

NYC Taxi Data Engineering - End-to-End Pipeline on Microsoft Fabric

MICROSOFT FABRIC

LAKEHOUSE

 PYSPARK

3.X

DELTA LAKE

ENABLED

 MLFLOW

TRACKING

POWER BI

ANALYTICS

A production-grade **end-to-end data engineering project** built on **Microsoft Fabric**, demonstrating enterprise-level data pipeline architecture using the **Medallion Architecture** pattern. This project processes **190+ million NYC Yellow Taxi trip records** through a complete data lifecycle from ingestion to business intelligence.



Table of Contents

- [Project Overview](#)
- [Architecture](#)
- [Data Pipeline](#)
- [Star Schema Design](#)
- [Machine Learning Models](#)
- [Power BI Dashboards](#)
- [Technology Stack](#)
- [Getting Started](#)
- [Project Structure](#)
- [Key Features](#)
- [Results & Insights](#)
- [Author](#)
- [License](#)

Project Overview

This project demonstrates a comprehensive data engineering solution that:

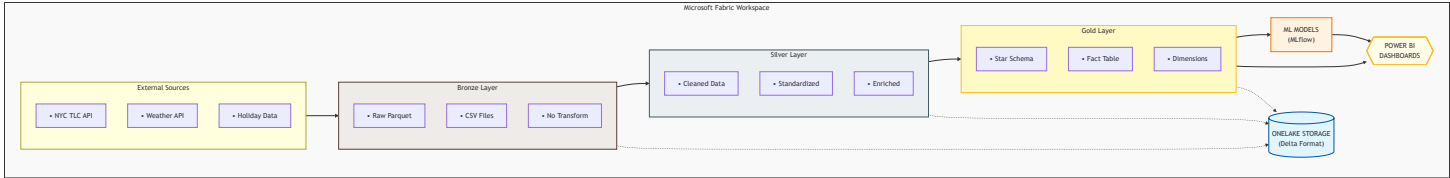
- **Ingests** real-world NYC Taxi data from NYC TLC (Taxi & Limousine Commission)
- **Processes** 190+ million records using PySpark on Microsoft Fabric
- **Transforms** raw data through Bronze → Silver → Gold medallion layers
- **Builds** a Star Schema optimized for analytical workloads
- **Trains** 3 Machine Learning models with MLflow experiment tracking
- **Delivers** interactive Power BI dashboards for business insights

Data Scale




Metric	Value
Total Records	189,523,210
Time Range	Jan 2021 - Nov 2025
Total Revenue	\$4.9 Billion
Unique Zones	265
Weather Records	1,826 days

