NYC Taxi Analytics: Data Warehouse Project

Diogo Santos (up202108747)
Manuel Alves (up201906910)
Rodrigo Esteves (up202403070)

7 January, 2025

Index

Introduction

Our objective is to develop a data warehouse for NYC Yellow Taxi operations, using a public dataset from the **Taxi and Limousine Commission (TLC)**. The main goal is to transform rawtrip data into a multidimensional schema that assures online analytical processing (OLAP) of demand, revenue, operational intensity, and behavioural patterns across temporal and geographic dimensions.

Planning

Planning begins by identifying the facts and their grains. This model has two fact tables, one analyzing each trip, and the other one a daily aggregated report. Below, is the Dimensional Bus Matrix that summarizes which dimensions and facts participate in each star schema

Data mart	Star / Dimension	Date	TimeOfDay	Vendor	Location	RateCode	PaymentType	TripCharacteristics	PassengerGroup
NYC Taxi Analytics	Trip	X	X	X	X	X	X	X	X
	Daily_ZoneVendor	X		X	X				

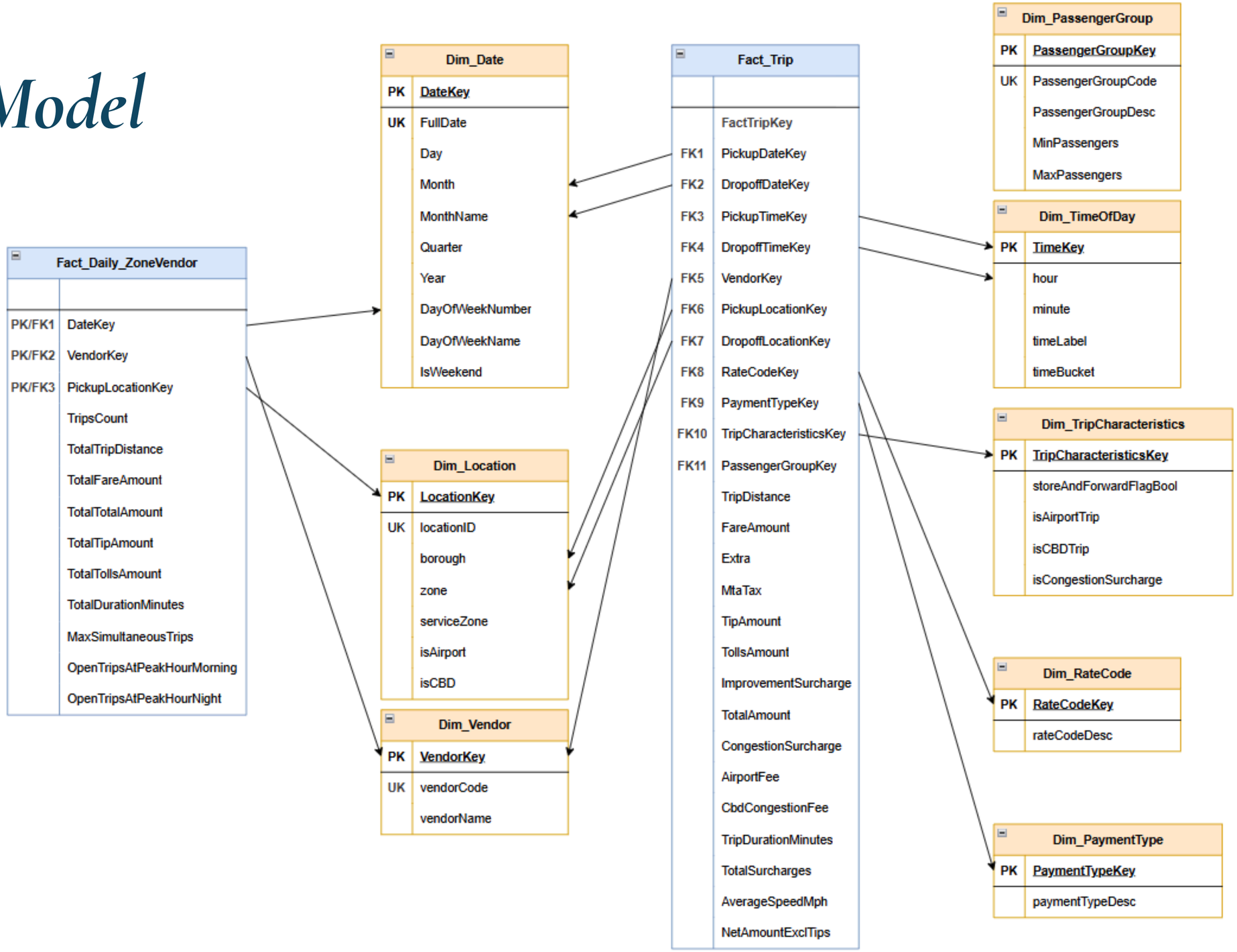
Dimensional Model

Two star schemas

- **Fact_Trip:** one row per trip (detailed analysis)
- **Fact_Daily_ZoneVendor:** daily snapshot by (Date, Vendor, Pickup Zone)

