**Data Pipeline and Sales Overview Dashboard Using Azure and Power BI**

1. **Introduction**

This project aims to develop a **Sales Overview Dashboard** that analyzes the **Brazilian E-Commerce Public Dataset** available on Kaggle. The purpose of this project is to transform raw data into meaningful insights that can assist stakeholders in **data-driven decision-making**.

The project simulates a real-world business scenario, covering the complete **data pipeline process** from data collection and preparation to visualization and insight generation. The final output is an **interactive Power BI dashboard** that evaluates and models the data, highlighting key metrics and patterns that can guide business strategies.

Although this is a relatively simple project, the focus is on demonstrating a structured **data processing workflow**, emphasizing how data moves from its raw form to a polished, decision supporting visualization.

1. **Data Pipeline and Methodology**

The project follows the **Medallion Architecture (Bronze–Silver–Gold model)**, which simulates how businesses integrate, process, and analyze data using cloud-based tools. Data was collected from multiple sources, including **GitHub repositories** and a **MySQL database**, then processed through successive transformation layers within the **Azure ecosystem**.

|  |
| --- |
|  |
| Figure 1 Azure Medallion Architecture (Bronze–Silver–Gold Data Pipeline) |
| This figure illustrates the Medallion Architecture implemented using Azure services, showing the data flow from raw sources through Azure Data Factory, Databricks, and Synapse to Power BI for visualization. |

* 1. **Data Ingestion (Bronze Layer – Raw Data)**

All raw data was ingested through **Azure Data Factory (ADF)** and stored in **Azure Data Lake Storage (ADLS)**. This layer serves as the **Bronze level** of the pipeline, where data from **GitHub** and **MySQL** is first loaded without modification. It ensures that the original, unprocessed data is preserved for **auditing, lineage, and recovery** purposes. **ADF pipelines** were used to automate the extraction and loading process, efficiently managing data movement between multiple sources and the data lake.

|  |
| --- |
|  |
| Figure 2. Azure Data Factory Pipeline for Data Ingestion |
| This pipeline automates data ingestion from GitHub and MySQL sources. The Lookup activity retrieves file metadata, while the ForEach loops iterate through source files and load them into Azure Data Lake Storage (Bronze layer) for raw data storage and further processing. |

* 1. **Data Transformation and Cleaning (Silver Layer – Processed Data)**

Once the raw data was stored in **Azure Data Lake Storage (ADLS)**, it was accessed and processed using **Azure Databricks**. This step represents the **Silver layer** of the Medallion Architecture and focuses on refining the data to make it consistent and analysis ready. The main tasks performed in this layer include:

* **Data cleaning** by removing duplicates and handling missing values.
* **Feature engineering** to create meaningful attributes for further analysis.
* **Data standardization** to ensure consistent formats and naming conventions.

After transformation, the cleaned and enriched data was saved back to the **silver zone** in ADLS.

* 1. **Data Modeling and Aggregation (Gold Layer – Analytics-Ready Data)**

After the Silver layer, the curated and refined data was loaded into **Azure Synapse Analytics**, where it was **modeled, aggregated, and optimized** for analytical use. This stage represents the **Gold layer** of the **Medallion Architecture**, where data becomes **analytics-ready** and aligned with the goals of the **Sales Overview Dashboard**.

Within **Azure Synapse**, the data was structured into a **Snowflake schema model**—a normalized design that supports efficient querying, scalability, and high-performance reporting. Additional transformations and aggregations were performed to extract **key business metrics**, including:

* **Sales performance** across time periods and regions.
* **Product category analysis** to highlight best- and worst-performing items.
* **Geographical insights** across different **Brazilian states**.
* **Delivery performance** based on shipping times and reliability.
* **Customer satisfaction trends** derived from review scores and feedback.

Once the data was aggregated and modeled, the **final analytics-ready dataset** was **stored back into the Gold zone of Azure Data Lake Storage (ADLS)**. This ensures centralized storage, consistency, and easy access for visualization.

The curated Gold data was then **connected to Power BI**, enabling the creation of an **interactive Sales Overview Dashboard** that visualizes performance insights, customer satisfaction, and delivery trends to support **data-driven business decisions**.

1. **Data Visualization and Analysis**

Finally, the **Gold-level data** stored in **Azure Data Lake Storage (ADLS)** was **connected directly to Power BI** for visualization and reporting. This connection allows Power BI to **access the curated, analytics-ready dataset** produced from the Gold layer of the **Medallion Architecture**.

