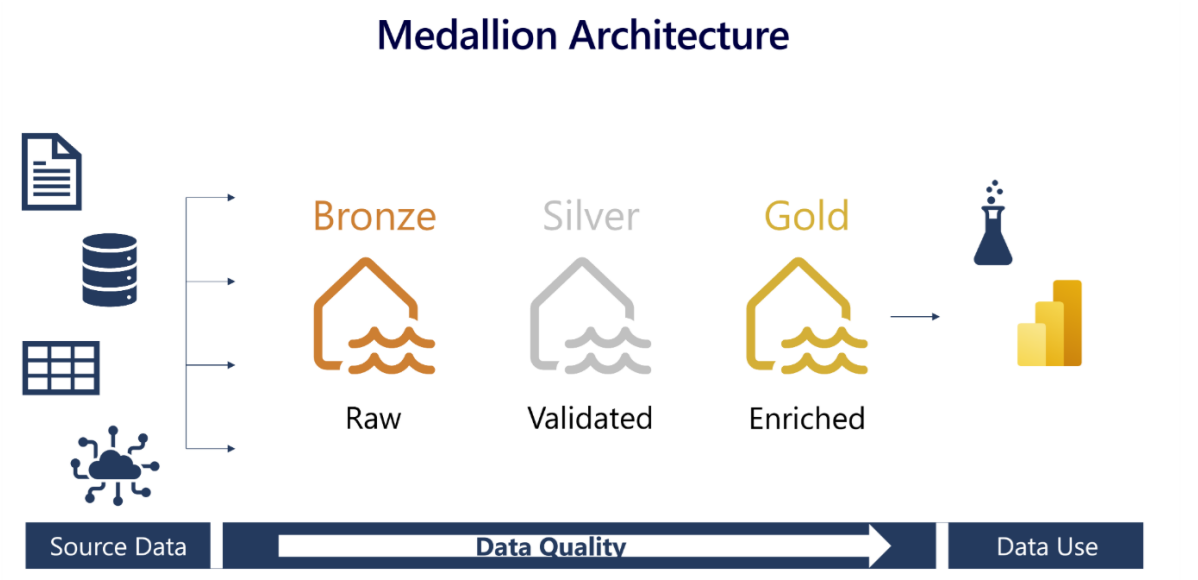
Medallion Architecture Explained

The Medallion Architecture is a layered data design pattern used to organize data processing in data lakes or lakehouses.
It enables a clean, scalable, and reliable flow of data, moving from raw to refined to business-ready stages.

Purpose

- Maintain data quality and traceability
- Enable modular and scalable data pipelines
- Support machine learning, BI, and LLM applications



Bronze Layer → Raw Data

What it contains: Ingested raw data from external sources
Format: Logs, JSON, CSV, Parquet, PDFs, etc.
Operations: Capture as-is data, add ingestion metadata (e.g., source, time)
Examples:

- Web-scraped pages
- Uploaded PDFs
- IoT sensor streams

Silver Layer → Cleaned & Structured Data

What it contains: Refined, deduplicated, and structured data
Operations:

- Data cleansing
- Normalization
- Deduplicated
- Joins

Examples:

- Extracted and cleaned text
- Chunked text with metadata

Gold Layer → Curated, Business-Ready or ML-Ready Data

What it contains: Final curated data for business use or ML/LLM tasks
Operations:

- Aggregations
- Embeddings
- QA generation
- Data Analysis

Examples:

- Vector embeddings
- ML models
- LLM fine-tuning datasets

How It Helps in RAG (Retrieval-Augmented Generation)

- **Bronze**: Store raw documents (PDFs, HTML)
- **Silver**: Extract and clean text, chunk content, add metadata
- **Gold**: Generate embeddings, store in vector DB, build indices

Benefits

- **Reusability**: Build once, reuse across analytics and ML
- **Modularity**: Easy to debug

- **Traceability**: Track data lineage
- **Scalability**: Handle millions of records efficiently

How They Are Connected: End-to-End Flow

1. Data Ingestion (ETL starts here)

- **Tools Used**: Azure Data Factory, Databricks Auto Loader, Kafka
- **Source Systems**: APIs, Databases, IoT Devices, Logs

2. Bronze Layer (Raw Layer, Extracted raw data)

- Stored in: **Data Lake** (e.g., **ADLS** on Azure)
- Format: JSON, CSV, Parquet, etc.
- Purpose: Capture raw, unprocessed data

3. Silver Layer (Clean Layer,Transformations begin for cleaning/enrichment)

- Stored in: **Delta Lake** on top of **ADLS**
- Format: Delta Tables
- Processed via: Spark, Databricks, Azure Synapse
- Purpose: Apply joins, filters, schema enforcement, validation

4. Gold Layer (Business Layer)

- Stored in: **Delta Lake** (also queried via **Lakehouse engines**)
- Use: Reporting, ML, BI, dashboards
- Data is now structured, enriched, trusted

5. Business Consumption Layer(Consumed by)

- **BI Tools**: Power BI, Tableau
- **ML Models**: Azure ML, Databricks ML
- **Analytics**: SQL endpoints via Lakehouse engine

6. Lakehouse Integration

- The **Lakehouse** acts as the central system that:
    - Uses **Delta Lake** as storage engine
    - Manages **Medallion layers**