NYC Taxi Data Analytics Using Microsoft Fabric

Problem Statement

New York City's Yellow Taxi service generates millions of ride records each month. However, this vast volume of raw data is scattered, unstructured, and updated monthly, making it difficult for stakeholders (city planners, transport analysts, vendors) to gain timely insights into ride patterns, revenue trends, and service performance.

The challenge was to build a scalable, automated analytics solution that could:
- Efficiently ingest and transform large volumes of ride data
- Update reports automatically as new monthly data arrives
- Provide actionable visual insights into trends such as revenue, passenger volume, vendor performance, and geographic demand

Tools & Technologies Used

- Microsoft Fabric (OneLake, Lakehouse, Pipelines, Dataflow Gen2, Warehouse, Semantic Models)
- Power BI
- SQL

Project Overview

This project leverages the end-to-end capabilities of Microsoft Fabric to analyze New York City Yellow Taxi trip data. The goal was to automate the ingestion, transformation, storage, and reporting process for monthly-updated public transportation data, and deliver interactive dashboards with key insights for stakeholders.

My Role

As a Data Analyst on this project, I was responsible for designing and building the entire analytics pipeline in Microsoft Fabric. My key contributions included:
- Setting up the Lakehouse and Warehouse environments
- Designing data pipelines for ingestion and transformation
- Creating Dataflow Gen2 transformations for data cleaning
- Building semantic models to support reporting
- Designing and publishing interactive Power BI dashboards

Project Workflow

1. Lakehouse Creation:
  - Set up a Fabric Lakehouse (`Project_Lakehouse`) to store raw files and datasets.

2. Warehouse Development:
  - Created a Fabric Warehouse (`Project_Warehouse`) for SQL-based querying and structured modeling.

3. Pipelines for Ingestion:
  - Built 4 key pipelines (orchestration, preprocessing, staging, and lookup).
  - Pipelines are scheduled monthly to handle new data.

4. Data Transformation:
  - Used Dataflow Gen2 (`df_pres_processing_nyctaxi`) to clean, format, and aggregate data.

5. Power BI Report:
  - Interactive report with KPIs:
    $159.53M Revenue
    5.97M Trips
    7.52M Passengers
  - Visuals include date filters, vendor/payment filters, borough heatmaps, and daily trends.
  - Dashboard auto-refreshes monthly based on pipeline schedule.

Business Insights Delivered

- Revenue distribution by day and payment method
- Vendor-wise trip performance
- Top pickup/dropoff routes
- Daily trip and revenue trends

Challenges & Solutions

- Monthly data updates required manual refresh Implemented automated pipelines with scheduled refresh
- Inconsistent formats in raw data Cleaned and standardized data using Dataflow Gen2
- Performance issues with large data sets in Power BI Used semantic models and optimized queries

Outcome / Impact

- Built a production-ready analytics solution using Microsoft Fabric.
- Automated refresh cycle reduces manual effort.
- Delivered strategic insights to support transportation planning.

