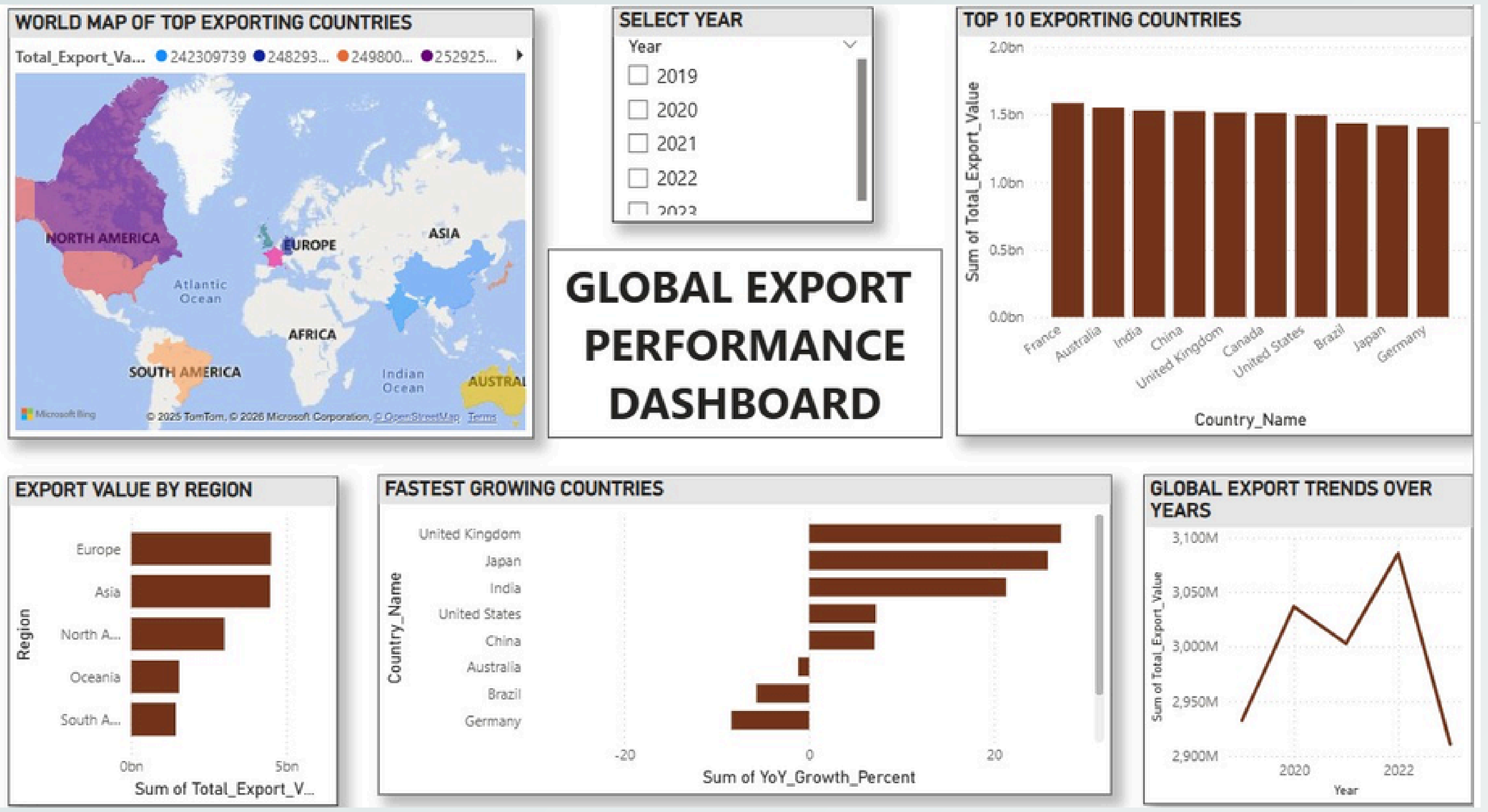
COUNTRY AND PRODUCT COMPARISION DASHBOARD

REGION VS PRODUCT MATRIX

Region	Agriculture	Automobile	Electronics	Food
Asia	44,30,91,567.00	46,45,41,731.00	80,46,65,643.00	47,
Europe	41,87,90,593.00	45,51,35,287.00	90,76,24,449.00	46,
North America	29,57,34,707.00	31,33,45,017.00	61,92,68,996.00	29,
Oceania	13,59,18,176.00	16,64,91,547.00	34,37,28,440.00	13,
South America	14,00,05,326.00	12,93,66,996.00	28,72,27,085.00	12,
Total	**1,43,35,40,369.00**	**1,52,88,80,578.00**	**2,96,25,14,613.00**	**1,50,1**

Country & Product Comparison Dashboard - This dashboard provides a detailed view of how different product categories and countries contribute to global exports, helping analysts understand trade composition, performance variation, and growth trends.

