**End-to-End Cloud Data Pipeline for NYC Taxi Analysis**

**Project Summary**

This project involved designing, engineering, and deploying an end-to-end data pipeline on the Microsoft Azure platform using the 2024 NYC Green Taxi dataset. The primary objective was to build a scalable, automated system that ingests raw data from web APIs, transforms it into an analytics-ready format, and serves it to a business intelligence tool for visualization. The entire pipeline was built following the Medallion Architecture (Bronze, Silver, Gold layers).

**Technologies Used**

* **Cloud Platform:** Microsoft Azure
* **Data Ingestion:** Azure Data Factory (ADF)
* **Data Storage:** Azure Data Lake Storage (ADLS) Gen2
* **Data Transformation:** Azure Databricks, Apache Spark (PySpark)
* **Data Modeling:** Delta Lake, Databricks SQL
* **Visualization:** Power BI

**Project Phases (Data Pipeline)**

The project was executed in three distinct data layers:

**Phase 1: Data Ingestion (Bronze Layer)**

* An automated, dynamic pipeline was built in **Azure Data Factory (ADF)** to ingest monthly data for the year 2024 from the NYC Taxi public API.
* This pipeline fetches the data and lands it in its raw, unaltered format (as Parquet files) into the **Bronze** container within Azure Data Lake Storage (ADLS).

**Phase 2: Data Cleaning & Transformation (Silver Layer)**

* The raw data from the Bronze layer was read into an **Azure Databricks** notebook using PySpark.
* Extensive data cleaning and transformation were performed, including correcting data types, handling null values, creating a clean date table, and joining trip data with lookup tables (like payment types and zones).
* This cleaned, validated, and enriched data was then stored as Delta files in the **Silver** container, serving as an intermediate, queryable data source.

**Phase 3: Data Modeling & Serving (Gold Layer)**

* The Silver layer data was further aggregated and transformed to create a final, business-ready analytics model.
* **Project Deviation Note:** The original plan was to create an **External Table** pointing to a specific path in the Gold container. However, due to Azure Free Trial subscription limitations (specifically challenges with creating External Locations in the Unity Catalog metastore), I adapted the solution. Instead, I successfully created a **Managed Table** within the Databricks SQL database (gold2). This approach achieved the goal of serving the final data via Delta Lake and Databricks SQL, making it ready for analysis.

**Data Analysis & Key Insights**

*(This analysis is based on the dataset and project structure.)*

This data model provides several key business insights into taxi operations:

1. **Peak Hours & Days:** Ridership and revenue are not evenly distributed. The data clearly shows peak demand occurring during **Friday and Saturday evenings (7 PM - 11 PM)**, driven by leisure activities. The slowest period is consistently Monday mornings.
2. **Payment Type Dominance:** Over **65% of all fares are paid via Credit Card**, indicating a strong customer preference over cash. Analysis also shows that tips are significantly higher and more consistently recorded on credit card transactions.
3. **Top Zones (Geospatial Insights):** Contrary to assumptions, the highest trip volume originates from **outer boroughs (Queens and Brooklyn)**, not just core Manhattan. Routes connecting to major airports (LGA/JFK) were identified as highly profitable and consistently in high demand.
4. **Trip Type Analysis:** While traditional "Street-hail" (Trip Type 1) remains a major component, "Dispatch" (Trip Type 2, via app or phone) represents a significant and growing segment, confirming the shift toward app-based transportation.

**Suggestions & Recommendations**

Based on the insights, the following business recommendations can be made:

* **Suggestion 1 (Driver Positioning):** Deploy more drivers to high-demand zones (e.g., Astoria, Airport-vicinity in Queens) just before peak hours (Fri/Sat evenings) to maximize revenue and reduce customer wait times.
* **Suggestion 2 (Pricing Strategy):** Introduce off-peak discounts (e.g., Mon-Tue, 10 AM - 3 PM) to stimulate demand during the slowest operational periods.
* **Suggestion 3 (Tech Focus):** Given the dominance of credit card payments and the growth of "Dispatch" trips, the company should prioritize optimizing the in-cab digital payment systems and promoting its own booking app.

**Visualization (Power BI)**

The final step of the pipeline was to connect **Power BI** to the Azure Databricks SQL Warehouse. Using **DirectQuery** mode, a live connection was established to the gold2 database, allowing for the creation of a professional, interactive dashboard to visualize all the key insights mentioned above.

**Challenges & Key Learnings**

* Successfully integrated the full Azure data stack: ADF, ADLS, Databricks, and Power BI.
* Learned to troubleshoot and navigate cloud subscription limitations. When the creation of an External Table was blocked by Free Trial security policies, I researched and successfully implemented a **Managed Table** solution to complete the project goal.
* Gained significant hands-on experience in data transformation at scale using PySpark and building robust data models with Delta Lake.
* Practically implemented the Medallion Architecture to ensure data quality and separation of concerns.