Eight Dimensions

- **Dim_Date**
- **Dim_Location**
- **Dim_Vendor**
- **Dim_PassengerGroup**
- **Dim_TimeOfDay**
- **Dim_TripCharacteristics**
- **Dim_RateCode**
- **Dim_PaymentType**



Facts

Fact_Trip

The main schema is centred on Fact_Trip, at the grain of one row per taxi trip, and links to role-playing dimensions for pickup and dropoff date, pickup and dropoff time-of-day, and pickup and dropoff location, as well as to Vendor, RateCode, PaymentType, PassengerGroup, and TripCharacteristics.

Star	Trip	*Version*	1.0
Granularity	One row per completed yellow taxi trip.		
Dimension	*Role in Fact Table*		
Date	PickupDate / DropoffDate (role-playing)		
TimeOfDay	PickupTime / DropoffTime (role-playing)		
Vendor	Vendor		
Location	PickupLocation / DropoffLocation (role-playing)		
RateCode	RateCode		
PaymentType	PaymentType		
TripCharacteristics	TripCharacteristics		
FactTripKey	FactTripKey		
PassengerGroup	PassengerGroup		
Measure	*Description*	*Additivity*	
TripDistance	Distance of the trip in miles.	Additive	
FareAmount	Metered fare (time & distance).	Additive	
Extra	Extras and surcharges.	Additive	
MtaTax	MTA tax applied to the trip.	Additive	
TipAmount	Tip amount (card tips only).	Additive	
TollsAmount	Total tolls paid.	Additive	
ImprovementSurcharge	Improvement surcharge at flag drop.	Additive	
TotalAmount	Total amount charged to passenger (no cash	Additive	
CongestionSurcharge	NYS congestion surcharge amount.	Additive	
AirportFee	Airport fee for JFK/LGA pickups.	Additive	
CbdCongestionFee	CBD congestion relief fee.	Additive	
TripDurationMinutes	Duration from pickup to dropoff, in minutes	Additive	
AverageSpeedMph	trip_distance / (duration_hours) (derived).	Non-additive	
TotalSurcharges	Sum of all surcharges and taxes (derived).	Additive	
NetAmountExclTips	total_amount − tip_amount (derived).	Additive	

Facts

Fact_Daily_ZoneVendor

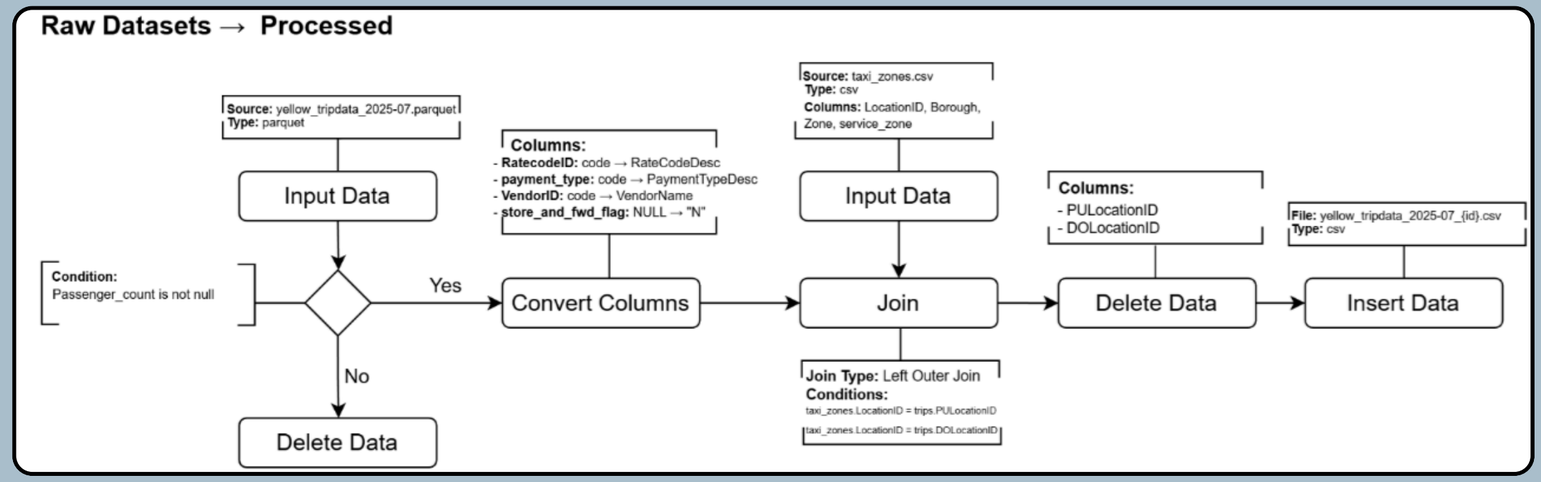
Fact_Daily_ZoneVendor is a daily snapshot at the grain (Date, Vendor, Pickup Location), implemented with a composite primary key over these foreign keys. It supports aggregated reporting through additive totals and includes semi-additive measures, such as the maximum number of simultaneous trips and the number of open trips at peak hours.

