Introduction to Delta Lake

Surendra Panpaliya

Agenda

What is Delta Lake?

Delta Lake operations on Databricks

Create a table.

Upsert to a table.

Read from a table.

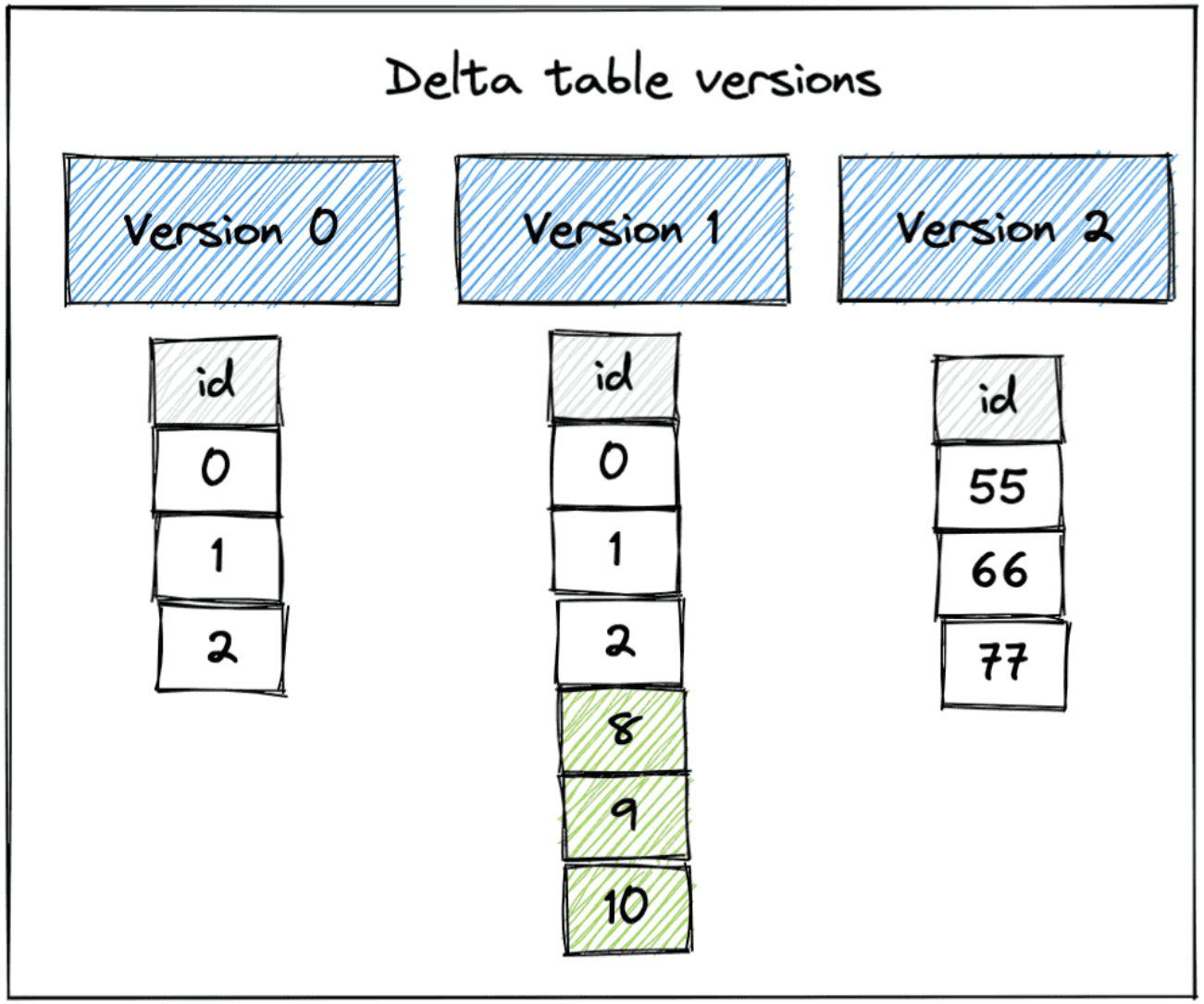
Display table history

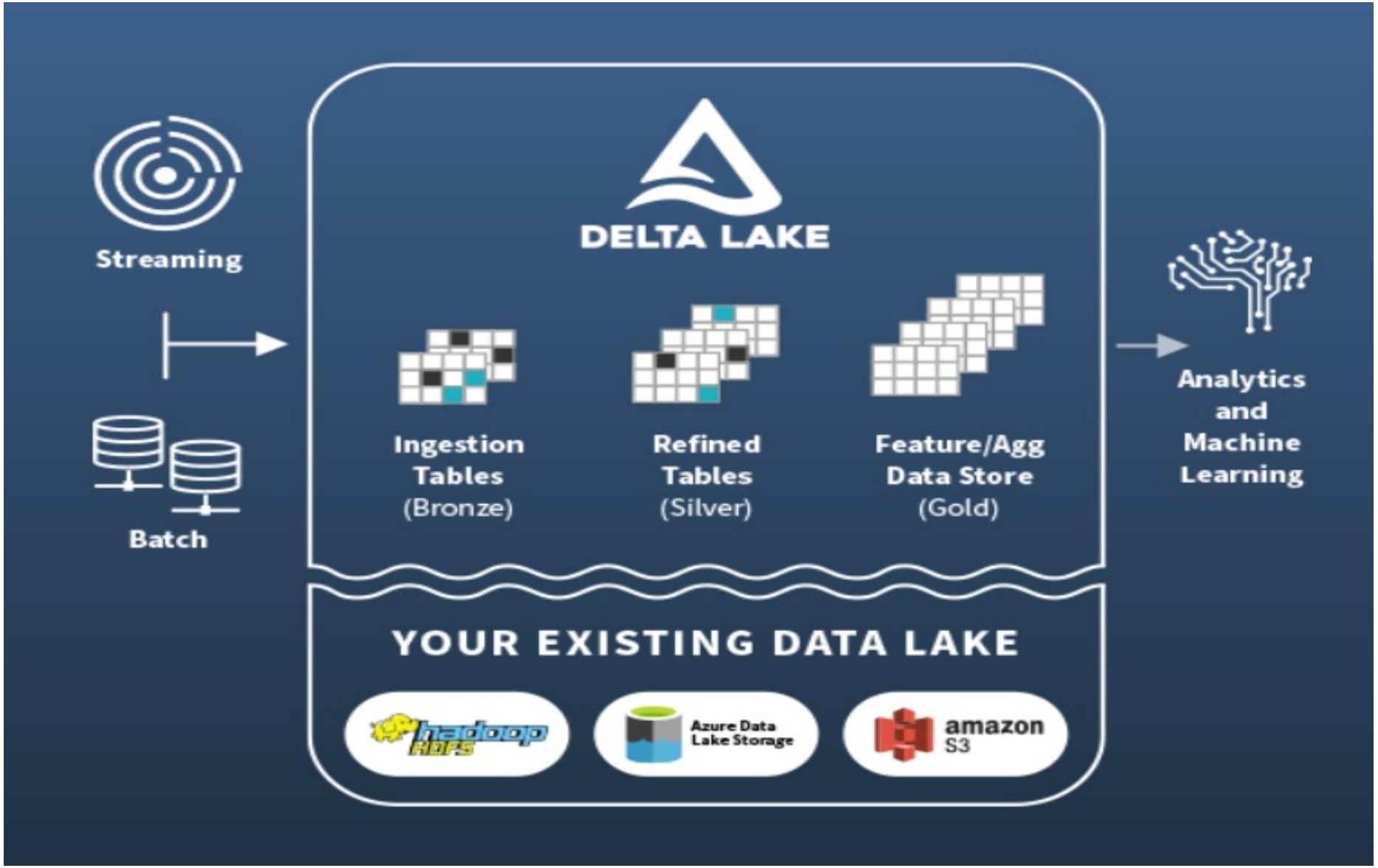
Agenda

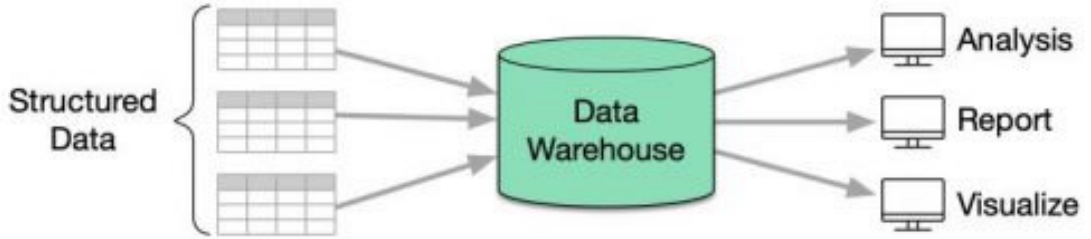
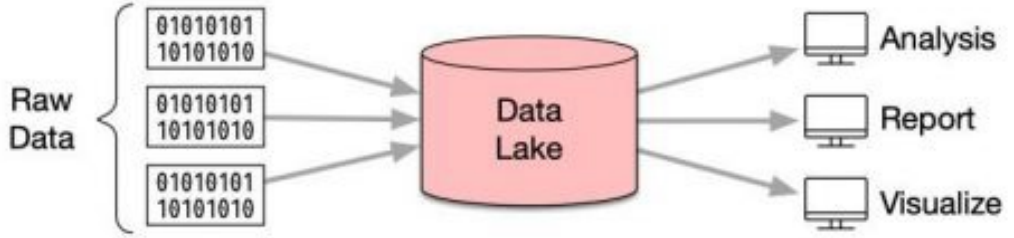
- Query an earlier version of a table.
- Optimize a table
- Add a Z-order index.
- Liquid Clustering
- Vacuum unreferenced files

What is Delta Lake?

- An open-source Storage Layer
- Provides ACID transactions
- Atomicity, Consistency, Isolation, Durability
- Scalable metadata handling
- Time travel capabilities to data lakes.


Delta table versions
Version 0
Version 1
Version 2
id
0
1
2
id
0
1
2
8
9
10
id
55
66
77


Streaming
Batch
DELTA LAKE
Ingestion Tables (Bronze)
Refined Tables (Silver)
Feature/Agg Data Store (Gold)
Analytics and Machine Learning
YOUR EXISTING DATA LAKE
hadoop HDFS
Azure Data Lake Storage
amazon S3

Data Warehouse	A large, structured repository of integrated data from various sources, used for complex querying and historical analysis	
Data Lake	A more focused, department-specific subset of a data warehouse providing quick data retrieval and analysis	

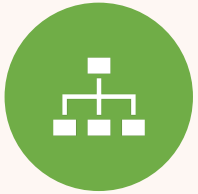
What is Data Warehouse?

A data warehouse is

A central repository

Integrated and structured data

From various sources



within an organization.

What is Data Warehouse?

Designed to support

Business intelligence (BI) and

Analytics activities by

Providing a unified

Consistent view of data.

What is Data Warehouse?

DW typically follow

A schema-on-write approach,

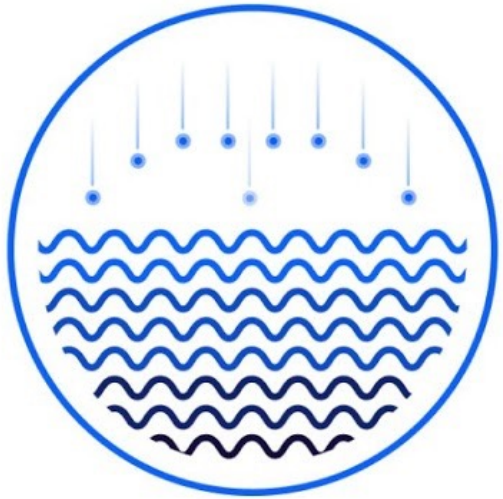
Where data is transformed and

Loaded into predefined schemas

Optimized for Reporting and Analysis.

