# Azure Data Pipeline for Trip Analytics using Delta Lake

## Project Overview

This project presents a complete Azure-based data engineering solution designed to analyze trip transaction data using cloud-native services. By leveraging Azure Data Factory, Azure Data Lake Storage Gen2, and Azure Databricks, the solution enables automated data ingestion, transformation, and analytics with a strong focus on pipeline reliability and structured data processing.

The primary objective is to build a scalable and resilient data pipeline that transforms raw trip data stored in SQL Server into actionable insights that support informed business decisions.

## Business Objective

The goal of this end-to-end Azure implementation is to utilize cloud technologies to streamline data movement and transformation. Azure Data Factory is used to ingest and replicate source data into a **Bronze layer**, where raw data is stored.

Azure Databricks then processes the data in Delta format, transforming it into a refined **Silver layer**, followed by an analytics-ready **Gold layer** stored within Azure Blob Storage. This layered Medallion Architecture ensures organized data flow, improved quality control, and structured analytics.

Through this pipeline, organizations can:

- Analyze customer behavior trends
- Identify top-performing drivers
- Examine trip frequency across time periods
- Evaluate operational and financial patterns

Additionally, the project explores advanced Delta Lake capabilities such as schema evolution, cloning, and time travel, demonstrating modern data lake enhancements for enterprise-grade data management.

## Technology Stack

**Programming Languages:** Python, SQL, Spark
**Framework / Package:** PySpark
**Azure Services Used:**

- Azure Data Factory (ADF)
- Azure Data Lake Storage Gen2 (ADLS Gen2)
- Azure Databricks
- Azure SQL Database
- Azure Logic Apps

## Core Components of the Solution

This project integrates multiple Azure services into a cohesive and automated data pipeline:

- **Azure SQL Database** – Serves as the primary source of trip transaction data
- **Azure Data Factory** – Handles orchestration, ingestion, and data movement
- **Azure Databricks** – Performs distributed data transformation using Spark
- **Delta Lake** – Ensures reliable storage with ACID compliance
- **Azure Logic Apps** – Automates notifications for pipeline execution status

Each component plays a vital role in delivering a streamlined and resilient analytics workflow.

**1. Azure Storage Implementation**

The project includes detailed configuration of Azure Storage services to support structured data layering.

Key tasks performed include:

- Creating and managing storage accounts
- Configuring access tiers and security keys
- Creating Blob containers for Bronze and Silver zones
- Organizing data storage to align with Medallion Architecture

Additionally, secure access was configured using Shared Access Signatures (SAS), enabling controlled external access and practical experience in secure cloud storage management.

**2. Azure Databricks Implementation**

Azure Databricks is utilized as the core transformation engine powered by Apache Spark.

Within this project:

- A Databricks workspace was configured and deployed
- Clusters were created and optimized for distributed processing
- Storage accounts were mounted to access ADLS Gen2
- PySpark notebooks were developed for data cleaning and transformation
- Jobs were executed and monitored for performance tracking

The solution also includes:

- Version control via Git integration
- Notebook export/import for collaboration
- CI/CD concepts for deployment readiness

