Azure Data Factory and Databricks End-to-End Project

Overview

Delta Lake is an open-source storage layer that sits on top of existing data lakes to provide ACID transactions, data versioning, and other advanced features. It was developed by Databricks, the company behind Apache Spark, and is designed to address some of the limitations and challenges of using traditional data lakes for big data processing and analytics. Delta Lake offers several key benefits for big data workloads, including:

- **ACID transactions**: Delta Lake provides atomicity, consistency, isolation, and durability (ACID) transactions for data lake tables, making it easier to write reliable data pipelines and ensure data consistency.
- **Data versioning**: Delta Lake allows for the versioning of data, which enables data engineers to keep track of changes made to the data over time and revert to previous versions if necessary.
- **Schema enforcement**: Delta Lake enforces schema on write, which helps ensure that data is clean and consistent.
- **Time travel**: Delta Lake enables time travel queries, which allow data engineers to query the data as it existed at a specific point in time.
- **Faster queries**: Delta Lake uses advanced indexing and caching techniques to optimize query performance and reduce query latency.

Delta Lake is compatible with Apache Spark and can be used with a variety of other big data tools and platforms. It is often used in conjunction with cloud storage services like Amazon S3 or Azure Data Lake Storage to provide a scalable and cost-effective solution for big data processing and analytics.

This project aims to implement Analytics/Insights on the trip transaction and ride-based source data from SQL Server, utilizing a combination of Azure Data Factory (ADF), Azure Blob Storage (ADLS Gen2), and Azure Databricks, in accordance with a Medallion Architecture approach. The project entails data ingestion into a Bronze Zone, data transformation through Azure Databricks, and data loading into Delta tables for the Gold Zone. Moreover, we plan to create a pipeline and schedule it using Azure Data Factory and leverage Logic Apps for pipeline resiliency in order to trigger emails.

Tech Stack

➔

Language: Python, SQL, Spark

➔

Package: PySpark

➔

Services: Azure Data Factory (ADF), Azure Blob Storage (ADLS Gen2), and Azure Databricks, Logic Apps, Azure SQL Database

Azure Data Factory (ADF):
Azure Data Factory is a cloud-based data integration service provided by Microsoft as part of its Azure cloud computing platform. It allows organizations to create, schedule, and orchestrate data workflows to ingest, prepare, transform, and move data from various sources to various destinations. Azure Data Factory supports a wide range of data sources and destinations, including relational databases, big data platforms, cloud storage, and other data services.

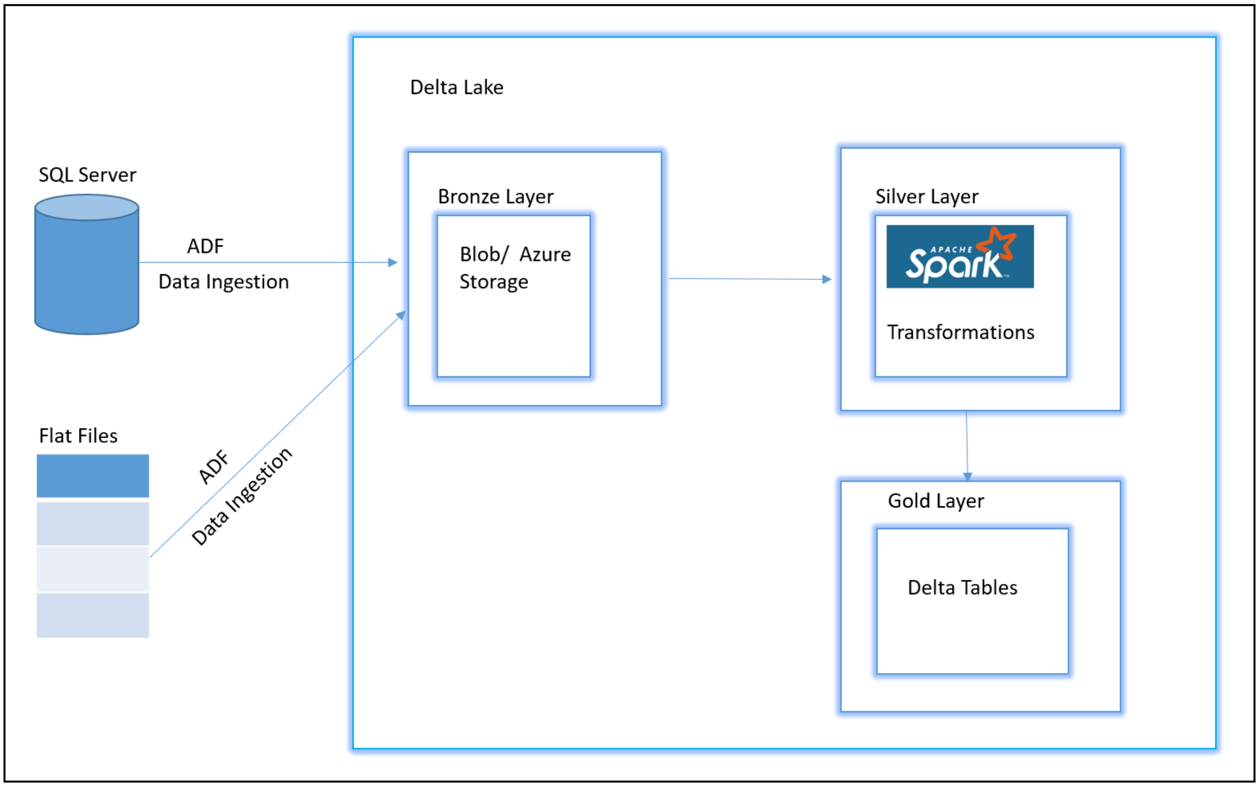
Azure Databricks:
Azure Databricks is a cloud-based big data and analytics service provided by Microsoft, which is powered by Apache Spark, an open-source big data processing framework. It provides a collaborative workspace for data scientists, data engineers, and analysts to process, analyze, and visualize large datasets using Spark's distributed computing capabilities.

Logic Apps:
Azure Logic Apps is a cloud-based service provided by Microsoft that allows users to create and run workflows to automate business processes and integrate with various systems and services. Logic Apps provides a visual designer and a wide range of connectors to enable organizations to create workflows without writing any code, making it accessible to both technical and non-technical users.

Key Takeaways

- Understanding the Trip transaction dataset
- Understanding the Features of Delta Lake
- Understanding the Evolution of Delta Lake from Data Lake
- Understanding the Medallion Architecture
- Overview of Azure Data Factory
- Creating Dataflow in Azure Data Factory
- Creating Pipelines in Azure Data Factory
- Creating Datasets in Azure Data Factory
- Transforming data using PySpark in Databricks notebooks
- Scheduling the Pipeline in Azure Data Factory
- Creating Logic Apps to trigger emails for pipeline resiliency
- Monitoring Sessions in Azure Data Factory

Architecture Diagram:


Delta Lake
SQL Server
ADF
Data Ingestion
Flat Files
ADF
Data Ingestion
Bronze Layer
Blob/ Azure Storage
Silver Layer
APACHE Spark
Transformations
Gold Layer
Delta Tables