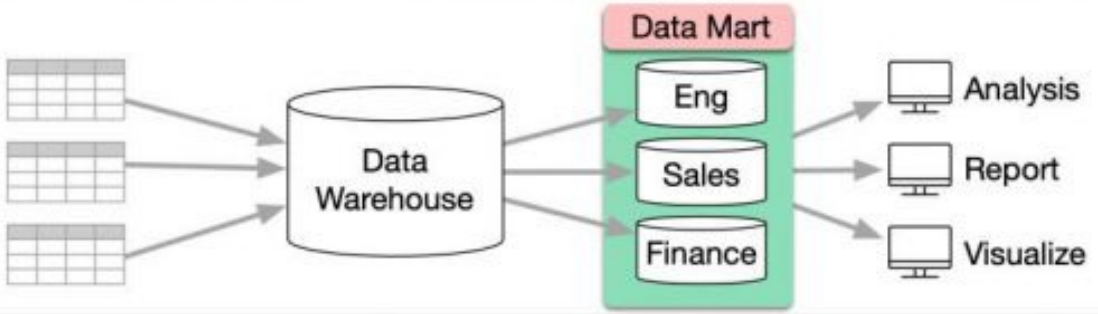
WAREHOUSE

VS



DATA LAKE

adservio

Data Mart	A vast pool of raw, unstructured data stored in its native format until it's needed for use	
Delta Lake	An open-source storage layer that brings reliability and ACID transactions to data lakes, unifying batch and streaming data processing	

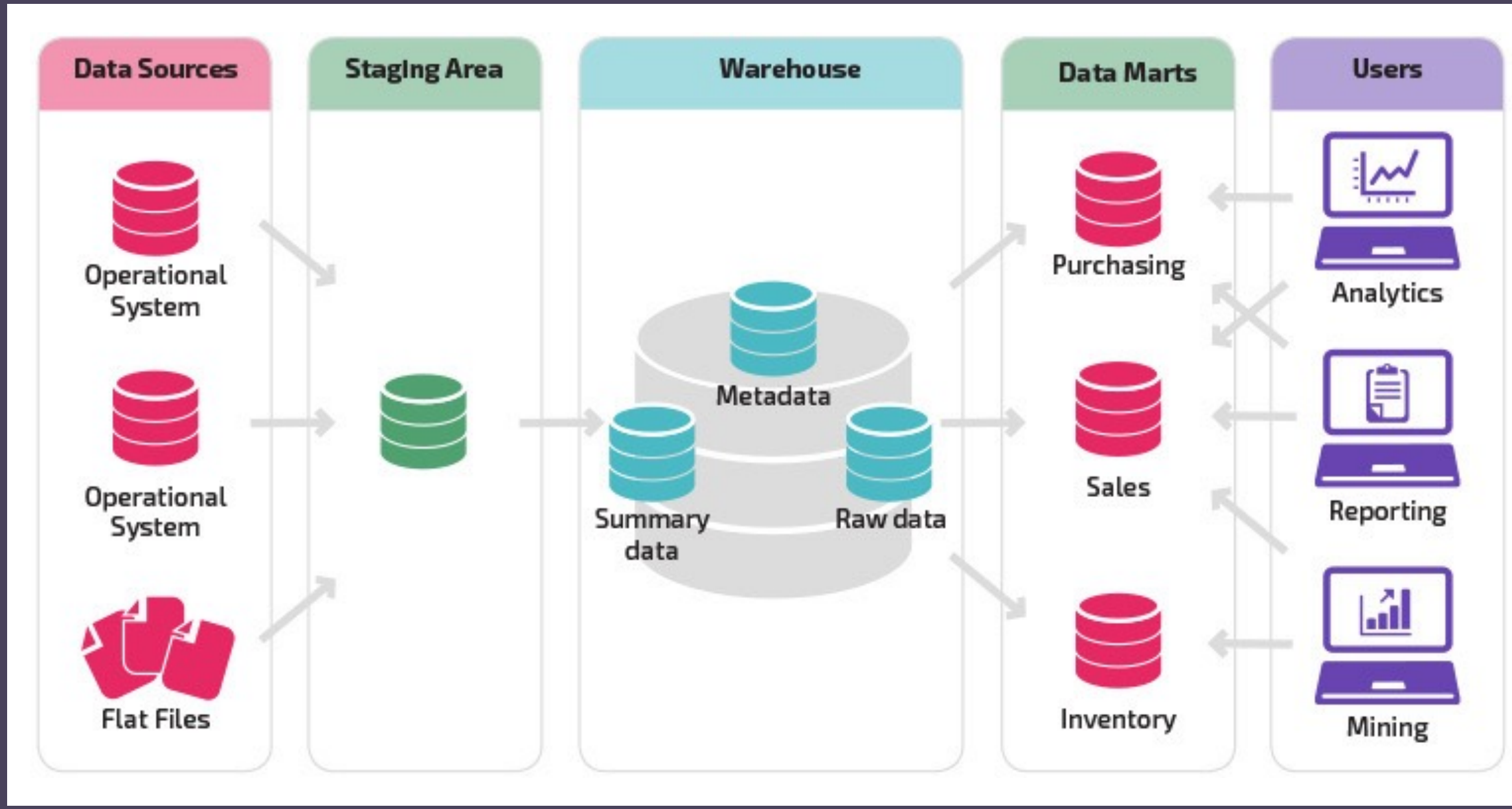
Data Mart

A data mart is a

subset of a data warehouse

focused on a

specific business function or department.


Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
Operational System
Flat Files
Metadata
Summary data
Raw data
Purchasing
Sales
Inventory
Analytics
Reporting
Mining

Data Mart

Contains a subset of data

Relevant to a particular group of users,

making it more

targeted and specialized.

Data Mart

Data marts are often designed

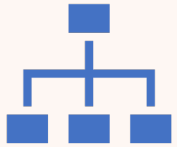
to provide faster and

more specific insights

compared to the broader

data warehouse.

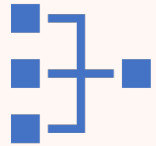
Data Mart



Can be structured

Similar to a data warehouse or

May use different

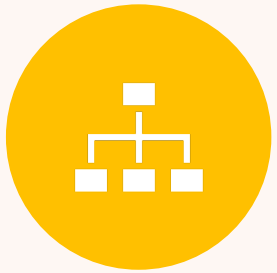


Schemas and models.

What is Data Lake?

A DATA LAKE IS A LARGE,

CENTRALIZED REPOSITORY



THAT STORES STRUCTURED, SEMI-STRUCTURED, AND

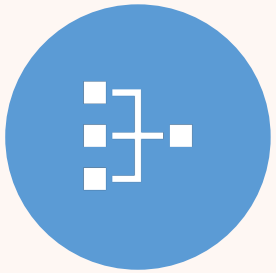
UNSTRUCTURED DATA IN ITS RAW AND UNPROCESSED FORM.

What is Data Lake?

ALLOWS ORGANIZATIONS

TO STORE VAST AMOUNTS OF DATA

FROM DIFFERENT SOURCES



WITHOUT PREDEFINED SCHEMAS.

What is Data Lake?

DATA LAKES ENABLE DATA EXPLORATION,

ADVANCED ANALYTICS, AND

MACHINE LEARNING BY PROVIDING FLEXIBILITY

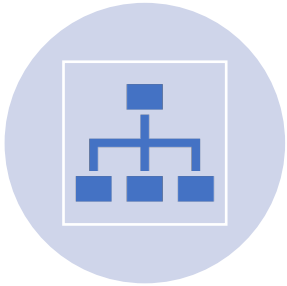
IN DATA PROCESSING AND ANALYSIS.

What is Data Lake?

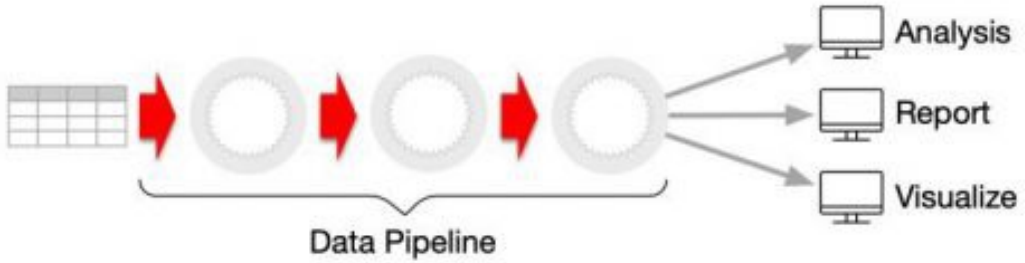
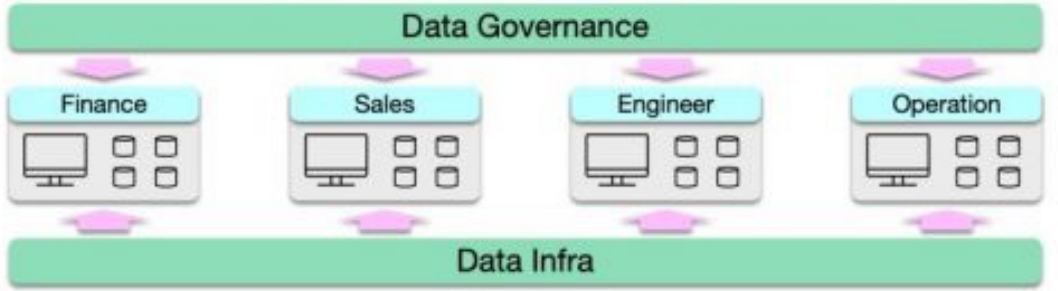
The data lake architecture typically

follows a schema-on-read approach,

where data is transformed and



structured at the time of analysis.