Project Screenshots

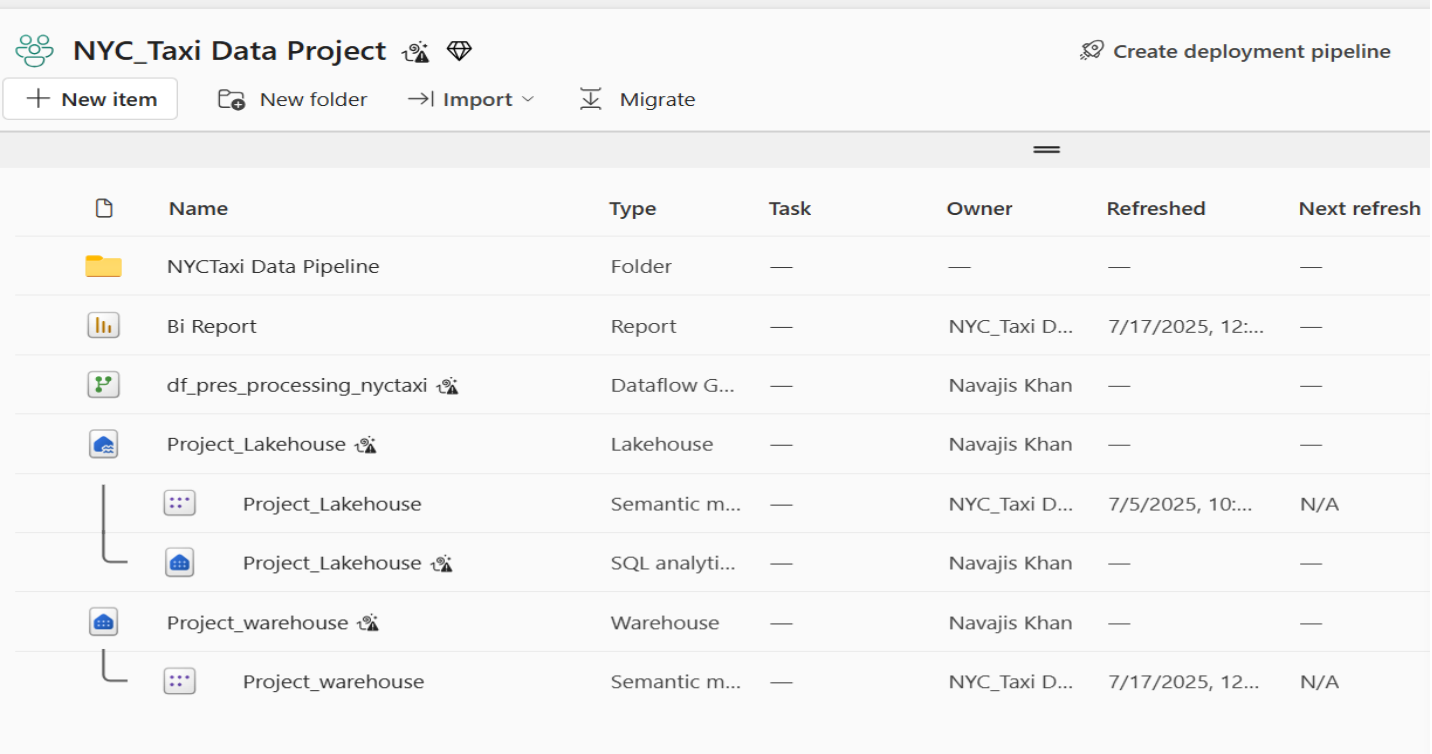

Image 1: Workspace overview and project structure.

