The big picture

Lesson goals

- Recall foundational concepts related to the contemporary big data landscape.
- Define the core challenges with building a big data architecture.
- Explain how the core components of a Lakehouse relate to the Inmon Architecture.
- Explain how Delta Lake can be used to build a Lakehouse.
- Describe the core components of Delta Lake.

The big data landscape

databricks

The big data problem

Volume

Velocity

Variety

Veracity

Value

At Moovio, your SLA includes:

moovio

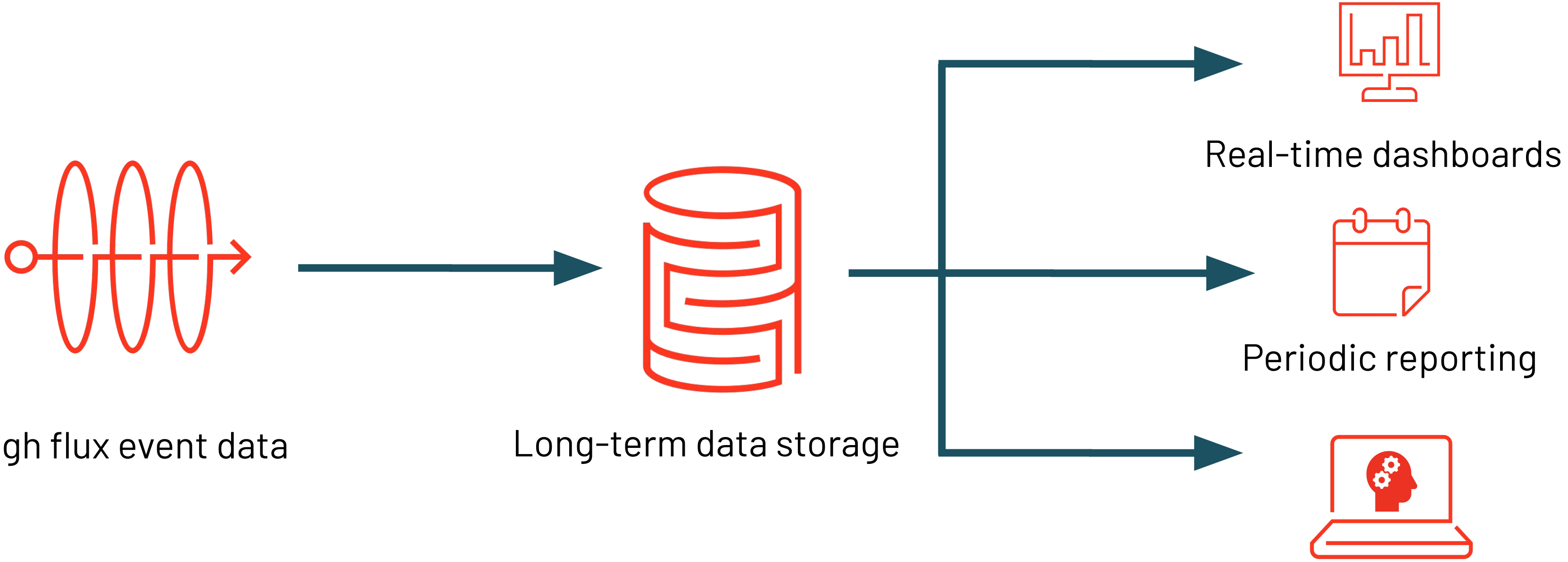
Data freshness

Query speed

Data reliability

Ease of use

Big data needs



High flux event data

Long-term data storage

Real-time dashboards

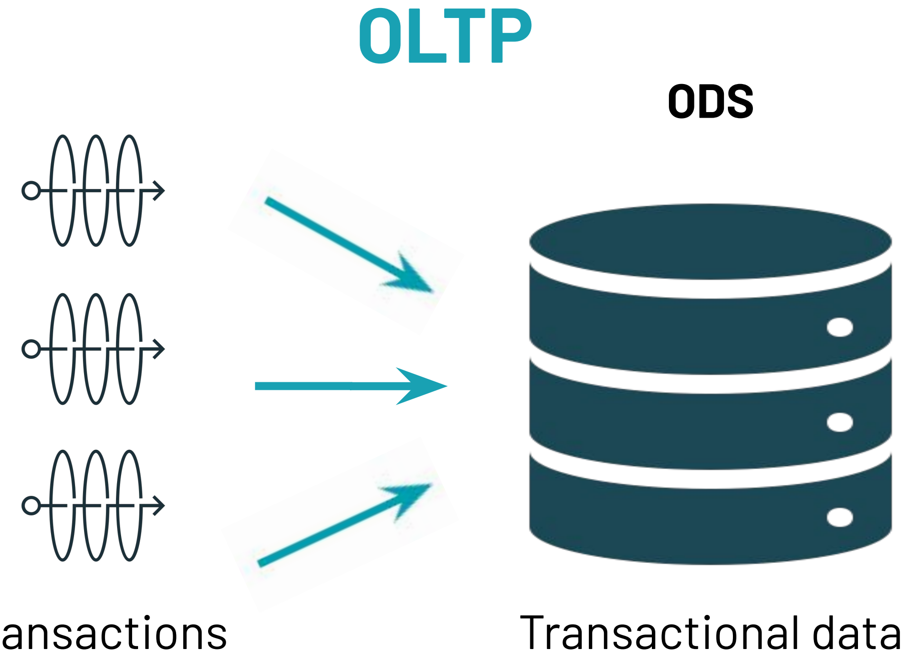
Periodic reporting

Artificial intelligence

A single source of truth

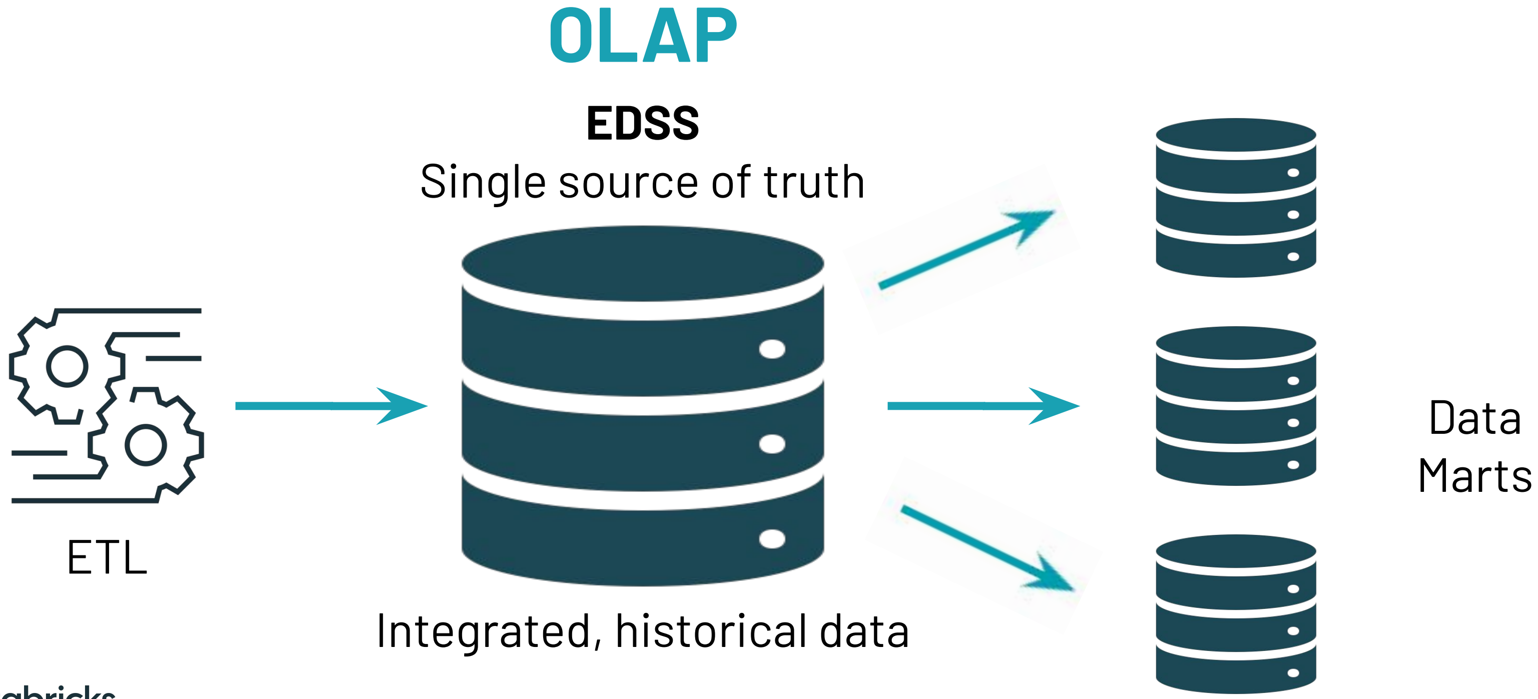
ODS and OLTP

Operational Data Store and Online Transaction Processing