Data Pipeline	A process that moves and transforms data from one system to another, often used to populate data warehouses and data lakes	
Data Mesh	An architectural and organizational approach where data ownership and delivery are decentralized across domain-specific, cross-functional teams	

Data Pipeline

A system or framework

Facilitates the automated

Extraction, Transformation, Loading (ETL)

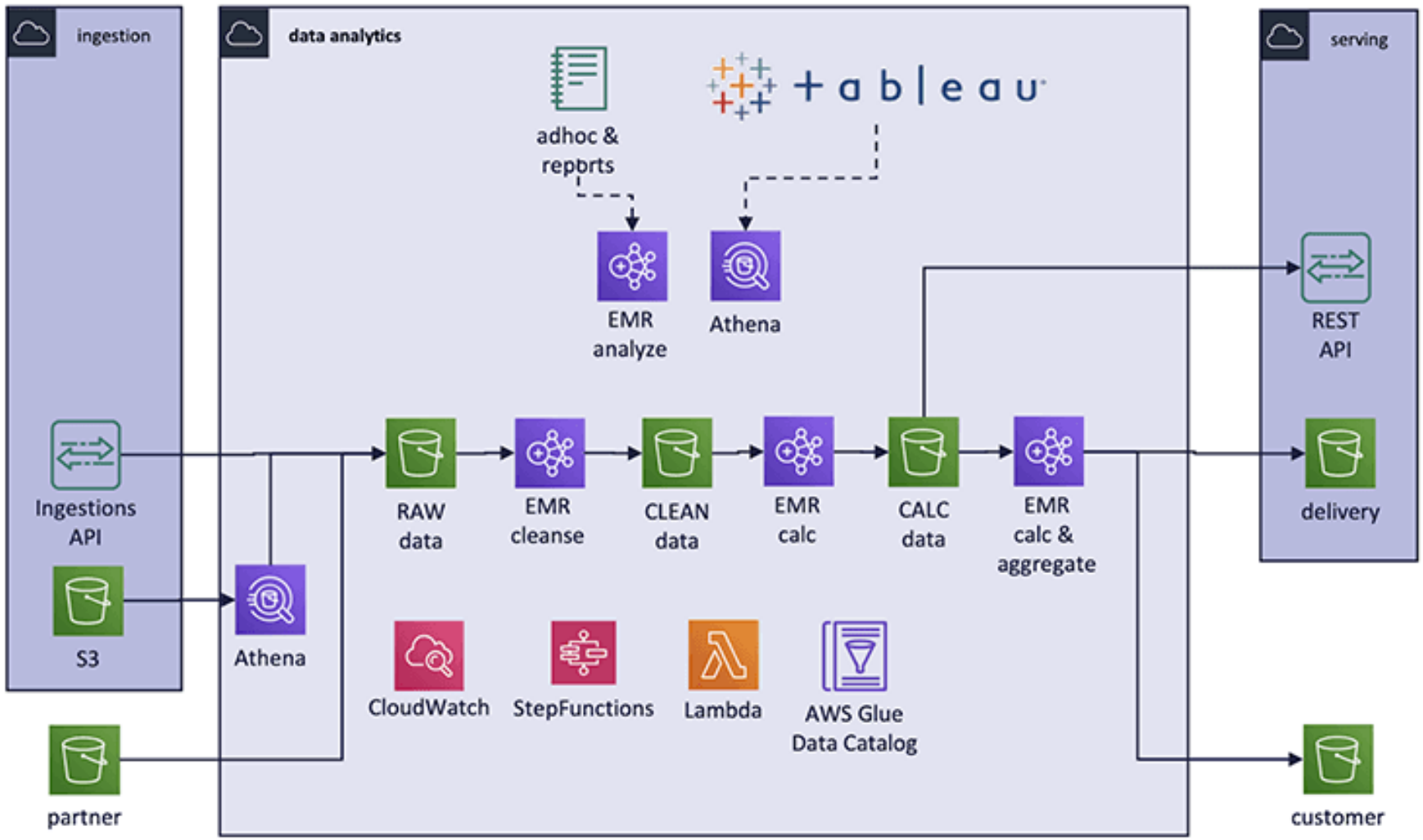
From various sources into

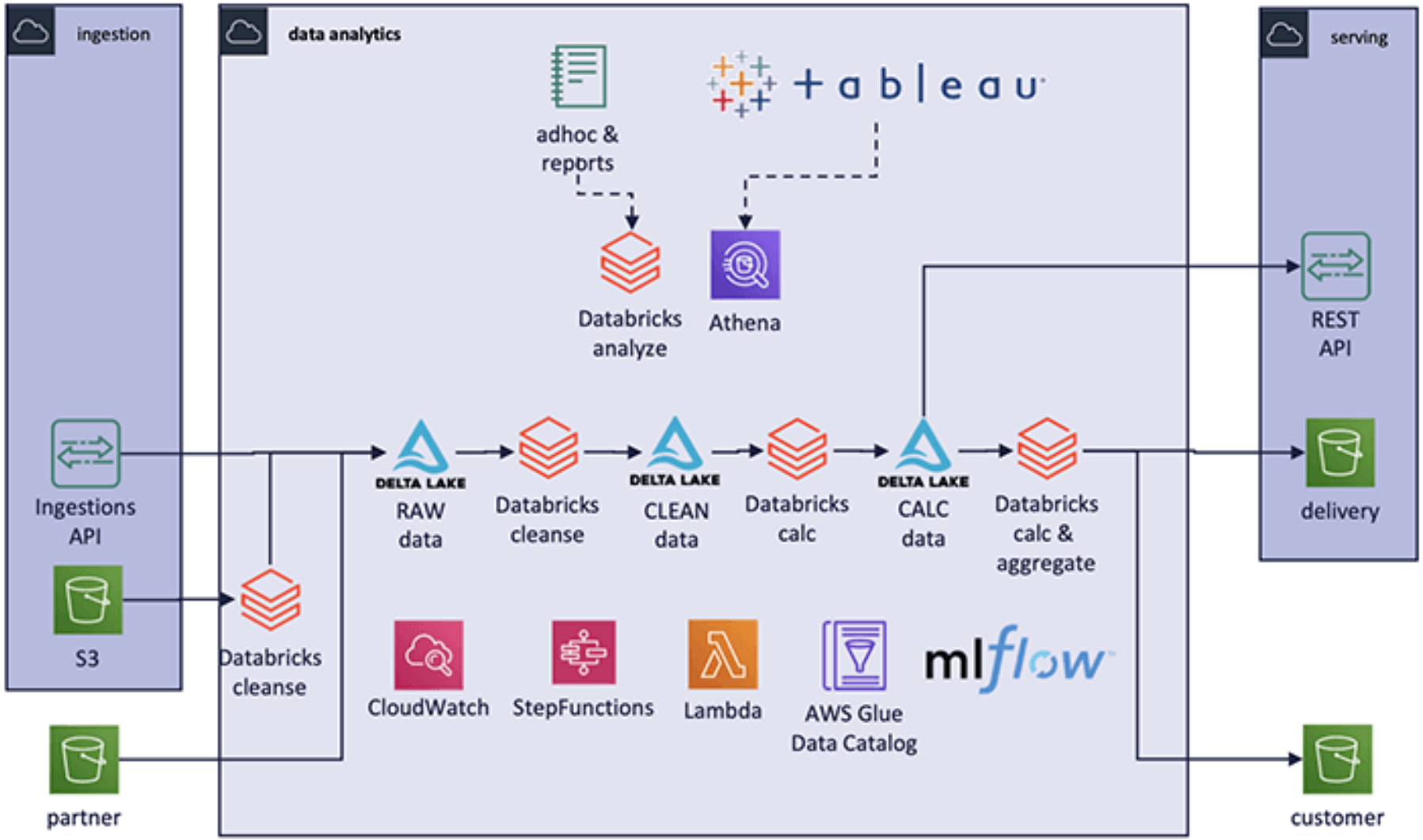
Target storage or processing system.


User Views / Clicks
Impressions
Conversions
User
Conversions
Amazon
Kinesis /
Kafka
databricks
ETL, Aggregation, Attribution
Real-Time Reporting
in BI Tools

High level architecture & tech stack

How our world looked before Databricks




ingestion
data analytics
serving
adhoc &
reports
tableau
Databricks
analyze
Athena
REST
API
Ingestions
API
DELTA LAKE
RAW
data
Databricks
cleanse
DELTA LAKE
CLEAN
data
Databricks
calc
DELTA LAKE
CALC
data
Databricks
calc &
aggregate
delivery
S3
Databricks
cleanse
CloudWatch
StepFunctions
Lambda
AWS Glue
Data Catalog
mlflow
partner
customer

Data Pipeline

ENABLES THE SMOOTH FLOW OF DATA

FROM SOURCE SYSTEMS

TO DATA WAREHOUSES,

DATA LAKES, OR

OTHER DESTINATIONS.

Data Pipeline

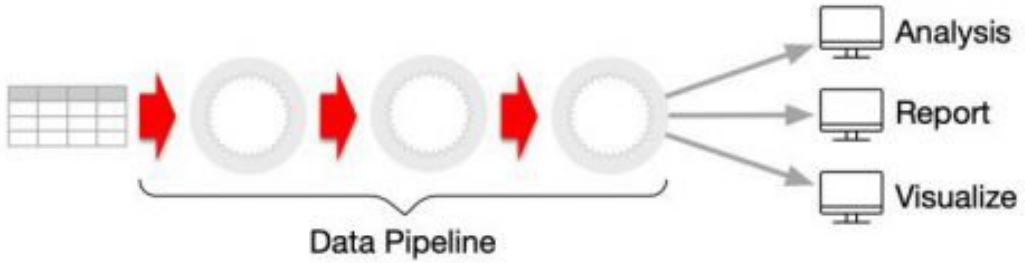
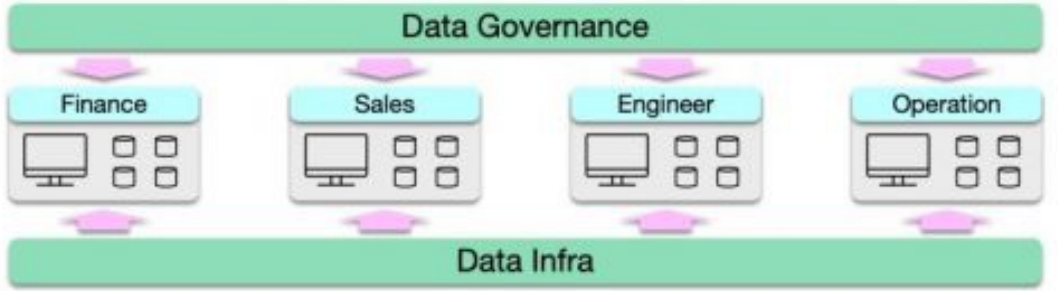
DATA PIPELINES HANDLE TASKS

DATA INGESTION,

DATA TRANSFORMATION,

DATA QUALITY CHECKS,

DATA DELIVERY.

Data Pipeline	A process that moves and transforms data from one system to another, often used to populate data warehouses and data lakes	
Data Mesh	An architectural and organizational approach where data ownership and delivery are decentralized across domain-specific, cross-functional teams	