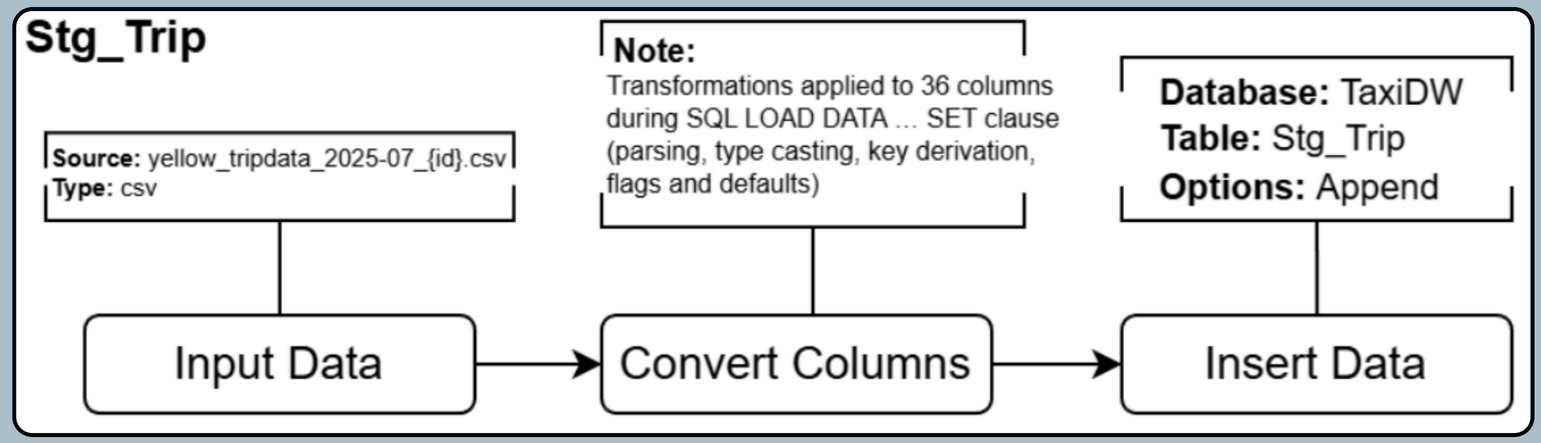
Star	Daily_ZoneVendor	*Version*	1.0
Granularity	One row per day, per vendor, per pickup location		
Dimension Name	***Role in Fact Table***		
Date	Date		
Vendor	Vendor		
Location	PickupLocation		
Measure	***Description***		***Additivity***
TripsCount	Number of trips in that day/vendor/location.		Additive
TotalTripDistance	Sum of trip_distance.		Additive
TotalFareAmount	Sum of fare_amount.		Additive
TotalTotalAmount	Sum of total_amount.		Additive
TotalTipAmount	Sum of tip_amount.		Additive
TotalTollsAmount	Sum of tolls_amount.		Additive
TotalDurationMinutes	Sum of trip_duration_minutes.		Additive
MaxSimultaneousTrips	Max Simultaneous Trips during day D, for vendor V at location L		Semi-Additive
OpenTripsAtPeakHourMorning	Open Trips At 09h during day D, for vendor V at location L		Semi-Additive
OpenTripsAtPeakHourNight	Open Trips At 18h during day D, for vendor V at location L		Semi-Additive

ETL

Pre-processing + Staging

- Raw TLC trips + zone lookup and standardized file.
- **Staging:** type casting + derived fields (duration, speed, keys)
- **Quality filter:** remove invalid/incomplete trips