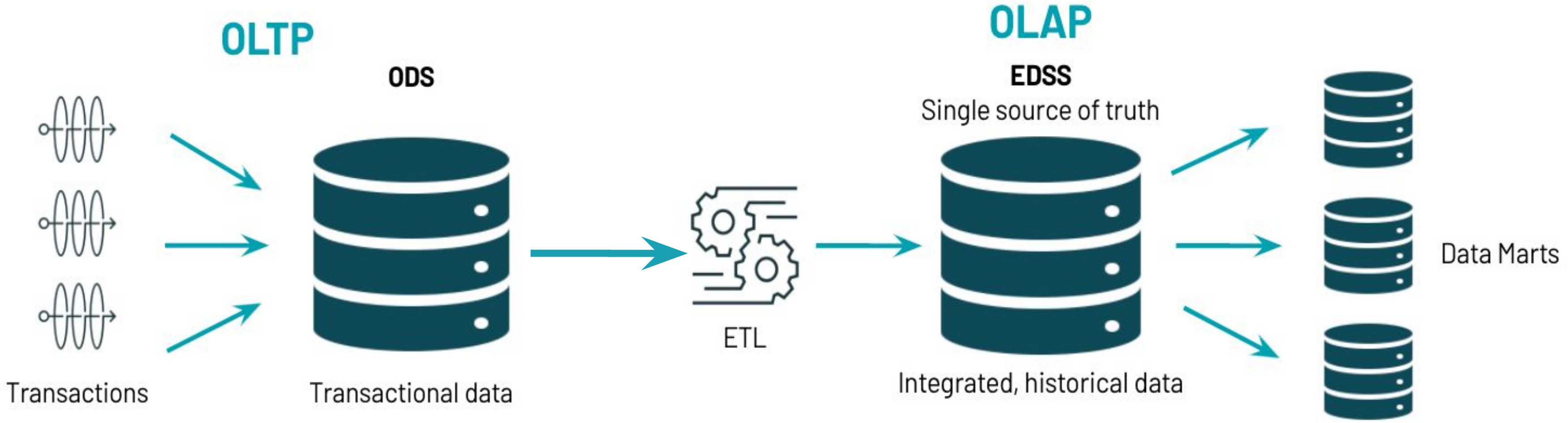
EDSS and OLAP

The Enterprise Decision Support System and Online Analytical Processing

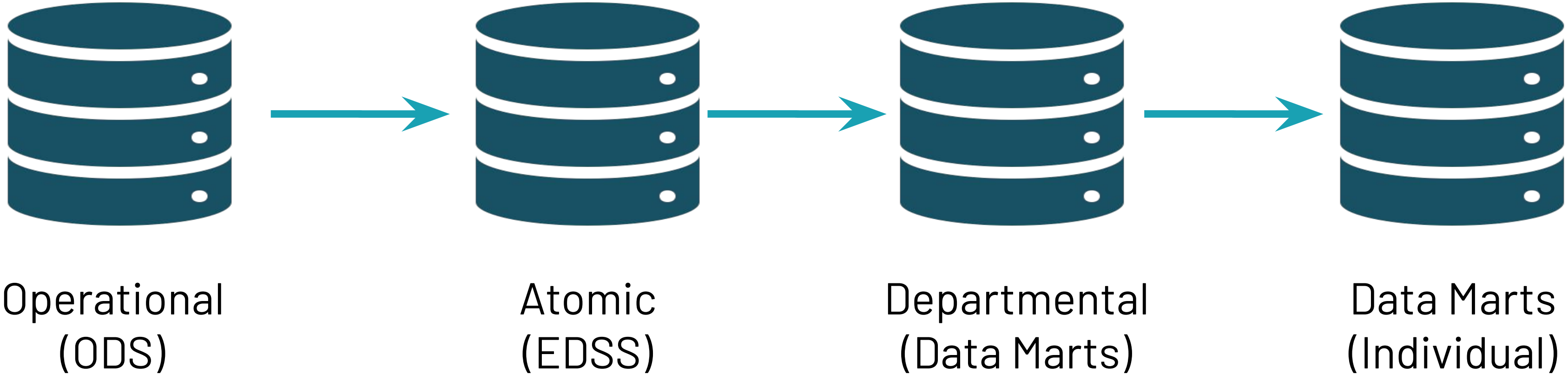


Complete data system

- An ETL process pulls data from the ODS to be loaded into the EDSS



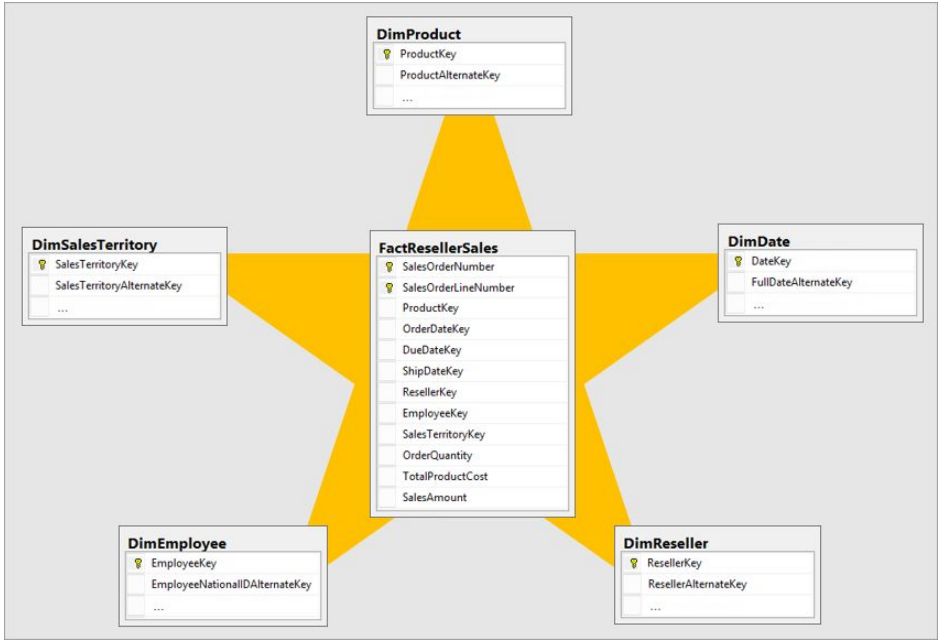
Levels of data



Fact and dimension tables

Dimensional modeling

- Fact tables
- Dimension tables
- Aggregate fact tables



DimProduct
ProductKey
ProductAlternateKey
