Yellow Taxis

Green Taxis

For-Hire-Vehicles (FHVs)

Operate in different ways

Attributes

- Pickup & drop time
- Pickup & drop location details
- Trip distance
- Passenger count
- Solo / shared ride
- Payment – Tips, tolls, tax etc.

Exponential increase in data volumes

Different schemas & formats

Comparison between all taxi services

Granular & new aggregated reports

Faster turnaround time

Raw data for analysis

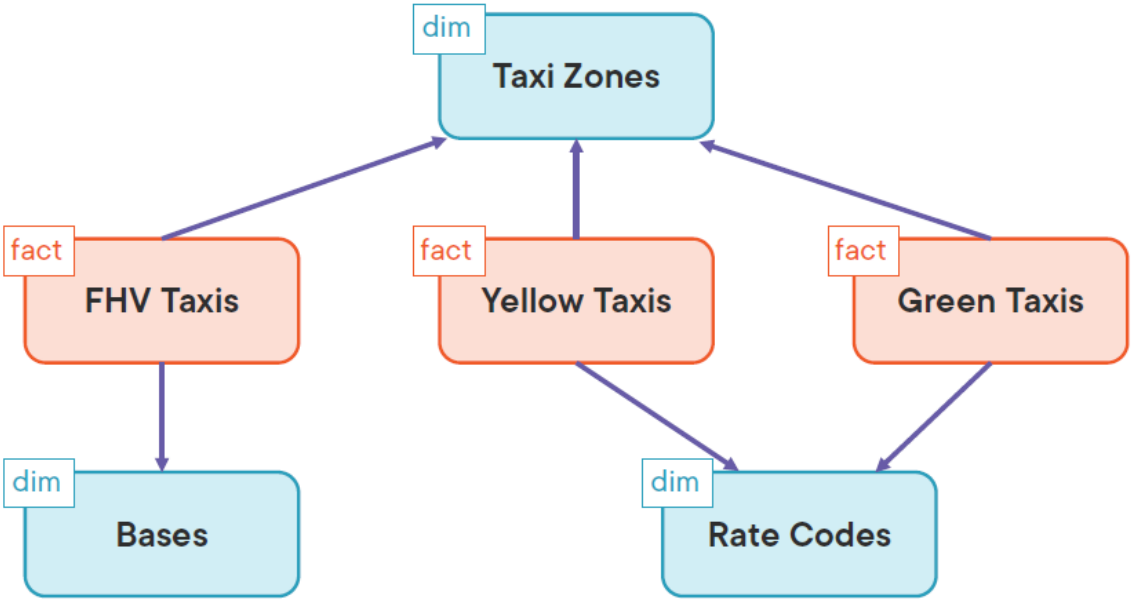
Problem Statement

Assume that your organization is responsible for processing New York City taxi service data. There are three types of taxis in New York, yellow taxis, green taxis, and for-hire vehicles, or FHVs. FHVs include app-based taxis like Uber, Via, and Lyft, and each of them dispatches more than 10,000 trips per day, and all of them operate in different ways. The company collects ride-related information like pickup and drop time, details of pickup and drop location, trip distance, passenger count as reported by the driver, if it's a solo or a shared trip, as well as payment-related information. Initially, the system was working well, but now the company is seeing exponential increase in data volumes, and each taxi type uses different schema and shares data in different file formats like CSV, JSON, and Parquet. The new requirement is to compare taxis against one another and build granular and new aggregated reports. And the users are looking for faster turnaround time in terms of data processing and new requirement implementation. And of course, raw data must also be available for analysis. That's why to address these challenges your organization has decided to build a data lakehouse.

Data Model

We have three facts, one for FHV taxis, second for yellow taxis, and third for green taxis. FHV taxi's fact is linked to bases dimension, which tells us the dispatch-based location of the taxi. On the other hand, yellow and green tax facts are linked to the dimension rate codes, which tells us if the trip is a shared trip, solo trip, or trip to any specific airport. And all the three facts are linked to taxi zones dimension, which will tell us the pickup location and drop location details for each trip. Implement the data warehouse using both spark pool and SQL pool and list the Pros and Cons of each implementation.

Data Model Design


dim
Taxi Zones
fact
FHV Taxis
fact
Yellow Taxis
fact
Green Taxis
dim
Bases
dim
Rate Codes