DATA LAKE ANALYTICS

TABLE OF CONTENTS

1. INTRODUCTION

1.1 PURPOSE

The purpose of this project is to design and implement a Data Lake Analytics solution for e-commerce data using Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), Azure Data Lake Analytics (ADLA), and Power BI.

1.2 BACKGROUND

E-commerce generates large volumes of transactional data, including customers, orders, products, and payments. To extract insights, the data must be centralized, cleaned, and transformed into analytics-ready formats. A layered Medallion Architecture (Bronze → Silver → Gold) was implemented to ensure data quality, scalability, and governance.

1.3 SCOPE

In-Scope:

- Design and implement data pipelines using ADF to ingest raw e-commerce data into ADLS Bronze.
- Transform and validate data into Silver and Gold layers using ADF & ADLA.
- Build Power BI dashboards to deliver business insights from curated datasets.

Out-of-Scope:

- Real-time streaming pipelines and advanced AI/ML models.
- Integration with Databricks or other third-party analytics tools.
- Enterprise-level governance and multi-region deployment.

2. PROBLEM STATEMENT

The organization struggles with fragmented e-commerce data stored across multiple sources, causing inconsistent metrics and delays in reporting. To resolve this, there is a need for an automated ETL pipeline that consolidates data, ensures quality, and maintains a single source of truth. This project leverages Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), and Azure Data Lake Analytics (ADLA) to provide clean, structured, and analytics-ready data for Power BI dashboards.

3. OBJECTIVES

- Create an automated data ingestion pipeline with ADF.
- Store data securely in Azure Data Lake Storage.
- Implement a Bronze → Silver → Gold data architecture.
- Enable Power BI dashboards for decision-making.
- Ensure scalability, automation, and governance.

4. TOOLS & TECHNOLOGIES USED

4.1 AZURE DATA FACTORY (ADF):

Used to orchestrate and automate pipelines for ingesting raw data from CSV files, databases, and APIs into Azure Data Lake. Also handled scheduling, triggers, and monitoring of workflows.

4.2 AZURE DATA LAKE STORAGE (ADLS GEN2):

Served as the centralized repository following the Medallion Architecture (Bronze → Silver → Gold). Stored raw data in Bronze, cleaned data in Silver, and business-ready datasets in Gold.

4.3 AZURE DATA LAKE ANALYTICS (ADLA):

Performed data processing, transformations, and aggregations on ingested datasets. Used to run SQL-based queries for cleaning, standardization, and preparing curated datasets for reporting.

