

# 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

 **NYC\_Taxi Data Project**  

+ New item

 New folder

 Import

 Migrate

Create deployment pipeline













	Name	Type	Task	Owner	Refreshed	Next refresh
	NYCTaxi Data Pipeline	Folder	—	—	—	—
	Bi Report	Report	—	NYC_Taxi D...	7/17/2025, 12:...	—
	df_pres_processing_nyctaxi 	Dataflow G...	—	Navajis Khan	—	—
	Project_Lakehouse 	Lakehouse	—	Navajis Khan	—	—
	Project_Lakehouse	Semantic m...	—	NYC_Taxi D...	7/5/2025, 10:...	N/A
	Project_Lakehouse 	SQL analyti...	—	Navajis Khan	—	—
	Project_warehouse 	Warehouse	—	Navajis Khan	—	—
	Project_warehouse	Semantic m...	—	NYC_Taxi D...	7/17/2025, 12...	N/A


Image 1: Workspace overview and project structure.


Project\_Lakehouse


Search


Trial: 8 days left


 2














Home


Lakehouse


Share

 Get data

 New semantic model

 Open notebook

 Manage OneLake data access (preview)

 Update all variables

Materialized lake views (preview) is now available in your lakehouse, try it today!

A SQL analytics endpoint for SQL querying was created with this item.

Explorer

Search tables

Project\_Lakehouse

- Tables
- Files
  - NYCTaxi\_lookup\_zones
  - NYCTaxi\_yellow**

Files > NYCTaxi\_yellow

Showing 5 items

Search files

Name	Date modified	Type	Size
 yellow_tripdata_2025-01.parquet	7/5/2025, 11:04:5...	parquet	56 MB
 yellow_tripdata_2025-02.parquet	7/5/2025, 11:04:5...	parquet	57 MB
 yellow_tripdata_2025-03.parquet	... 7/5/2025, 11:07:2...	parquet	66 MB
 yellow_tripdata_2025-04.parquet	7/5/2025, 11:07:2...	parquet	64 MB
 yellow_tripdata_2025-05.parquet	7/5/2025, 11:08:4...	parquet	74 MB

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

Project\_warehouse

Search

Home Reporting Management Help

Get data New SQL query SQL templates Query activity Model layouts Download SQL database project Open in Copilot

This warehouse has a default Power BI semantic model. To automatically add objects, go to warehouse settings. To manually add objects, use Manage default semantic model. [Learn more](#)

Explorer

Warehouses

Views

Functions

Stored Pr...

INFORMATIO...

metadata

Tables

processir

Views

Functions

Stored Pr...

queryinsights

stg

processing\_log

SQL query 5

Data preview - processing\_log

Showing 1000 rows

Search

	ABC pipeline_run_id	ABC table_processed	123 rows_processed	latest_processed_pickup	processed_datetime
1	418556d4-9305-44f6-abce-2b9e6...	presentation_nyctaxi_yellow	11197940	2025-03-31 00:00:00.000000	2025-07-15 10:45:40.076667
2	0e01553f-e73e-46dc-b80a-a86bf...	presentation_nyctaxi_yellow	7052716	2025-02-28 00:00:00.000000	2025-07-12 12:34:41.653333
3	3d402cda-1168-47e5-881c-5db0...	staging_nyctaxi_yellow	4145224	2025-03-31 23:59:59.000000	2025-07-15 10:42:42.523333
4	b4890d6d-025e-4b88-b098-7491...	staging_nyctaxi_yellow	3577512	2025-02-28 23:59:59.000000	2025-07-12 12:32:33.980000
5	d6eac29-7bb8-4c1e-a8df-b9994...	presentation_nyctaxi_yellow	3475204	2025-01-31 00:00:00.000000	2025-07-12 12:23:10.993333
6	58546452-123d-4349-a38f-42a00...	staging_nyctaxi_yellow	3475204	2025-01-31 23:59:59.000000	2025-07-12 12:20:40.116667

Copy SQL connection string Succeeded (7 sec 870 ms)