Data Mesh

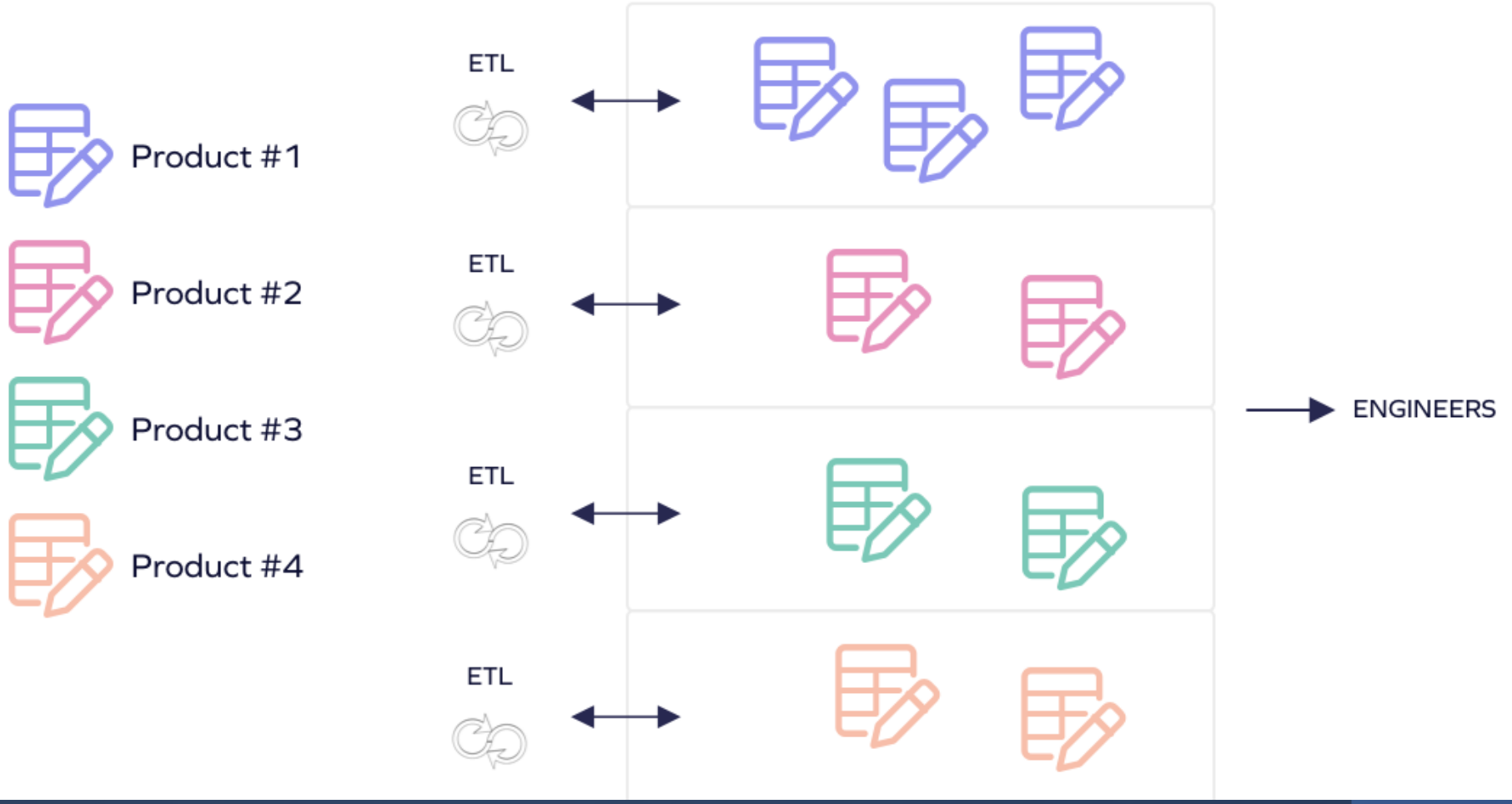
Data Mesh is a

decentralized approach

to data architecture and

management.

Data Mesh Architecture



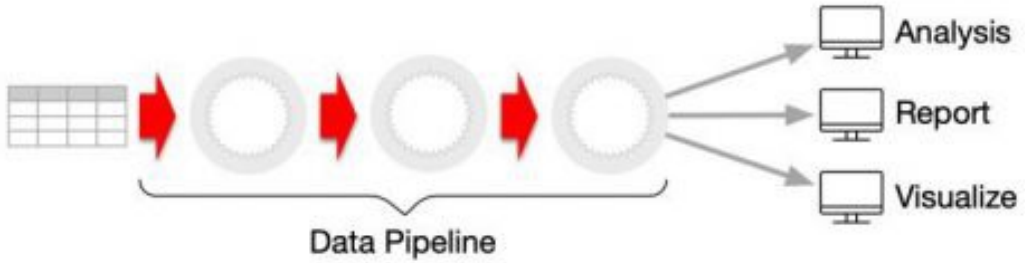
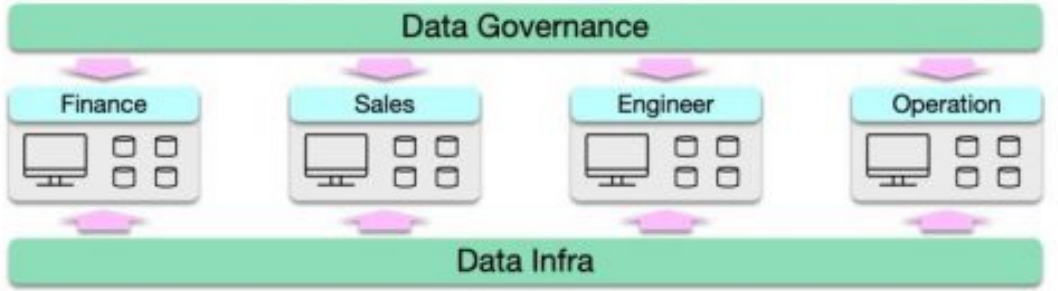
Data Mesh

IT EMPHASIZES DOMAIN-ORIENTED,

SELF-SERVE DATA PRODUCTS,

WHERE CROSS-FUNCTIONAL TEAMS

TAKE OWNERSHIP OF THEIR DATA DOMAINS.

Data Pipeline	A process that moves and transforms data from one system to another, often used to populate data warehouses and data lakes	
Data Mesh	An architectural and organizational approach where data ownership and delivery are decentralized across domain-specific, cross-functional teams	

Data Mesh

DATA MESH AIMS TO EMPOWER TEAMS

WITH DATA AUTONOMY, ALLOWING THEM

TO DEFINE THEIR DATA MODELS,

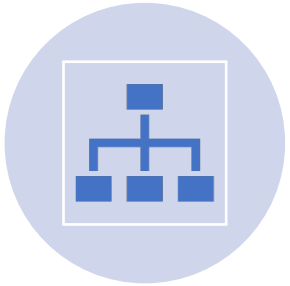
DATA PIPELINES, AND DATA PRODUCTS.

Data Mesh

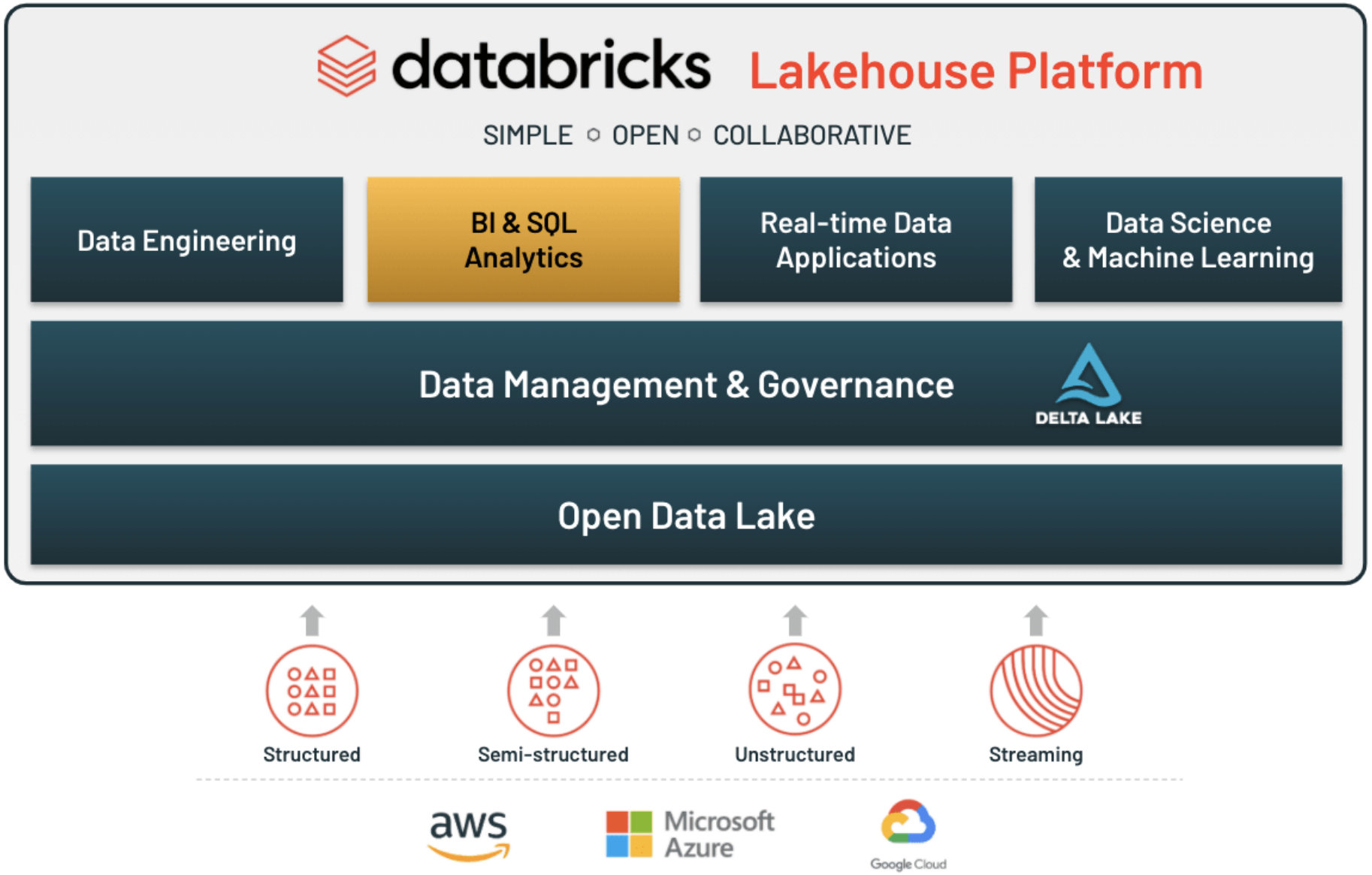
It promotes the idea of

treating data as a product and

advocates for data democratization and



data collaboration across an organization.


databricks
Lakehouse Platform
SIMPLE ○ OPEN ○ COLLABORATIVE
Data Engineering
BI & SQL Analytics
Real-time Data Applications
Data Science & Machine Learning
Data Management & Governance
DELTA LAKE
Open Data Lake
Structured
Semi-structured
Unstructured
Streaming
aws
Microsoft Azure
Google Cloud

What is Delta Lake?

DELTA LAKE IS AN OPEN-SOURCE STORAGE LAYER

BUILT ON TOP OF A DATA LAKE,

OFTEN USED IN CONJUNCTION

WITH APACHE SPARK.

What is Delta Lake?

Provide Atomicity, Consistency, Isolation, Durability

Transactions,

Data versioning

Data reliability

Improve the quality

Reliability of data in a data lake

What is Delta Lake?