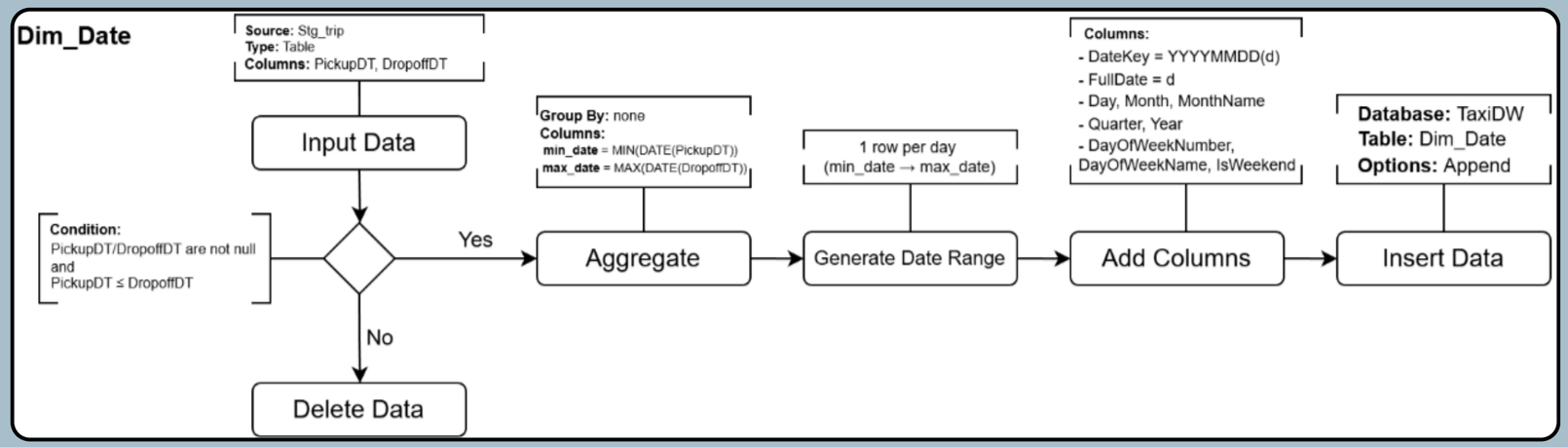


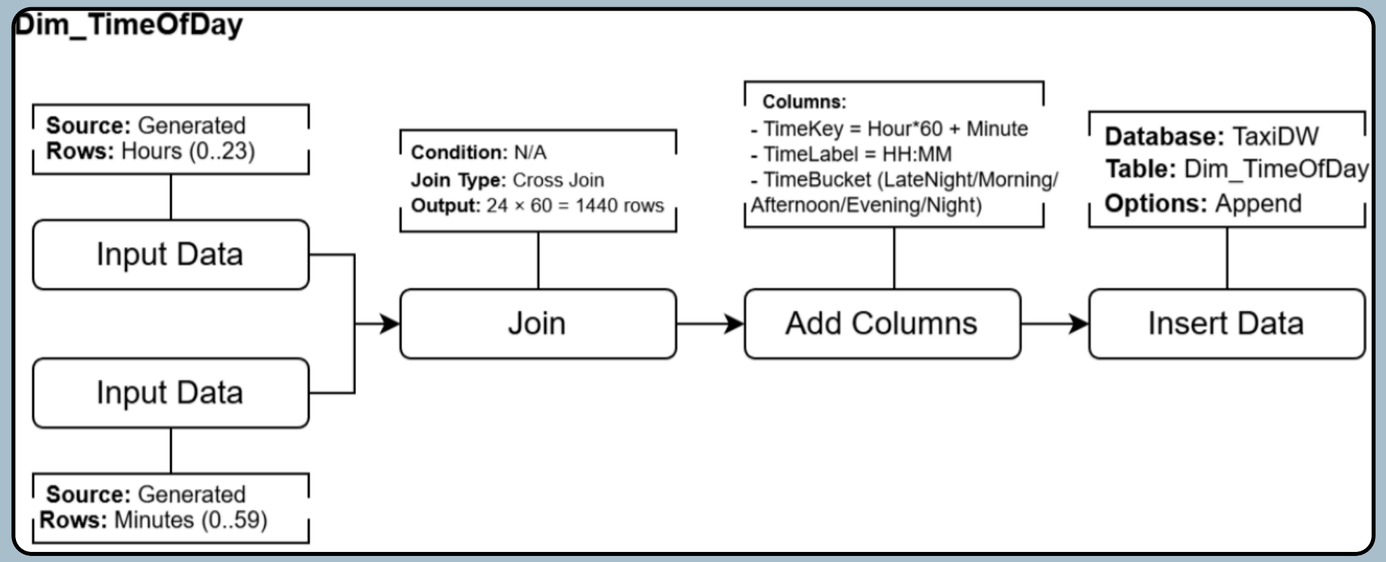


ETL

Time dimensions

- Date and time are conformed and role-played (pickup vs dropoff)
- Enables drill-down: day - hour bucket (and weekday patterns)

