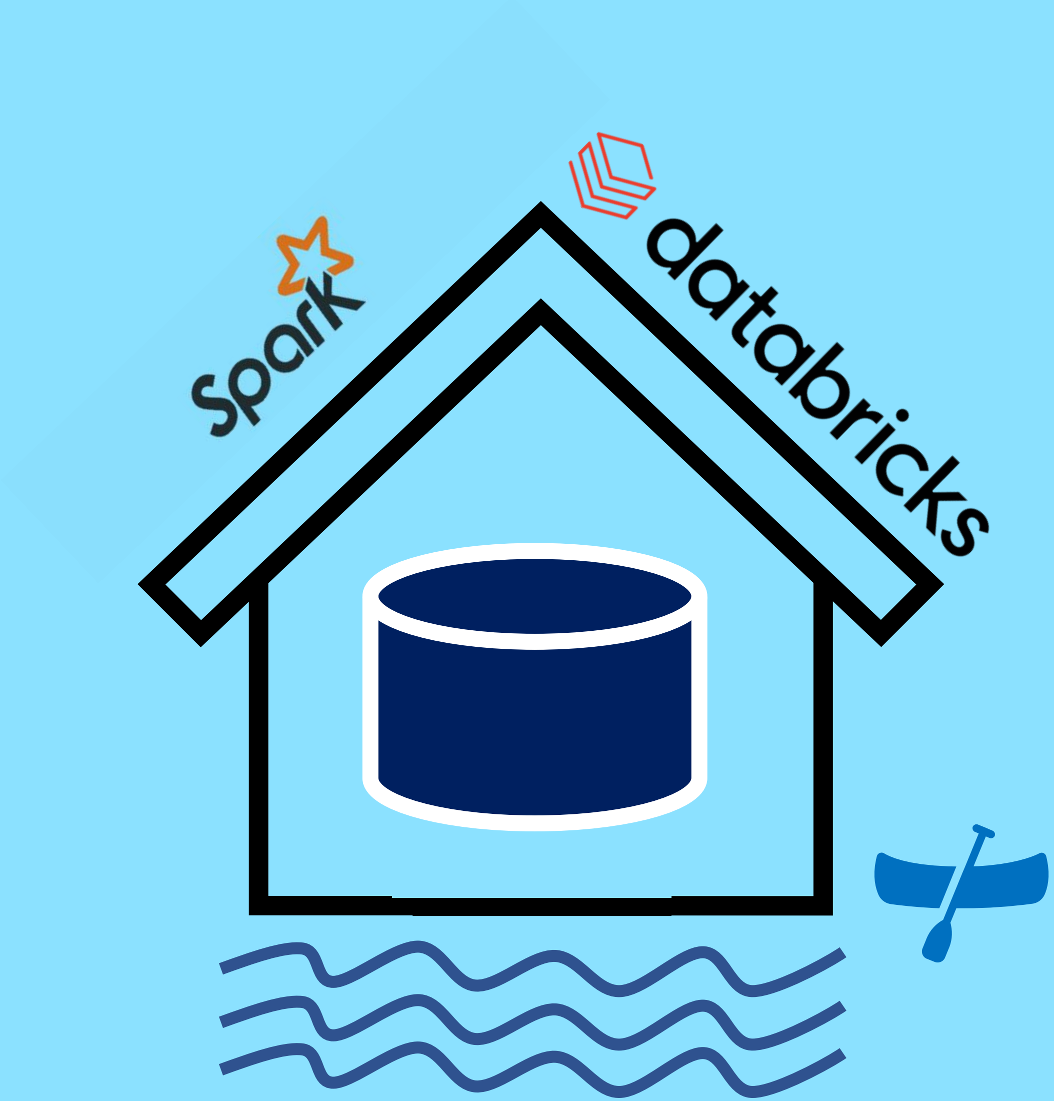
Understanding Data Lakehouse

patreon



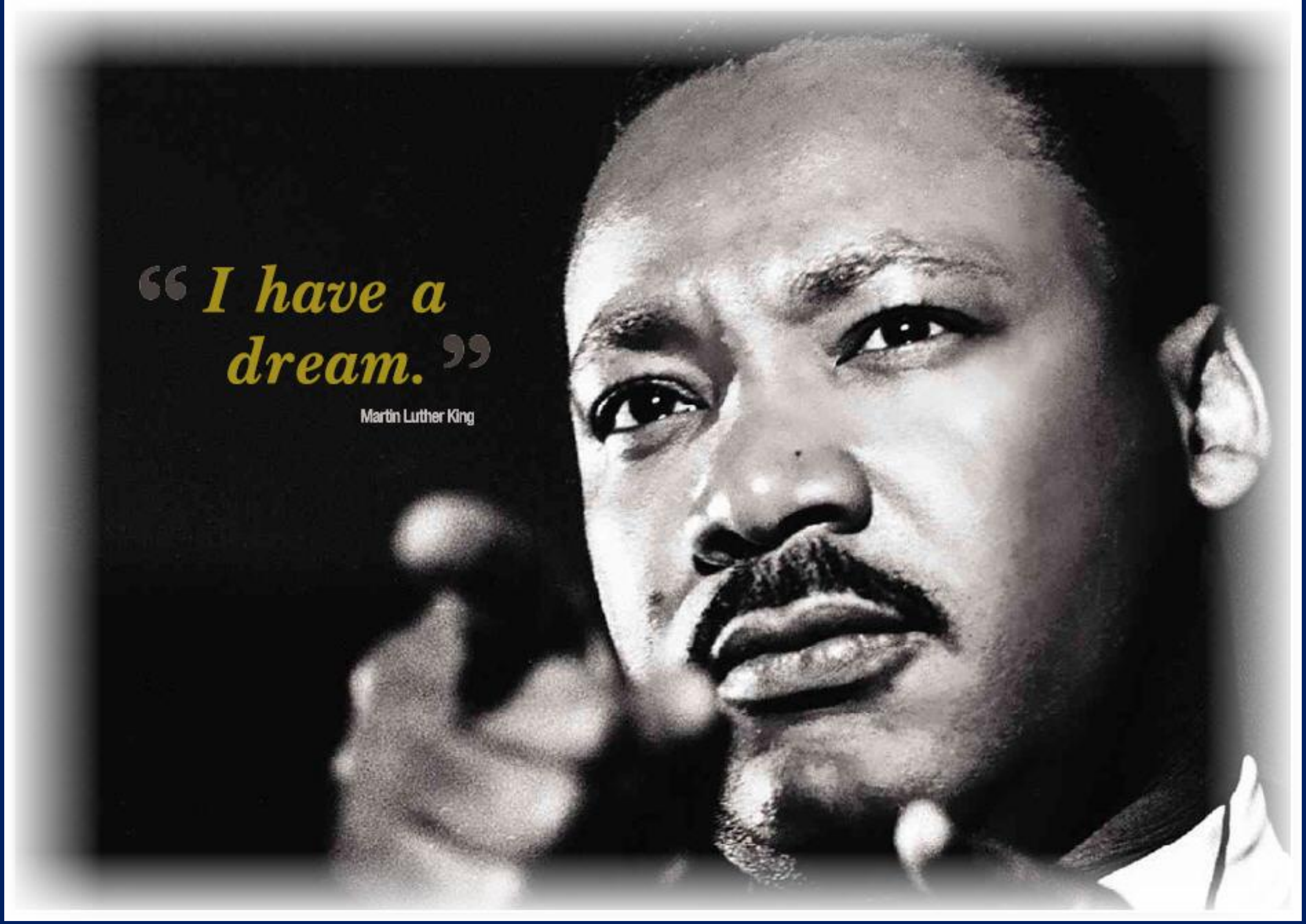
Bryan Cafferky Copyright 2023

Where Are We Going?

Laying the Conceptual Foundation

- I Have a Dream
- Is/Is Not
- What You Must Understand!

It Starts With a Vision



The Data Lakehouse Vision

I have a dream that all data will flow automatically to where it is needed and, in a format easily consumed seamlessly providing data driven decision making.