Columns: 5 Rows: 6

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

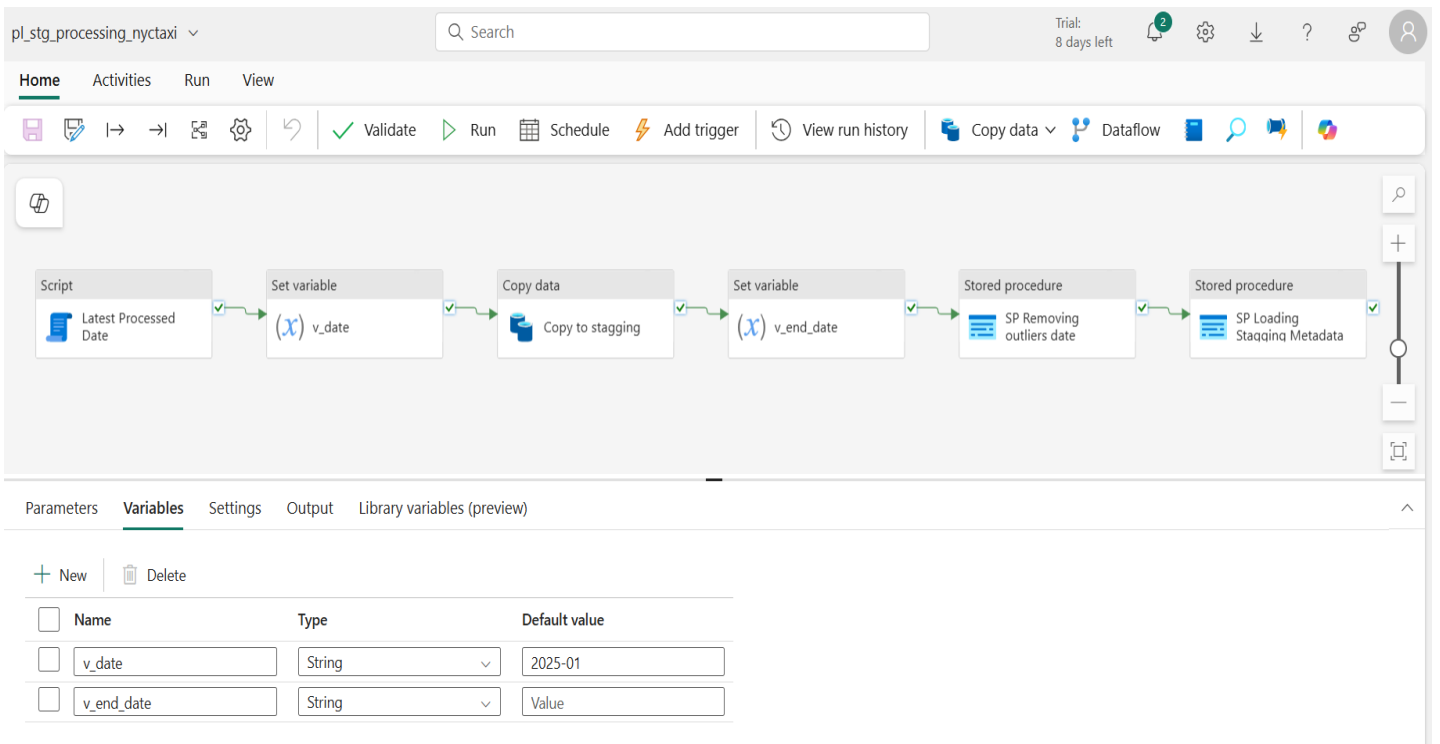


Image 4: Data pipeline flow designed to automate the ingestion and transformation of monthly ride data.