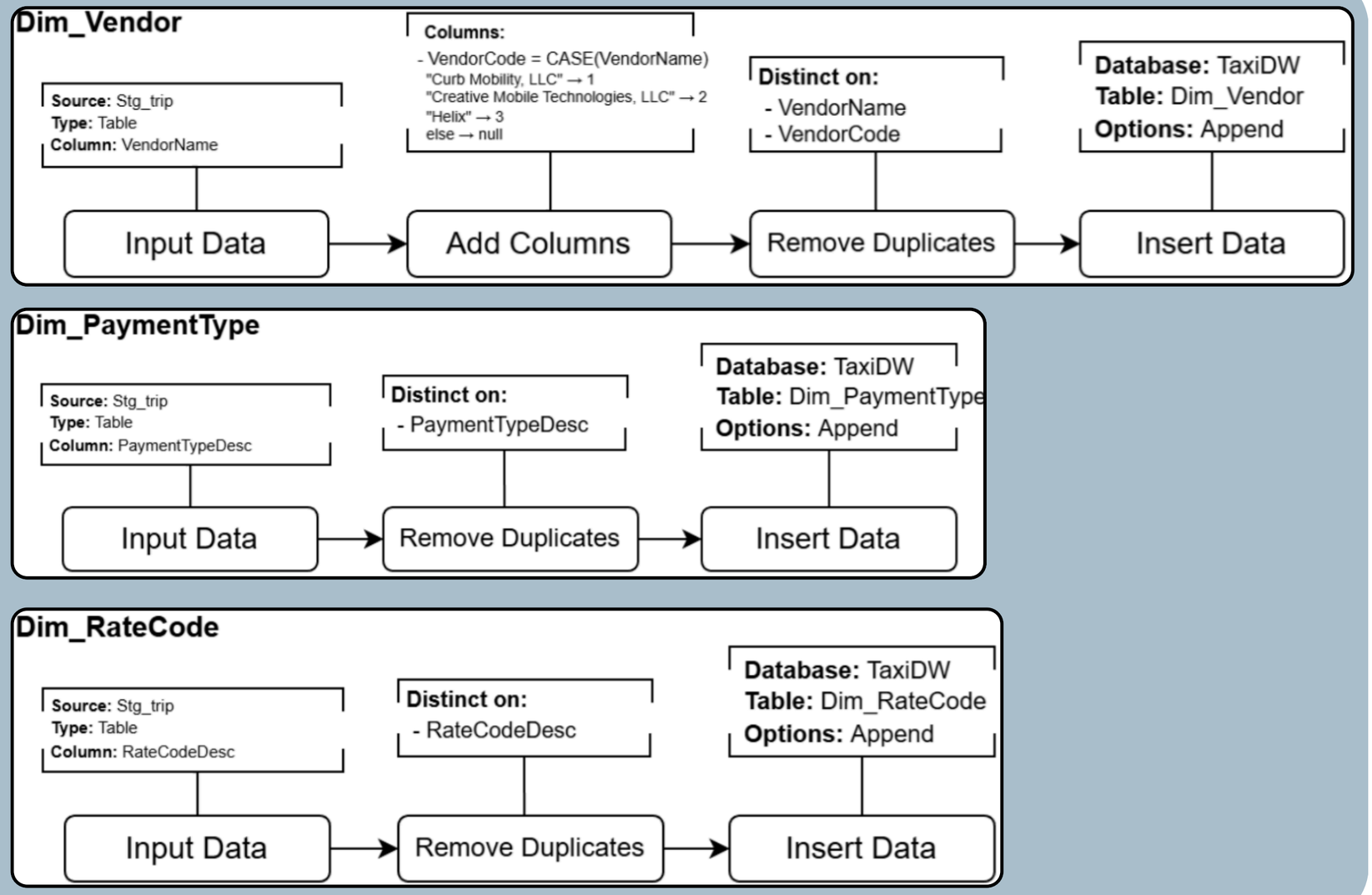




ETL

Reference dimensions

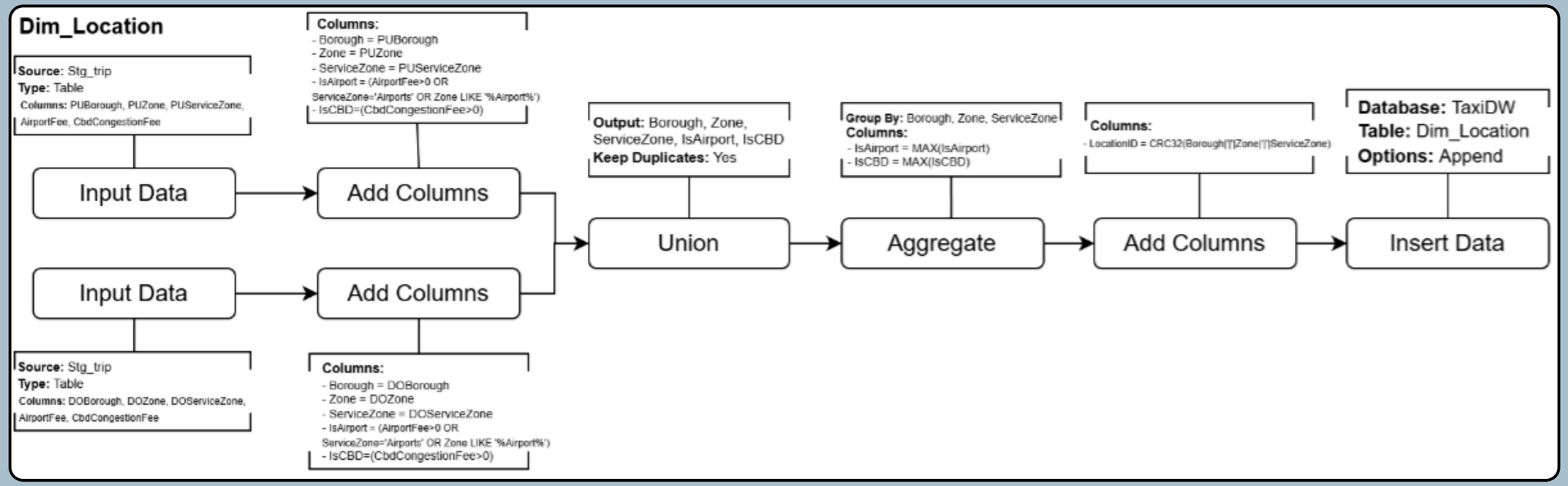
- Static reference tables generated from distinct codes + labels