4.4 POWER BI:

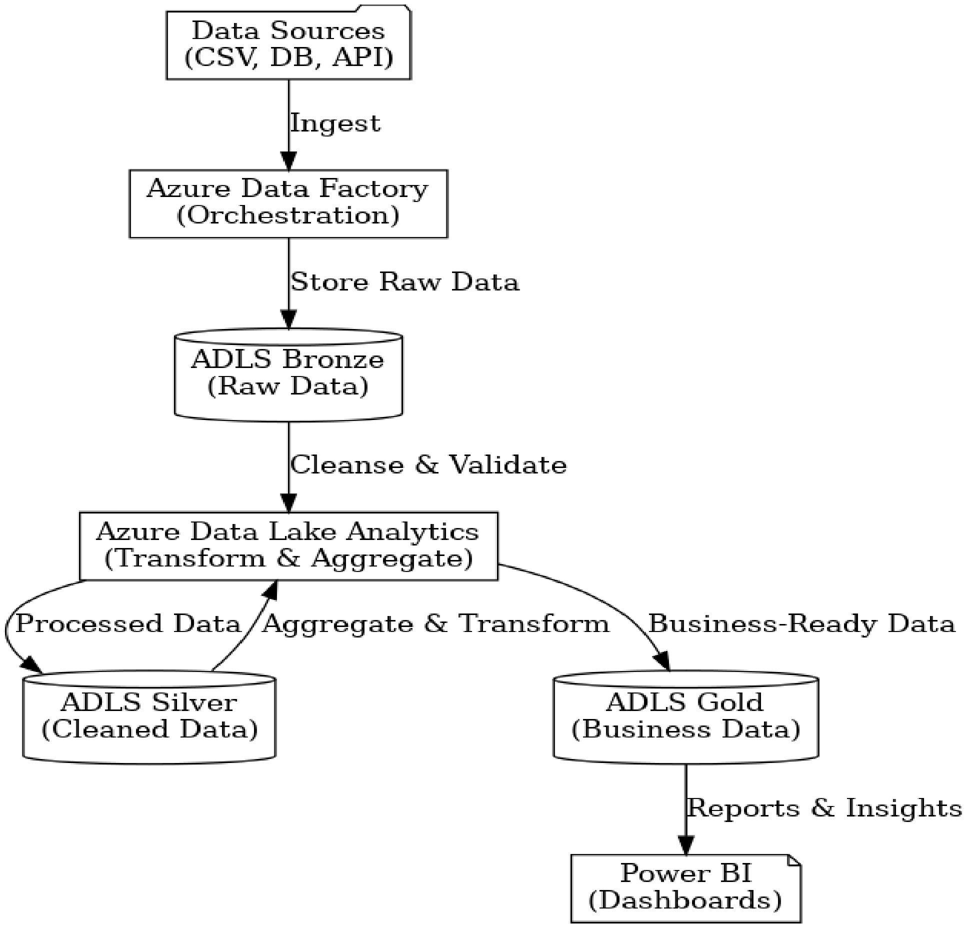
Connected to the Gold Layer to build interactive dashboards and reports, delivering insights such as sales trends, customer behavior, and revenue growth patterns.

5. SOLUTION ARCHITECTURE

5.1 HIGH-LEVEL DESIGN

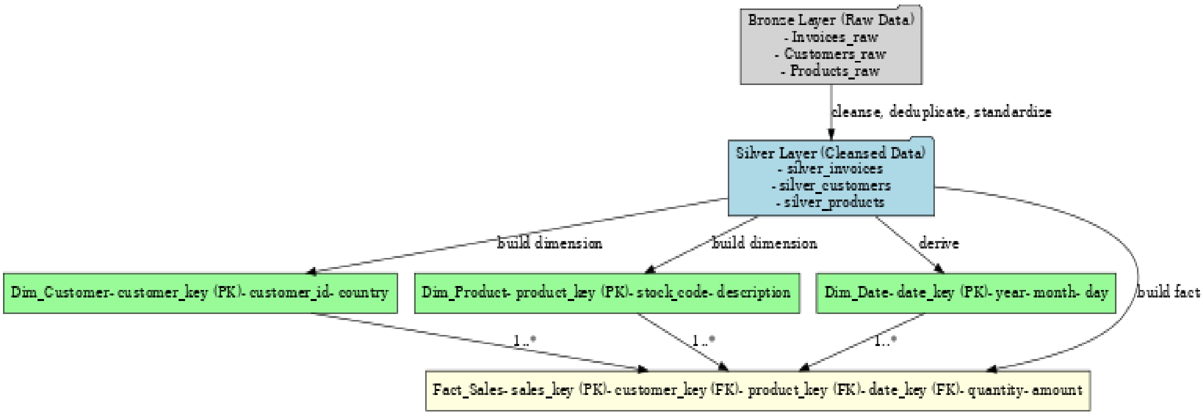
The solution follows the Medallion Architecture (Bronze → Silver → Gold) for structured data processing. ADF pipelines ingest raw data into ADLS (Bronze), then clean and standardize it into Silver, and aggregate it into Gold. Power BI dashboards consume Gold layer data to provide business-ready insights and reporting.

