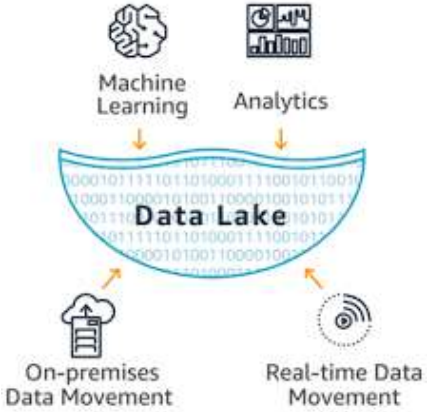

Data Lake

Data Lake

- Organizations embark on building data lakes primarily to get maximum value out of the data landscape that they have.
- A data lake basically is a centralized repository to store largescale structured and unstructured data from all the sources that are relevant to the enterprise sometimes including external sources.
- The data is stored as-is, that is without first structuring the data and most often not pre-defining the end uses.
- The data lake can be leveraged by running different types of analytics – from dashboards & visualizations to big data processing, real-time analytics and building machine learning models.

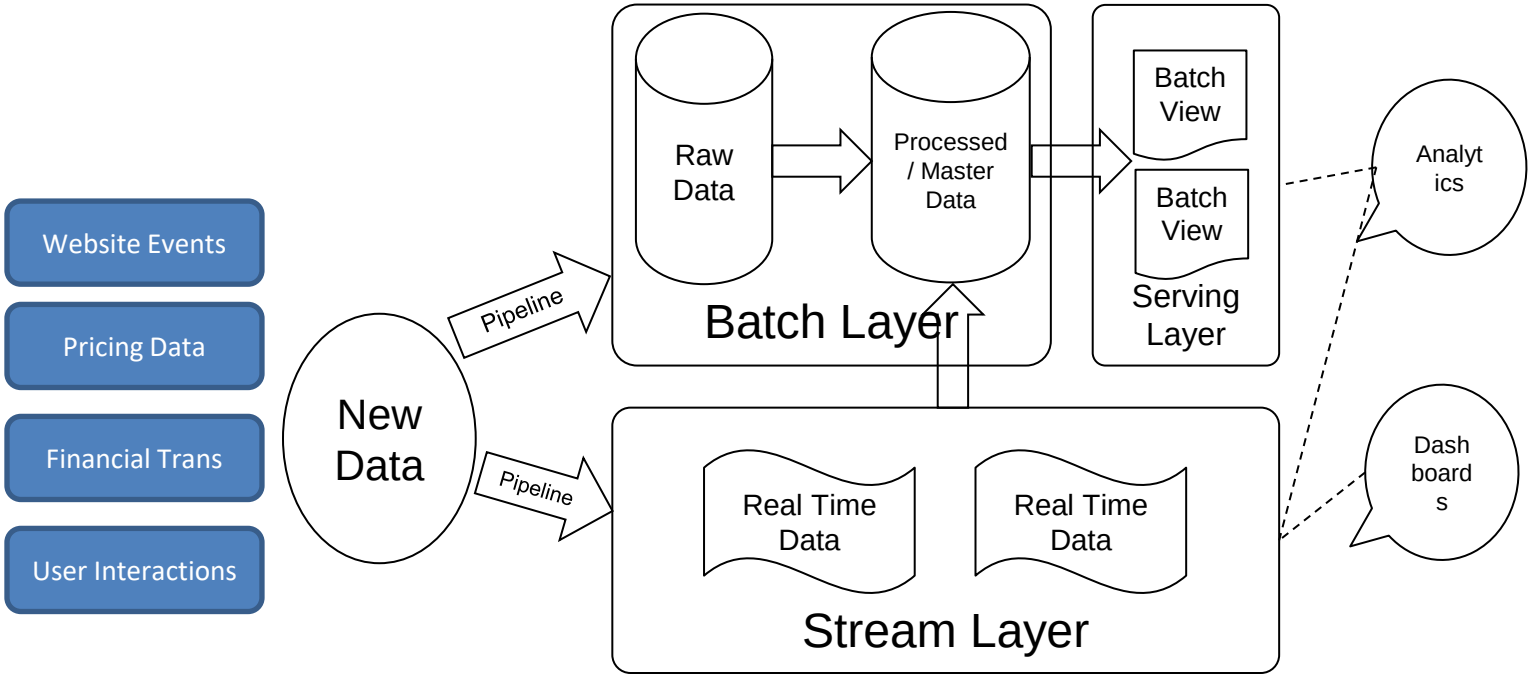
Data Lake

- Data lakes are built to get maximum value from the enterprise data landscape
- Centralized repository of largescale structured and unstructured data from all the sources relevant