ETL

Location

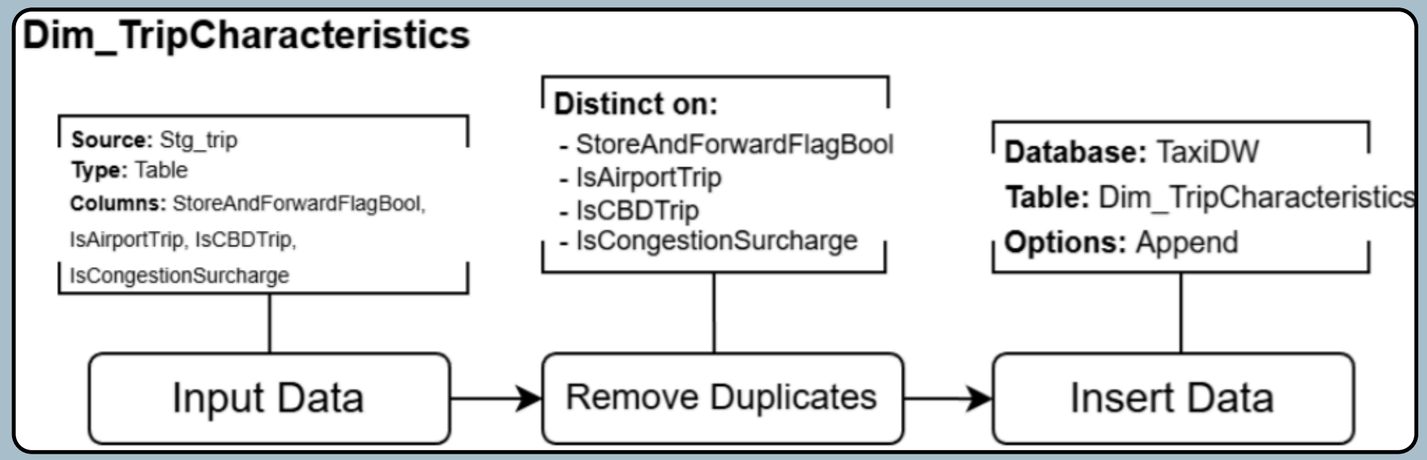
- Enriches location IDs with zone/borough
- Central for spatial analysis (maps, top zones, revenue concentration)

