

|  |
| --- |
| **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%, NULL
* **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

Each record represents one taxi trip. This will be the finest level of detail. We will be analyzing this in this assignment.

#### Dimensions and Facts

The granularity of the datasets is defined at the level of individual trips. Each row in the fact table represents one taxi trip: vendor\_id, customer\_id, booking\_id, payment\_id, transmission\_id, tarife\_id, promo\_id, trip\_id

* vendor

The taxi company transports people from point A to point B

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| vendor\_id | unique identifier for each row PK | int |
| vendor\_name | vendor name | varchar |
| street | part of full address | varchar |
| house | part of full address | varchar |
| city | part of full address | varchar |
| country | part of full address | varchar |
| postal code | address postal code | varchar |
| vendor\_telephone | telephone number | varchar |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| vendor\_id | vendor\_name | street | house | city | country | postal code | vendor\_telephone |
| 3 | Green Taxi | Beaver Street | 33 | New York | USA | NY 10004 | +1 212-639-9675 |

Example with filled data

* customer

A Customer is a client who can be either an individual or a business.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| customer\_id | unique identifier for each row PK | int |
| customer\_type | Business/Individual | varchar |
| passenger\_count | count of passengers | int |
| customer\_telephone | telephone number/NULL | varchar |

|  |  |  |  |
| --- | --- | --- | --- |
| customer\_id | customer\_type | passenger\_count | customer\_telephone |
| 101 | Business | 3 | +1 212-639-9675 |
| 1 | Individual | 1 | NULL |

Example with filled data

* booking

Green Taxi can be ordered from the street and via phone. Yellow Taxi can only be ordered from the street, therefore, the call is only made on the street.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| booking\_id | unique identifier for each row PK | int |
| booking\_type | Phone/Street | varchar |
| booking\_time | time of booking/NULL | datetime |

|  |  |  |
| --- | --- | --- |
| booking\_id | booking\_type | booking\_time |
| 52 | Phone | 3/26/2016 13:18 |

Example with filled data

* payment

For ride can by paid card or cash

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| payment\_id | unique identifier for each row PK | int |
| payment\_type | Card/Cash | varchar |
| payment\_time | payment time/NULL | datetime |

|  |  |  |
| --- | --- | --- |
| payment\_id | payment\_type | payment\_time |
| 1015 | Card | 6/12/2016 0:57 |
| 1158 | Cash | NULL |

Example with filled data

* data\_transmision

Data transmission is necessary to indicate whether the trip record was sent to the vendor immediately or not.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| transmition\_id | unique identifier for each row PK | int |
| store\_and\_fwd\_flag | Y/N | varchar |
| description\_store\_and\_fwd\_flag | description text | text |

|  |  |  |
| --- | --- | --- |
| transmition\_id | store\_and\_fwd\_flag | description\_store\_and\_fwd\_flag |
| 1016 | Y | Yes - The trip record was held in vehicle memory before sending to the vendor because there was no connection. |
| 1101 | N | No - The trip record was sent to the vendor immediately |

Example with filled data

* tarifes

Taxi tarifes base fare plus rate per mile.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| tarife\_id | unique identifier for each row PK | int |
| base\_fare | base fare | decimal(10,2) |
| rate\_per\_mile | rate per mile | decimal(10,2) |

|  |  |  |
| --- | --- | --- |
| tarife\_id | base\_fare | rate\_per\_mile |
| 888 | 2.50 | 1.50 |

Example with filled data

* promotion

Taxi promo codes. With discount: 10%, 20%.

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| promo\_id | unique identifier for each row PK | int |
| promo\_code | promo code with discount | varchar |
| discount\_percentage | discount percentage | int |

|  |  |  |
| --- | --- | --- |
| promo\_id | promo\_code | discount\_percentage |
| 666 | Promo10 | 10 |

Example with filled data

* trip

Information about taxi trip from point A to point B

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| trip\_id | unique identifier for each row (PK) | varchar |
| pickup\_longitude | pickup longitude | float |
| pickup\_latitude | pickup latitude | float |
| dropoff\_longitude | dropoff longitude | float |
| dropoff\_latitude | dropoff latitude | float |
| pickup\_datetime | pickup datetime | datetime |
| dropoff\_datetime | dropoff datetime | datetime |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| trip\_id | pickup\_longitude | pickup\_latitude | dropoff\_longitude | dropoff\_latitude | pickup\_datetime | dropoff\_datetime |
| id18133569 | -73.98041534 | 40.73856354 | -73.9994812 | 40.73115158 | 6/12/2016 0:43 | 6/12/2016 0:54 |

Example with filled data

* fact\_taxi\_trip

Fact table with references on all dimensional tables. Plus measurements: distance, duration, amount

|  |  |  |
| --- | --- | --- |
| Column name | Description | Data Type |
| vendor\_id | vendor id (FK) | int |
| customer\_id | customer id (FK) | int |
| booking\_id | booking id (FK) | int |
| payment\_id | payment id (FK) | int |
| transmision\_id | transmission id (FK) | int |
| tarife\_id | tarife id (FK) | int |
| promo\_id | promotion id (FK) | int |
| trip\_id | trip id (FK) | int |
| distance\_miles | measure; trip distance | float |
| amount | measure; trip amount | decimal(10,2) |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| vendor\_id | customer\_id | booking\_id | payment\_id | transmision\_id | tarife\_id | promo\_id | trip\_id | distance\_miles | amount |
| 1 | 101 | 5001 | 2001 | 3001 | 5 | 7001 | 9001 | 12.5 | 25.75 |

Example with filled data

# Business Layer 3NF

# Business Layer Dimensional Model

# Logical Scheme

# Data Flow

# Fact Table Partitioning Strategy