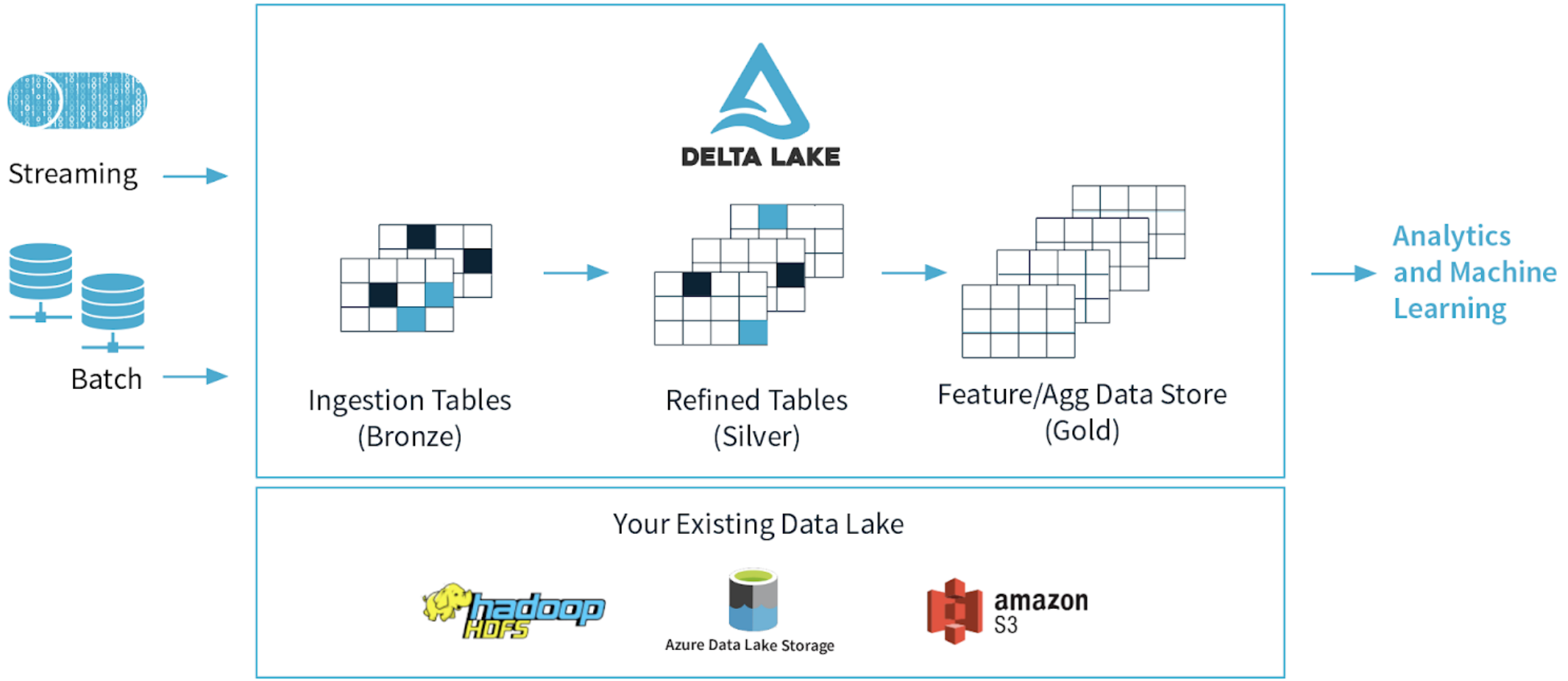
Combines the Scalability

Flexibility of a data lake

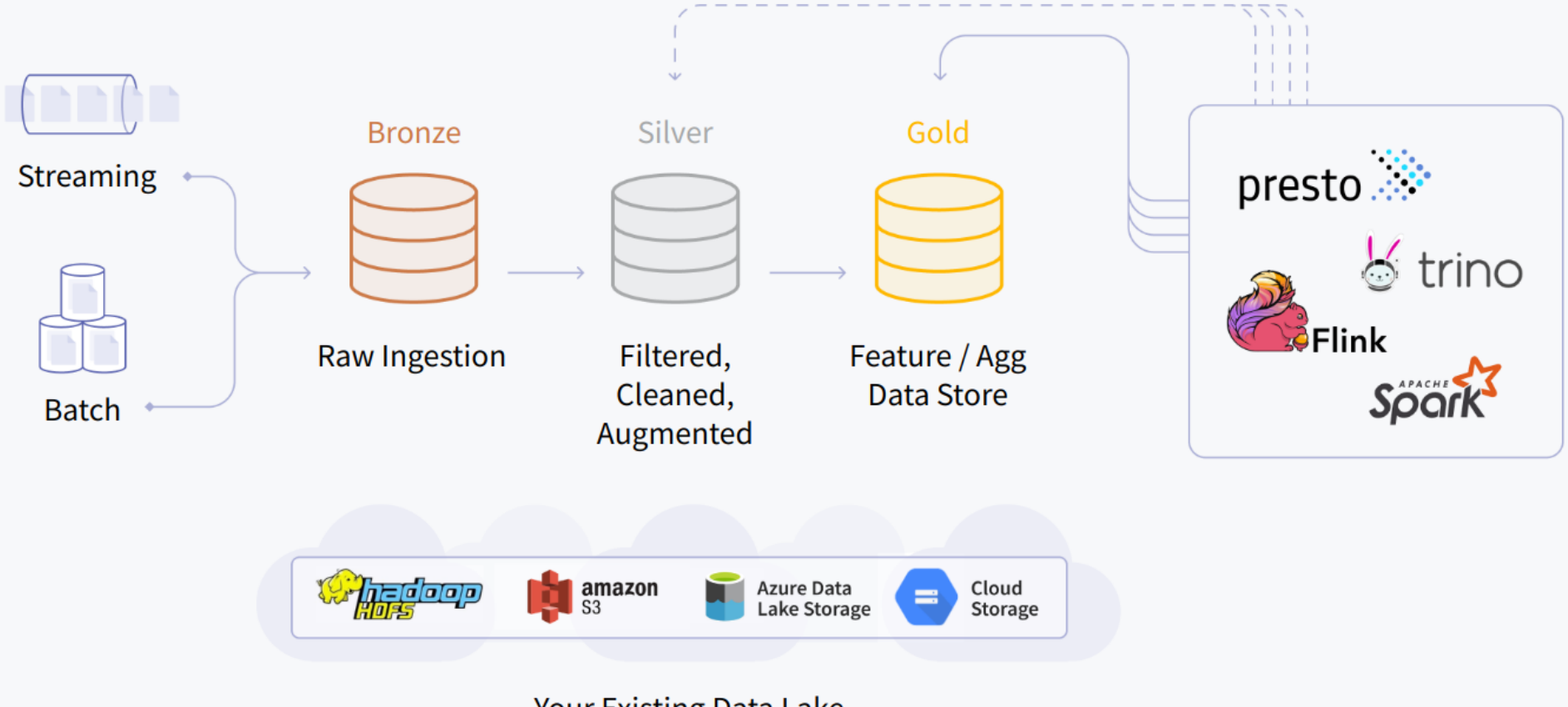
With the Reliability

Transactional capabilities

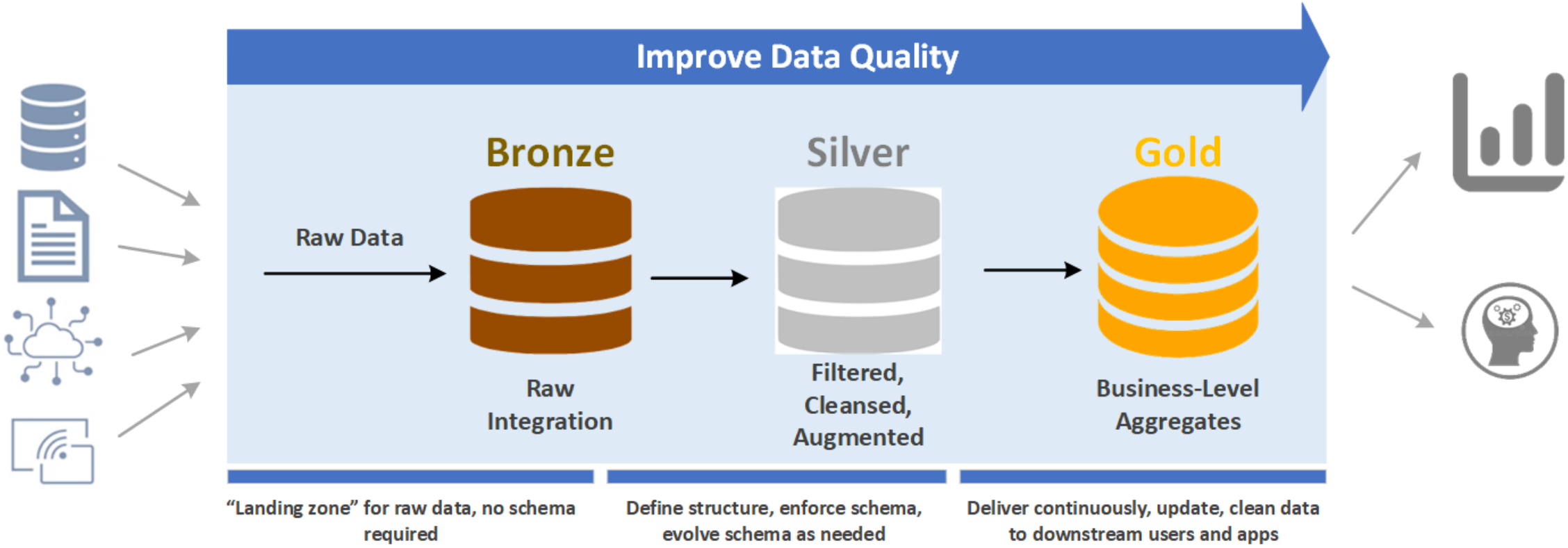
of a Data Warehouse.


Streaming
Batch
DELTA LAKE
Ingestion Tables
(Bronze)
Refined Tables
(Silver)
Feature/Agg Data Store
(Gold)
Analytics
and Machine
Learning
Your Existing Data Lake
hadoop
HDFS
Azure Data Lake Storage
amazon
S3

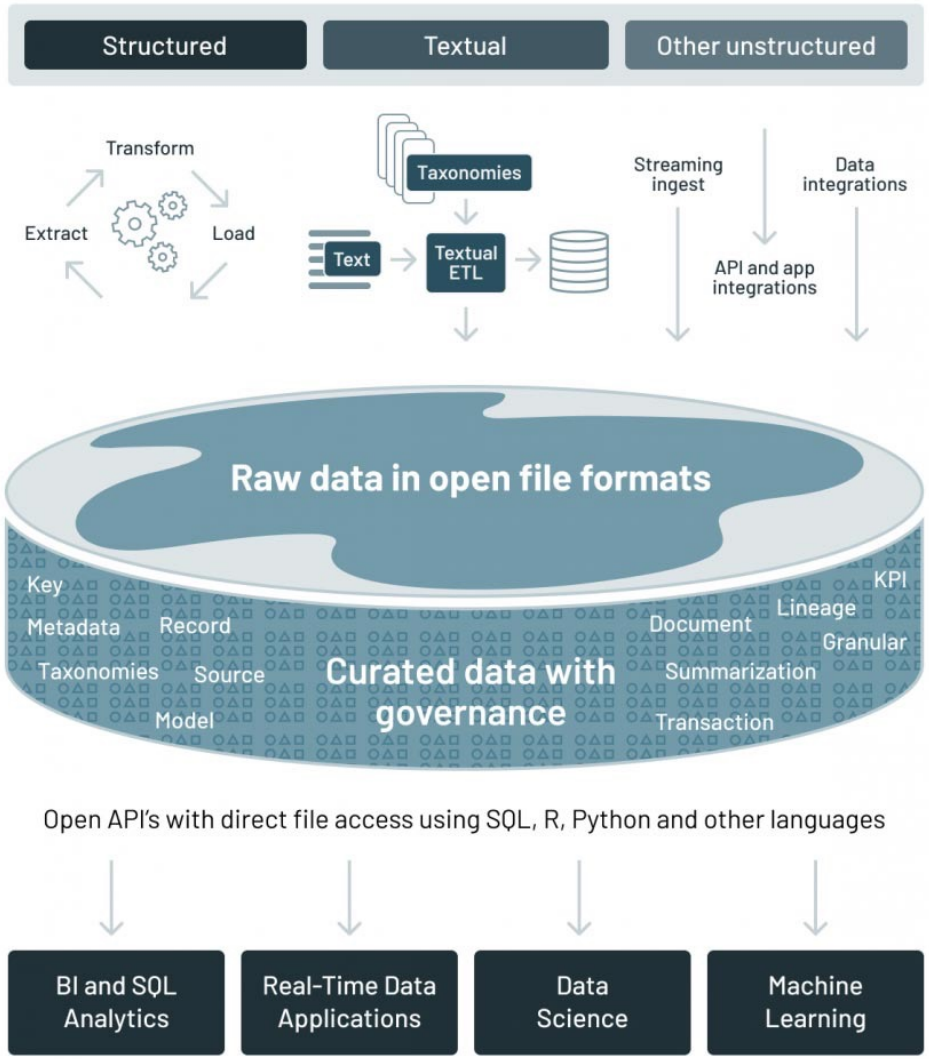
DELTA LAKE



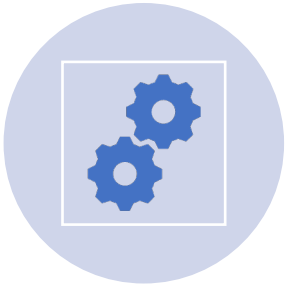
BIGDATA


Improve Data Quality
Bronze
Silver
Gold
Raw Data
Raw
Integration
Filtered,
Cleansed,
Augmented
Business-Level
Aggregates
"Landing zone" for raw data, no schema required
Define structure, enforce schema, evolve schema as needed
Deliver continuously, update, clean data to downstream users and apps

Data Lakehouse



What is Delta Lake?