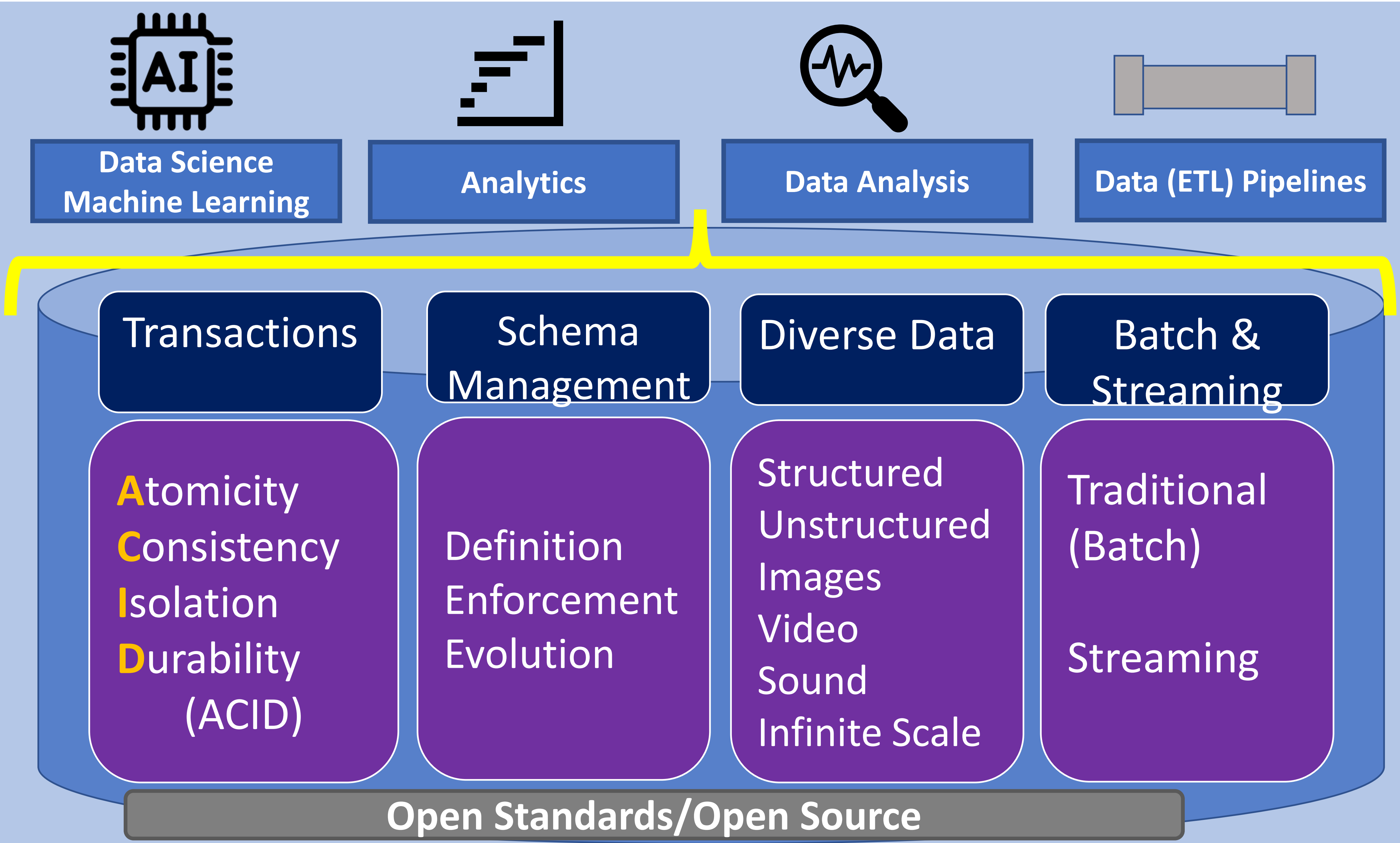
"a Data Lakehouse is an architecture that enables efficient and secure Artificial Intelligence (AI) and Business Intelligence (BI) directly on vast amounts of data stored in Data Lakes."

https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html

Is/Is Not

What Data Lakehouse Is	What Data Lakehouse Is Not
A Vision and Roadmap to the Future Data Warehouse	A specific technology
An Evolving Set of Ideas and Objectives	Fait Complete! It Ain't Done!
Open to Many Implementations	Specific to a Vendor or Product
Broad Outline or Specification	Dictatorship
Implementation Requires Many Levels: • Frameworks • Technical Services and Libraries • Standards • Patterns	Not One Thing

Data Lakehouse Goals



Data Science Machine Learning

Analytics

Data Analysis

Data (ETL) Pipelines

Transactions

Atomicity
Consistency
Isolation
Durability
(ACID)

Schema Management

Definition
Enforcement
Evolution

Diverse Data

Structured
Unstructured
Images
Video
Sound
Infinite Scale

Batch & Streaming

Traditional (Batch)

Streaming

Open Standards/Open Source

Wrapping Up

Laying the Conceptual Foundation

- I Have a Dream
- Is/Is Not
- What You Must Understand!

Thank You!

Wrapping Up

Laying the Conceptual Foundation

- I Have a Dream
- Is/Is Not
- What You Must Understand!

ACID Transactions
Schema Governance
Diverse Data
Batch & Streaming

What is Data Lakehouse?

