databricks

Introduction to Data Engineering in Databricks

Delta Lake Review

Delta Lake delivers open, reliable, and scalable data management for the Lakehouse, empowering you to ingest data from external sources and efficiently manage it across Bronze (raw), Silver (cleaned), and Gold (curated) layers—all with full ACID transactions, time travel, schema enforcement, and support for both batch and streaming workloads
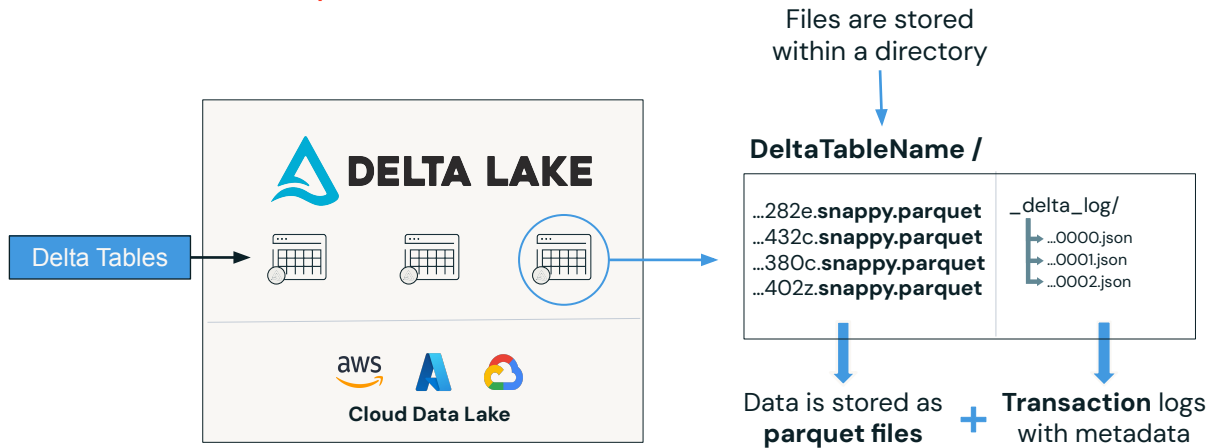
Delta Lake Review

Ingesting Data Into Delta Lake
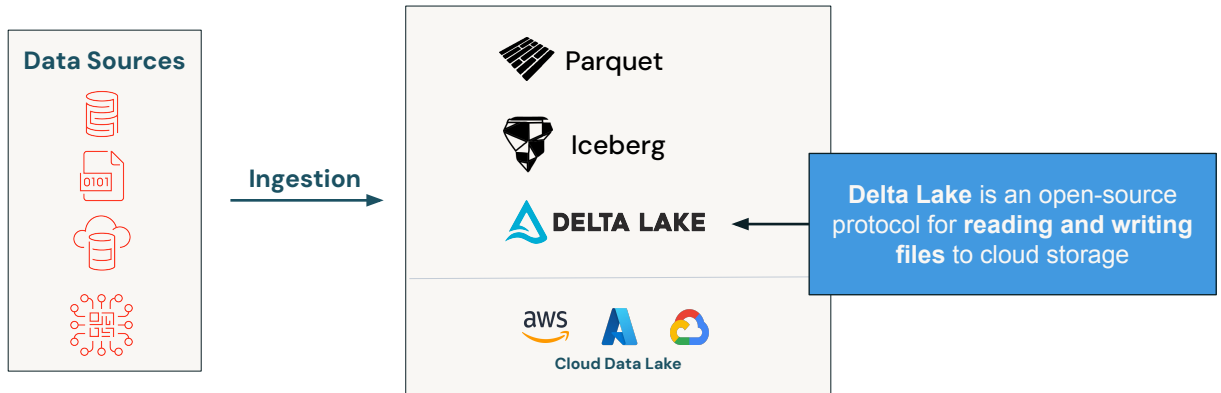
Data Sources

Ingestion →

Parquet

Iceberg

DELTA LAKE ←

Delta Lake is an open-source protocol for **reading and writing files** to cloud storage

aws A G
Cloud Data Lake

Let's quickly review Delta Lake.

The goal is to ingest files from external data sources like cloud object storage into Delta Lake as Delta tables. Remember, Delta Lake is simply an open-source protocol for reading and writing files to cloud storage. Delta tables offer an open table format that supports the Lakehouse architecture.

Delta Lake Review
Delta Table Components Overview

Files are stored within a directory

DeltaTableName /

...282e.**snappy.parquet**
...432c.**snappy.parquet**
...380c.**snappy.parquet**
...402z.**snappy.parquet**

_delta_log/
→ ...0000.json
→ ...0001.json
→ ...0002.json

Data is stored as **parquet files** + **Transaction** logs with metadata

Delta Tables

DELTA LAKE

aws

Cloud Data Lake

Within Delta Lake you will work with Delta tables.