...

DimSalesTerritory
SalesTerritoryKey
SalesTerritoryAlternateKey
...

FactResellerSales
SalesOrderNumber
SalesOrderLineNumber
ProductKey
OrderDateKey
DueDateKey
ShipDateKey
ResellerKey
EmployeeKey
SalesTerritoryKey
OrderQuantity
TotalProductCost
SalesAmount

DimDate
DateKey
FullDateAlternateKey
...

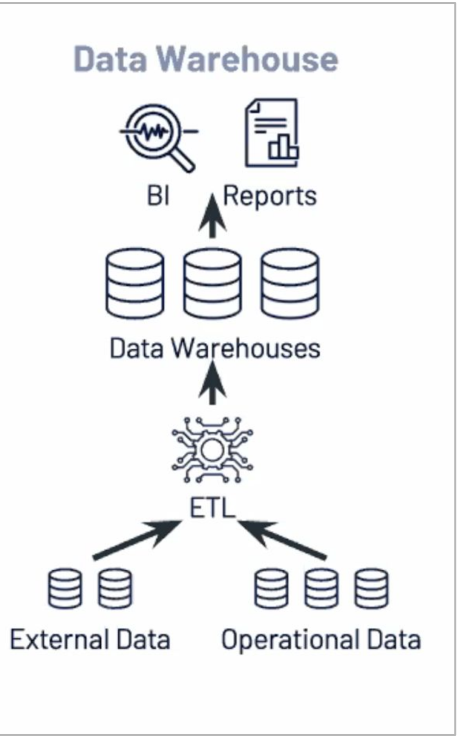
DimEmployee
EmployeeKey
EmployeeNationalIDAlternateKey
...

DimReseller
ResellerKey
ResellerAlternateKey
...

