databricks

Medallion Architecture

Aug 2024

Agenda

Medallion Architecture

- Introduction
- Layers of the architecture
- Benefits and drawbacks
- Different flavors of the architecture
- Q&A

Introduction

What it is the Medallion Architecture

Introduction

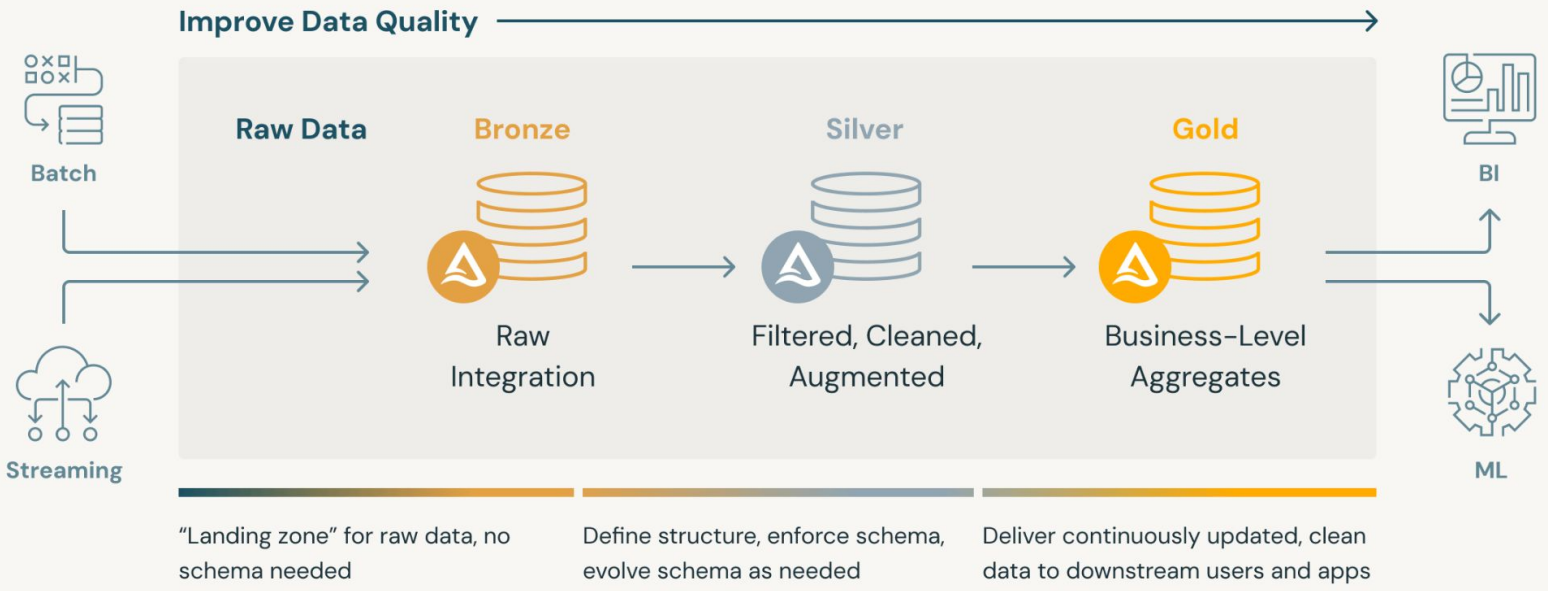
- Logically organize data in a lakehouse into multiple layers

> **Goal:** incrementally and progressively improving the structure and quality of data as it flows through each layer of the architecture.

- More of design pattern than an architecture.
- Sometimes referred to as "multi-hop" architectures
- Not new, exists since 80s/90s (staging, integration, data marts)

Layers of Medallion Architecture

Introduction



Bronze Layer

Bronze Layer

Description

- **Purpose:** Land all raw data from external systems as fast as possible into the data lake.
- **Data Characteristics:** Raw data with minimal to no transformation, even if it is dirty. Usually loaded incrementally, growing over time.
- **Storage:** Stored in a columnar format such as Parquet or Delta.
- **Operations:** Append only, full overwrite, or possibly partitioned by date.

Bronze Layer

Data Model

- **Schema:** "As is" along with possible additional metadata columns that may capture load time, source file name, etc...
- **Ingestion options:**
  - A 1 to 1 mapping of source dataset to bronze table
  - A many to 1 mapping of sources dataset to 1 bronze table

Bronze Layer

Advantages

- **Raw Data:** By ingesting data "raw", we avoid bugs in the system or processing logic. We have the data as it originally existed and can always "go back" to it for both current and future projects.
- **Historical Archive:** Provide a historical archive of source data (cold storage) with possible CDC changes from source system.
- **Lineage and Auditing:** Ensure data lineage, auditability, and reprocessing if needed without rereading the data from the source system.

