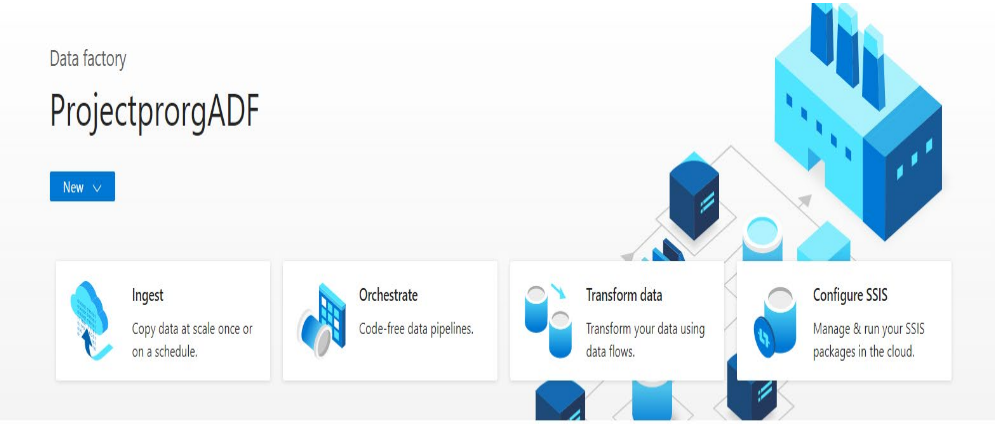
ADF-Azure Data Factory(Pipelines, Dataflow, Triggers, Email triggers for Pipeline resiliency), Delta Lake, Medallion Architecture



Data factory

ProjectprorgADF

New ∨

Ingest

Copy data at scale once or on a schedule.

Orchestrate

Code-free data pipelines.

Transform data

Transform your data using data flows.

Configure SSIS

Manage & run your SSIS packages in the cloud.

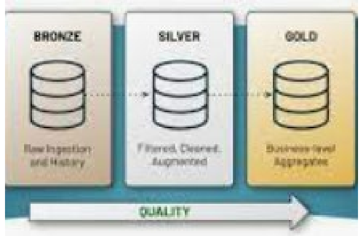
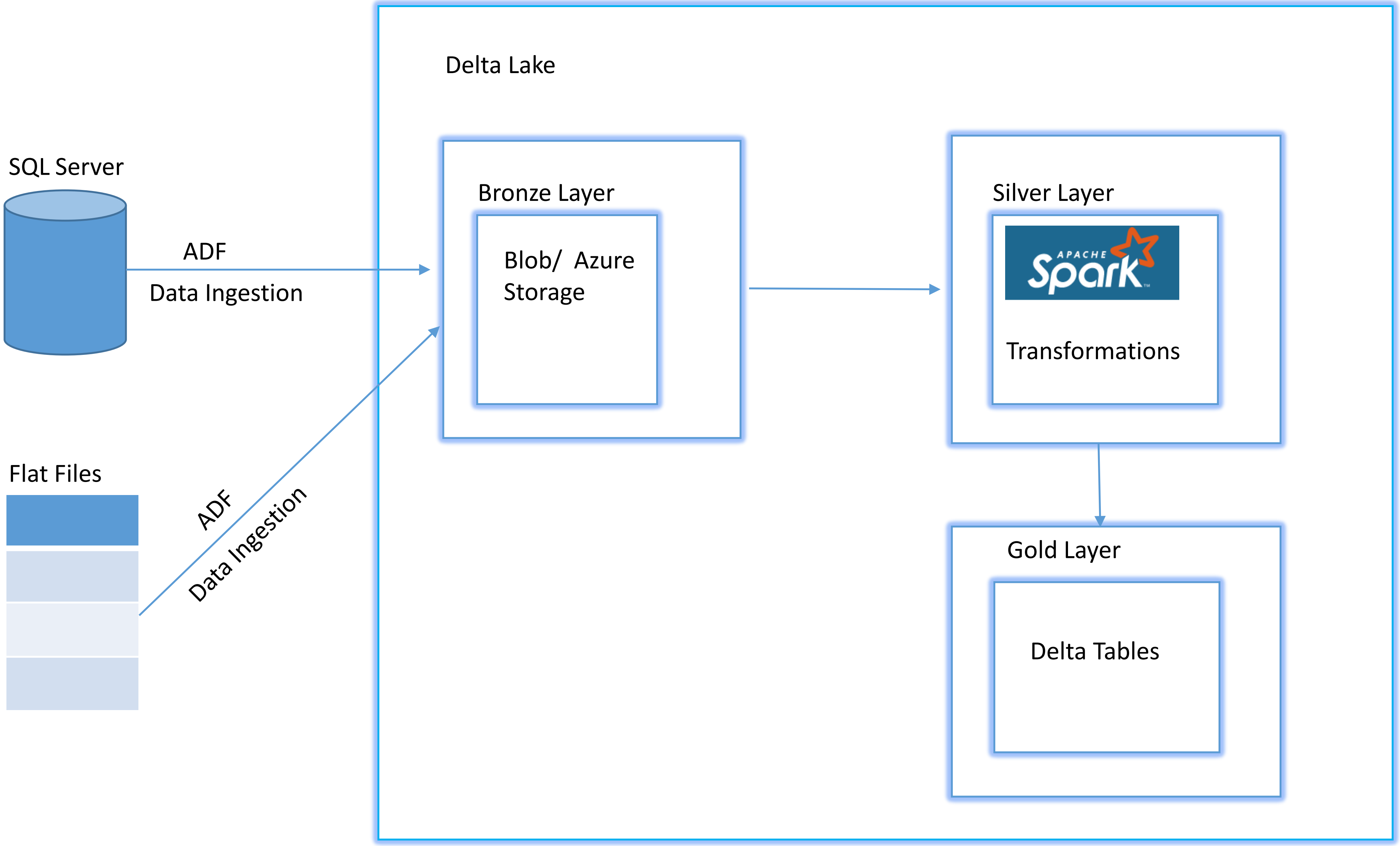