Under the hood, Delta tables store data within a folder directory. Within the directory, data is stored as parquet files, and what Delta adds is delta logs stored as JSON files alongside the parquet files. The Delta logs keep track of all of the transactions on the data (parquet files) and table versions.

The transaction logs provide a wide array of functionality to the Delta table. With the transaction log, we now have the concept of table states, so if you insert, delete or update data in your table, Delta basically adds a transaction (the log file) and your table stays updated and managed. So, with the transaction log you are able to easily get consistent views of your data and you're actually able to travel back in time! We will dive deeper into the inner workings of Delta tables later.

Traditionally in a data lake, modifying and managing your data was a difficult task. For example, if you wanted to insert, update or delete records from a data lake file, you would typically need to manually recreate the file with the modifications and keep track of the updates. With Delta lake it's much more performant and very easy to use!

Delta Lake Review

Delta Tables Key Features Review

- ACID Transactions
- Data Manipulation Language (DML)
- Time Travel
- Schema Evolution and Enforcement
- Many more!

Delta Tables provide a variety of key features in our cloud data lake.

Let's quickly review some of those key features.

- ACID transactions (atomicity, consistency, isolation, and durability) for all operations, allowing multiple users to read and write data concurrently without conflicts.
- Supports Data Manipulation Language (DML) operations such as INSERT, UPDATE, DELETE, and MERGE, enabling flexible data management.
- Time travel allows users to query and revert to previous versions of data, facilitating auditing and recovery.
- Enforces a defined schema for data integrity while allowing schema evolution, enabling changes to the structure without breaking existing workflows.

Delta Lake also provides many other features like unified batch and streaming processing, optimization and performance, and scalability.

Delta Lake is also open source!

Delta Lake Review
Medallion Architecture (Multi Hop) Review
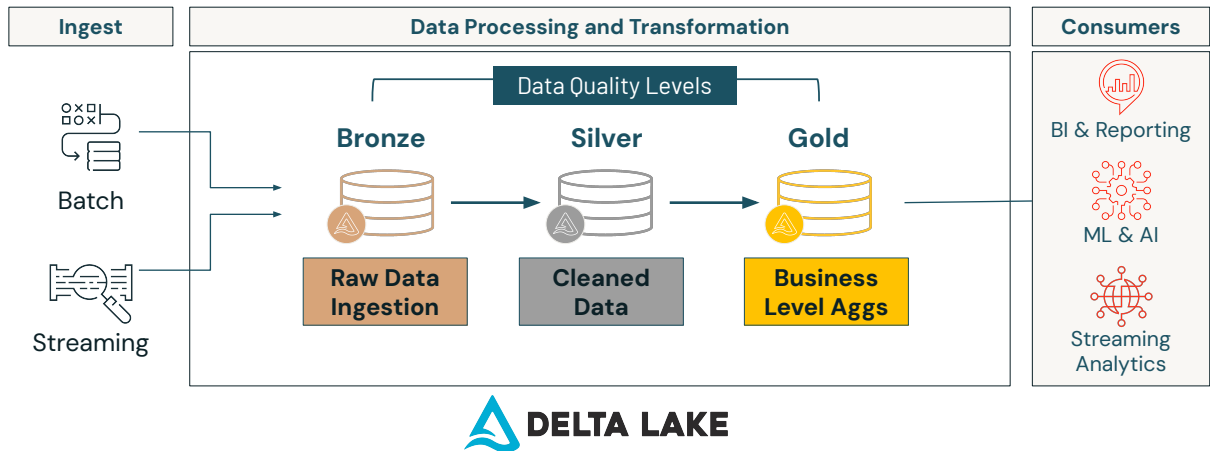
As you ingest data into your Delta Lake through batch or streaming methods, or both, you can begin processing and transforming your data in Databricks.

The goal is to incrementally and progressively improve the structure and quality of data as it moves through each layer: Bronze, Silver and Gold.

It begins with the Bronze layer, the raw data ingestion layer. This layer ingests raw, unprocessed data from various sources 'as is', serving as the foundational storage for all data.

In the Silver layer, the data is cleaned, transformed, and enriched, providing a more refined dataset that is suitable for analysis.

Lastly the Gold layer contains curated, aggregated, and high-quality data, optimized for reporting and advanced analytics, often used for business intelligence applications

Thank you for completing this lesson and continuing your journey to develop your skills with us.