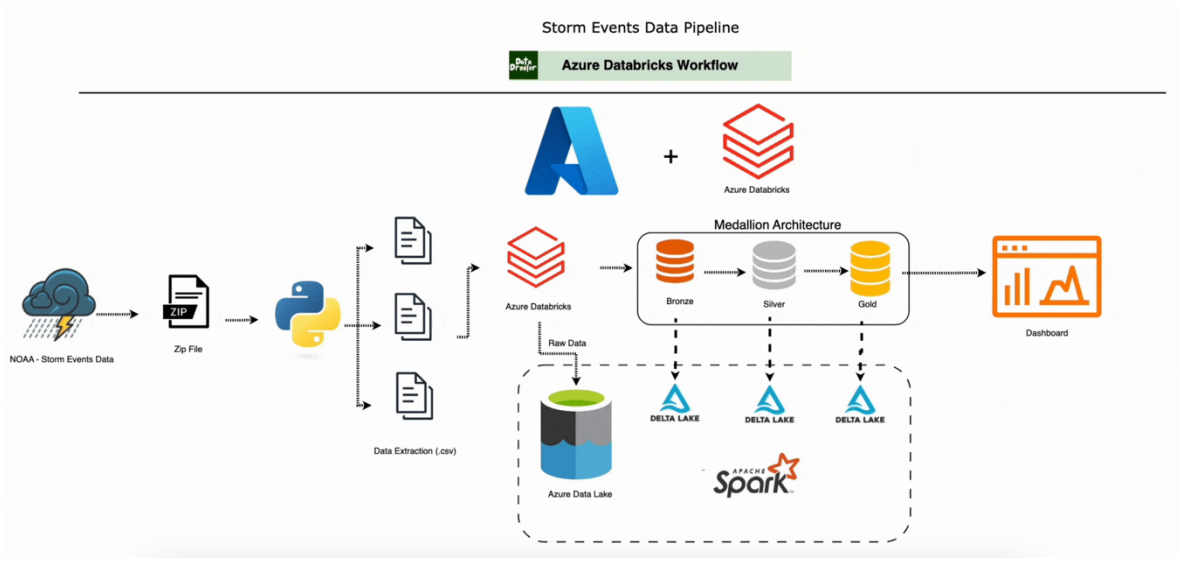
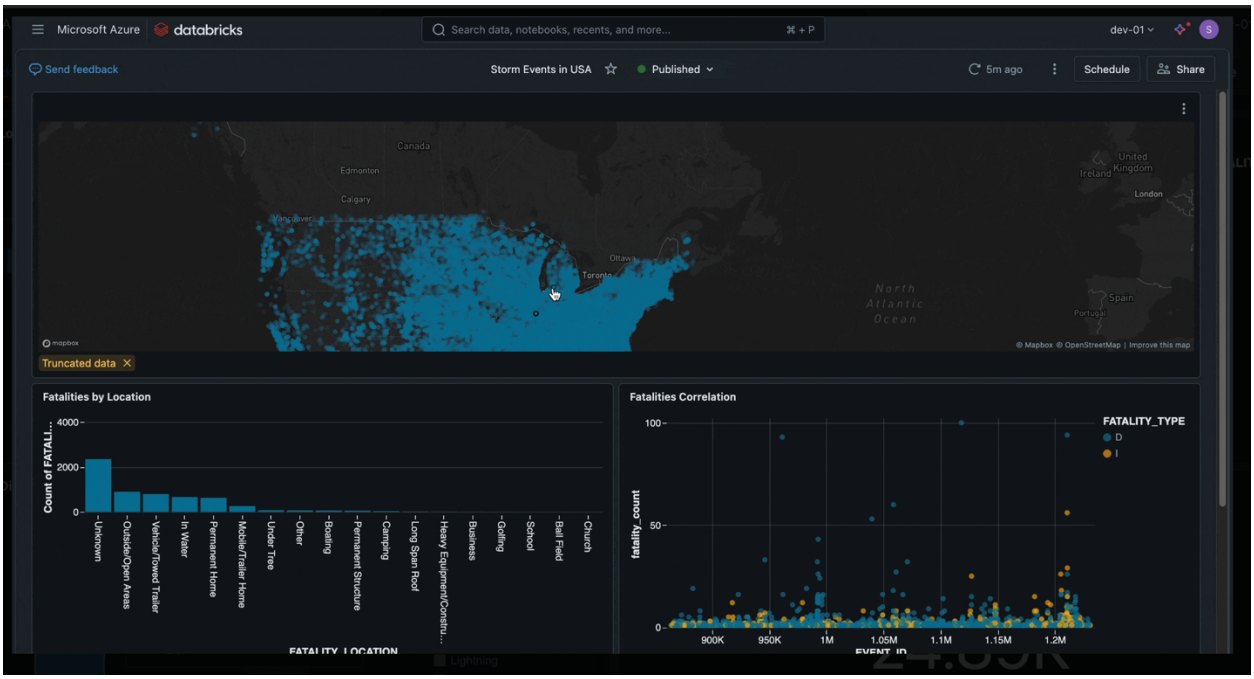
Building Data Pipelines on Azure

