### **Project Report: End-to-End Cloud Data Pipeline & NYC Taxi Analysis**

#### **Project Summary**

This project demonstrates a complete, end-to-end data engineering pipeline built on the **Microsoft Azure** platform. Using the 2024 NYC Green Taxi dataset, this solution automates the process of ingesting raw data from web APIs, cleaning and transforming it at scale with Azure Databricks, modeling it in a Delta Lake, and finally, serving it to a live, interactive Power BI dashboard for analysis. The entire pipeline was built following the industry-standard **Medallion Architecture** (Bronze, Silver, Gold layers).

#### **Technologies Used**

Azure Data Factory | Azure Databricks | PySpark | Delta Lake | ADLS Gen2 | Power BI | DAX

### **Project Phases**

1. **Data Ingestion (Bronze Layer):** An automated **Azure Data Factory (ADF)** pipeline was built to dynamically ingest monthly 2024 NYC Taxi data from a public API, landing the raw files in Azure Data Lake Storage (ADLS).
2. **Data Transformation (Silver Layer):** **Azure Databricks** notebooks with **PySpark** were used to clean, validate, and enrich the raw data. This involved handling nulls, correcting data types, and joining trip data with lookup tables. The clean data was stored in the Silver layer as Delta files.
3. **Data Modeling (Gold Layer):** The final, business-ready analytics model was created in the **Gold** layer. Due to Azure Free Trial limitations, a **Managed Table** solution was successfully implemented in the Databricks SQL database, making the data available for querying.

### **DAX Measures Used in Power BI**

To build the final dashboard, the following key DAX measures were created:

Total Trips = COUNT('trip\_main\_table'[PULocationID])

Average Fare = AVERAGE('trip\_main\_table'[fare\_amount])

Average Trip Distance = AVERAGE('trip\_main\_table'[trip\_distance])

Total Revenue = SUM('trip\_main\_table'[total\_amount]

Total Dropoff Trips =

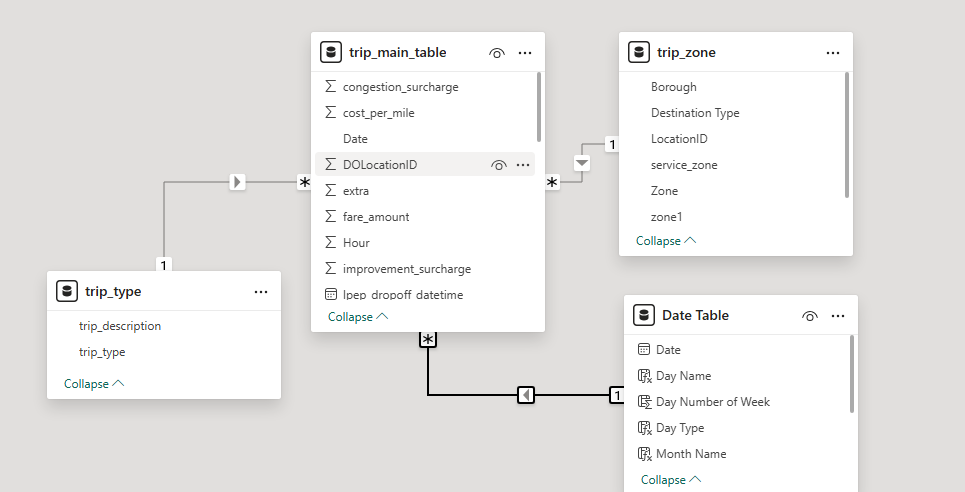
CALCULATE(

[Total Trips],

USERELATIONSHIP('trip\_main\_table'[DOLocationID], 'trip\_zone'[LocationID])

)

### **Model view**



**Data Model Overview:**

"The Power BI data model illustrates the established relationships between the fact table (trip\_main\_table) and the dimension tables (trip\_type, trip\_zone, and Date Table), ensuring proper data filtering and analytical capabilities."