5.2 DATA FLOW DIAGRAM



5.3 ENTITY RELATIONSHIP DIAGRAM

- Customers → CustomerKey (PK), CustomerID, Country
- Products → ProductKey (PK), StockCode (ProductID), Description, UnitPrice
- Orders (Invoices) → OrderKey (PK), InvoiceNo, InvoiceDate, CustomerKey (FK)
- OrderDetails → OrderDetailKey (PK), OrderKey (FK), ProductKey (FK), Quantity, UnitPrice, LineTotal


Bronze Layer (Raw Data)
- Invoices_raw
- Customers_raw
- Products_raw
cleanse, deduplicate, standardize
Silver Layer (Cleansed Data)
- silver_invoices
- silver_customers
- silver_products
build dimension
build dimension
derive
build fact
Dim_Customer- customer_key (PK)- customer_id- country
Dim_Product- product_key (PK)- stock_code- description
Dim_Date- date_key (PK)- year- month- day
1..*
1..*
1..*
Fact_Sales- sales_key (PK)- customer_key (FK)- product_key (FK)- date_key (FK)- quantity- amount

6. EXECUTION OVERVIEW

6.1 ENVIRONMENT SETUP

Provisioned Azure Data Factory (ADF), Azure Data Lake Storage (ADLS), and Azure Data Lake Analytics (ADLA). Created Bronze, Silver, and Gold containers in ADLS. Configured linked services, datasets, and secured access with SAS tokens.

6.2 DATA INGESTION

Implemented ADF pipelines to ingest raw CSVs and API data into the Bronze Layer of ADLS. Handled credential issues by updating linked services and configuring Key Vault-based authentication.

6.3 DATA TRANSFROMATION

Used ADF Mapping Data Flows and ADLA queries to clean nulls, remove duplicates, standardize formats, and apply business rules. Curated outputs were stored in the Silver Layer for structured, validated datasets.

6.4 DATA LOADING

Aggregated and transformed Silver datasets into business-ready fact and dimension tables (e.g., dim_customer, dim_product, dim_date, fact_sales) stored in the Gold Layer for analytics.

6.5 ANALYTICS & POWER BI

Connected Power BI to the Gold Layer datasets to create interactive dashboards and reports. Visualizations included KPIs such as total sales, revenue growth trends, top products, and customer segmentation, enabling stakeholders to make data-driven decisions.

7. TASKS PERFORMED

7.1 ENVIRONMENT SETUP TASKS

- Created Azure resource group, provisioned ADF, ADLS, and ADLA services.
- Configured Bronze, Silver, and Gold containers in ADLS.
- Established linked services, datasets, and secured access using SAS tokens and Key Vault.

7.2 DATA INGESTION TASKS

- Designed ADF pipelines to ingest raw CSV files and API data into the Bronze Layer.
- Handled credential issues by reconfiguring linked services and authentication.
- Scheduled pipelines for automated batch ingestion.

7.3 TRANSFORMATION TASKS

- Implemented ADF Mapping Data Flows and ADLA queries to clean nulls, remove duplicates, and apply data validation rules.
- Standardized data formats and ensured schema consistency.
- Stored curated outputs in the Silver Layer for analytics.

