

|  |
| --- |
| Business Template  **NYC taxi trip** |
|  |



Contents

[1 Business Description 3](#_Toc155614186)

[1.1 Business background 3](#_Toc155614187)

[1.2 Problems because of poor data management 3](#_Toc155614188)

[1.3 Benefits from implementing a Data Warehouse 3](#_Toc155614189)

[1.4 DATASETS DESCRIPTION 3](#_Toc155614190)

[1.5 GRAIN / DIM / FACT 4](#_Toc155614191)

[2 Business Layer 3NF 4](#_Toc155614192)

[3 Business Layer Dimensional Model 4](#_Toc155614193)

[4 Logical Scheme 4](#_Toc155614194)

[5 Data Flow 4](#_Toc155614195)

[6 Fact Table Partitioning Strategy 4](#_Toc155614196)

# 

# Business Description

## Business background

Taxis are a vital part of the transportation infrastructure in any large city, including New York. In this industry, there are many operators offering taxi services for various types of trips, from short city rides to long trips to airports. Competition in this sector is high, so successful companies must effectively manage their data to improve service and meet customer needs.

## Problems because of poor data management

Poor data management hampers effective business operations due to insufficient information for decision-making. Without using data analysis tools, companies cannot develop a strategy focused on customer needs and operational optimization. This can lead to decreased market competitiveness and lower quality of services provided.

## Benefits from implementing a Data Warehouse

Implementing a data warehouse can help solve the described problems and answer the following questions:

* What times of day are most in demand for taxi rides?
* Which areas of the city (by longitude and latitude) are most frequently served?
* How does demand for taxis change depending on the day of the week or month?
* Which types of trips (business, private) are more popular?

## DATASETS DESCRIPTION

**The First Dataset: Yellow Taxi Data**

Information about Yellow Taxi trips:

* **Vendor Information**:
  + vendor\_name: Vendor name
  + vendor\_address: Vendor address
  + vendor\_telephone: Vendor telephone
* **Payment Information**:
  + payment\_type: card or cash
  + amount: price per trip
  + payment\_time: payment time
* Data transmission information:
  + store\_and\_fwd\_flag: represents a flag that indicates whether the trip information was temporarily stored in the taxi's memory before being sent to the server
  + data\_description: description about record sended to served immediately, or not
* **Tarife Information**:
  + base\_fare: base fare
  + rate\_per\_mile: rate per mile
* **Trip Information**:
  + trip\_id: Trip identifier
  + pickup\_longitude: pickup longitude
  + pickup\_latitude: pickup latitude
  + dropoff\_longitude: dropoff longitude
  + dropoff\_latitude: dropoff latitude
  + pickup\_datetime: Pickup date and time
  + dropoff\_datetime: Dropoff date and time
* **Trip details Information**:
  + distance\_miles: distance\_miles
  + trip\_duration: trip duration

**The Second Dataset: Green Taxi Data**

Information about Green Taxi trips:

* **Vendor Information**:
  + vendor\_name: Vendor name
  + vendor\_address: Vendor address
  + vendor\_telephone: Vendor telephone
* **Customer Information**:
  + customer\_type: individual, or business
  + passenger\_count: how many people on board
  + customer telephone: telephone number used to place the ordrer
* **Booking**:
  + booking\_type: phone, or street
  + booking time: booking time
* **Payment Information**:
  + payment\_type: card or cash
  + amount: price per trip
  + payment\_time: payment time
* **Tarife Information**:
  + base\_fare: base fare
  + rate\_per\_mile: rate per mile
* **Promotion Information**:
  + promo code: promo code
  + discount\_percentage: 10%, 20%, none
* **Trip Information**:
  + trip\_id: Trip identifier
  + vendor\_id: Vendor identifier
  + vendor\_name: Vendor name
  + pickup\_datetime: Pickup date and time
  + dropoff\_datetime: Dropoff date and time
  + passenger\_count: Number of passengers
  + trip\_duration: Duration of the trip
* **Trip details Information**:
  + distance\_miles: distance\_miles
  + trip\_duration: trip duration

## GRAIN / DIM / FACT

#### Grain

The granularity of the datasets is defined at the level of individual trips. Each row in the fact table represents one taxi trip.

#### Dimensions and Facts

1. **Yellow Taxi Data**:
   * **Dimensions**:
     + Vendor: vendor\_id, vendor\_name
     + Time: datetime\_id, pickup\_datetime, dropoff\_datetime, day\_of\_week, month, year, hour, minute
     + Location: location\_id, pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude
     + Booking\_Type: trip\_id, booking\_type
     + Trip\_Flags: trip\_id, store\_and\_fwd\_flag, description\_store\_and\_fwd\_flag
   * **Facts**:
     + Trip: trip\_id, vendor\_id, pickup\_datetime, dropoff\_datetime, passenger\_count, trip\_duration, base\_fare, cost\_per\_mile, trip\_cost, payment\_type
2. **Green Taxi Data**:
   * **Dimensions**:
     + Vendor: vendor\_id, vendor\_name
     + Time: datetime\_id, pickup\_datetime, dropoff\_datetime, day\_of\_week, month, year, hour, minute
     + Location: location\_id, pickup\_longitude, pickup\_latitude, dropoff\_longitude, dropoff\_latitude
     + Customer\_Type: trip\_id, customer\_type
     + Promo\_Code: trip\_id, promo\_code\_used
   * **Facts**:
     + Trip: trip\_id, vendor\_id, pickup\_datetime, dropoff\_datetime, passenger\_count, trip\_duration, base\_fare, cost\_per\_mile, trip\_cost, payment\_type

# Business Layer 3NF

# Business Layer Dimensional Model

# Logical Scheme

# Data Flow

# Fact Table Partitioning Strategy