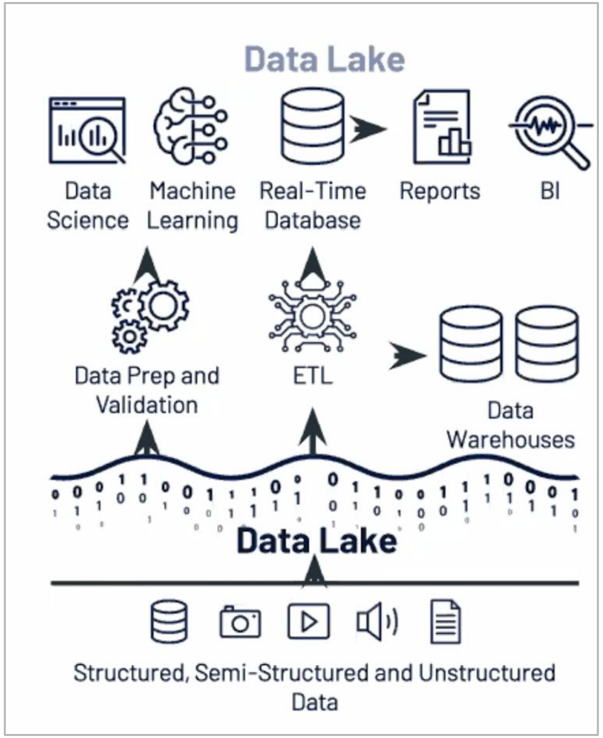
The Lakehouse

databricks

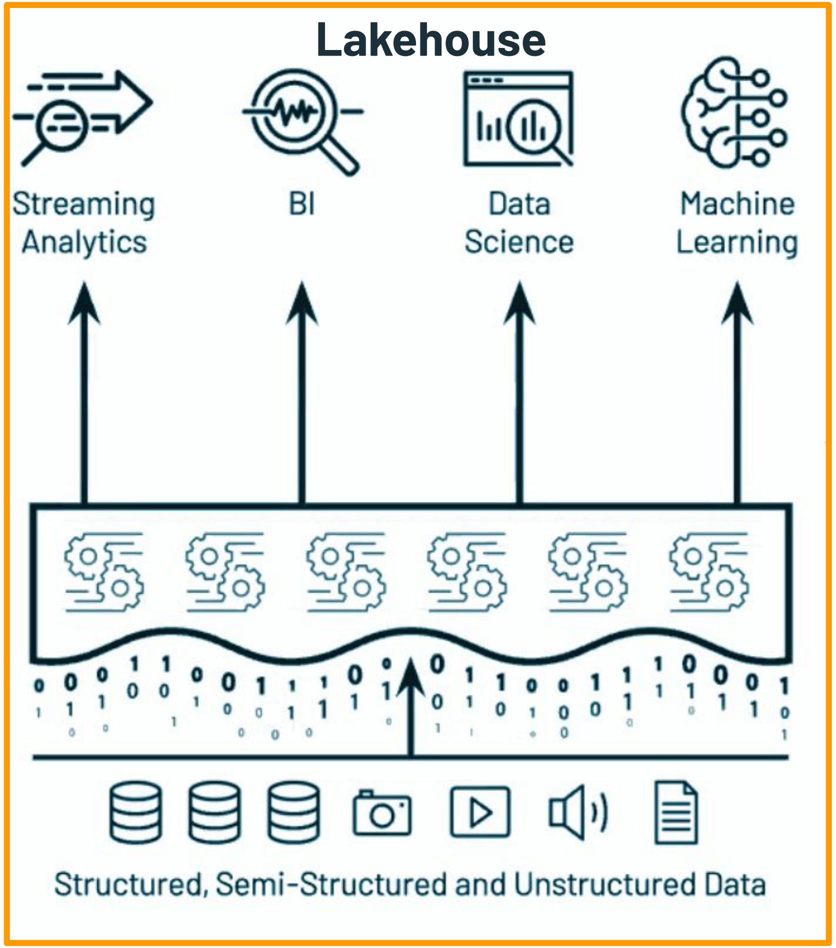
What is a Lakehouse?