A Conceptual Framework

Set of Conventions, aka patterns

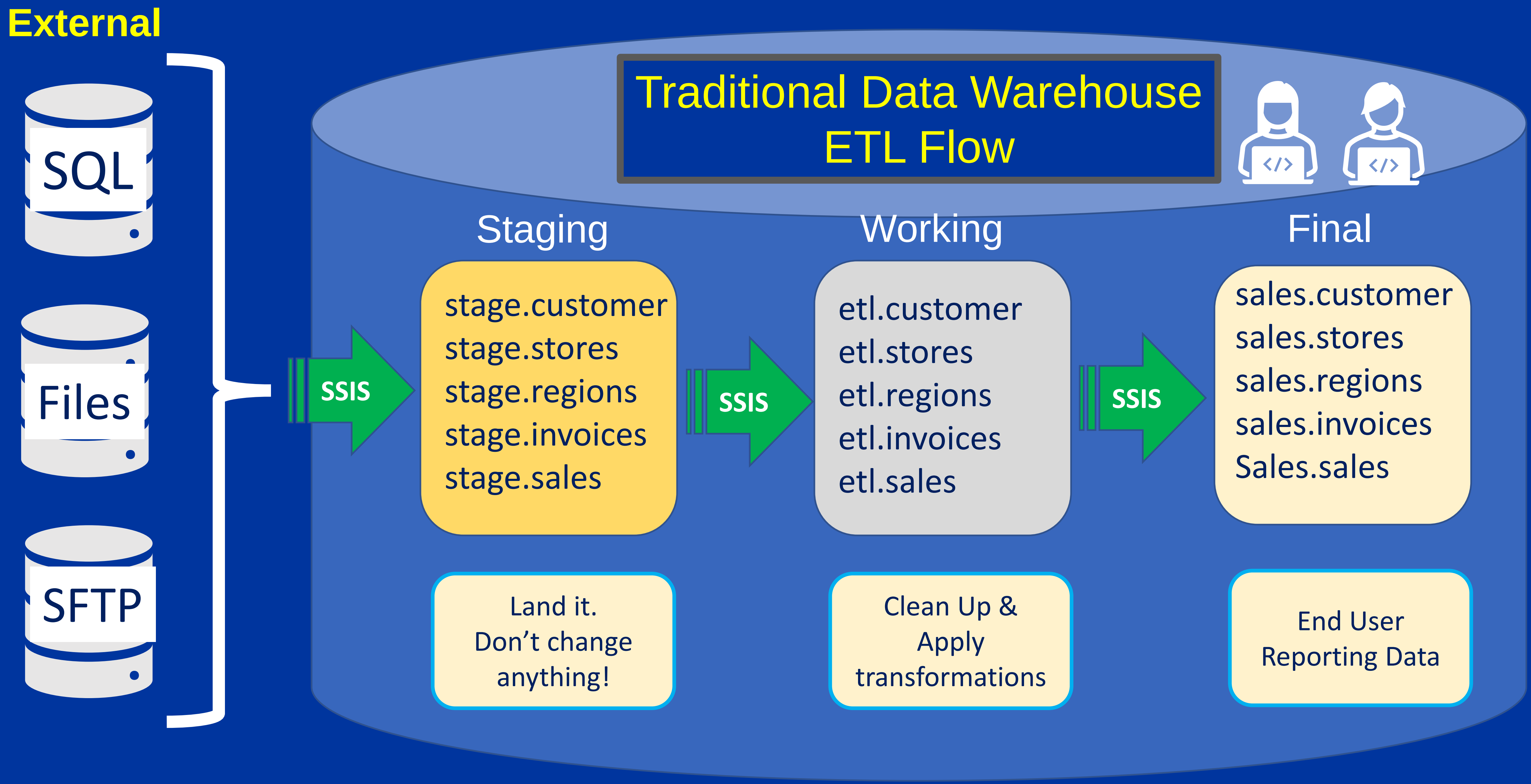
Leverages Spark Delta Framework provided as a Service

Leverages the Delta Table API

Delta is an Enhancement to Parquet

Open Standards and Open Source

Traditional Data Warehouse ETL Flow



External

SQL

Files

SFTP

SSIS

Staging

stage.customer
stage.stores
stage.regions
stage.invoices
stage.sales

Land it.
Don’t change anything!