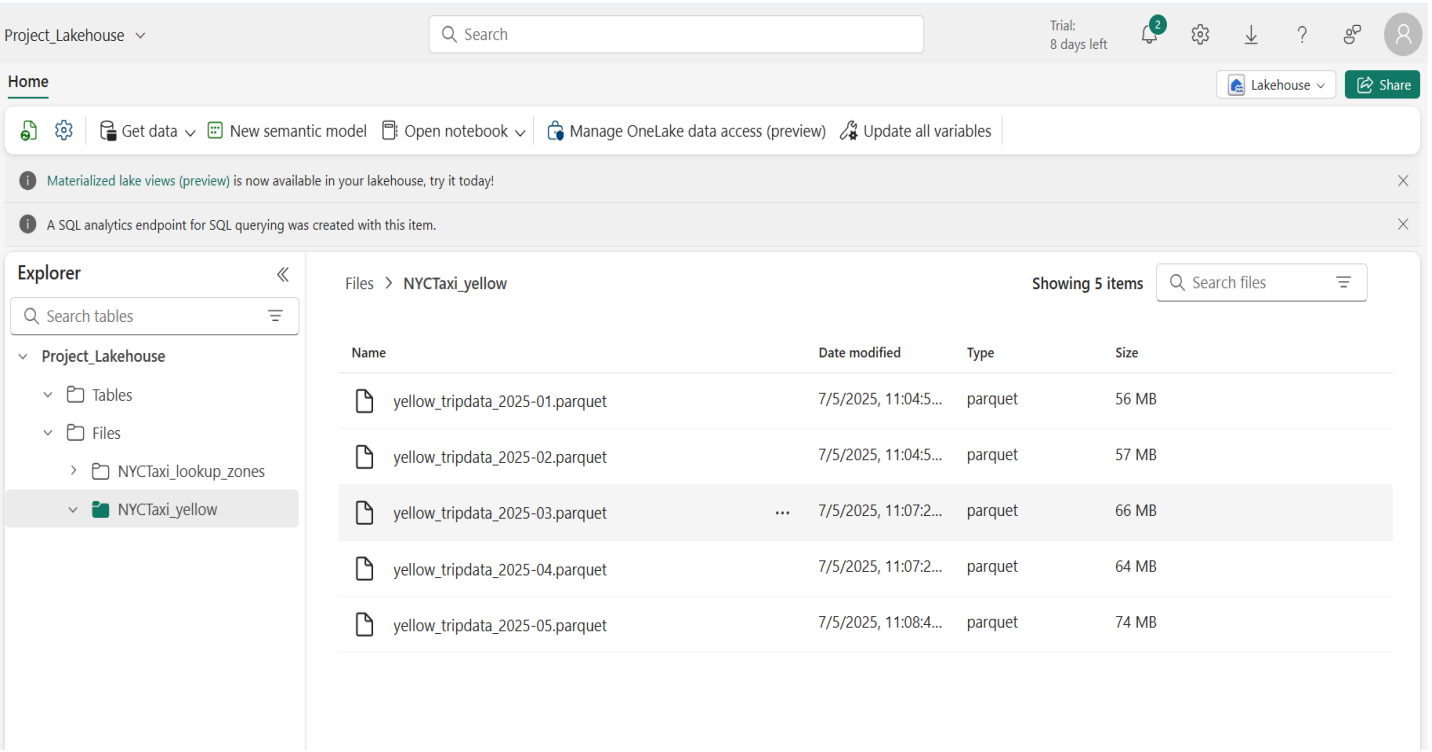

Image 2: Lakehouse environment showing ingested NYC taxi datasets organized in folders and tables.