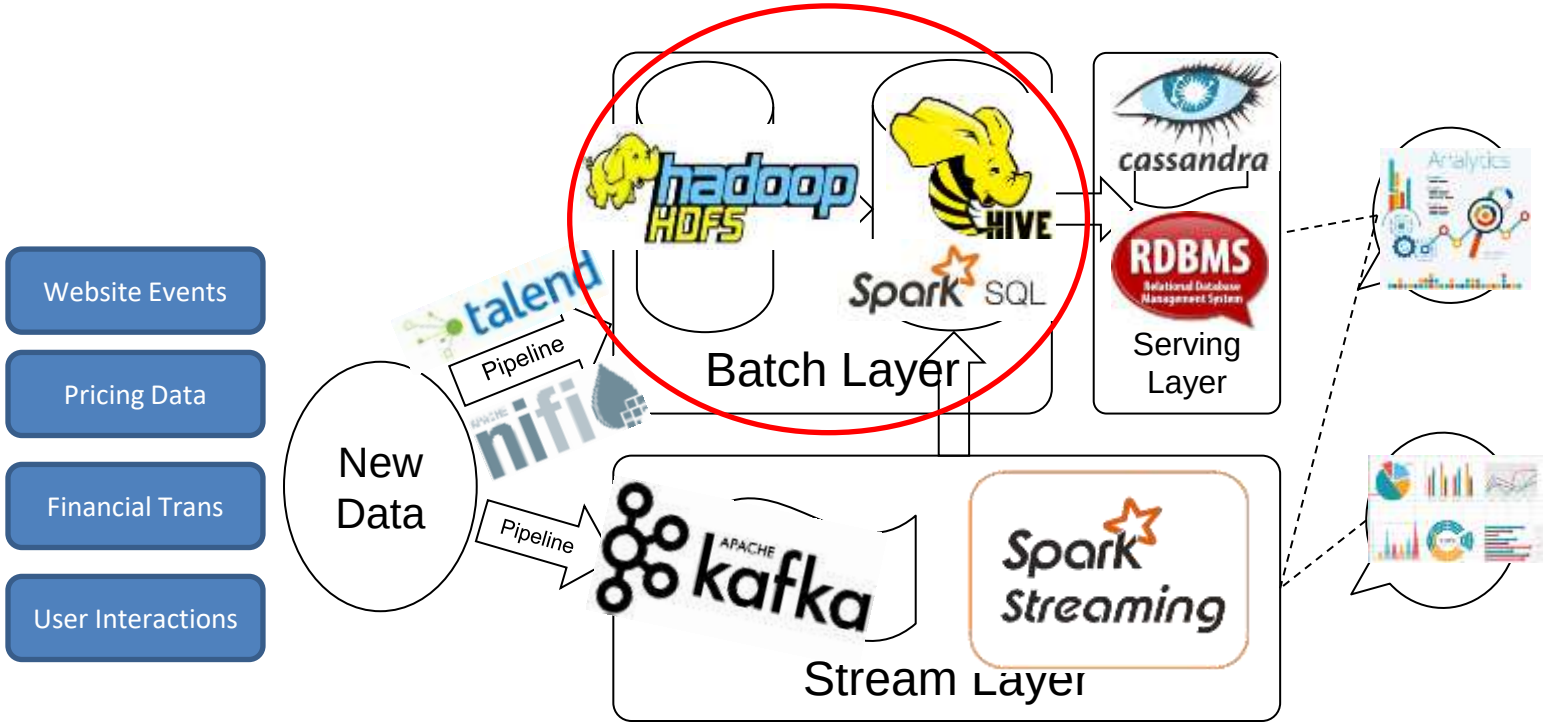


- Store data as-is
- Typically do not pre-define end uses
- Leverage Data Lake by running analytics
  - Dashboards and Visualizations
  - Big Data Processing
  - Real-time Analytics
  - Machine Learning

Data Lake


Website Events
Pricing Data
Financial Trans
User Interactions
New Data
Pipeline
Pipeline
Raw Data
Processed / Master Data
Batch Layer
Batch View
Batch View
Serving Layer
Real Time Data
Real Time Data
Stream Layer
Analyt ics
Dash board s

Data Lake


Website Events
Pricing Data
Financial Trans
User Interactions
New Data
Pipeline
Pipeline
talend
nifi
hadoop
HDFS
HIVE
Spark SQL
Batch Layer
cassandra
RDBMS
Serving Layer
Analytics
APACHE kafka
Spark Streaming
Stream Layer

Data Lakehouse

- A data lakehouse is a data management system that combines the benefits of data lakes and data warehouses.
- Databricks uses *Delta Lake* which is an open-source storage framework to store data to achieve the above benefits.
- Data engineers, data scientists, analysts, and production systems can all use the data lakehouse as their single source of truth, allowing timely access to consistent data and reducing the complexities of building, maintaining, and syncing many distributed data systems.
- The key features such as ACID transaction support, Audit History of data and Time travel i.e. Access/revert to earlier versions of data for audits or to reproduce are listed below.