Architecture

High-Level Architecture

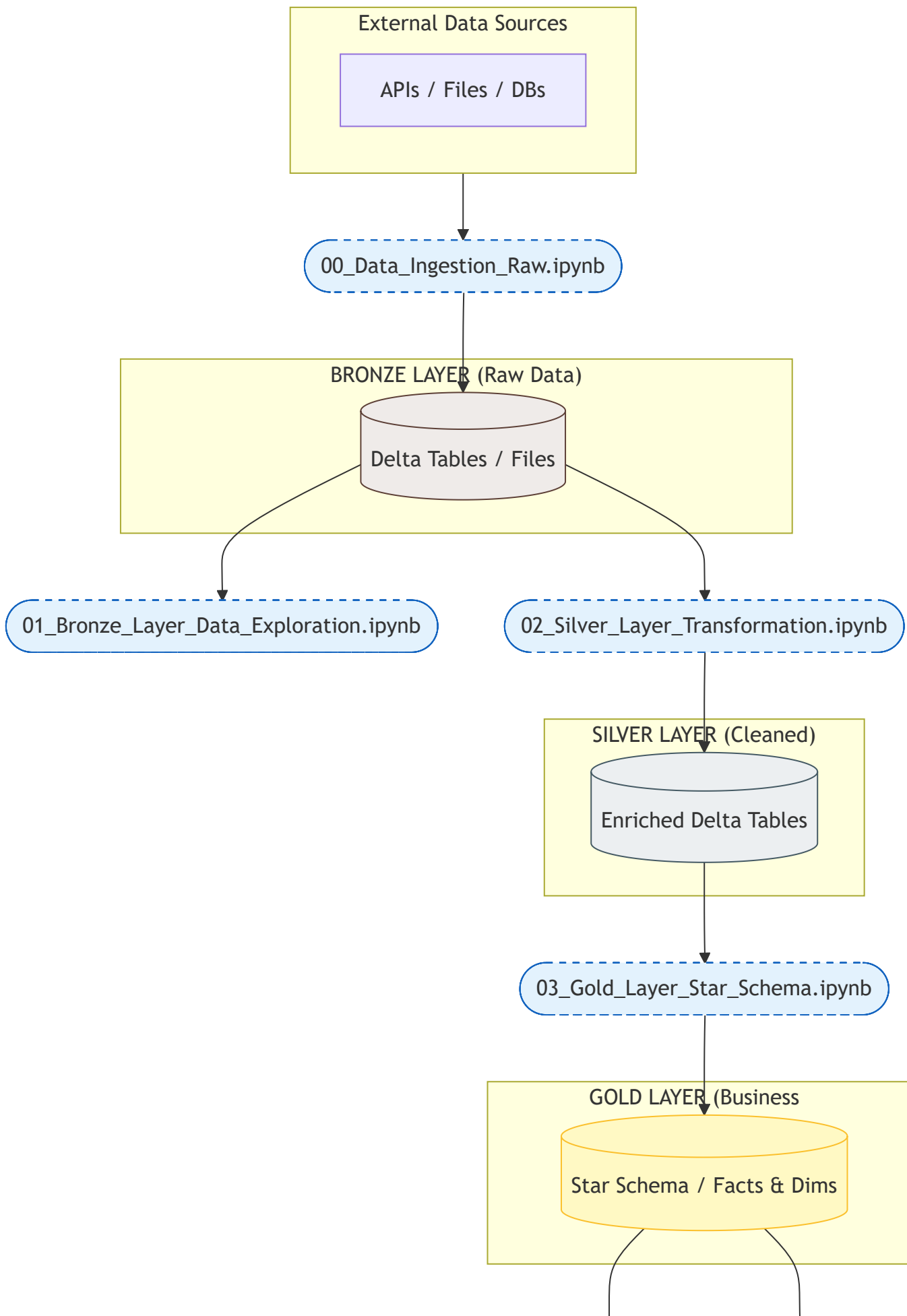


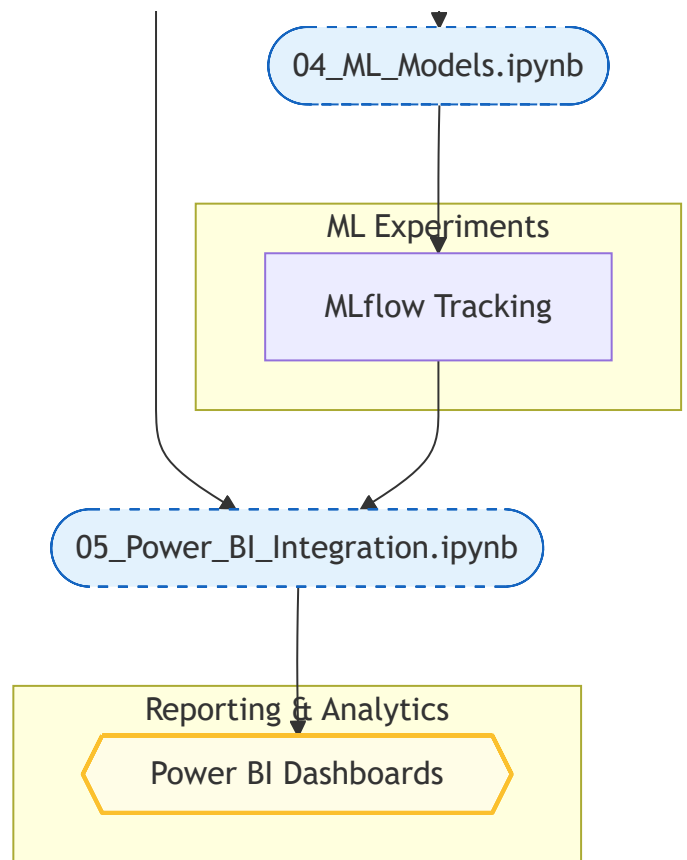
Medallion Architecture

Layer	Purpose	Storage Format	Tables
 Bronze	Raw data ingestion, schema validation	Parquet/CSV	Raw taxi trips, zones, weather, holidays
 Silver	Data cleansing, standardization, enrichment	Delta Lake	Cleaned trips, enriched zones, calendar
 Gold	Business-ready analytics, star schema	Delta Lake	Fact table + 7 dimensions

Data Pipeline

Notebook Execution Flow





Data Sources

Source	Description	Records	Format
NYC TLC	Yellow Taxi Trip Data (2021-2025)	189M+	Parquet
Taxi Zones	NYC Geographic Zones	265	CSV
Weather	Daily NYC Weather Data	1,826	CSV
US Holidays	Federal Holiday Calendar	75	CSV

★ Star Schema Design

The Gold layer implements a classic **Star Schema** optimized for Power BI analytics:

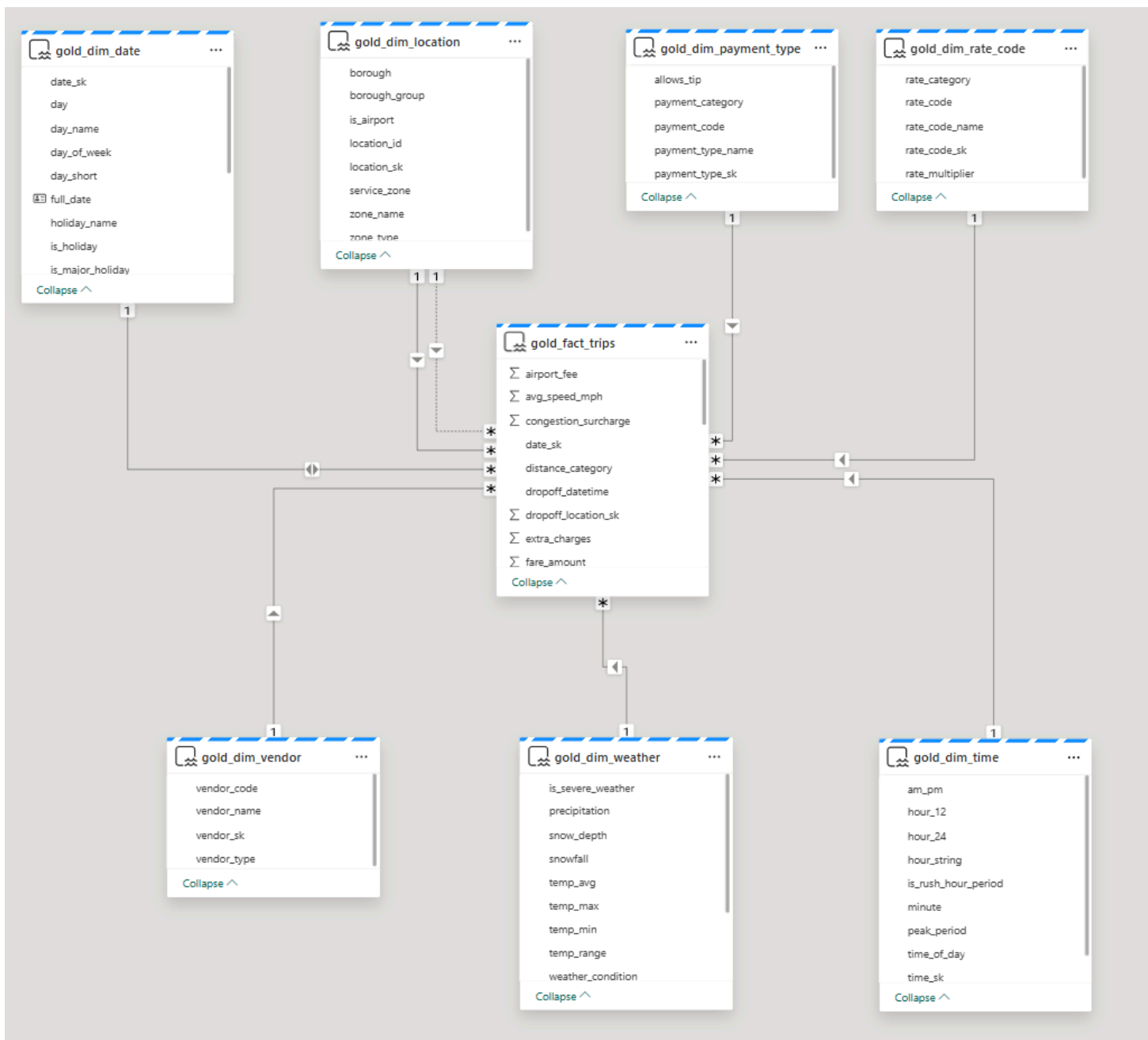


Table Details




Table	Type	Records	Key Columns
gold_fact_trips	Fact	189M	fare_amount, trip_distance, tip_amount, surcharges
gold_dim_date	Dimension	1,826	date_sk, year, month, day_name, is_holiday
gold_dim_time	Dimension	1,440	time_sk, hour_24, am_pm, is_rush_hour

Table	Type	Records	Key Columns
gold_dim_location	Dimension	265	location_sk, zone_name, borough, is_airport
gold_dim_weather	Dimension	1,826	weather_sk, temp_avg, precipitation, weather_condition
gold_dim_vendor	Dimension	3	vendor_sk, vendor_name, vendor_code
gold_dim_payment_type	Dimension	7	payment_type_sk, payment_type_name, allows_tip
gold_dim_rate_code	Dimension	7	rate_code_sk, rate_code_name, rate_multiplier

Machine Learning Models

Three ML models are trained using **PySpark MLlib** with **MLflow** for experiment tracking and model versioning:

Model Performance

Model	Algorithm	Metric	Score
 Fare Prediction	GBT Regressor	RMSE	\$5.45
		MAE	\$3.00
		R ²	0.886
 Tip Classification	Random Forest	AUC	0.619
		Accuracy	77.5%
		F1 Score	0.687
 Demand Forecasting	GBT Regressor	RMSE	36.42
		MAE	24.78
		R ²	0.919

MLflow Tracking

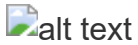
- **Experiment Name:** NYC_Taxi_ML_Models
- **Auto-versioning:** Models automatically increment version numbers