Global Export Performance Dashboard - This dashboard provides a comprehensive analytical overview of global export performance, helping policymakers and analysts understand how different countries and regions are contributing to international trade.

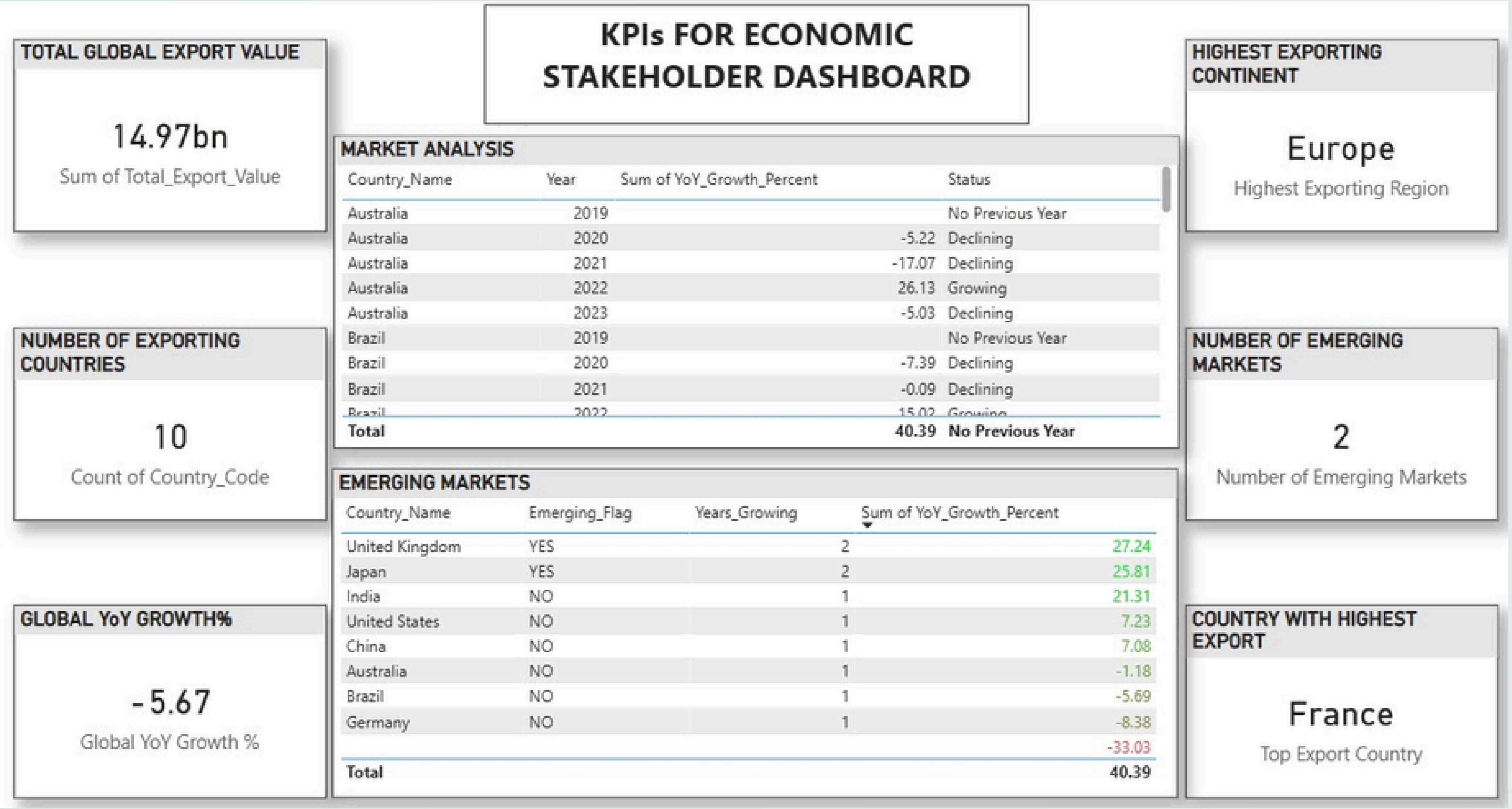


SELECT YEAR

Year

□ 2019
□ 2020
□ 2021
□ 2022
□ 2023

GLOBAL EXPORT PERFORMANCE DASHBOARD



KPIs FOR ECONOMIC STAKEHOLDER DASHBOARD

TOTAL GLOBAL EXPORT VALUE

14.97bn
Sum of Total_Export_Value

NUMBER OF EXPORTING COUNTRIES

10
Count of Country_Code

GLOBAL YoY GROWTH%

-5.67
Global YoY Growth %

MARKET ANALYSIS

Country_Name	Year	Sum of YoY_Growth_Percent	Status
Australia	2019		No Previous Year
Australia	2020	-5.22	Declining
Australia	2021	-17.07	Declining
Australia	2022	26.13	Growing
Australia	2023	-5.03	Declining
Brazil	2019		No Previous Year
Brazil	2020	-7.39	Declining
Brazil	2021	-0.09	Declining
Brazil	2022	15.02	Growing
Total		**40.39**	**No Previous Year**

EMERGING MARKETS

Country_Name	Emerging_Flag	Years_Growing	Sum of YoY_Growth_Percent
United Kingdom	YES	2	27.24
Japan	YES	2	25.81
India	NO	1	21.31
United States	NO	1	7.23
China	NO	1	7.08
Australia	NO	1	-1.18
Brazil	NO	1	-5.69
Germany	NO	1	-8.38