References

- https://docs.databricks.com/en/lakehouse/index.html
- https://docs.databricks.com/en/lakehouse/acid.html

Data Warehouses

- ACID compliant transactions
- Define a schema up front
- Good at analyzing structured data
- Cannot work with unstructured data
- Expensive
- Do not support Data Science / AI

Source: https://www.databricks.com/glossary/data-lakehouse

Data Lakes

- Can work with Structured, Unstructured, Semi-Structured Data
- Suitable for big data, analytics and Data Science / AI workflows
- Cheaper than Data Warehouses
- Not ACID compliant

Source: https://www.databricks.com/glossary/data-lakehouse

Data Lakehouse

- Combines the best elements of a Data Lake and Data Warehouse
- ACID transaction support

DELTA LAKE

Data Lakehouse

Source: https://www.databricks.com/glossary/data-lakehouse

References:

https://delta.io/

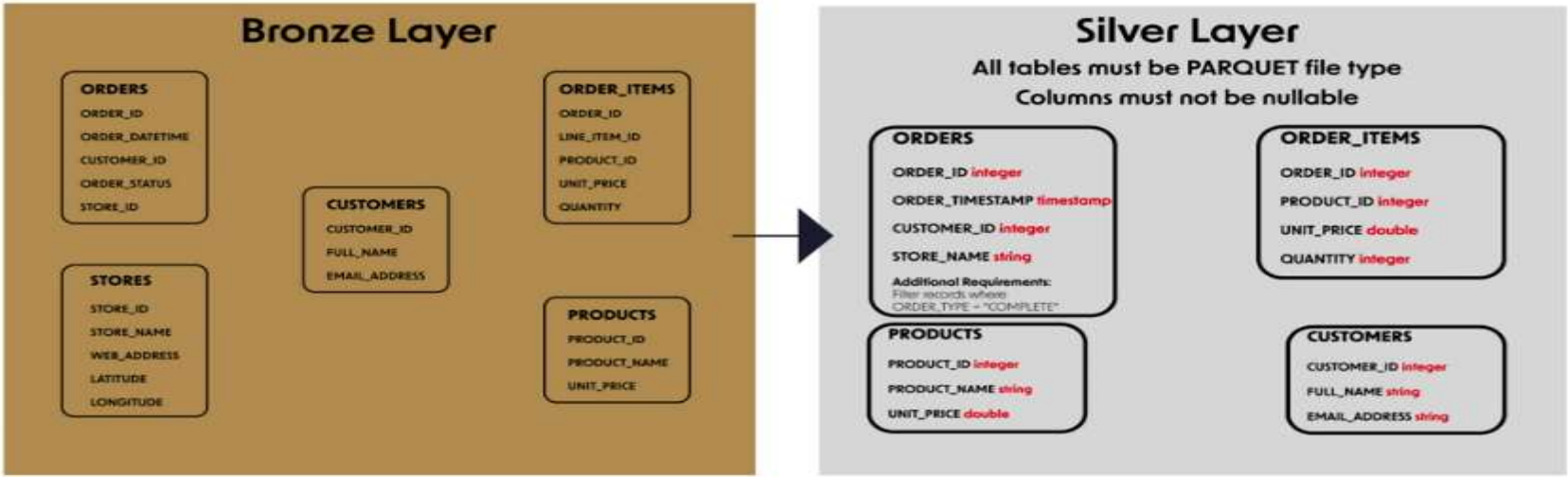
https://docs.databricks.com/en/lakehouse/medallion.html