Using this data, the **Sales Overview Dashboard** was developed to deliver clear, interactive insights across multiple business dimensions. The dashboard enables stakeholders to explore:

* **Sales performance** across different regions and time periods.
* **Product category comparisons** to identify top-selling and underperforming product categories.
* **State-level insights** highlighting geographical variations in sales and customer behavior.
* **Delivery performance metrics**, including shipping and delivery rates.
* **Customer satisfaction levels** derived from review scores.
* **Key performance indicators (KPIs)** such as total revenue, Profit and Cost.

|  |
| --- |
| **Note:** To improve visualization consistency and performance, minor adjustments were made to the dataset during dashboard preparation. These modifications were implemented solely for visualization purposes and do not affect the overall data integrity. For analytical or research use, it is recommended to refer to the original dataset available on Kaggle. |

This **end-to-end data pipeline**—from collection to visualization—demonstrates a complete **ETL and analytics architecture**, ensuring that every stage contributes to **accurate, insightful, and actionable business intelligence**.

|  |  |
| --- | --- |
| **Bronze Level – Raw Operational Data** | **Gold Level – Analytics-Ready Data** |
|  |  |
| Figure 3. Bronze and Gold Levels of the Medallion Architecture (Star Schema Model) | |
| This diagram illustrates the transition from the Bronze Level, where raw relational tables such as Orders, Customers, and Products are stored, to the Gold Level, where data is modelled into a Star Schema. The central FactOrders table connects directly to dimension tables (DimDate, DimProducts, DimCustomers, DimSellers, etc.), enabling efficient querying and reporting in Power BI. | |

1. **Tools and Technologies Used**

* **Azure Data Factory (ADF):** Orchestrated the ingestion process from GitHub and MySQL into Azure Data Lake (Bronze layer) using Lookup and ForEach activities for automated extraction and loading.
* **Azure Data Lake Storage (ADLS):** Provided multi-layer storage for raw, cleaned, and modeled data (Bronze, Silver, Gold).
* **Azure Databricks:** Used for data cleaning, handling duplicates, performing feature engineering, and applying transformations using **PySpark**.
* **Azure Synapse Analytics:** Used for **SQL-based data modeling**, creating **fact and dimension tables**, and implementing the **Snowflake schema** for efficient querying and aggregation.
* **Power BI:** Connected to the **Gold data in ADLS** for visualization. Additional **DAX measures, calculated columns,** and **custom tables** were created to improve analytical capability and visual performance.

1. **Project Architecture Summary**

The architecture follows the Medallion (Bronze–Silver–Gold) design, enabling modular and maintainable data flow across Azure services.

* **Bronze Layer (ADF + ADLS):** Raw data ingestion from GitHub and MySQL sources.
* **Silver Layer (Databricks):** Data cleaning, deduplication, and enrichment through PySpark notebooks.
* **Gold Layer (Synapse + ADLS):**
  + SQL queries were used to **aggregate key business metrics** such as total revenue, average delivery time, and review averages.
  + Relationships were defined between the **Fact table** and multiple **Dimension tables**, following a **Snowflake structure** to support normalized and efficient data modeling.
  + The final curated data was stored back in **ADLS (Gold zone)** for Power BI access.

1. **Conclusion**

This project demonstrates an end-to-end implementation of the Azure Medallion Architecture for e-commerce analytics, integrating ADF, Databricks, Synapse, ADLS, and Power BI.  
The inclusion of SQL-based modeling in Synapse and DAX-driven calculations in Power BI ensures data consistency, optimized performance, and meaningful interactivity across dashboards.  
The approach highlights a practical, scalable framework for building enterprise-grade analytics solutions capable of supporting real-time decision-making.

1. **References**

Olist (2018) *Brazilian E-Commerce Public Dataset by Olist*. Kaggle. Available at: [https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce](https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce?utm_source=chatgpt.com) (Accessed: 25 September 2025)

Microsoft. (2024) *OPENROWSET (Transact-SQL)*. Microsoft Learn. Available at: [https://learn.microsoft.com/en-us/sql/t-sql/functions/openrowset-transact-sql?view=sql-server-ver17](https://learn.microsoft.com/en-us/sql/t-sql/functions/openrowset-transact-sql?view=sql-server-ver17&utm_source=chatgpt.com) (Accessed: 2 October 2025)

Microsoft. (2025) *Secure a data lakehouse with Azure Synapse Analytics*. Microsoft Learn. Available at: [https://learn.microsoft.com/en-us/azure/architecture/example-scenario/analytics/secure-data-lakehouse-synapse](https://learn.microsoft.com/en-us/azure/architecture/example-scenario/analytics/secure-data-lakehouse-synapse?utm_source=chatgpt.com) (Accessed: 5 October 2025)