by Sunjana Ramana
Last edited 6 days ago

Azure Databricks Workflow Implementation Guide | Notion





What is Data Lake and Lakehouse

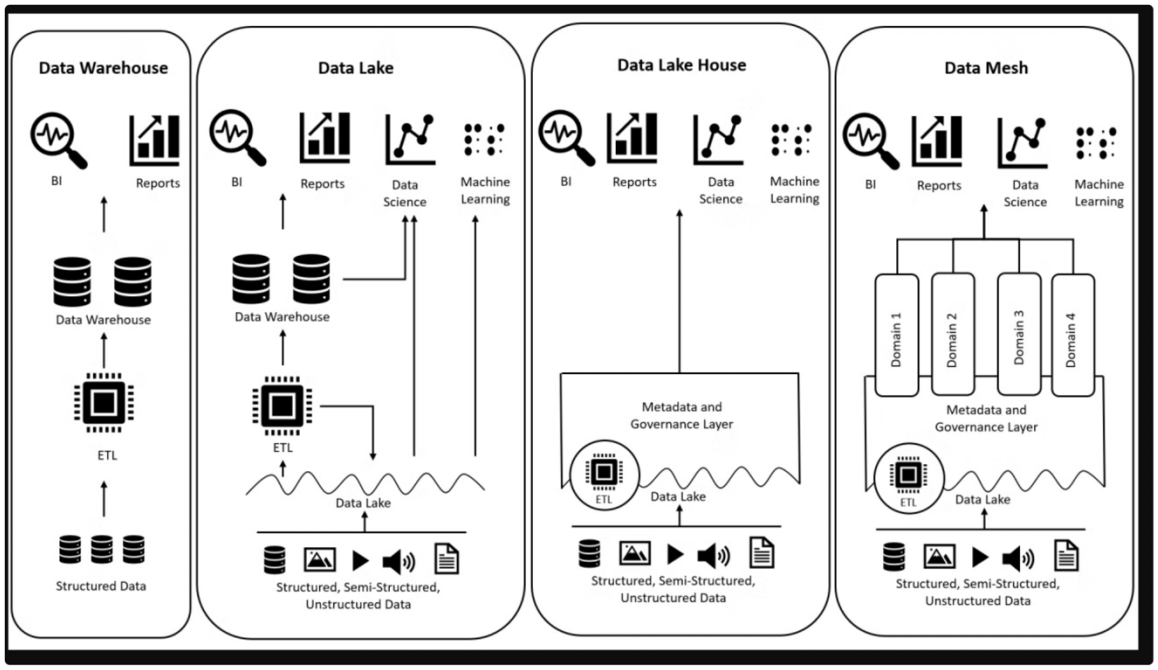
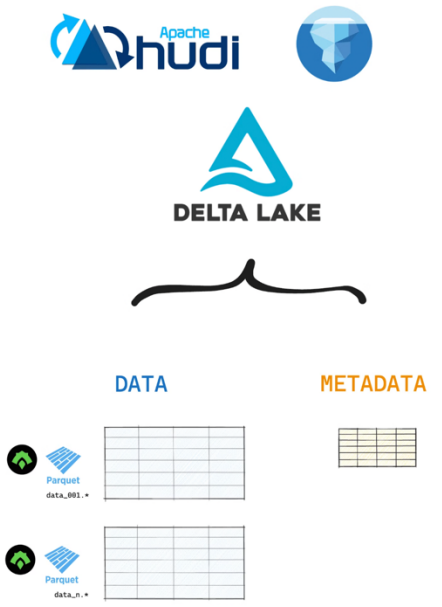


Image Credit - Web

1. **Data Warehouse**: Stores organized data for easy analysis.
2. **Data Lake**: Stores all types of raw data without organization.
3. **Data Lakehouse**: Combines the best of data lakes and warehouses for better analysis.
4. **Data Mesh**: Distributes data management across teams for better control and scalability.