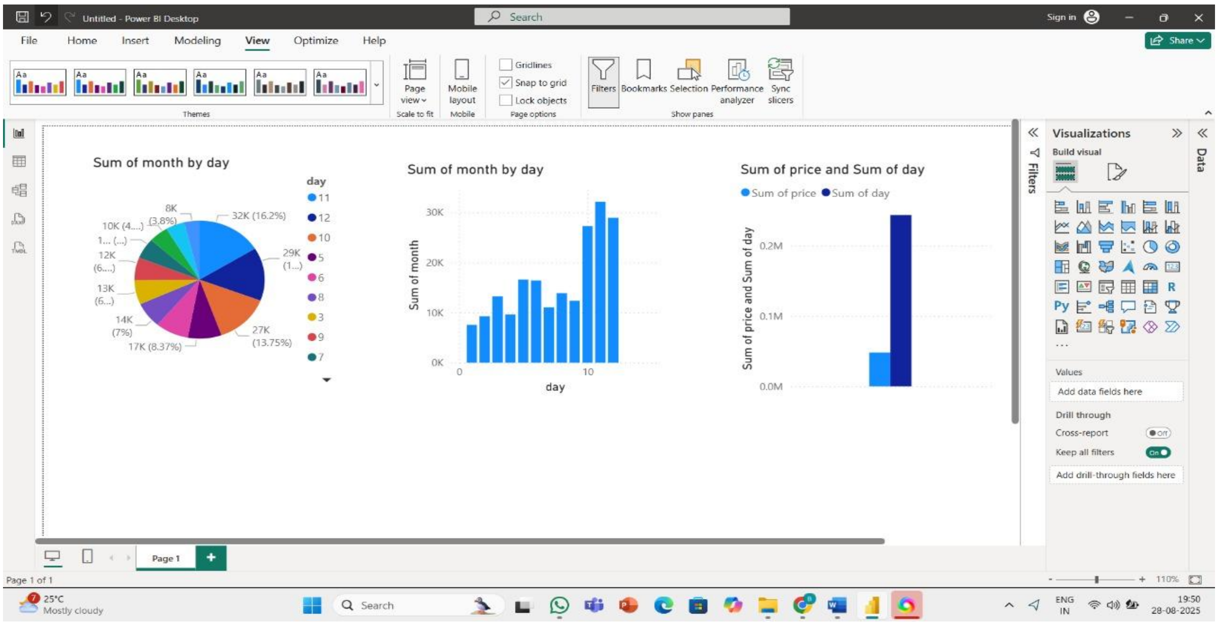
7.4 DATA LOADING TASKS

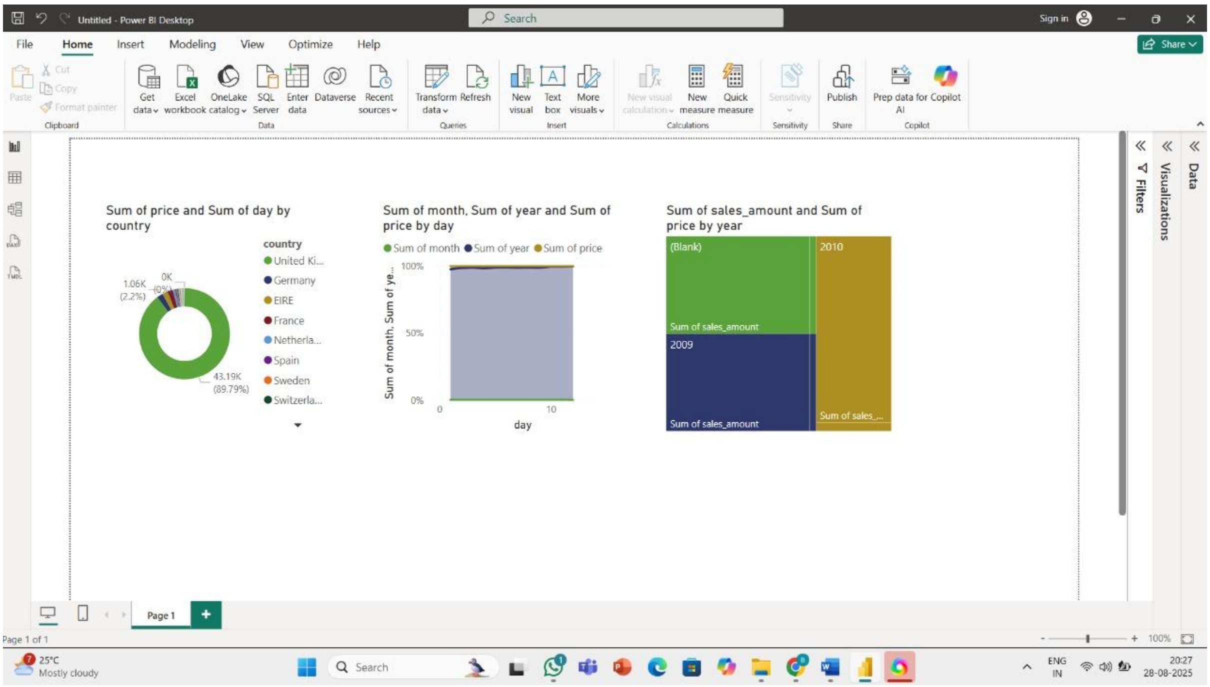
- Designed fact and dimension tables (e.g., dim_customer, dim_product, dim_date, fact_sales).
- Aggregated Silver data into the Gold Layer for business reporting.
- Optimized data structures for Power BI consumption.