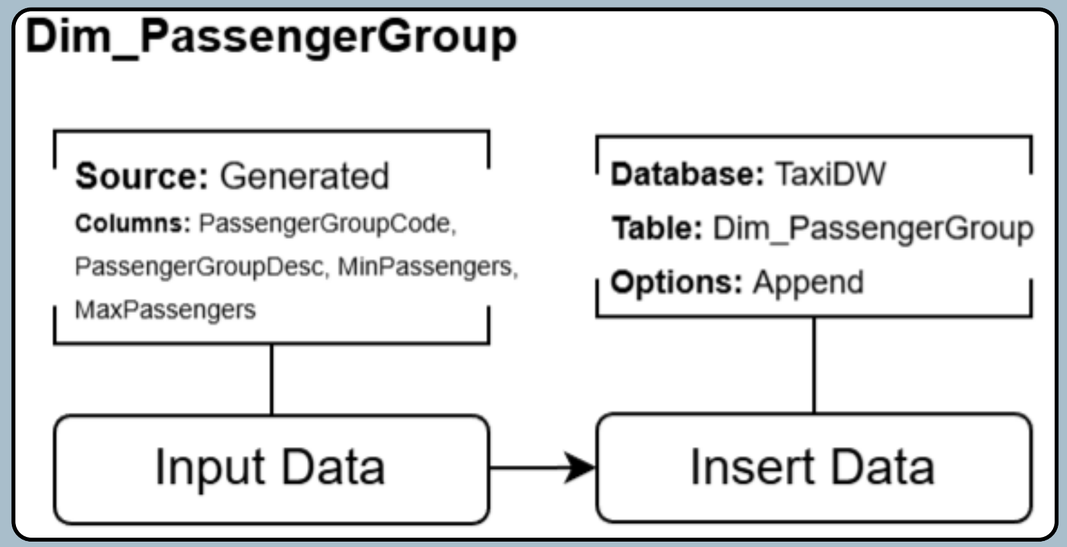


ETL

Grouping dimensions

- Derived categorical groupings for stable slicing
- Improves readability and avoids high-cardinality noise