Open Table Formats



Open table formats are open-source technologies designed to manage tabular data stored in data lakes.

They provide advanced features like **ACID transactions**, **schema evolution**, **time travel**, and **metadata management**, enabling database-like functionality on top of file formats like Parquet or ORC.

These formats optimize data storage and querying while maintaining compatibility with various analytics tools.

Popular Open Table Formats:

1. **Apache Iceberg**:
   - Optimized for large-scale analytics with hidden partitioning and snapshot isolation.
   - Supports schema evolution and time travel.
2. **Delta Lake**:
   - Offers ACID compliance, schema enforcement, and unified batch/streaming processing.
   - Ideal for Spark-based workloads.
3. **Apache Hudi**:
   - Specializes in incremental processing with record-level updates/deletes.
   - Best suited for streaming and write-heavy use cases.

Open table formats bridge the gap between raw data storage in lakes and advanced analytics, making them essential for modern data architectures like lakehouses.

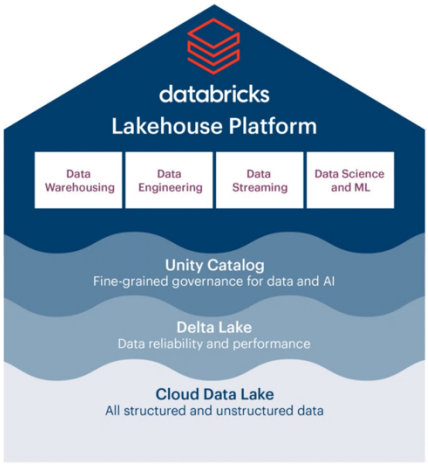
What is Azure Databricks

[Azure Databricks documentation]

- Lakehouse Management Solution
- Uses Delta Lake as the default open source management layer

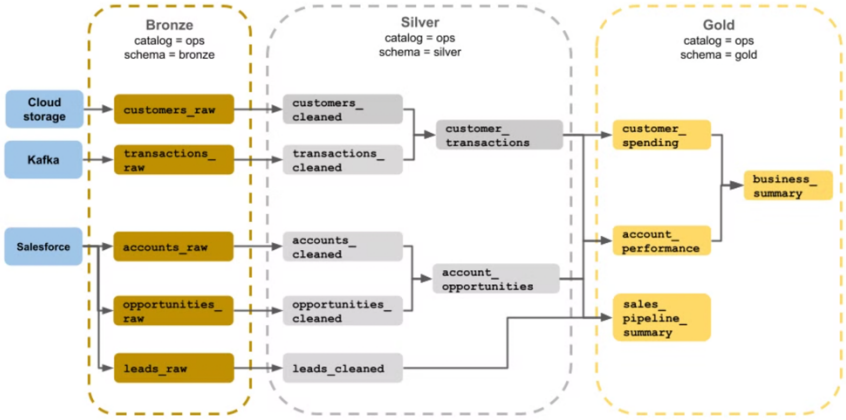
[Delta Lake] is an open-source storage layer that brings reliability to data lakes by adding a transactional storage layer on top of data stored in cloud storage (on AWS S3, Azure Storage, and GCS). It allows for ACID transactions, data versioning, and rollback capabilities. It allows you to handle both batch and streaming data in a unified way.

Delta tables are built on top of this storage layer and provide a table abstraction, making it easy to work with large-scale structured data using SQL and the DataFrame API.



Medallion Architecture

[What is the medallion lakehouse architecture? - Azure Databricks]



Bronze - Preserves the source format

Silver - Filtering, cleanup, transformation, aggregation

Gold - Extracted for business process