SSIS

Working

etl.customer
etl.stores
etl.regions
etl.invoices
etl.sales

Clean Up & Apply transformations

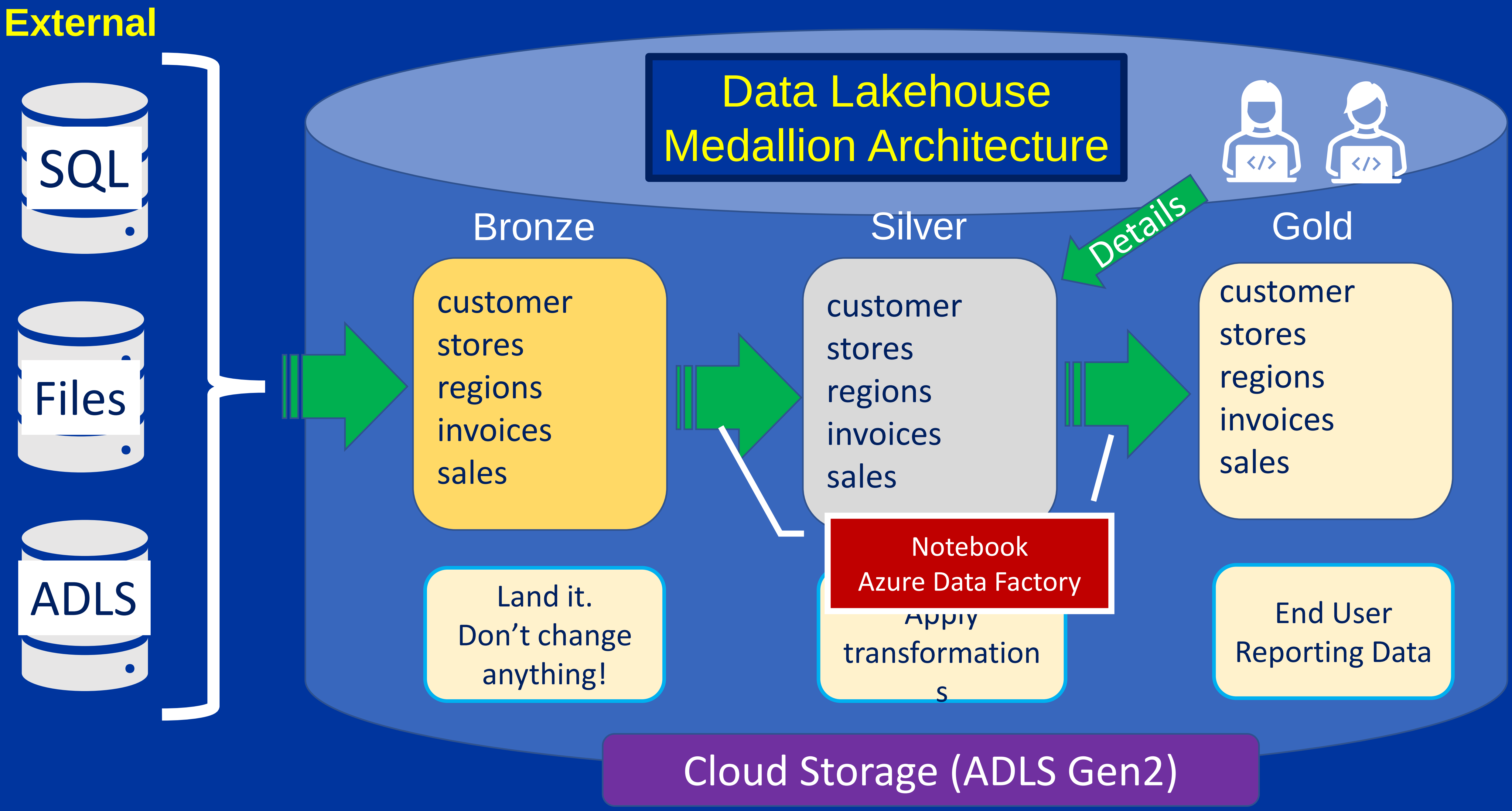
SSIS

Final

sales.customer
sales.stores
sales.regions
sales.invoices
Sales.sales

End User Reporting Data

Data Lakehouse Medallion Architecture



External

SQL

Files

ADLS

Bronze

- customer
- stores
- regions
- invoices
- sales

Land it. Don't change anything!

Silver

- customer
- stores
- regions
- invoices
- sales

Notebook
Azure Data Factory

Apply transformations

Gold

- customer
- stores
- regions
- invoices
- sales

End User Reporting Data

Details

Cloud Storage (ADLS Gen2)

Partitioned Views Demo

1) Determine what the partition key will be? Example: Sales Year.
2) Create a separate table to hold the set of data for each key value.
3) Add a constraint to each partition table that limits the partition key to the required value.
4) Load the data into each partition table, i.e. Sales.Sales2012.
5) Create a view that queries each partition table and does a UNION ALL between each query.

Links

Lakehouse Features

What Is a Lakehouse? - The Databricks Blog

https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html