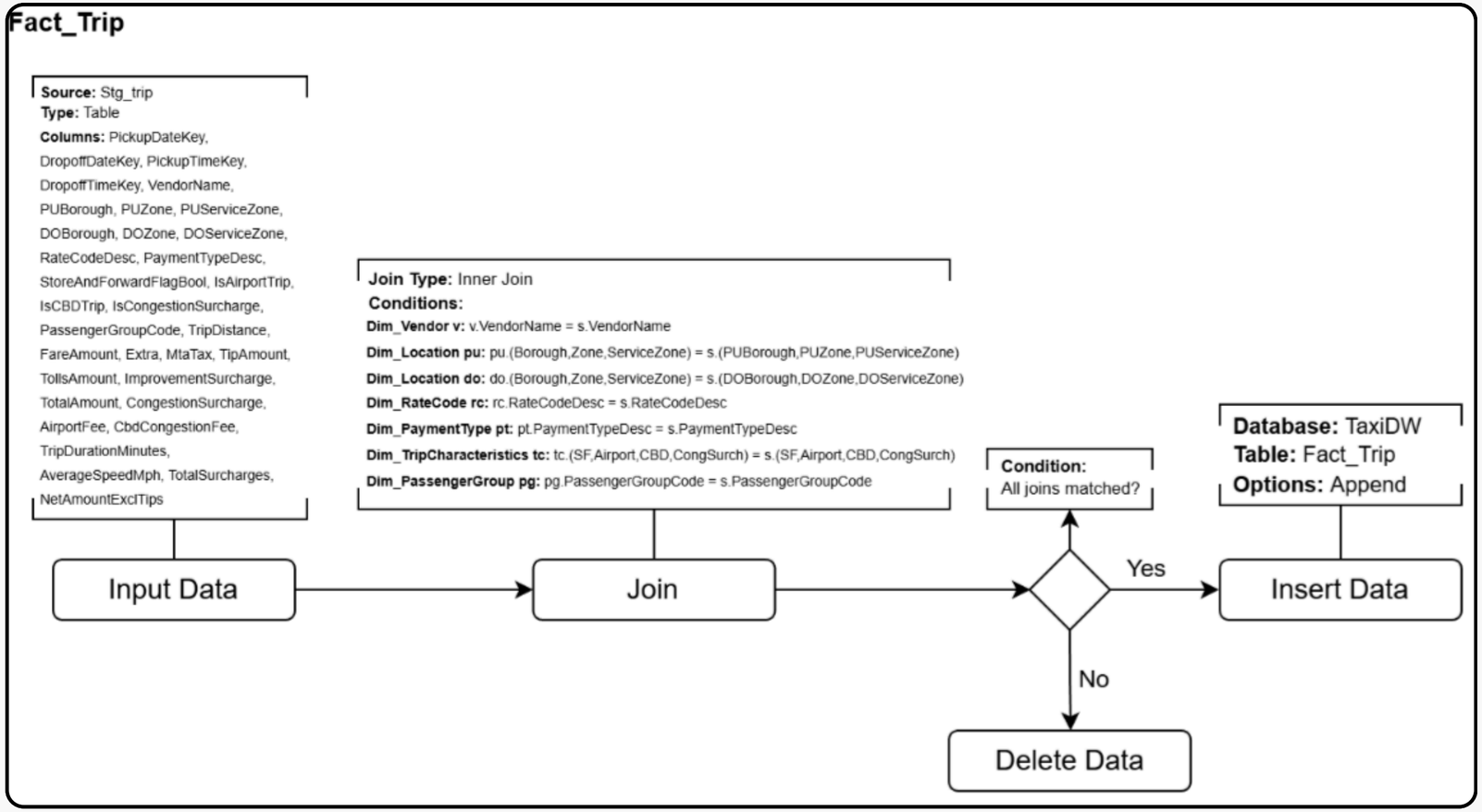




ETL

Fact_Trip

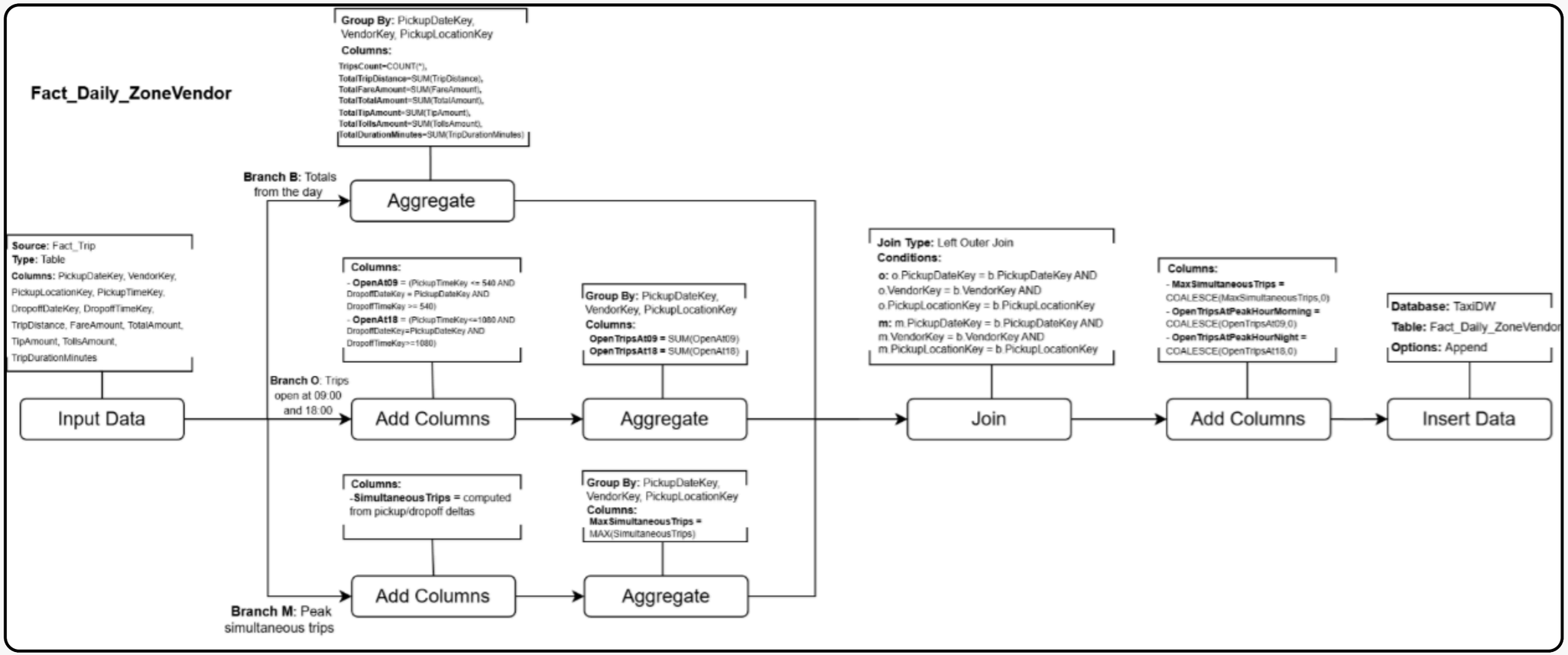
- **Grain:** 1 row per trip
- Fully keyed to pickup/dropoff time and location + vendor/payment/rate
- **Measures:** fare components, totals, distance, duration (and non-additive metrics)



ETL

Fact_Daily_ZoneVendor

- Grain: (Date, Vendor, Pickup Zone)
- Additive totals + semi-additive operational indicators (snapshots/max)



Querying and Data Analysis

Conclusion

This project shows why a separate **Data Warehouse** is ideal for this kind of analysis. In **OLTP** systems, complex queries (summaries, rankings, overlap checks) are expensive and hard to run, while a **dimensional model** makes them faster and easier to understand. Heavy calculations were moved to the **ETL**, including pre-built summaries and special metrics. This lets the **Power BI** dashboard display results instantly. With this setup, we found that demand and revenue follow strong time patterns and are concentrated in a few zones. Next steps include scaling the pipeline to the full dataset with incremental updates, stronger data validation.

Questions?