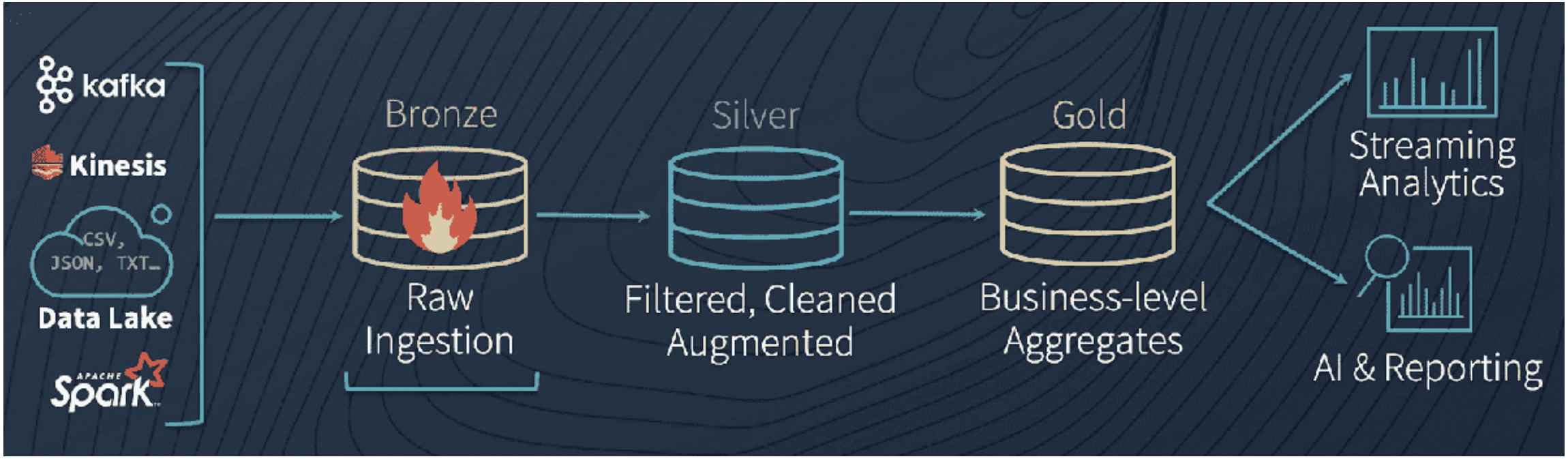


Designed to enhance the reliability,

Performance, and

Operational ease of data processing

Analytics on large-scale datasets.


kafka
Kinesis
CSV,
JSON, TXT…
Data Lake
APACHE Spark
Bronze
Raw
Ingestion
Silver
Filtered, Cleaned
Augmented
Gold
Business-level
Aggregates
Streaming
Analytics
AI & Reporting

What is Delta Lake?

Delta Lake is tightly integrated

with the Databricks platform and

offers significant advantages

for managing and analyzing data

within Databricks.

What is Delta Lake?

Serves as a powerful addition

to your data processing toolkit.

How Delta Lake functions within Databricks?

Surendra Panpaliya

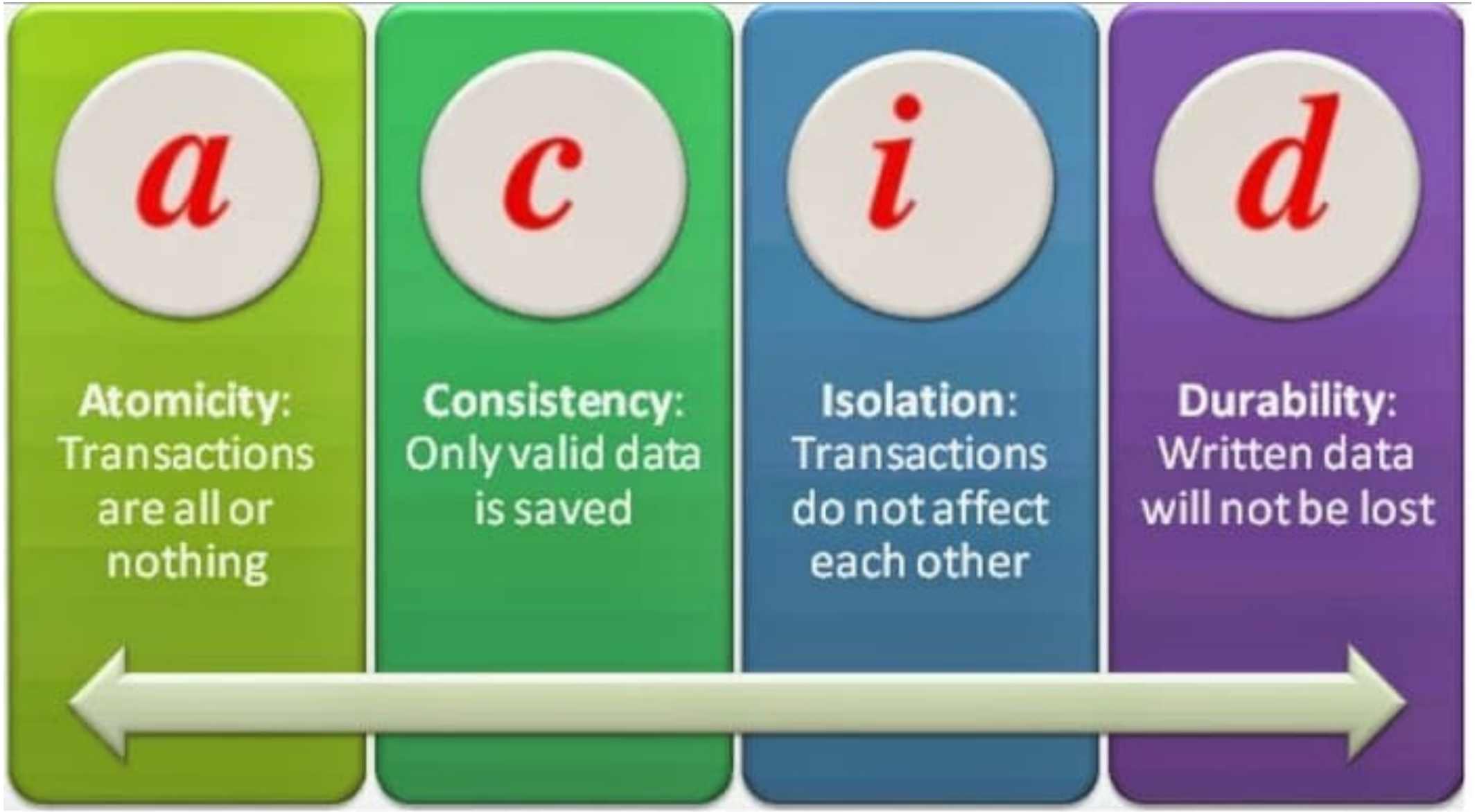
ACID Transactions

DL provides Atomicity, Consistency,

Isolation, and Durability (ACID) guarantees for data operations.

Perform updates, inserts, and deletes on your data

while ensuring data integrity and consistency.


a
Atomicity:
Transactions
are all or
nothing
c
Consistency:
Only valid data
is saved
i
Isolation:
Transactions
do not affect
each other
d
Durability:
Written data
will not be lost

Scalable Metadata Handling

Delta Lake efficiently manages Metadata

Utilizing a transaction log

Captures all changes to the data.

Scalable Metadata Handling

Ensures Quick access

to Metadata,

Regardless of

Data volume

Time Travel

Delta Lake's time travel capabilities

Allow you to access and query

historical versions of your data.

Time Travel

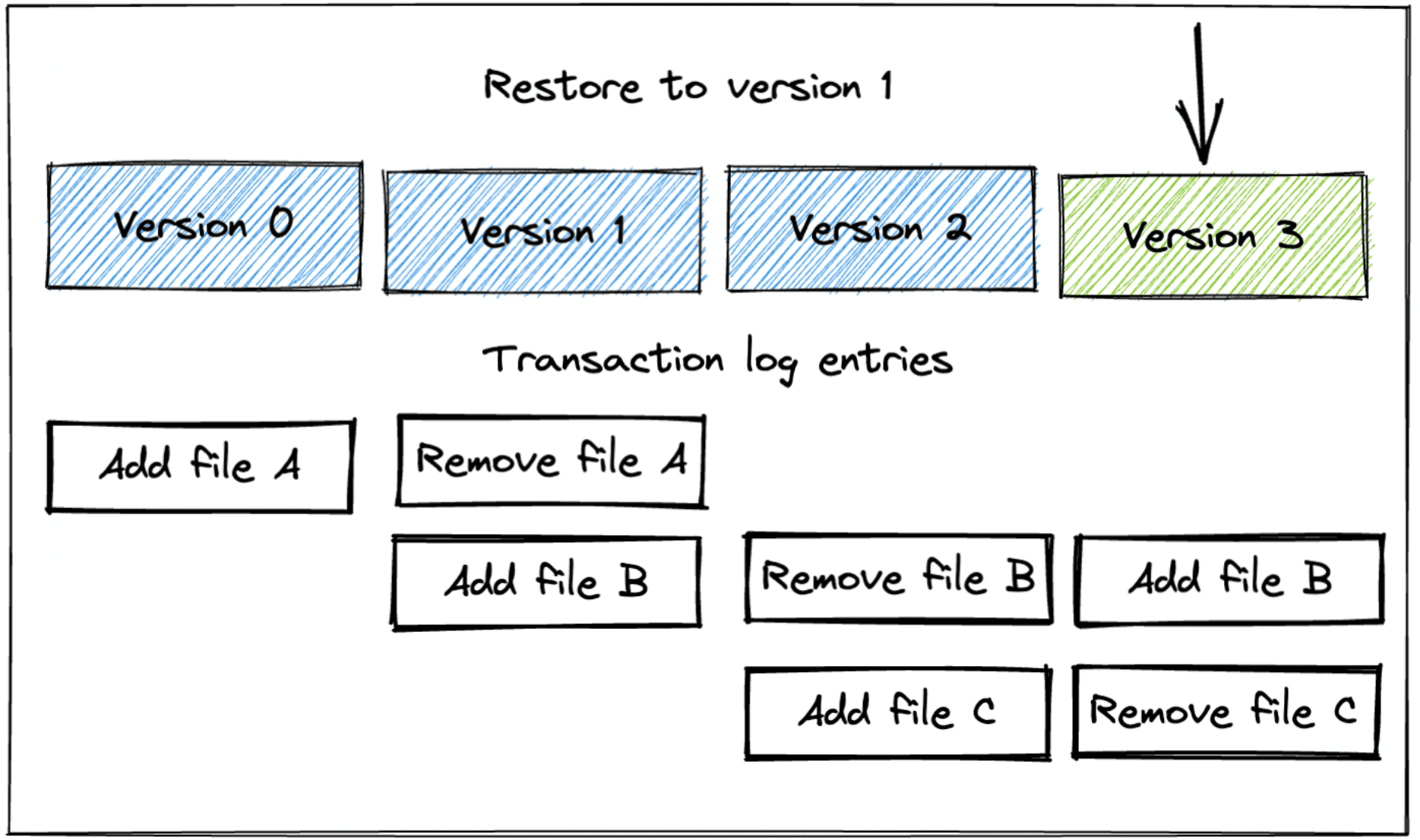
Can query the data

as it appeared at different points in time,

Aiding in Debugging,

Auditing

Historical analysis.