Silver Layer

Silver Layer

Description

- **Purpose:** Transform data from the Bronze layer by matching, merging, conforming, and cleansing it ("just-enough") to provide a single source of truth "Enterprise view" of all key business entities, concepts, and transactions.
- **Data Characteristics:** Data is validated, filtered, cleaned, augmented, deduplicated, and enriched. Errors are fixed, business data is added, and minimal business rules are applied to improve quality.

Silver Layer

Description

- **Storage:** Ideally stored in Delta format.
- **Operations:**
  - Schema enforcement and evolution are applied. Data validation rules ensure no nulls, uniqueness, correct type and format, and logical consistency.
  - Data from different source systems is generally not joined together yet but may be enriched with reference data.

Silver Layer

Data Model & Advantages

- **Data Model:** More 3rd-Normal Form like data models.
- **Advantages:**
  - Acts as a “single source of truth” for the enterprise across many projects.
  - Enables self-service analytics for ad-hoc reporting, advanced analytics, and ML.

Gold Layer

Gold Layer

Description

- **Purpose:** Organize data in consumption-ready "project-specific" databases.
- **Data Characteristics:** Data is transformed for specific use cases and business-level aggregation is applied. Data from different source files or systems may also be joined together.
- **Storage:** Uses more de-normalized and read-optimized data models with fewer joins.
- **Operations:** The final layer of data transformations and data quality rules are applied.

Gold Layer

Data Model & Advantages

- **Data Model:** Kimball style star schema-based data models or Inmon style Data marts fit.
- **Advantages:**
  - Consumption and reporting.
  - Create projects and analysis to answer business problems via enterprise and departmental data projects.

Benefits of Medallion Architecture

Improved Scalability and Data Quality

Benefits of Medallion Architecture

- **Segregation of Raw Data:** Raw data is stored in the Bronze layer, ensuring it is preserved in its original form.
- **Transformations and Validations:** Occur in the Silver layer, where data is cleaned, standardized, and enriched.
- **Quality Assurance:** The Gold layer ensures that only high-quality, aggregated data is available for business use.
- **Independent Scaling:** Each layer can be scaled independently based on data volume and processing needs.

Flexibility & Enhanced Performance

Benefits of Medallion Architecture

- **Supports Diverse Workloads:** Handles various data processing and analytics tasks, from simple reporting to complex machine learning.
- **Adaptability:** Can be tailored to meet specific business requirements and changing needs.
- **Improved Query Performance:** By organizing data into optimized layers, the architecture ensures faster data retrieval and analysis.

Simplified Data Governance

Benefits of Medallion Architecture

- **Schema Enforcement:** Ensures data adheres to predefined schemas, preventing ingestion of erroneous data.
- **Data Lineage:** Tracks data flow and transformations across layers, providing visibility and auditability.
- **Access Controls:** Granular permissions and security measures for each layer ensure data security and compliance.

Drawbacks of Medallion Architecture

Drawbacks of Medallion Architecture

Cons of the architecture

- **Increased Storage Needs:** Data is stored in multiple layers each retaining different versions and transformations of the data.
- **Complexity and Maintenance:** Managing multiple layers of data, each with its own set of transformations and validations, can increase the complexity of the data pipeline. Requiring robust data engineering practices to ensure data integrity and consistency across layers.
- **Performance Overheads:** The additional processing required to move data through these layers can introduce performance overheads

Drawbacks of Medallion Architecture

Cons of the architecture

- **Not a One-Size-Fits-All Solution:** The architecture may not be suitable for all organizations or use cases. Strictly adhering to a three-tiered structure might not align with the unique needs of every business.
- **Potential for Over-Simplification:** The classification of data into just three layers can sometimes oversimplify the complexity of real-world data processing needs, hindering the flexibility required.

Different Flavors of the Architecture

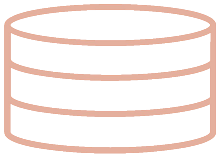
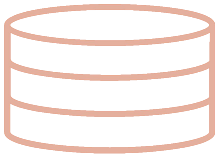
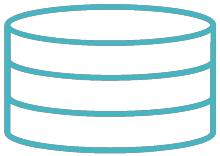
Different Names

Additional examples

- Raw > Validated > Enriched
- Raw > Base > Curated
- Raw > Stage > Curated
- Bronze > Silver > Gold > Platinum
- Bronze > Silver > Gold > Diamond