Data Warehouse

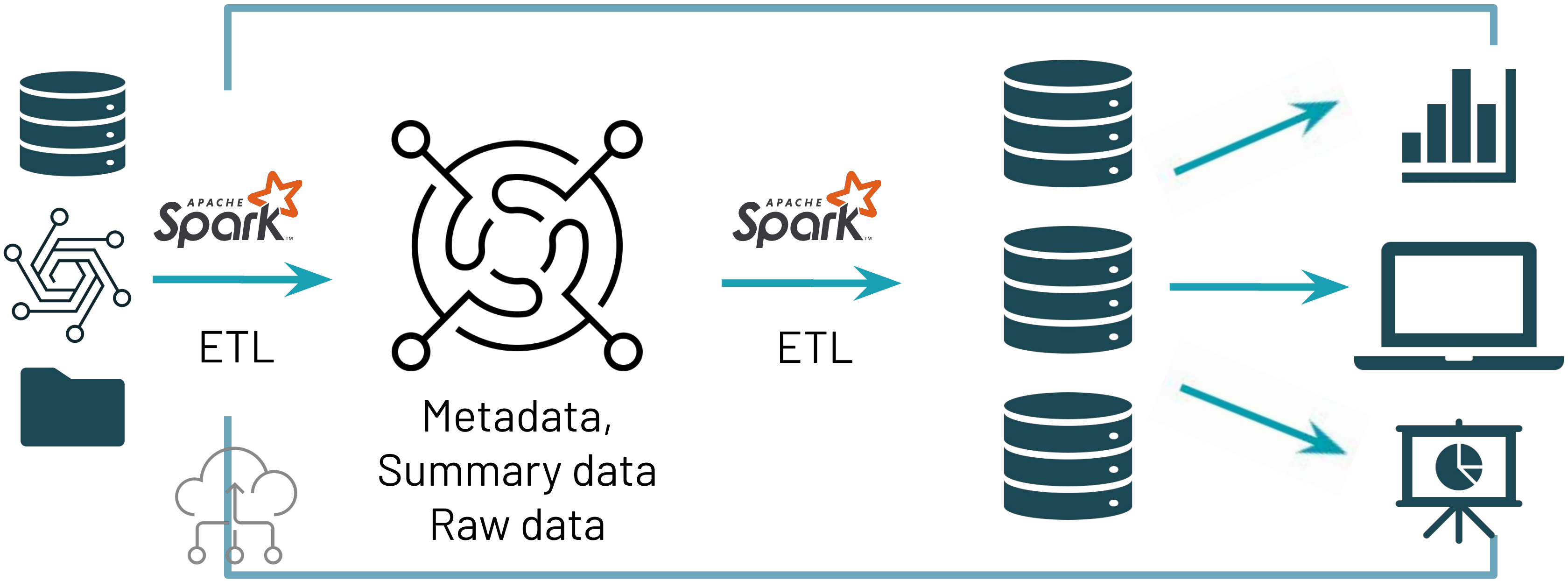


Data Lake



Lakehouse

Levels of data in a Lakehouse



ETL

Metadata,
Summary data
Raw data

ETL

Operational
ODS

Atomic
EDSS
Data Lake

Departmental
Data Marts
Data warehouse

Individual
End users

Benefits of a Lakehouse

- Separation of compute and storage
- Infinite storage capacity
- Leverage best aspects of a data warehouse
- Low data gravity
- High data throughput
- No limits on data structure
- Mix batch and streaming workloads

The Delta architecture

databricks

What is **Delta Lake**?

- Technology designed to be used with Apache Spark to build robust data lakes

Delta Lake features

- ACID transactions on Spark
- Scalable metadata handling
- Streaming and batch unification
- Schema enforcement
- Time travel
- Upserts and deletes
- Fully configurable/optimizable
- Structured streaming support

Delta Lake components

Delta Lake storage layer

Delta tables

Delta Engine

Delta Lake components

Delta Lake storage layer

Delta tables

Delta Engine

Delta Lake storage layer

- Highly performant and persistent
- Low-cost, easily scalable object storage
- Ensures consistency
- Allows for flexibility

Delta Lake components

Delta Lake storage layer

Delta tables

Delta Engine

Delta table components

- Data in Parquet/Delta files
- Transaction log
- Registered in metastore (optional)

Data - Parquet files

- File format for tabular data stored as columns
- Fast and powerful
- Delta files = Parquet + versioning + metadata

Transaction log

- Record of all transactions on a Delta table
- Prevents read conflicts
- Commits - ordered, atomic, json files
- Created automatically in the *_delta_log* subdirectory