Restore to version 1
Version 0
Version 1
Version 2
Version 3
Transaction log entries
Add file A
Remove file A
Add file B
Remove file B
Add file B
Add file C
Remove file C

Schema Evolution

Supports schema evolution

Enabling you to modify the data schema

Without requiring complex ETL processes.

Easier to adapt to changing

business requirements.

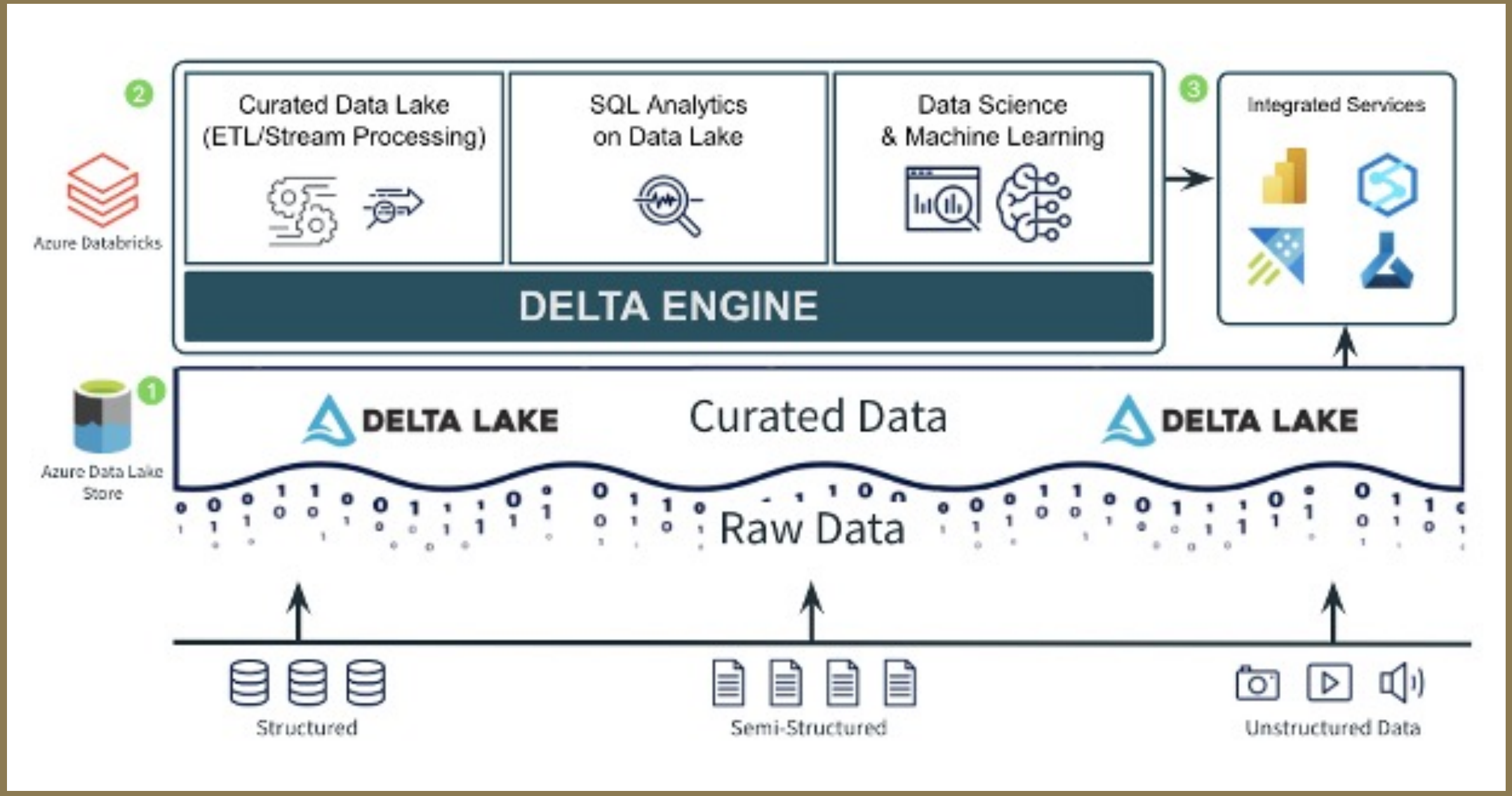
Delta Lake schema evolution

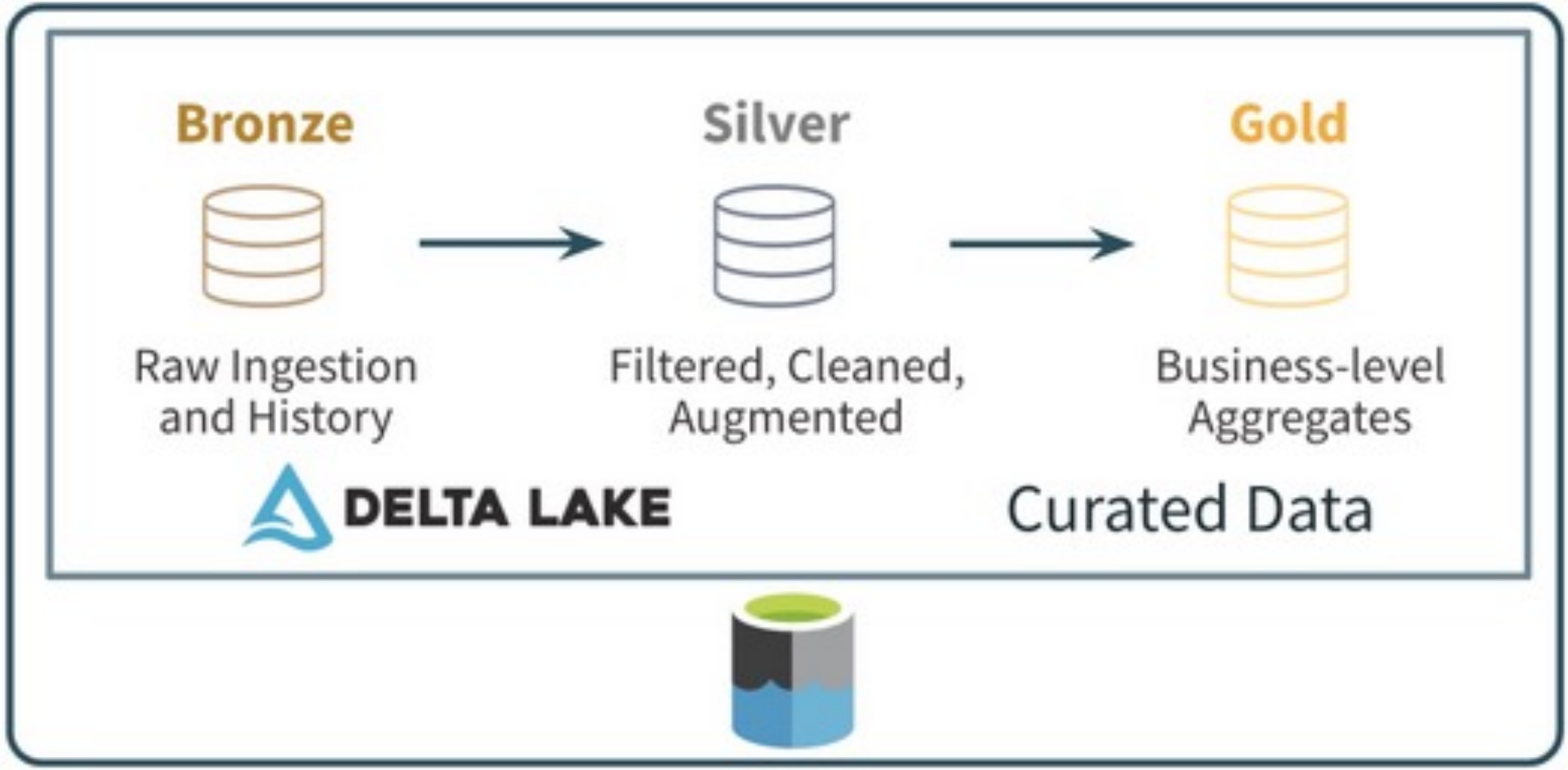
name	age
bob	3
sue	5

name	age	country
bob	3	usa
sue	5	uk

Data Consistency Checks

- Automatically performs
- Data consistency checks,
- Including Data Integrity,
- Schema validation, and Statistics validation.
- Ensures that your data is accurate and reliable.


2
Curated Data Lake
(ETL/Stream Processing)
SQL Analytics
on Data Lake
Data Science
& Machine Learning
3
Integrated Services
Azure Databricks
DELTA ENGINE
1
Azure Data Lake
Store
DELTA LAKE
Curated Data
DELTA LAKE
Raw Data
Structured
Semi-Structured
Unstructured Data


Bronze
Silver
Gold
Raw Ingestion and History
Filtered, Cleaned, Augmented
Business-level Aggregates
DELTA LAKE
Curated Data

Improved Performance

Delta Lake optimizes data storage

Using Advanced Indexing

Data Skipping Techniques,

Leading to improved query performance

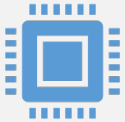
Reduced I/O overhead.

Unified Batch and Streaming

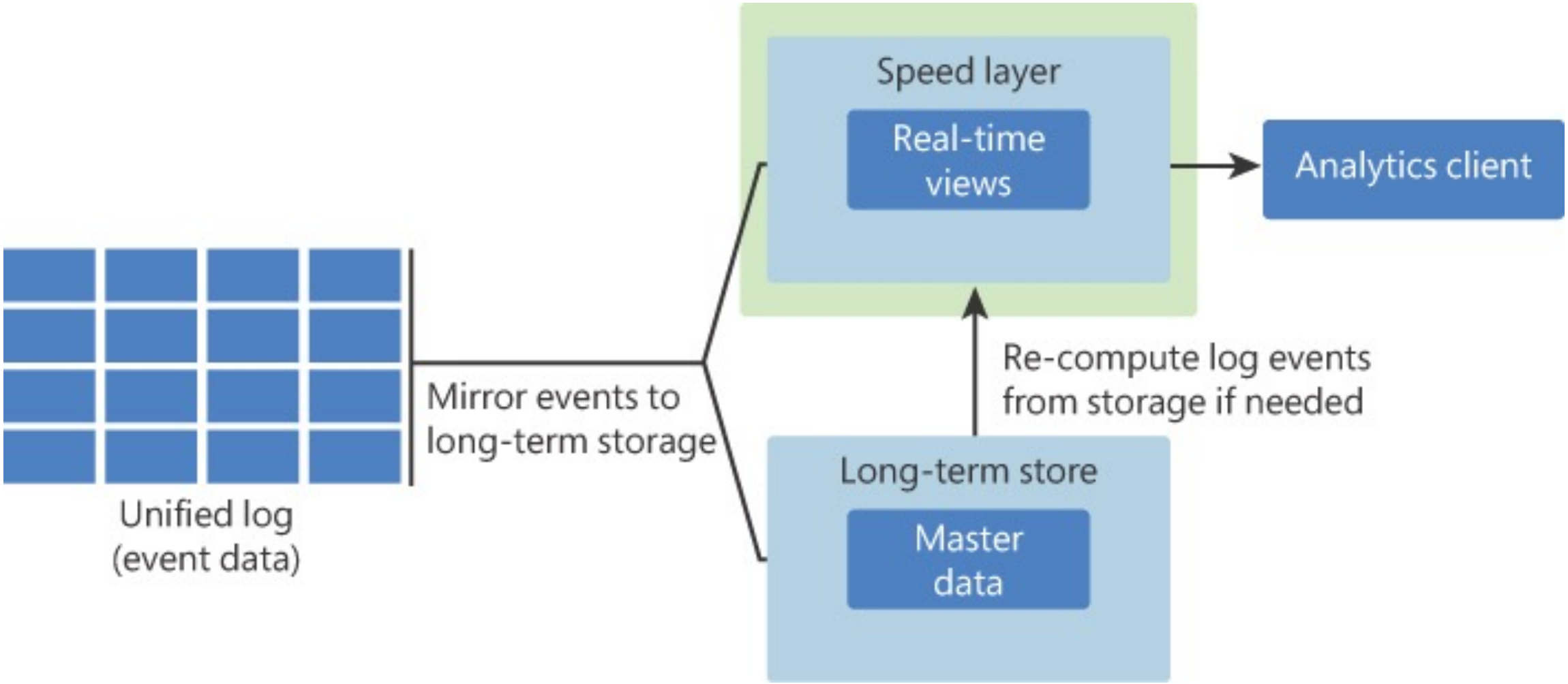
Seamlessly integrate

batch and structured streaming workloads using Delta Lake.