- **Artifact Storage:** Files/ml_models/{model_name}/{version}/
- **Model Registry:** gold_ml_model_registry table



Power BI Dashboards

Interactive dashboards providing business insights across 6 analytical dimensions:

1. Executive Dashboard



Key metrics at a glance:

- **189.52M** Total Trips
- **\$4.9B** Total Revenue
- **\$17.86** Average Fare
- **20.35%** Average Tip Percentage

2. Geographic Analysis



- Top 10 Pickup Zones
- Top Routes Table
- Airport vs Non-Airport Distribution
- Borough Comparison

3. Time Analysis



- Hourly Demand Pattern
- Day of Week Distribution
- Monthly Trend by Year
- Hour vs Day Heatmap

4. Weather Impact Analysis

Weather Impact

- Trips by Weather Condition
- Temperature vs Trips Correlation
- Speed vs Weather Relationship
- Severe Weather Loss Estimation

5. Financial Performance

Financial Performance

- Revenue YTD with YoY Growth
- Revenue Trend with Forecast
- Revenue by Payment Type
- Revenue per Mile Analysis

6. ML Model Performance

ML Performance

- Model Metrics Summary
- Actual vs Predicted Visualization
- Model Version History
- Experiment Tracking

🔧 Technology Stack

Category	Technologies
Platform	Microsoft Fabric, OneLake
Processing	Apache Spark, PySpark
Storage	Delta Lake, Parquet
ML/AI	PySpark MLlib, MLflow
Visualization	Power BI