7.5 POWER BI TASKS

- Connected Power BI to the Gold Layer datasets.
- Built interactive dashboards and reports showing KPIs such as total sales, revenue trends, top products, and customer segmentation.
- Configured scheduled refresh to keep reports up to date for decision-making.

8. RESULTS AND INSIGHTS





- The United Kingdom contributed nearly 90% of total sales, making it the dominant market compared to other countries.

- Day-wise analysis showed a steady increase in sales, with peak activity observed around the middle of the month.

- Year-wise comparison indicated that 2010 recorded significantly higher sales volumes than 2009, reflecting strong growth.

- A few top-performing days accounted for a large share of monthly revenue, suggesting seasonal or event-driven demand spikes.

9. KEY LEARNINGS

- Gained hands-on experience in implementing the Medallion Architecture (Bronze → Silver → Gold) for structured data processing.
- Understood the importance of data quality checks such as handling nulls, duplicates, and schema mismatches before analytics.
- Learned how ADF pipelines automate ETL workflows and how ADLA queries handle transformations efficiently at scale.
- Developed practical skills in Power BI dashboarding, converting raw datasets into KPIs and actionable business insights.
- Realized that automation, monitoring, and validation are critical for building a reliable and trustworthy data platform.

10. CHALLENGES AND SOLUTIONS

10.1 DATA ACCESSIBILITY ERRORS

Faced issues linking ADF pipelines with ADLS; resolved by reconfiguring linked services and access permissions.

10.2 CREDENTIAL & DEPENDENCY CHALLENGES

Initial ingestion pipeline failures due to dependency misconfigurations; resolved by sequencing pipelines with triggers and ensuring proper service principal access.

10.3 SCHEMA MISMATCHES

Source data had inconsistent formats; solved by applying ADF Mapping Data Flows and schema validation in ADLA.

11. DELIVERABLES

- Requirement Document – Defined objectives, scope, and technical requirements.
- Project Overview Document – Outlined architecture, tools, and methodologies.
- Execution Overview Document – Detailed pipelines, transformations, and loading processes.
- Results & Insights Document – Captured business KPIs and Power BI dashboards.
- Final Project Report – Consolidated all documentation with diagrams and outcomes.
- PowerPoint Presentation – Summarized approach, architecture, results, and key learnings.

12. FUTURE SCOPE

- Implement real-time ingestion using Azure Event Hubs or Stream Analytics.
- Extend with AI/ML models for churn prediction, recommendations, and forecasting.
- Enhance data governance with role-based access, lineage tracking, and compliance.
- Improve scalability for enterprise-wide, multi-region deployments.
- Integrate with Azure Synapse or Purview for advanced analytics and governance.

13. CONCLUSION

The Data Lake Analytics project successfully delivered a centralized, automated, and scalable data platform using ADF, ADLS, ADLA, and Power BI. Raw data was ingested into the Bronze Layer, cleansed and validated into Silver, and aggregated into Gold for analytics. With Power BI dashboards, the organization gained valuable insights into sales performance, customer behaviour, and revenue trends. The solution improved data quality, reporting efficiency, and decision-making, while laying a solid foundation for future AI/ML and real-time analytics use cases.