			-33.03
Total			**40.39**

HIGHEST EXPORTING CONTINENT

Europe
Highest Exporting Region

NUMBER OF EMERGING MARKETS

2
Number of Emerging Markets

COUNTRY WITH HIGHEST EXPORT

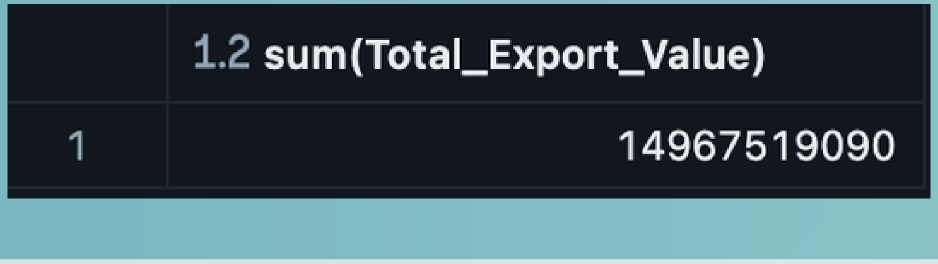
France
Top Export Country

KPI Dashboard for Economic Stakeholders - This dashboard provides a strategic overview of global export performance through key KPIs, growth indicators, and emerging market insights. It helps policymakers and trade authorities evaluate economic strength, identify risk, and recognize emerging opportunities in global trade.

Step 3: Validate Dashboards

Total global export value

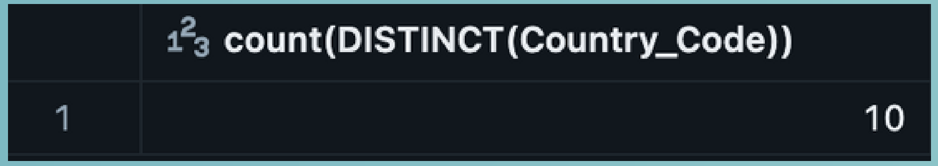
```
select sum(Total_Export_Value) from
gold_country_performance;
```



	1.2 sum(Total_Export_Value)
1	14967519090

Number of exporting countries

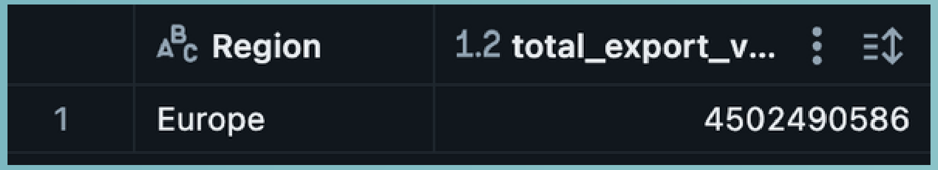
```
select count(distinct(Country_Code)) from
gold_country_growth_trends;
```



	count(DISTINCT(Country_Code))
1	10

Highest Exporting Continent

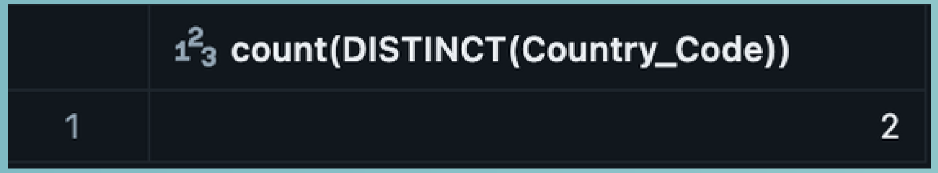
```
select Region, sum(Total_Export_Value) AS
total_export_value from gold_region_performance
group by Region order by total_export_value DESC
limit 1;
```