Write streaming data into Delta Lake tables



query them using both batch and streaming processing.


Speed layer
Real-time views
Analytics client
Mirror events to long-term storage
Re-compute log events from storage if needed
Long-term store
Master data
Unified log (event data)

Integration with Databricks

Delta Lake is closely integrated with Databricks,

making it straightforward

to create, manage, and

analyze Delta Lake tables

directly within the Databricks environment.

Integration with Databricks

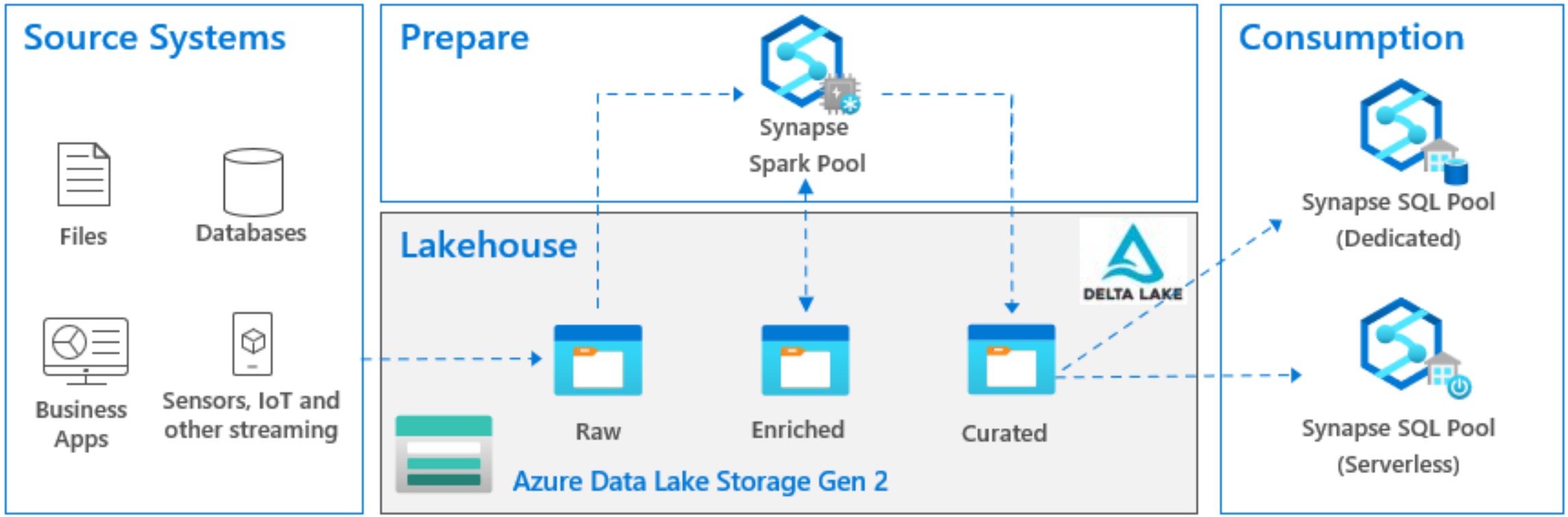
Databricks provides built-in commands and

features to work with

Delta Lake efficiently.

	Delta Lake	Lake ETLs
Lock-in	**High** Need to change ingestion and query interfaces to Delta, no support for reliable concurrent writes.	**Low**. No change in interfaces and no proprietary metadata Lake ETL vendor can be replaced with home-grown ETL.
Ingestion Performance	**Low**. ACID transactions and indexes.	**High**. Append-only writes.
Entry barrier	**High.** Requires non-trivial expertise in Spark coding	**Low** Visual interface and SQL
Ease-of-use	**Medium**. ACID operations replace ETLs but all ingestion and query interfaces need to be migrated to Delta. Delta requires a DBA for operations like Vacuum and Optimize	**Depends** on ETL platform Upsolver offers a turn-key solution, automating ETLs.

DATA WAREHOUSE	vs.	DATA LAKE
structured, processed	DATA	structured / semi-structured / unstructured, raw
schema-on-write	PROCESSING	schema-on-read
expensive for large data volumes	STORAGE	designed for low-cost storage
less agile, fixed configuration	AGILITY	highly agile, configure and reconfigure as needed
mature	SECURITY	maturing
business professionals	USERS	data scientists et. al.


Source Systems
Files
Databases
Business Apps
Sensors, IoT and other streaming
Prepare
Synapse Spark Pool
Lakehouse
DELTA LAKE
Raw
Enriched
Curated
Azure Data Lake Storage Gen 2
Consumption
Synapse SQL Pool (Dedicated)
Synapse SQL Pool (Serverless)

Summary

Delta Lake in Databricks enhances your ability

to build reliable and efficient data pipelines,

enables historical analysis, and

simplifies the process of managing large-scale data lakes.

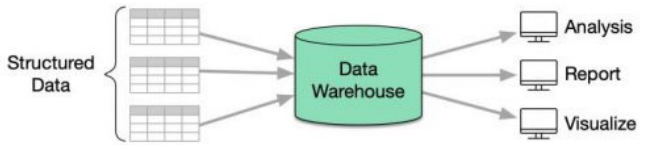
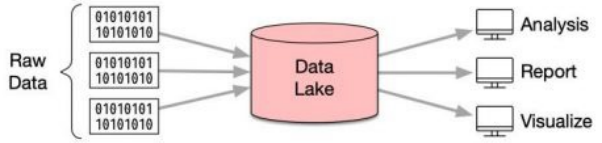
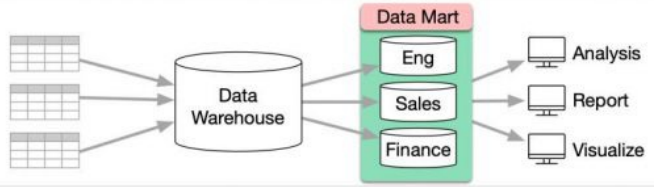
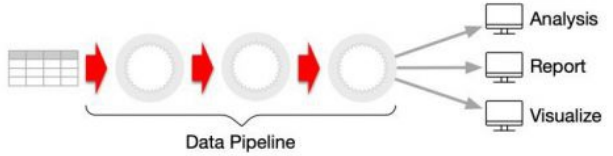
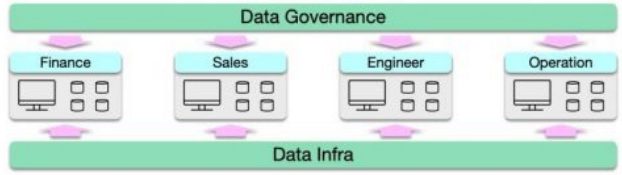
Summary

It's a valuable tool

for organizations seeking

to leverage the full potential of their data

within the Databricks platform.

Data Warehouse	A large, structured repository of integrated data from various sources, used for complex querying and historical analysis	
Data Lake	A more focused, department-specific subset of a data warehouse providing quick data retrieval and analysis	
Data Mart	A vast pool of raw, unstructured data stored in its native format until it's needed for use	
Delta Lake	An open-source storage layer that brings reliability and ACID transactions to data lakes, unifying batch and streaming data processing	
Data Pipeline	A process that moves and transforms data from one system to another, often used to populate data warehouses and data lakes	
Data Mesh	An architectural and organizational approach where data ownership and delivery are decentralized across domain-specific, cross-functional teams	

Data Warehouse vs Data Mart

Data Warehouse:

- A large, structured repository of integrated data
- from various sources, used for complex querying and historical analysis.

Data Mart:

- A more focused, department-specific subset of a data warehouse
- providing quick data retrieval and analysis.

Data Lake vs Delta Lake

Data Lake

- A vast pool of raw, unstructured data
- stored in its native format until it’s needed for use.

Delta Lake: An open-source storage layer

- that brings reliability and ACID transactions to data lakes,
- unifying batch, and streaming data processing.

Data Pipeline vs Data Mesh

Data Pipeline

A process that moves and transforms data from one system to another, often used to populate data warehouses and data lakes.

Data Mesh:

An architectural and organizational approach where data ownership and delivery are decentralized across domain-specific, cross-functional teams.

Delta Lake Hands On Session

Surendra Panpaliya

Delta Lake operations

Perform a variety of operations on Delta Lake tables

To manage, query, and transform your data.

Delta Lake offers capabilities for data manipulation,

schema evolution, time travel

Create a Delta Lake Table

create a new Delta Lake table from existing data

Such as Parquet files or structured data.

from pyspark.sql import SparkSession

Create a Spark session

spark = SparkSession.builder.appName("DeltaLakeExample").getOrCreate()

Create a Delta Lake Table

```
# Read data from a source (e.g., Parquet files)
```

```
data_df = spark.read.parquet("/path/to/source/data")
```

```
# Write data to a Delta Lake table
```

```
data_df.write.format("delta").save("/path/to/delta-table")
```

Z-Ordering

Z-Ordering is a technique

to optimize query performance

by co-locating related data

in the same files.

Z-Ordering

You can use the OPTIMIZE command

to optimize the table's layout

based on specified columns

delta_table.optimize("column_name")

Cache Metadata

Caching the metadata of a

Delta Lake table can

improve query performance.

delta_table.cache()

Update Statistics

Delta Lake maintains statistics

Help improve query optimization.

Periodically updating these statistics

can lead to better query plans

delta_table.generate("STATISTICS")

Liquid Clustering in Delta Lake Databricks

Available in Public Preview in Databricks Runtime 13.2 and above.

Delta Lake liquid clustering

Replaces table partitioning &

ZORDER to simplify data layout decisions

Optimize query performance.

Liquid Clustering in Delta Lake Databricks

Liquid clustering provides flexibility

to redefine clustering keys

without rewriting existing data,

allowing data layout

to evolve alongside analytic needs over time.

Liquid Clustering in Delta Lake Databricks

- https://docs.databricks.com/en/delta/clustering.html