A more practical design

Different Flavors of the Architecture

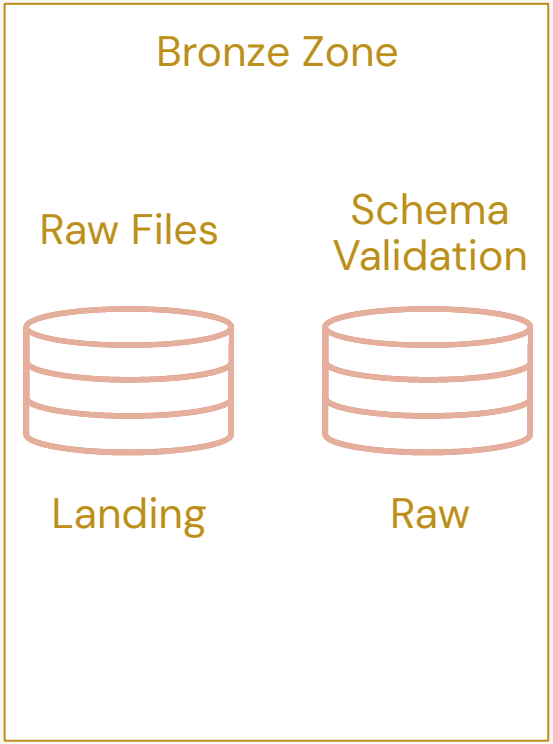
Bronze Zone		Silver Zone		Gold Zone	
Raw Files	Schema Validation	Cleansed & Validated	Conformed Clean Data	Analytical Models	Reporting Models
					
Landing	Raw	Base	Enriched	Curated	Semantic

A more practical design - Bronze Zone

Different Flavors of the Architecture

- **Landing:** Raw files
  - Data as is from the source, as files (CSV, JSON, XML, etc..) in storage.
  - Decouples from the source system to reduce load and be able to reload the data without affecting the source and allows the source system to have more control on their end for ingestion.
  - Usually not a lot of history

- **Raw:** Schema validated
  - Load the landing files and check against defined schema.
  - Ensure data ingested meets defined structural validation checks.
    - Might include garbage data, test data or duplicate but it still ingested if it fits in the structure defined
  - Usually append only and immutable
  - Might be enriched with extra metadata, for debugging or auditing.



A more practical design – Silver Zone

Different Flavors of the Architecture

- **Base:** Cleansed and validated
  - Apply data cleansing, validation, and cleaning rules.
  - Possible to track history, e.g., SCD2.
  - Closer to how the raw data is.
  - Can be considered as ODS layer
  - Same data as in the source system but cleaned and optionally enriched with some auditing and tracking metadata

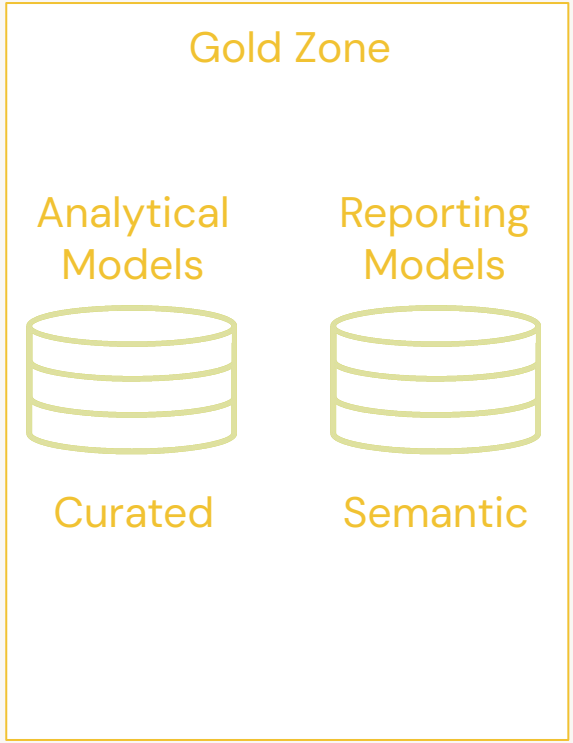
- **Enriched:** Conformed clean data
  - Conform data from different systems to have the same labels, column names, etc…and create a homogenous schema
  - Data cleaning and fixing of data quality.
  - Applying same grain or level of detail and reworking relationships

A more practical design - Gold Zone

Different Flavors of the Architecture

- **Curated:** Analytical models
  - Integrate across sources
  - Kimball style data modeling, fact, dimensions.
  - Could have star schema, with big fact tables and aggregated summary tables.
  - Data is highly governed and well documented

- **Semantic:** Reporting models
  - Reporting layer with calculations and metrics defined.
  - Cut and slice style ready tables.



Conclusion

The Medallion Architecture

Designing and structuring the medallion architecture should be done to fit your organization and use case needs.

- Don’t be restricted with the base layers only
- Select layer names that clearly convey the state and context of the data, making it easy for users to understand and query based on the layer name.

The focus of the architecture is about building an **operating model** of how data gets from its original state to being business-ready in a **trustworthy** and **repeatable** way

Q&A