Category	Technologies
Languages	Python, SQL, DAX
Version Control	Git, GitHub

Getting Started

Prerequisites

- Microsoft Fabric workspace with Lakehouse capacity
- Power BI Pro license (for publishing reports)
- Git (for cloning repository)

Installation

1. Clone the repository

```
git clone https://github.com/pqnghiep1354/nyc-taxi-fabric-pipeline.git
cd nyc-taxi-fabric-pipeline
```

2. Create Fabric Lakehouse

- Navigate to your Microsoft Fabric workspace
- Create a new Lakehouse named `TaxiAnalytics_Lakehouse`

3. Upload Notebooks

- Import all `.ipynb` files from the `notebooks/` directory
- Attach notebooks to the Lakehouse

4. Execute Pipeline

Run notebooks in order:
`00 → 01 → 02 → 03 → 04 → 05`

5. Create Power BI Report

- Follow instructions in `05_Power_BI_Integration.ipynb`
- Create Semantic Model from Gold layer tables
- Build dashboards using provided DAX measures

Project Structure

```
nyc-taxi-fabric-pipeline/
|
├── 📁 notebooks/
|   ├── 00_Data_Ingestion_Raw.ipynb           # Data source ingestion
|   ├── 01_Bronze_Layer_Data_Exploration.ipynb # Schema validation & exploration
|   ├── 02_Silver_Layer_Transformation.ipynb  # Data cleansing & enrichment
|   ├── 03_Gold_Layer_Star_Schema.ipynb       # Star schema creation
|   ├── 04_ML_Models.ipynb                    # ML training with MLflow
|   └── 05_Power_BI_Integration.ipynb         # BI layer setup
|
├── 📁 reports/
|   └── NYC_Taxi_Analytics.pbix                # Power BI report file
|
├── 📁 docs/
|   ├── images/                              # Dashboard screenshots
|   ├── NYC_TLC_Data_Dictionary.md            # Column definitions
|   └── DAX_Measures_Library.md               # DAX formula reference
|
├── 📄 README.md                             # This file
└── 📄 LICENSE                               # MIT License
```



Key Features

Data Engineering






- ✓ **Medallion Architecture** - Bronze/Silver/Gold layer separation
- ✓ **Schema Evolution** - Handling NYC TLC schema changes across years
- ✓ **Incremental Loading** - Support for new data additions
- ✓ **Data Quality Rules** - Comprehensive validation and cleansing
- ✓ **Delta Lake Optimization** - ACID transactions, time travel

Machine Learning

- ✓ **MLflow Integration** - Experiment tracking and model registry
- ✓ **Auto-versioning** - Automatic model version management

-  **Multiple Models** - Regression, classification, forecasting
-  **Feature Engineering** - Time-based, weather, and geographic features

Business Intelligence

-  **Star Schema** - Optimized dimensional modeling
-  **Pre-aggregated Tables** - Performance-optimized for Power BI
-  **42 DAX Measures** - Complete measure library
-  **6 Dashboard Pages** - Comprehensive analytical views
-  **Forecasting** - Built-in trend analysis



Results & Insights

Key Business Insights

1. **Manhattan Dominance:** 88.39% of all trips originate in Manhattan
2. **Peak Hours:** Evening rush (5-7 PM) shows highest demand
3. **Weather Impact:** Rainy days reduce trip volume by 0.7%
4. **Payment Shift:** Credit card usage at 76.9% (vs 13.1% cash)
5. **Airport Traffic:** 7.4% of trips are airport-related

Model Insights



- **Fare Prediction:** Distance and time of day are strongest predictors
- **Tip Classification:** Payment type and trip distance impact tip likelihood
- **Demand Forecasting:** Strong weekly seasonality patterns detected



Author

Phạm Quốc Nghiệp

Last update: 2026-01-13

-  GitHub: [@phamquocnghiep](#)
-  LinkedIn: [Phạm Quốc Nghiệp](#)



License

This project is licensed under the MIT License - see the [LICENSE](#) file for details.



Acknowledgments

- **NYC Taxi & Limousine Commission** - For providing open taxi trip data
- **Microsoft Fabric Team** - For the excellent data platform
- **Apache Spark Community** - For the powerful processing engine



★ Star this repository if you found it helpful!