BRONZE
SILVER
GOLD
Raw Ingestion and History
Filtered, Cleaned, Augmented
Business-level Aggregates
QUALITY

Table of Contents

1. Arch Diagram
2. Features of Delta Lake
3. Evolution of Delta Lake from Data Lake
4. Medallion Architecture
5. Introduction to ADF(Azure Data Factory)
6. Dataflow, Pipelines in ADF – Creating a Delta bronze zone
7. Databricks – Py spark notebook for transformations
8. Scheduling the Pipelines in ADF, Triggers
9. Logic Apps , Email trigger for pipeline resiliency
10. Monitoring sessions



SQL Server

ADF

Data Ingestion

Flat Files

ADF

Data Ingestion

Delta Lake

Bronze Layer

Blob/ Azure Storage

Silver Layer

APACHE Spark

Transformations

Gold Layer

Delta Tables

Delta Lake

Key features of Delta Lake:

1. ACID transactions on Spark
2. Unified Batch and Stream Processing
3. Schema Enforcement
4. Time Travel

ACID transactions on Spark:

Delta Lake stores a transaction log to keep track of all the commits made to the table directory to provide ACID transactions. It provides Serializable isolation levels to ensure the data consistent across multiple users.

Unified Batch and Stream Processing :

In Data lake, data coming in as Stream (maybe from Kafka) or any historical data you have (say HDFS) is the same table. It gives an unified view of both these 2 different paradigms. Streaming data ingest, batch historic backfill, and interactive queries work just out of the box without much of the extra effort.

Time Travel :

Data in the data lake will be versioned and snapshots are provided so that you can query them as if that snapshot was the current state of the system. This helps us to revert to older versions of our data lake for Audits, rollbacks and stuff like that.

Evolution of Delta Lake from Data Lake

Data Lake
Structured, Semi-structured and Unstructured Data

- Effective Transactional control
- ACID compliance
- Schema evolution
- Delta format files vs Parquet

Medallion Architecture

The medallion architecture describes a series of data layers that denote the quality of data stored in the lake house. Databricks recommends taking a multi-layered approach to building a single source of truth for enterprise data products. This architecture guarantees atomicity, consistency, isolation, and durability as data passes through multiple layers of validations and transformations before being stored in a layout optimized for efficient analytics. The terms bronze(raw), silver (validated), and gold (enriched) describe the quality of the data in each of these layers.

- **Bronze:**

The bronze layer contains un-validated data. Data ingested in the bronze layer typically contains :Maintains the raw state of the data source, Is appended incrementally and grows over time & Can be any combination of streaming and batch transactions.

