ARCHITECTURE DOCUMENT

Project Overview

The objective of this project is to design and implement an **enterprise-grade data engineering pipeline** that ingests raw transactional data, processes it incrementally, enforces data quality and calibration rules, and delivers analytics-ready datasets for business reporting.

The solution is built using:

- **Azure Databricks** for distributed data processing
- **Apache Spark** for scalable transformations
- **Delta Lake** for reliable storage with ACID guarantees
- **Delta Live Tables (DLT)** for declarative pipeline orchestration
- **Power BI** for data visualization and analytics

Architecture Pattern

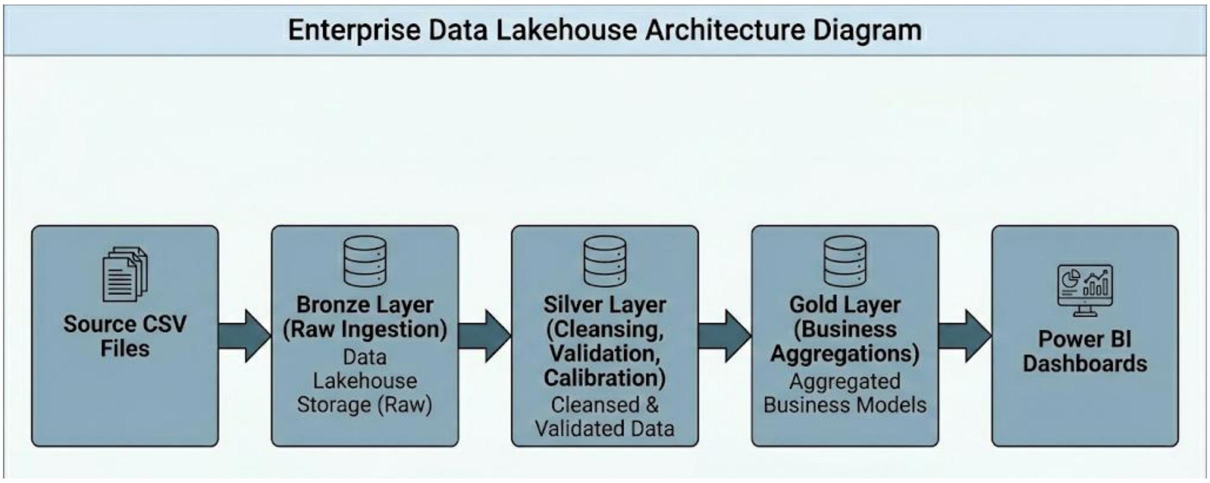
The project follows the **Medallion Architecture**, which separates concerns into three logical layers:

Bronze → Silver → Gold

This layered design ensures:

- Traceability of raw data
- Progressive data quality improvement
- Separation of engineering and analytics concerns

High-Level Architecture Flow


Enterprise Data Lakehouse Architecture Diagram
Source CSV Files
Bronze Layer (Raw Ingestion)
Data Lakehouse Storage (Raw)
Silver Layer (Cleansing, Validation, Calibration)
Cleansed & Validated Data
Gold Layer (Business Aggregations)
Aggregated Business Models
Power BI Dashboards

Component Responsibilities

Azure Databricks

- Provides scalable Spark execution environment
- Hosts DLT pipelines
- Manages clusters and job scheduling

Delta Live Tables (DLT)

- Declaratively defines data pipelines
- Manages dependencies between tables
- Handles incremental processing automatically
- Provides built-in event logging and monitoring

Delta Lake

- Ensures ACID transactions
- Supports schema enforcement and evolution
- Enables Time Travel and rollback
- Prevents partial writes and corruption

Power BI

- Consumes Gold datasets
- Provides interactive dashboards
- Enables business users to explore insights

Architectural Benefits

- **Fault tolerance**: Pipeline failures do not corrupt data
- **Scalability**: Handles large volumes of data
- **Auditability**: Raw data preserved in Bronze
- **Performance**: Optimized analytics using Gold datasets
- **Maintainability**: Clear separation of layers