Delta Lake components

Delta Lake storage layer

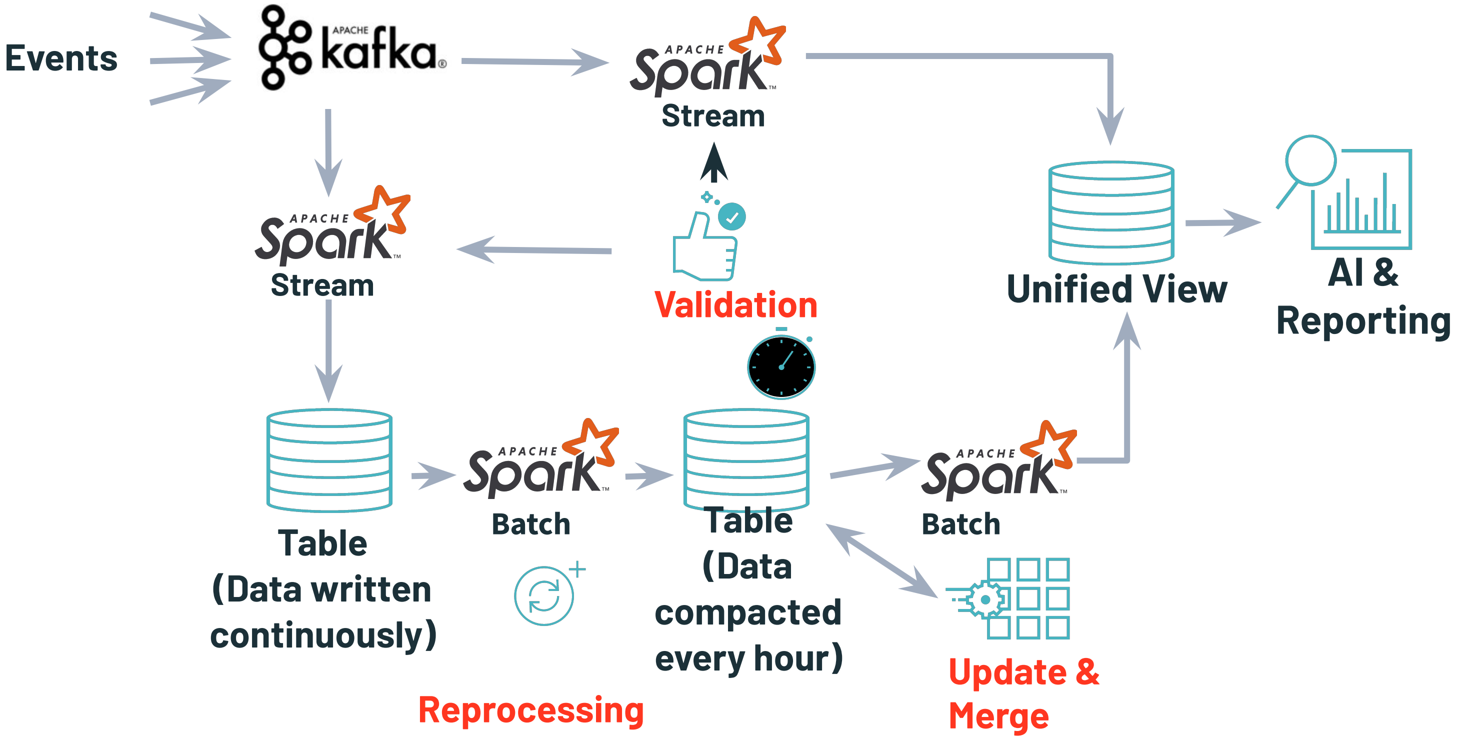
Delta tables

Delta Engine

Delta Engine

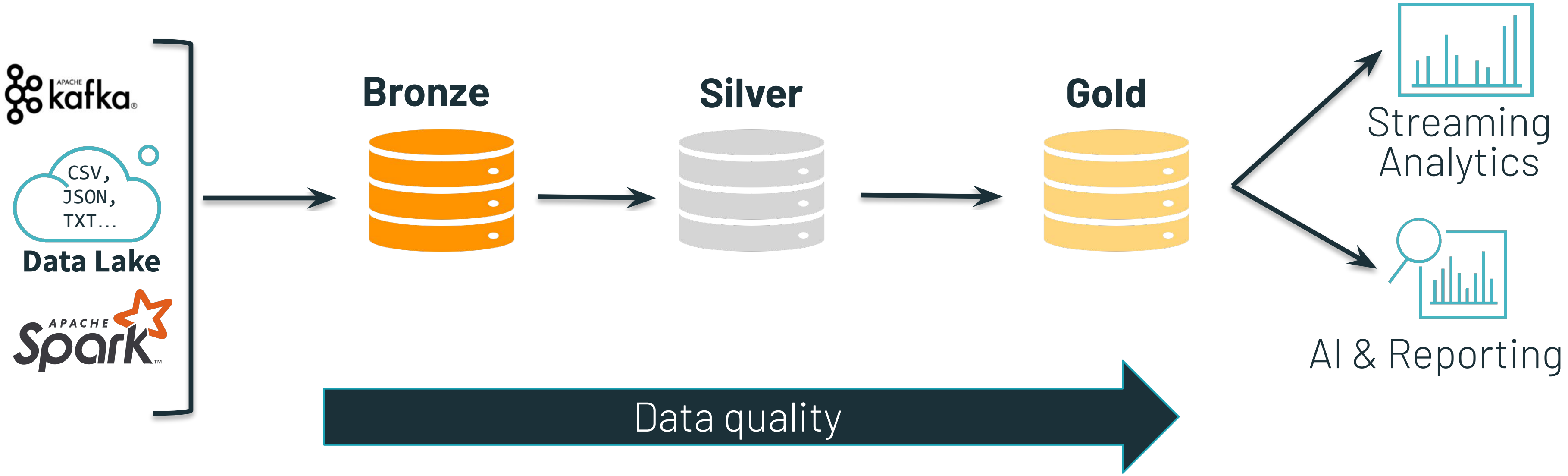
- File management optimizations
- Auto-optimized writes
- Performance optimization via Delta caching

The goal of a data engineer



databricks

The Delta architecture design pattern



Data Lake

CSV,
JSON,
TXT...

