HLD

High-Level Design (HLD)

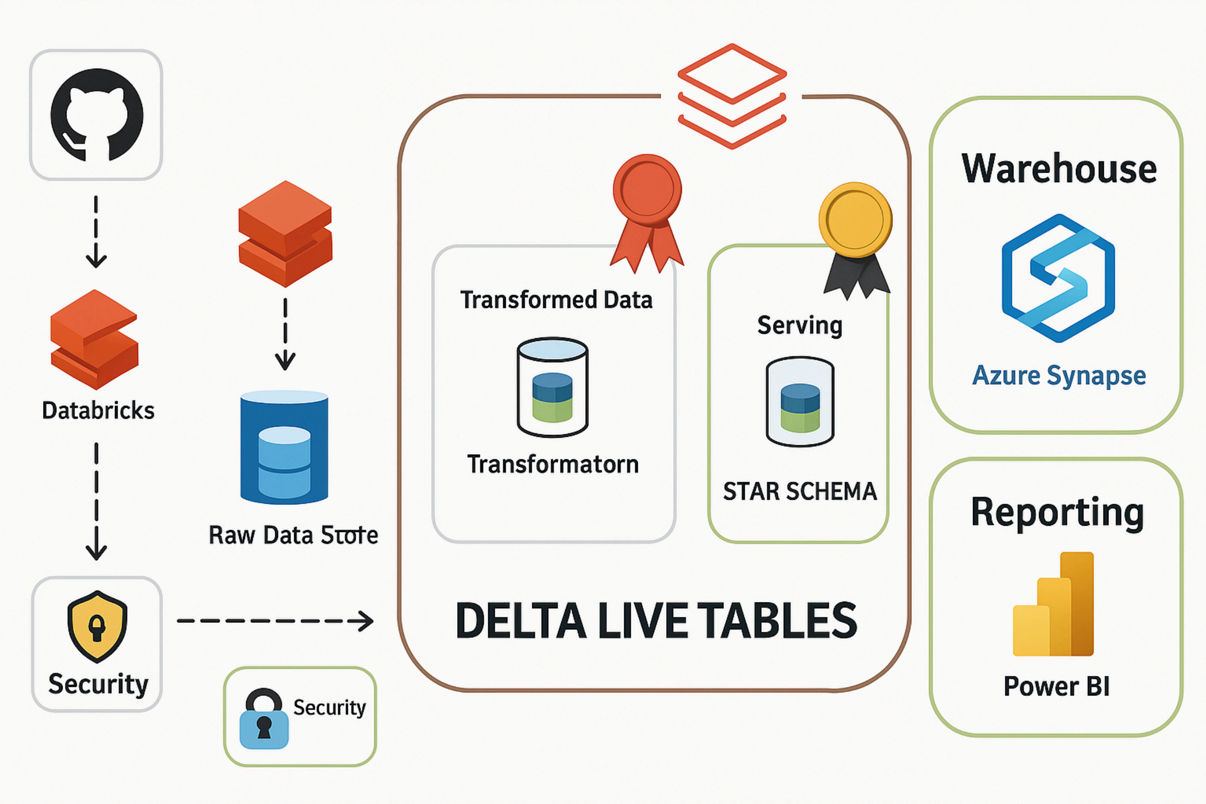
Overview

The system is an automated, scalable data pipeline that processes Netflix data from GitHub, transforming it through multiple layers and making it available for analytics. It leverages Azure cloud services and Databricks for real-time and batch processing, ensuring data quality and extensibility.

Objectives

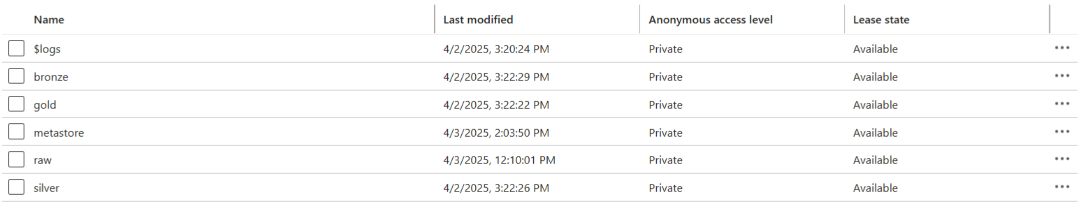
- Ingest Netflix dataset (CSVs) from GitHub into Azure Data Lake.
- Process and transform data incrementally across bronze, silver, and gold layers.
- Enforce data quality and deliver analytics-ready datasets.
- Enable integration with BI tools for reporting.

Architecture Diagram



Components

1. **Data Source**: Netflix dataset (CSVs) from GitHub (e.g., titles, cast, directors, countries, categories).
2. **Ingestion Layer**:
   - **Azure Data Factory (ADF)**: Orchestrates data ingestion from GitHub to Data Lake.
   - **Dynamic Pipeline**: Parameterized to handle multiple files with validation.
3. **Storage Layer**:
   - **Azure Data Lake**: Stores raw (bronze), processed (silver), and refined (gold) data in Delta format and metadata for Delta tables (metastore).



	Name	Last modified	Anonymous access level	Lease state	
□	$logs	4/2/2025, 3:20:24 PM	Private	Available	···
□	bronze	4/2/2025, 3:22:29 PM	Private	Available	···
□	gold	4/2/2025, 3:22:22 PM	Private	Available	···
□	metastore	4/3/2025, 2:03:50 PM	Private	Available	···
□	raw	4/3/2025, 12:10:01 PM	Private	Available	···
□	silver	4/2/2025, 3:22:26 PM	Private	Available	···

4. **Processing Layer**:
   - **Databricks**: Core compute with Unity Catalog for governance.
   - **Autoloader**: Streams bronze data incrementally.
   - **PySpark Notebooks**: Transform data in the silver layer with workflows.
   - **Delta Live Tables (DLT)**: Refines gold layer with quality checks.
5. **Orchestration Layer**:
   - **Databricks Workflows**: Manages task dependencies and conditional execution.
6. **Output Layer**:
   - **Power BI/Synapse**: Connects via Partner Connect for analytics.

Data Flow

1. **Ingestion**: ADF pulls CSVs from GitHub into the raw layer.
2. **Bronze Layer**: Autoloader streams data into Delta tables.

3. **Silver Layer**: PySpark notebooks process data with dynamic workflows and conditional logic (e.g., Sunday-only runs).
4. **Gold Layer**: DLT creates streaming tables/views with data quality enforcement.
5. **Output**: Gold data exposed for BI tools via Unity Catalog.

Key Features

- **Scalability**: Incremental loading and streaming for large datasets.
- **Flexibility**: Parameterized workflows for multiple files.
- **Data Quality**: DLT expectations ensure clean data.
- **Real-Time**: Supports conditional and streaming processing.