### 3. Azure Data Factory (ADF) Orchestration

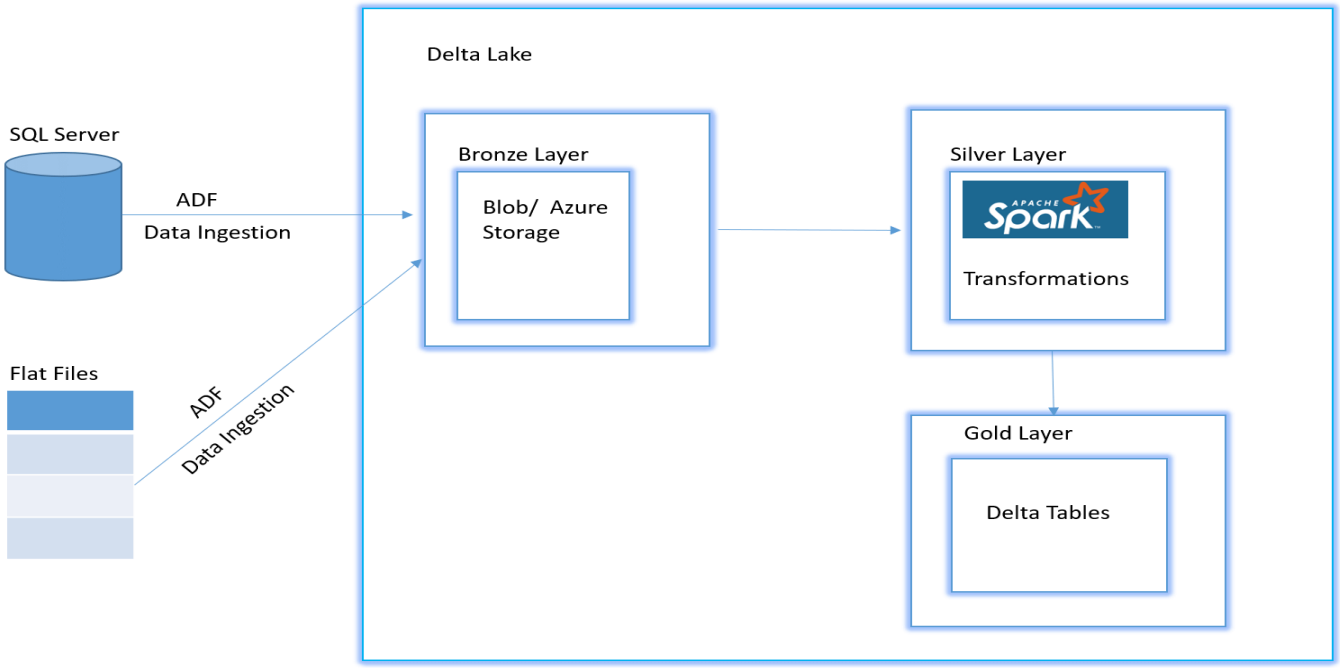
Azure Data Factory acts as the orchestration layer of the pipeline.

Key implementations include:

- Creating pipelines for ingesting data from Azure SQL Database
- Configuring datasets and linked services
- Replicating source data into the Bronze layer
- Designing data flows using low-code/no-code capabilities
- Scheduling and triggering pipeline executions
- Monitoring pipeline runs for performance and reliability

Integration with GitHub was configured to support version control and automated deployment workflows.

ADF ensures structured orchestration of data ingestion, transformation sequencing, and workflow dependency management.



## Pipeline Architecture Summary

The solution follows a **Medallion Architecture**:

**Bronze Layer**

Stores raw, ingested data directly from the source system.

**Silver Layer**

Contains cleaned, validated, and transformed datasets.

**Gold Layer**

Provides aggregated and analytics-ready datasets for reporting and business insights.

**Business Value Delivered**

This solution enables:

- Automated and structured data movement from SQL to cloud storage
- Reliable data processing using Delta Lake's ACID transactions
- Improved visibility into customer and driver analytics
- Reduced manual intervention through scheduled orchestration
- Resilient monitoring with automated email notifications via Logic Apps

The architecture supports scalable, maintainable, and analytics-ready data processing.

## Conclusion

This end-to-end Azure data engineering project demonstrates the implementation of a scalable cloud-based analytics pipeline. By integrating Azure Data Factory, Databricks, Delta Lake, and Logic Apps, the solution ensures reliable ingestion, transformation, storage, and monitoring of trip transaction data.

The project highlights modern data engineering best practices, including Medallion Architecture, distributed processing, orchestration automation, and pipeline resiliency—providing a robust foundation for enterprise analytics initiatives.