https://www.databricks.com/glossary/medallion-architecture

Medallion Architecture

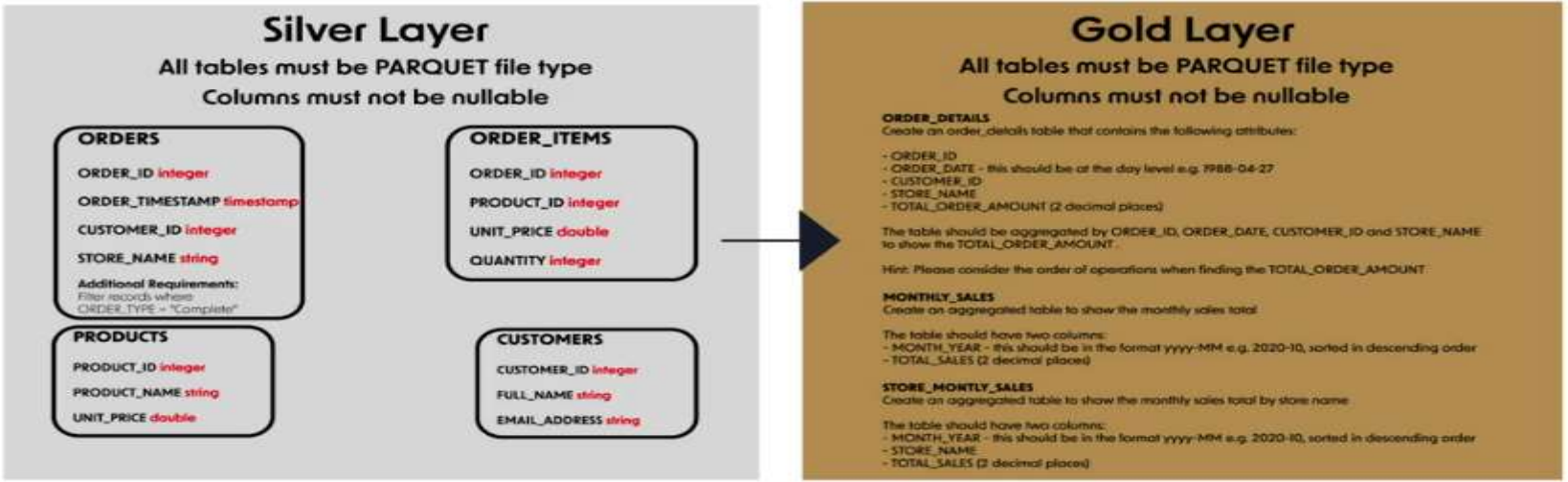
- Medallion Architecture is a data design pattern that logically organizes data in the data lakehouse.
- Databricks recommends taking a multi-layered approach to building a single source of truth for enterprise data products. There are typically 3 layers in the architecture.
- The goal of the medallion architecture is to incrementally and progressively improve the structure and quality of data as it flows through each layer of the architecture.

Medallion Architecture



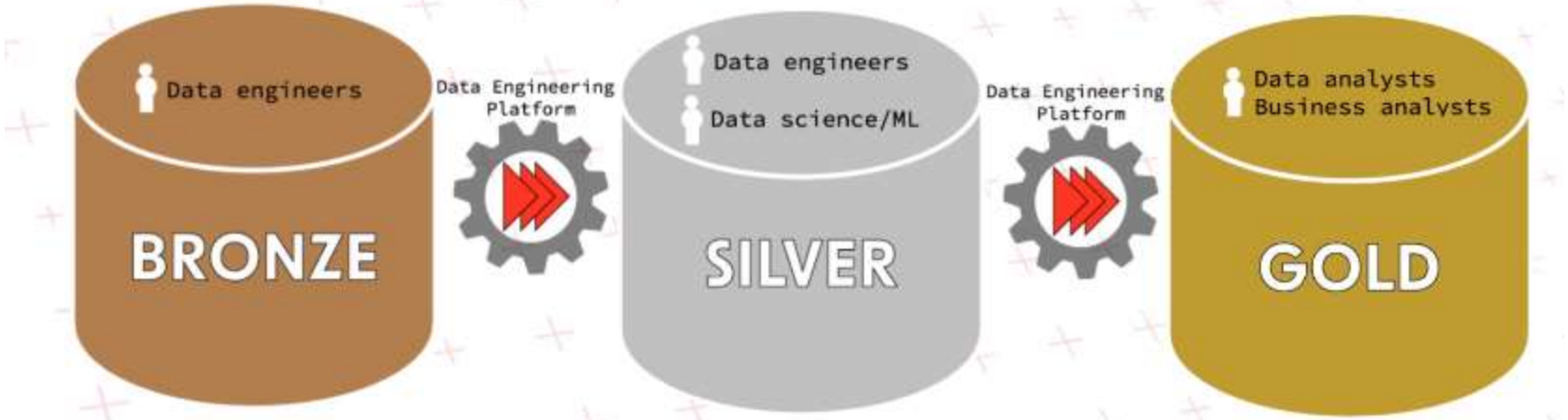
- The bronze layer is the ingestion layer. It contains unvalidated data.
- Data ingested in the bronze layer typically maintains the raw state of the data source.
- In the silver layer of the data lakehouse, the data from the bronze layer is cleansed sufficiently it provides an enterprise view of the key business entities and transactions.

Medallion Architecture



- The data in the gold layer is refined and aggregated, containing data that is used for reporting, analysis and production applications.
- While all tables in the lakehouse should serve an important purpose, gold tables represent data that is directly usable for end uses of reporting and analytics purposes.

Medallion Architecture



ingestion layer

maintains the raw state of the data source

filtered, cleaned augmented layer

data is conformed and cleansed "just-enough"

business level aggregates

curated business-level tables

Medallion Architecture

Raw Data

Validated Data

Enriched Data