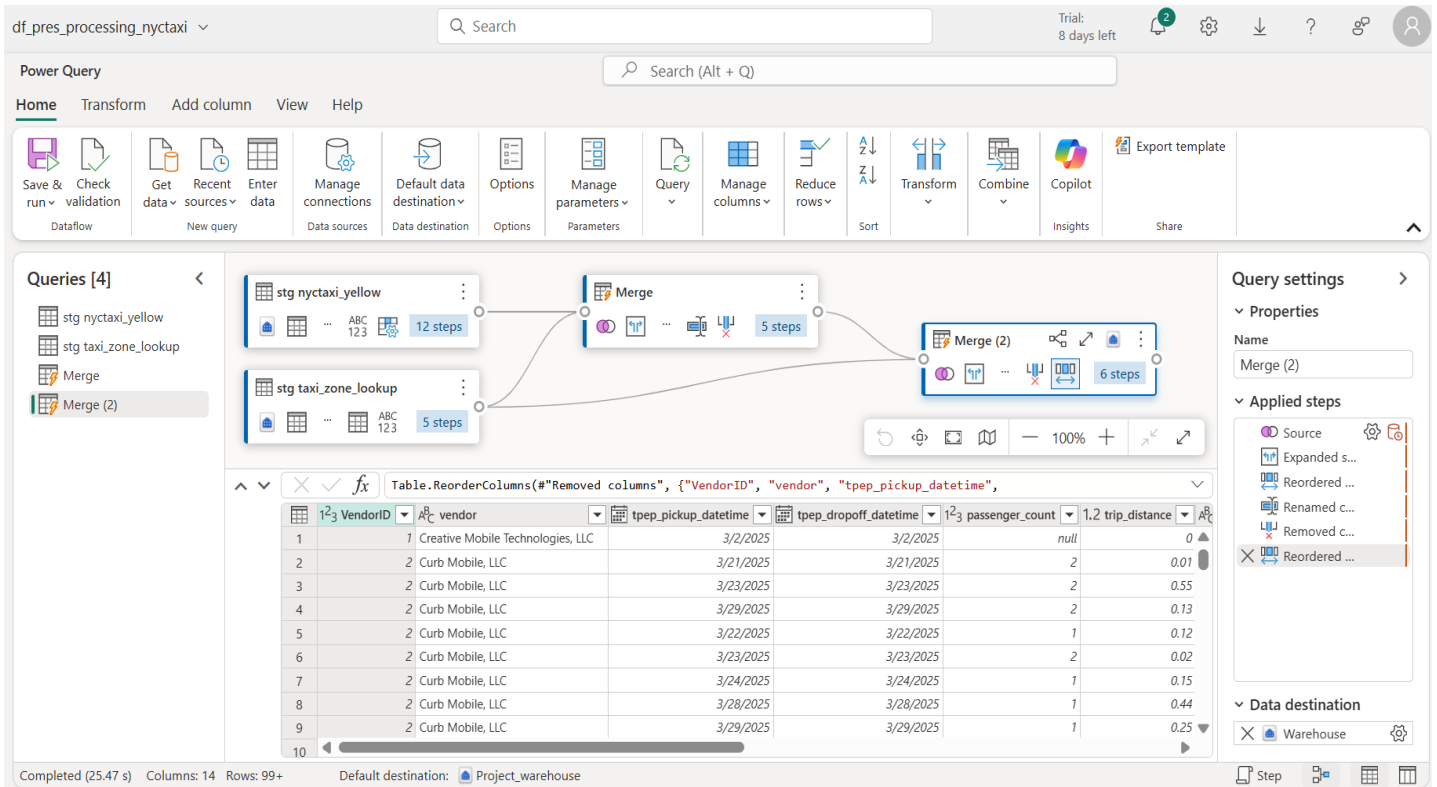


Image 5: Dataflow Gen2 displaying applied transformation steps for cleaning and enriching raw taxi data.

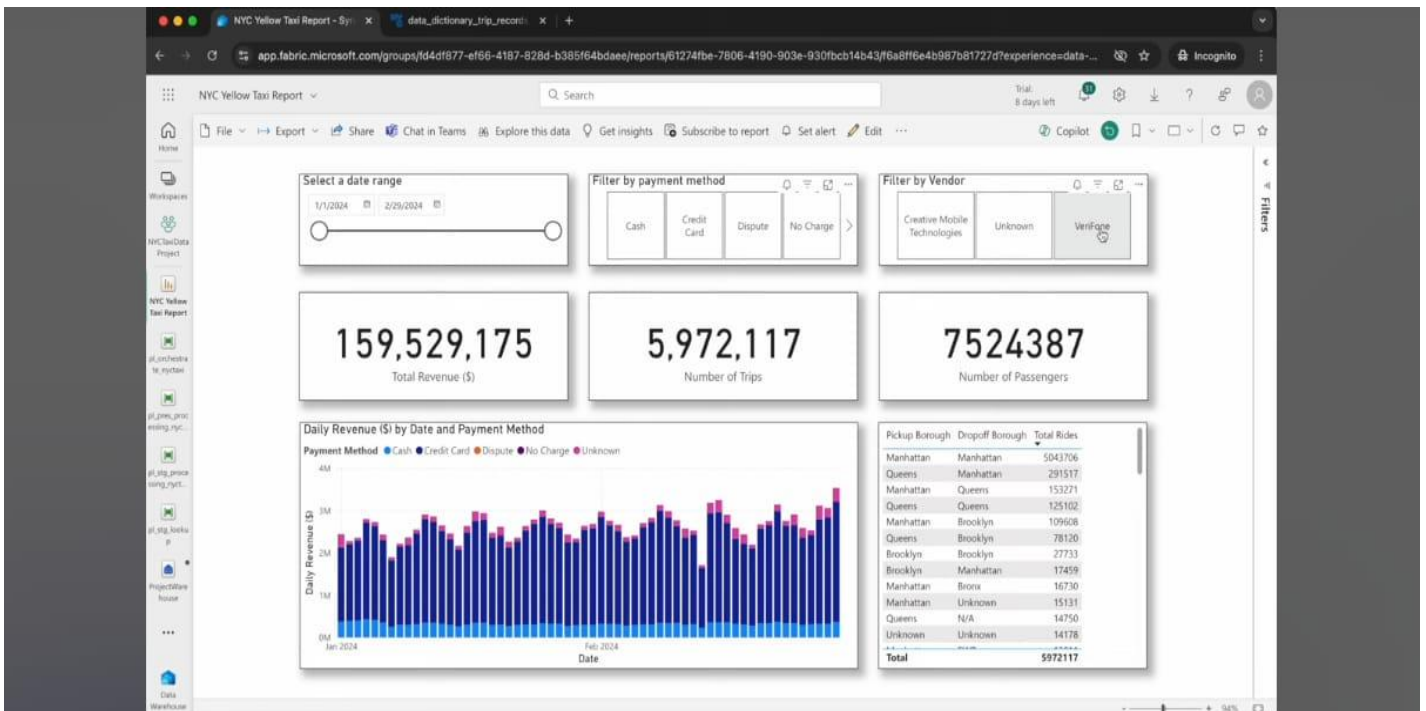


Image 6: Interactive Power BI dashboard visualizing revenue, trips, payment methods, and vendor activity.