Bronze

Silver

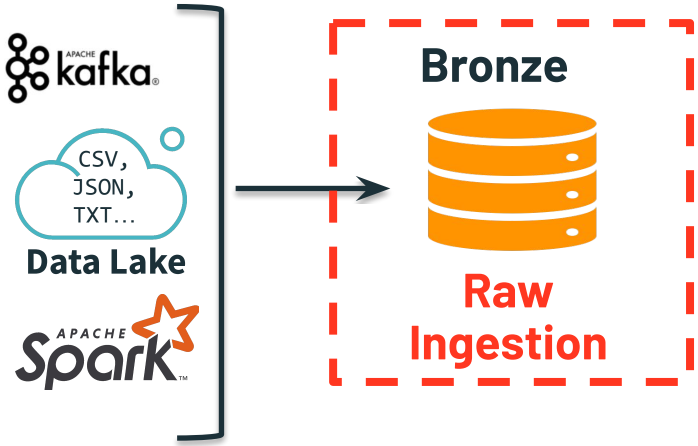
Gold

Data quality

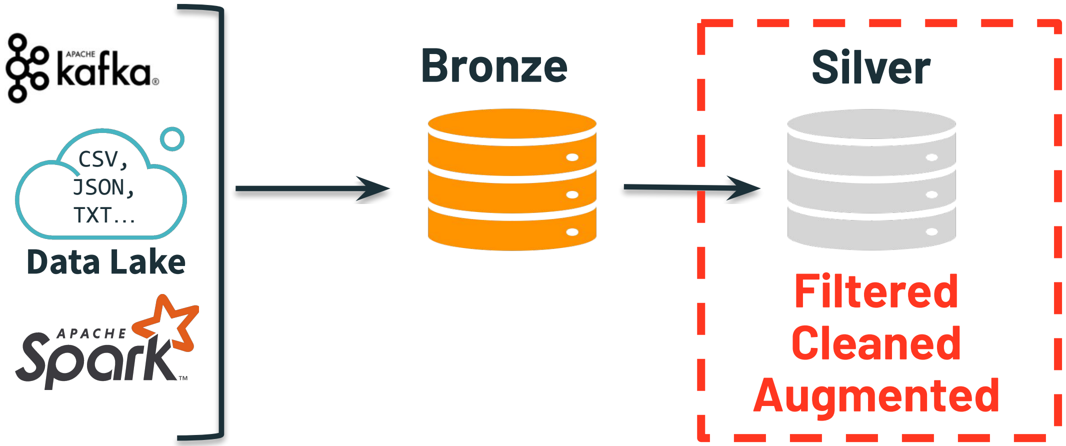
Streaming Analytics

AI & Reporting

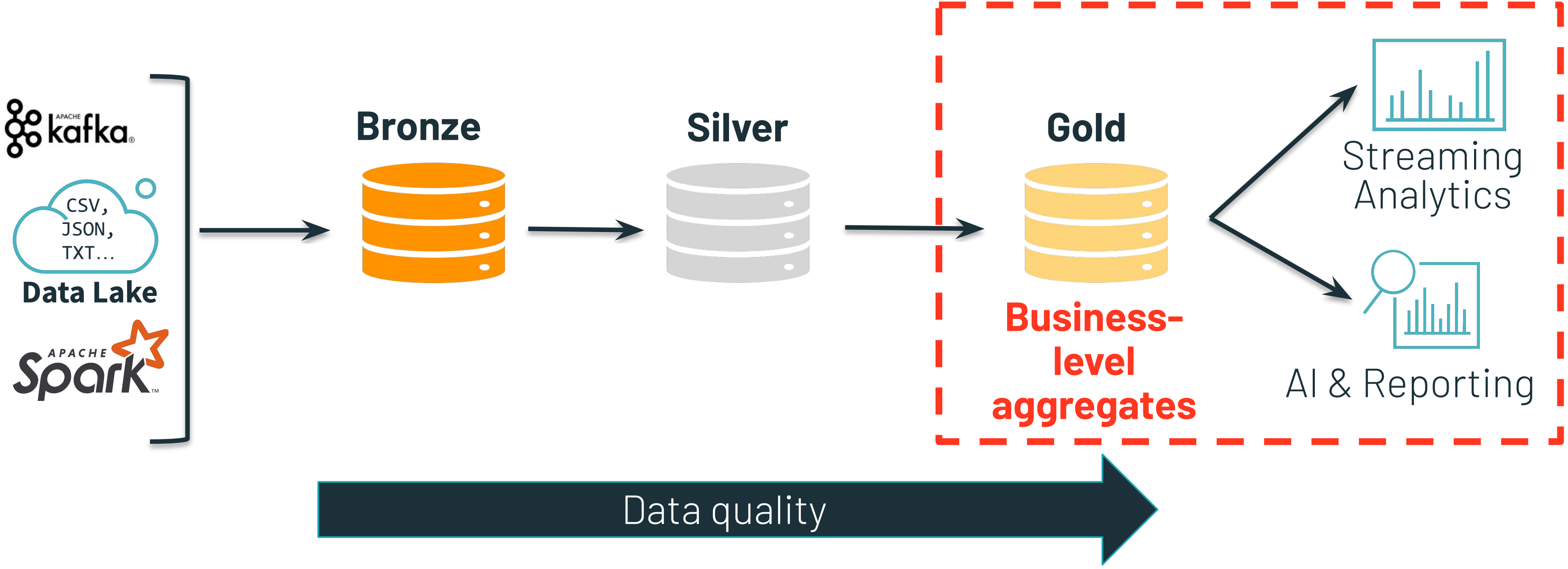
Delta architecture - Bronze



Delta architecture - Silver



Delta architecture - Gold



databricks

databricks