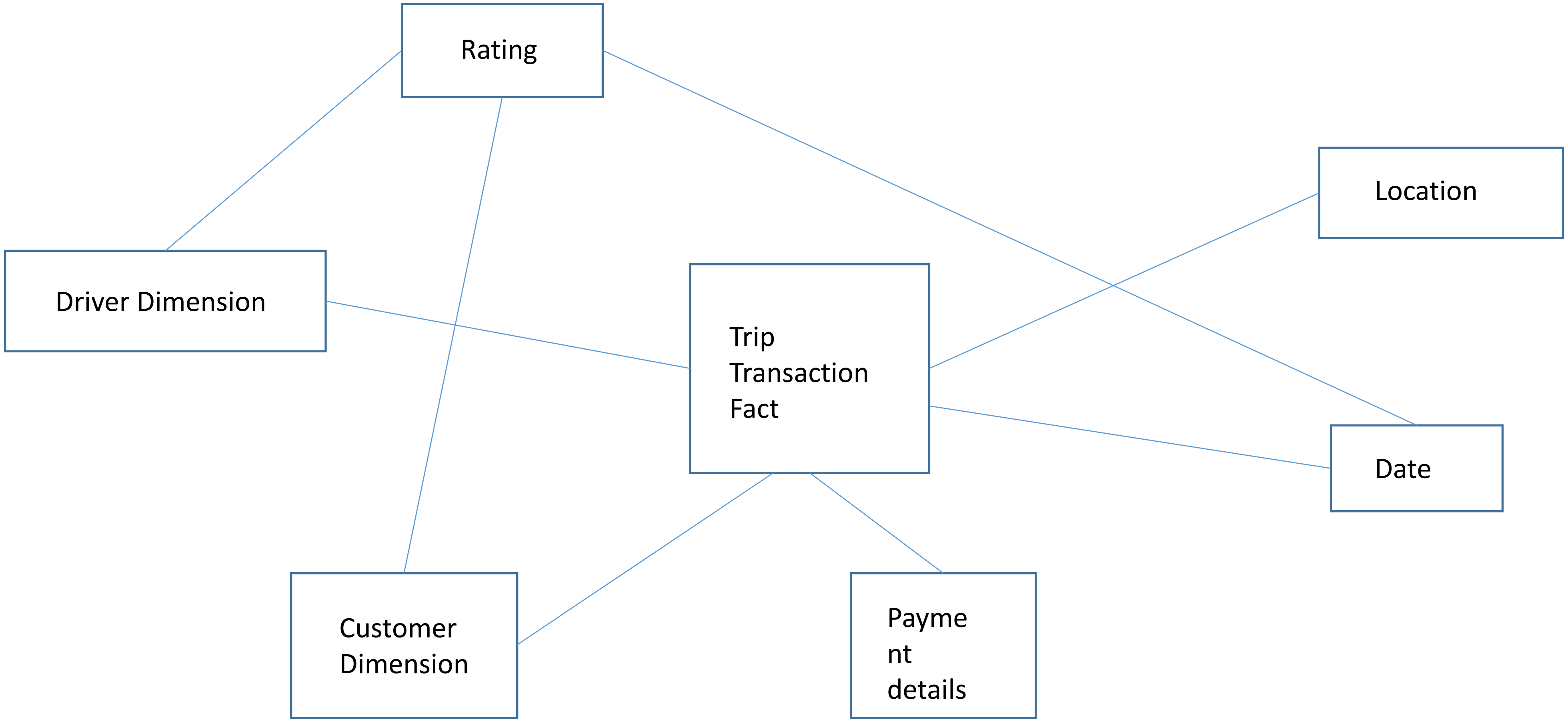
- **Silver:**

Recall that while the bronze layer contains the entire data history in a nearly raw state, the silver layer represents a validated, enriched version of our data that can be trusted for downstream analytics.

- **Gold:**

This gold data is often highly refined and aggregated, containing data that powers analytics, machine learning, and production applications. While all tables in the lake house should serve an important purpose, gold tables represent data that has been transformed into knowledge, rather than just information.

Fact- Dimension Data Modelling


Rating
Location
Driver Dimension
Trip Transaction Fact
Date
Customer Dimension
Payme nt details