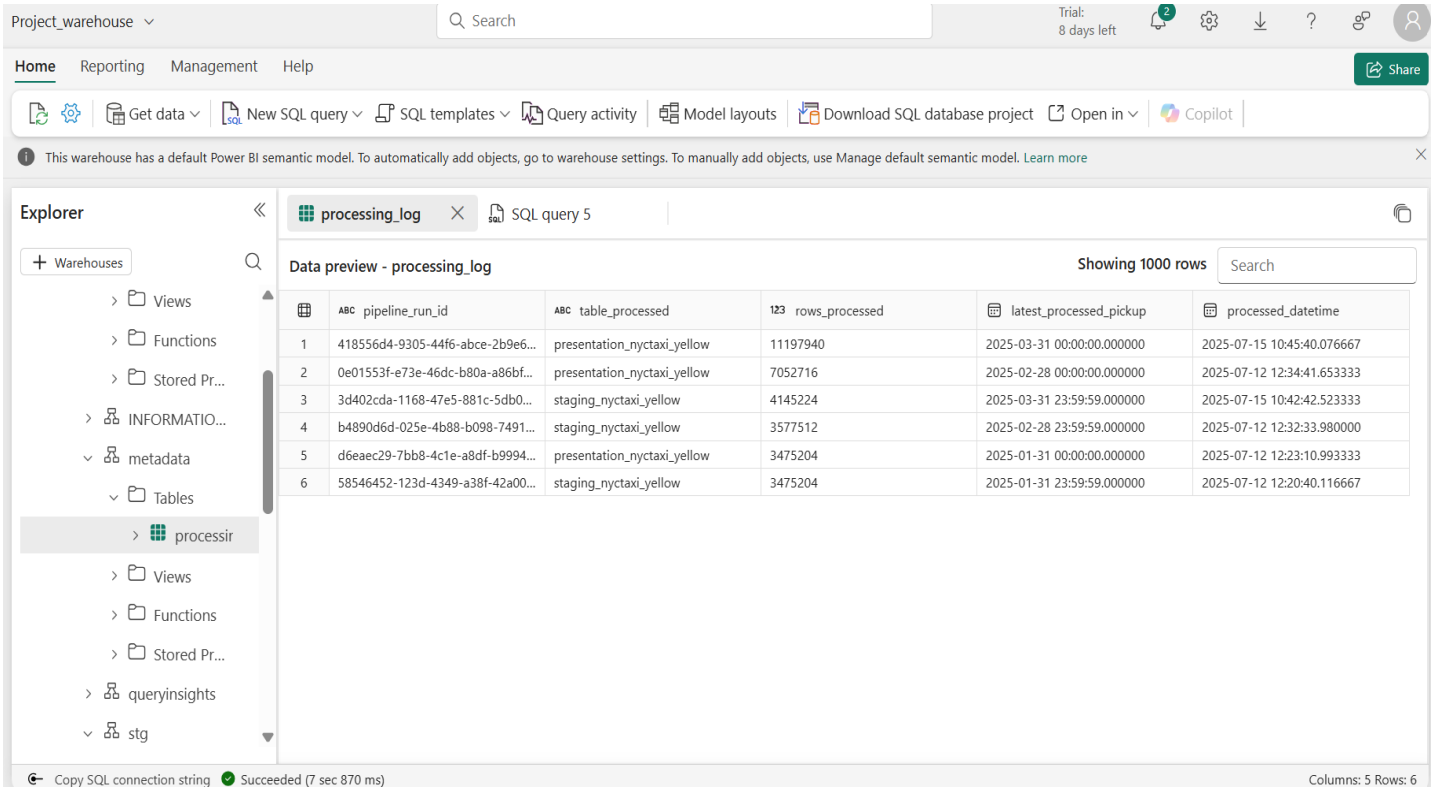

Image 3: Schema view of tables in the Microsoft Fabric Warehouse used for structured querying.