	Region	1.2 total_export_v...
1	Europe	4502490586

Number of emerging markets

```
select count(distinct(Country_Code)) from
gold_emerging_markets where
EMERGING_FLAG='YES';
```



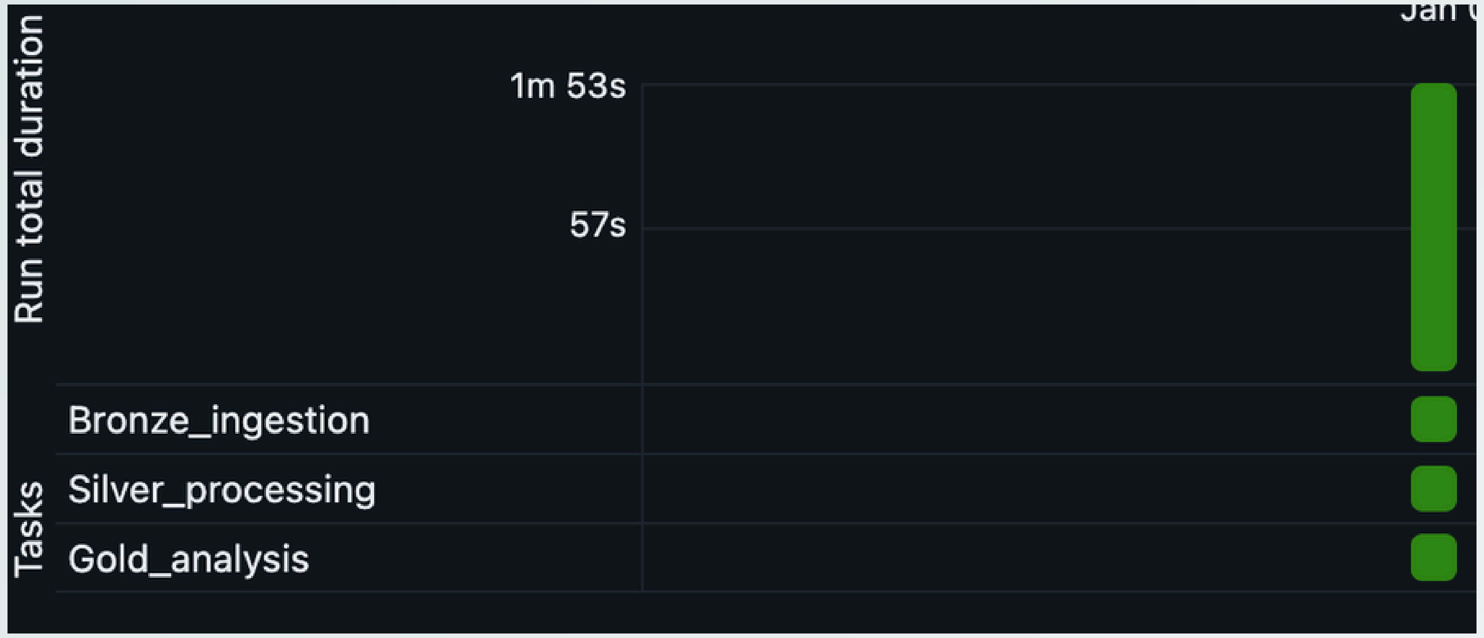
	count(DISTINCT(Country_Code))
1	2

Results & Conclusion

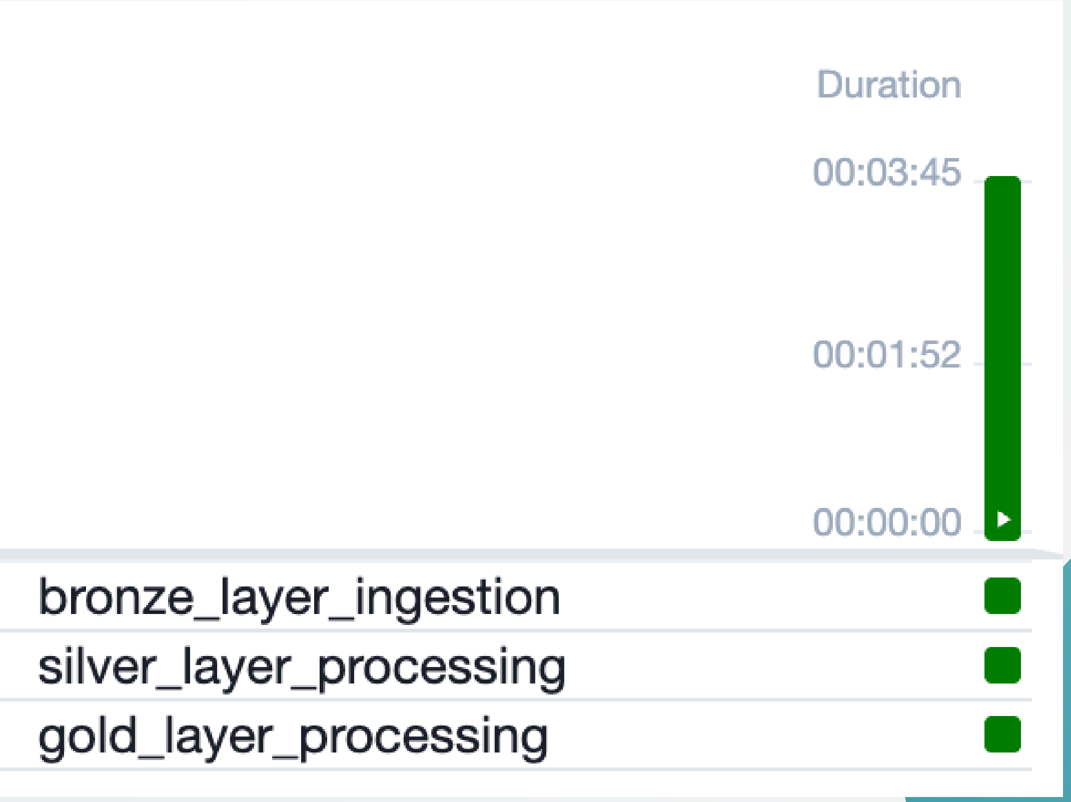
When the DAG is triggered in the Apache Airflow, the workflow starts running by accessing the notebooks from Databricks. Automatically, the jobs also runs in Databricks when airflow job is triggered without manual triggering. Hence, the workflow is automated.