### **Key Insights from the Dashboard**

Analysis of the **593K trips** in the final dashboard revealed several key operational patterns:

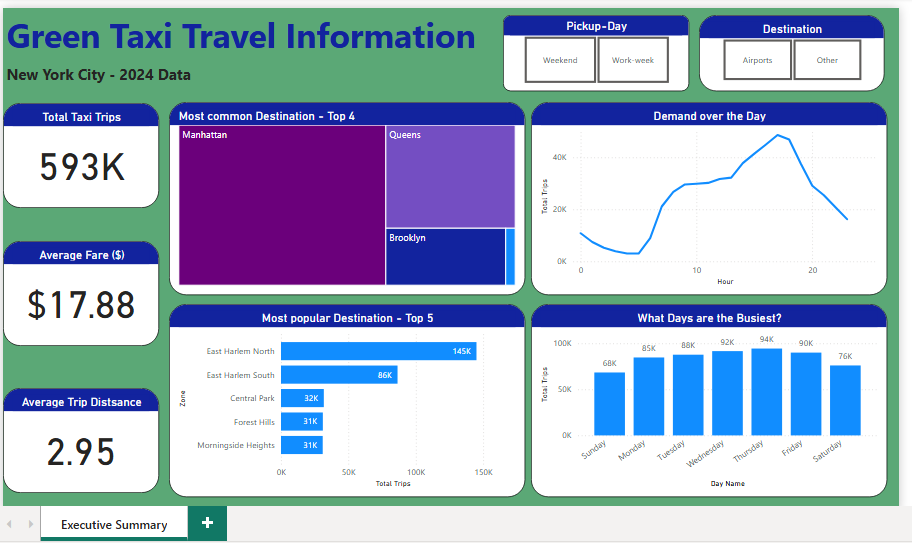
1. **Weekly Demand Peaks:** The busiest days are consistently **Friday** and **Saturday**, indicating strong demand driven by weekend leisure activities. **Monday** is the slowest day of the week.
2. **Daily Commuter Patterns:** A clear "double-peak" trend was observed throughout the day. The morning rush hour (around 8 AM) shows significant activity, but the **evening peak (around 5 PM / 17:00)** is substantially higher, representing the main commuter rush.
3. **Geospatial Hotspots:** **Manhattan** and **Queens** are the most dominant boroughs for Green Taxi trips. The bar chart reveals that **East Harlem North** is the single most popular pickup zone, making it a critical hub for operations.
4. **Core Metrics:** Across all trips, the **average fare is $17.88**, and the **average trip distance is 2.95 miles**, suggesting that short-to-medium distance trips are the most common.

### **Business Recommendations**

Based on these insights, the following data-driven recommendations can be made:

* **Optimize Driver Deployment:** Increase the number of available taxis in key zones like **East Harlem** and other parts of **Manhattan/Queens** during the weekday evening peak (4 PM - 7 PM) and on Friday/Saturday nights to maximize revenue and reduce customer wait times.
* **Targeted Marketing:** Introduce special offers or flat-rate fares during the slowest periods, such as **Monday mornings** and mid-day, to stimulate demand and improve vehicle utilization.
* **Route & Service Planning:** Analyze the most popular routes originating from the top zones (like East Harlem) to ensure optimal service and potentially create dedicated airport transfer packages.

### **Final Outcome (Visualization)**



### Visualization (Power BI)

The final Gold-layer database in Databricks was successfully connected to **Power BI** using **Import mode**. This approach allowed the full dataset to be loaded into Power BI's in-memory engine, ensuring maximum performance and responsiveness for the interactive dashboard.

### Challenges & Key Learnings

* Successfully integrated the full Azure stack: ADF, ADLS, Databricks, and Power BI.
* Learned to troubleshoot cloud subscription limitations and adapted the solution by implementing a Managed Table instead of an External Table (bypassing Unity Catalog limitations in the Free Trial).
* Gained significant hands-on experience in PySpark data transformation and the practical application of the Medallion Architecture.