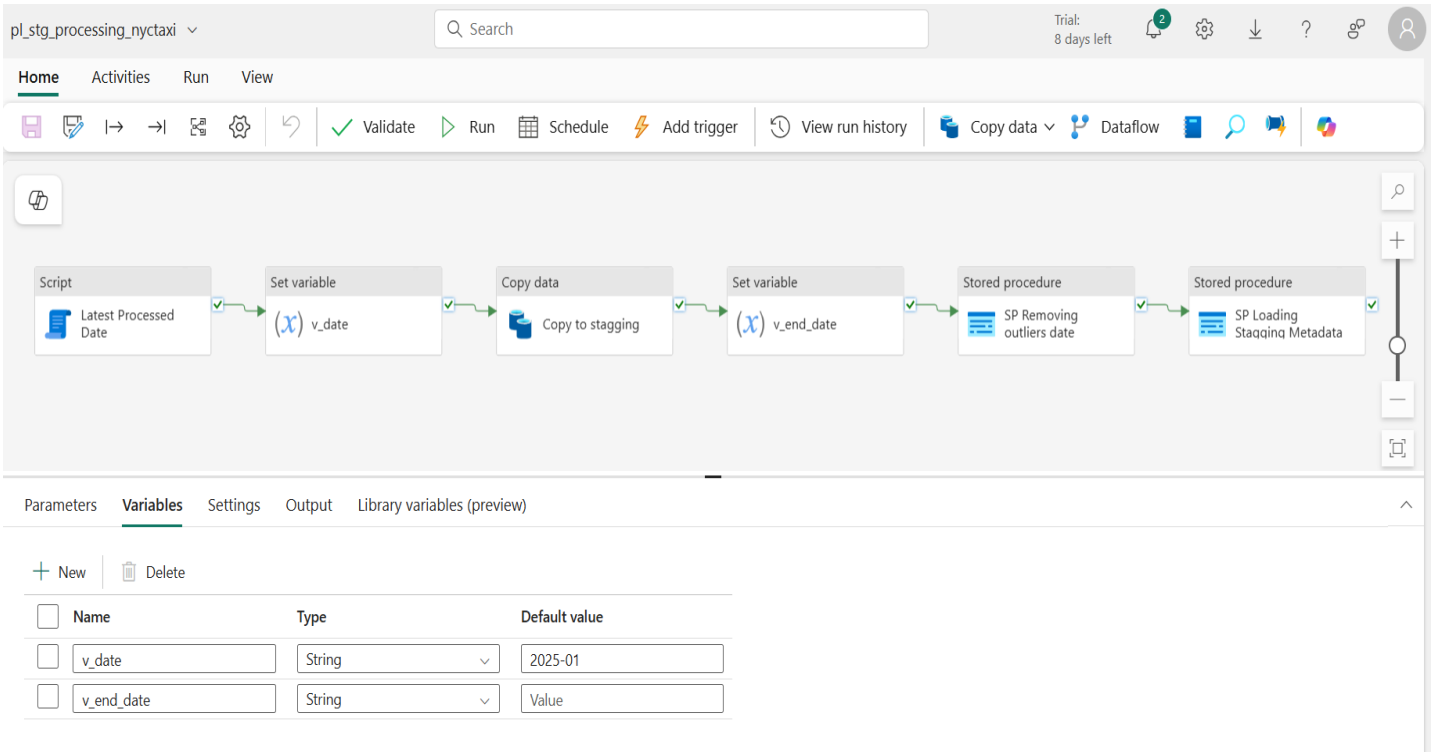

Image 4: Data pipeline flow designed to automate the ingestion and transformation of monthly ride data.