Through this project, we have succesfully automated an enterprise level ETL process that efficiently ingests, cleans, and transforms global trade data using Databricks and orchestrates the workflow throughApache Airflow. The processed Gold-layer data was then utilized to build insightful and interactive Power BI dashboards to represent data visually. Overall, this project demonstrates how modern tools can support data-driven business decisions.

Databricks UI

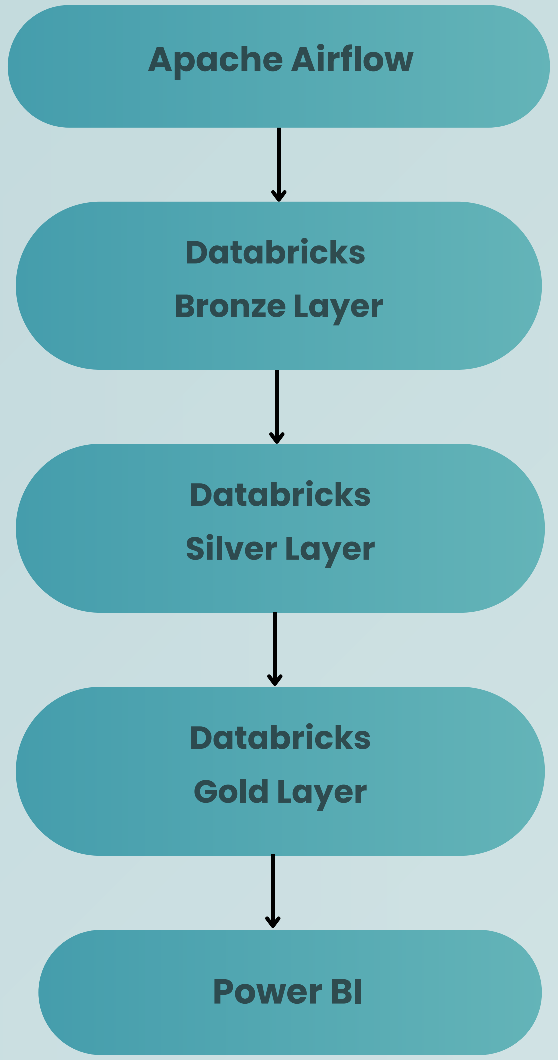


Airflow UI



Results & Conclusion

END - TO- END PIPELINE FLOW



Apache Airflow

Databricks
Bronze Layer

Databricks
Silver Layer

Databricks
Gold Layer

Power BI

Thank You