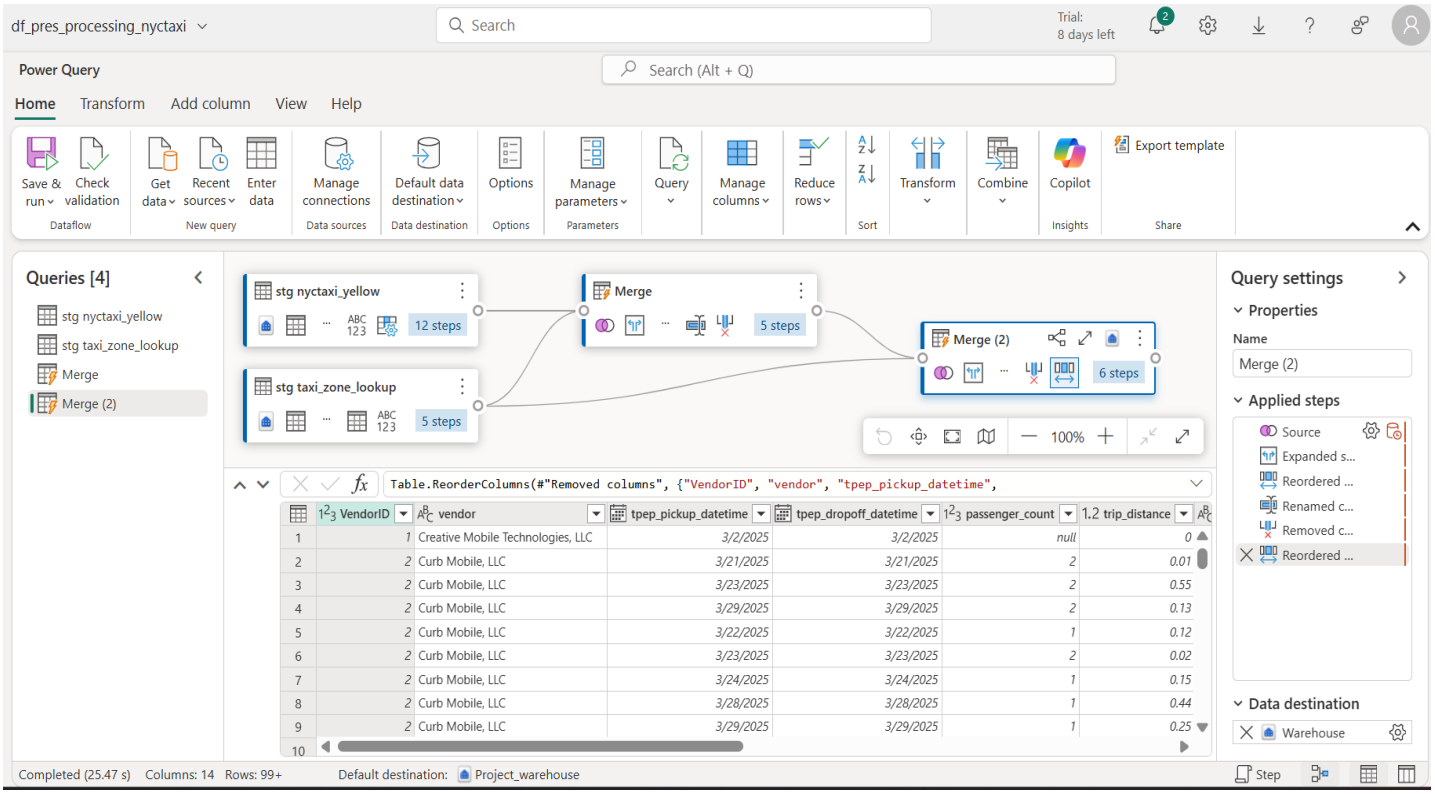

Image 5: Dataflow Gen2 displaying applied transformation steps for cleaning and enriching raw taxi data.

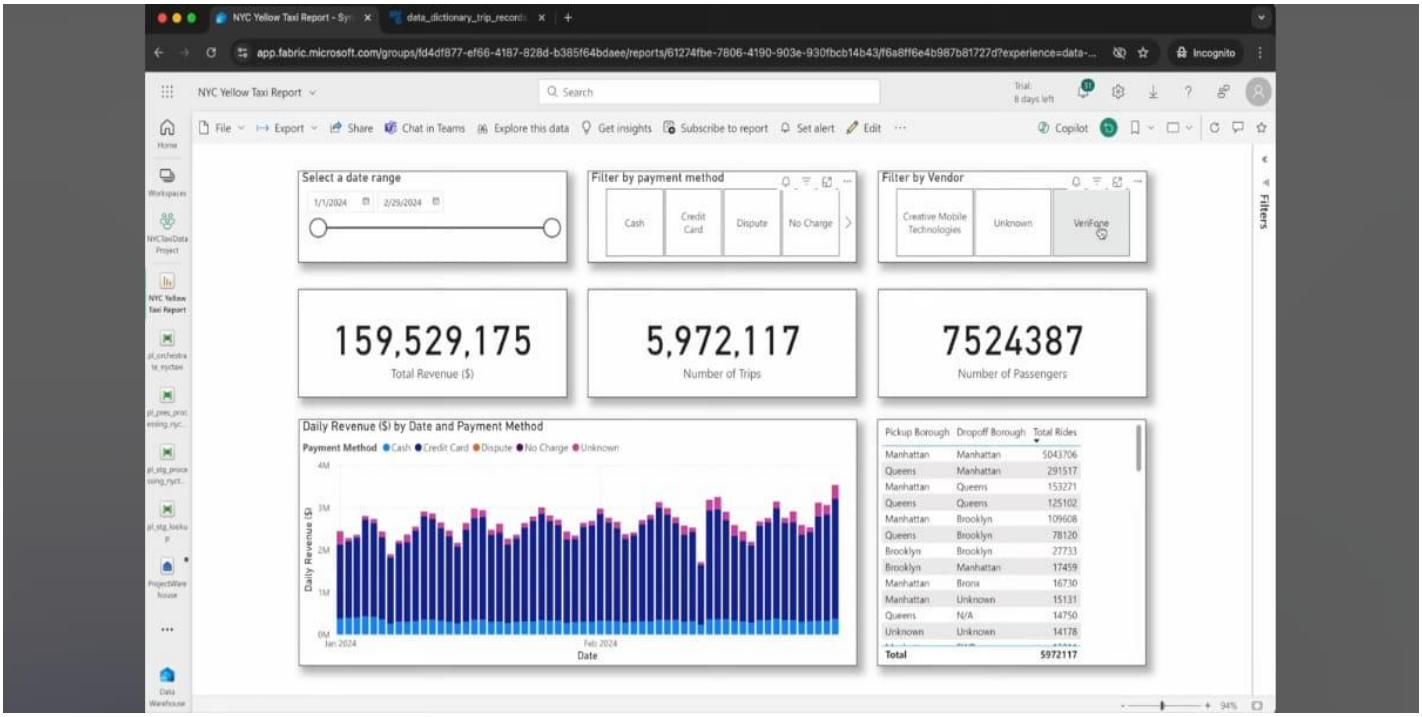

Image 6: Interactive Power BI dashboard visualizing revenue, trips, payment methods, and vendor activity.