Caso práctico: almacén de datos para el análisis de los registros de viajes de la New York City Taxi and Limousine Commission (NY taxis)

Solución PR1 – Análisis y diseño del *data warehouse*

Autores: **David Díaz Arias**
Carles Llorach i Rius
Maria Nerea Sevilla Marchena

Índice

1. Análisis de los requisitos

El análisis de los requisitos se basa en identificar las necesidades específicas que tiene una organización particular respecto al análisis de la información. Normalmente, en esta fase, hay que ser previsor y pensar más allá de las necesidades actuales para poder cubrir las necesidades futuras.

La necesidad principal de la organización encargada del análisis de los registros de viajes de la New York City Taxi and Limousine Commission (NYC TLC), es disponer de la información integrada para su análisis y posterior difusión mediante las herramientas de inteligencia de negocio. Estas ayudarán a facilitar la toma de decisiones a todos los usuarios potenciales para garantizar el cumplimiento, entre otros, de los siguientes objetivos:

- **Analizar los viajes de los vehículos de alquiler en Nueva York**.
- **Analizar el transporte de taxis en Nueva York**.

Estos objetivos subrayan la importancia del análisis de datos para comprender y mejorar el sistema de transporte de taxis y vehículos de alquiler en Nueva York, desde la perspectiva de la demanda y del impacto ambiental, que serán los requisitos a los que debe responder nuestro sistema.

El diseño de un almacén de datos incluye la creación y la implementación de un modelo dimensional o multidimensional, el diseño y la implementación de procesos de ETL y del modelo OLAP y, por último, el diseño de las consultas establecidas en el enunciado.

Teniendo en cuenta los objetivos generales a cubrir y las fuentes de datos disponibles, se indican las necesidades específicas:

- Analizar los viajes de vehículos de alquiler en Nueva York:
  - Por fechas y horas de recogida.
  - Por fechas y horas de devolución.
  - Por distritos y zonas de recogida.
  - Por distritos y zonas entrega.
  - Por tipo de licencia base.

- Analizar el transporte de taxis amarillos en Nueva York:
  - Por fechas y horas de recogida.
  - Por fechas y horas de devolución.
  - Por distritos y zonas de recogida.
  - Por distritos y zonas entrega.
  - Por tipo de tarifa.
  - Por tipo de pago.
  - Por distancia de viaje.
  - Por zonas con congestión.
  - Por número de pasajeros.
  - Por duración del viaje.

Si se tiene en cuenta toda esta información, el sistema podrá responder a múltiples preguntas y, de este modo, satisfacer las necesidades de los usuarios potenciales.

A continuación, se indican de manera específica las cuestiones que, como mínimo, el sistema debe ser capaz de responder:

- Mostrar la evolución de viajes de vehículos de alquiler iniciados en la zona de Times Square. ¿Cuántos viajes en vehículo de alquiler han iniciado viaje en la zona de Times Square durante 2023? Se desea conocer la evolución mes a mes.

- ¿Cuántos viajes en taxis amarillos se han iniciado en la zona del aeropuerto de LaGuardia durante el último trimestre de 2023?

- ¿Cuáles son las diez zonas de Nueva York (*top* diez) con mayor cantidad de viajes iniciados por vehículos de alquiler durante todo el periodo del que se disponen datos?

- Mostrar un listado de zonas de entrega de taxis en Nueva York durante el año 2023, ordenado de mayor a menor por número de pasajeros transportados.

- Calcular el promedio diario de la distancia recorrida en los viajes iniciados (recogidos) por taxis amarillos en octubre de 2023. El resultado se deberá mostrar redondeado a dos decimales.

- Mostrar un listado con el total de viajes finalizados por taxis amarillos en 2024. El listado se deberá mostrar ordenado alfabéticamente por método de pago, y agrupado por método de pago y tipo de tarifa.

- ¿Cuáles son las cinco zonas *service zone* y zona de recogida con menor duración total (*top* cinco) de los viajes en vehículos de alquiler durante el año 2023?

2. Análisis de las fuentes de datos

En este apartado se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato y qué datos deben ser cargados. A continuación, se puede ver un análisis detallado.

- **payment_type.xls**. Contiene información relativa a los diferentes modos de pago ofrecidos a los pasajeros en los taxis de la ciudad de Nueva York. Es un fichero Excel (.xlsx) de dos columnas y con el primer registro para las cabeceras, con la siguiente estructura:

Nombre de campo	Descripción	Tipo	Ejemplo
id	Código del tipo de pago.	Texto	1
Payment_type	Tipo de pago.	Texto	Credit card

Total de registros: seis.

- **fhv_tripdata-001.zip**. Este fichero comprimido con extensión .zip contiene cuatro ficheros planos con extensión .csv (ver volumetría). Cada uno de ellos almacena información de los viajes realizados por los vehículos de alquiler en la ciudad de Nueva York durante el último trimestre de 2023 y enero de 2024. Son ficheros de texto plano de siete columnas, con la coma (,) como delimitador de campos y el primer registro para las cabeceras. Cada uno de este tipo de fichero .csv tiene la siguiente estructura:

Nombre de campo	Descripción	Tipo	Ejemplo
dispatching_base_num	Número de licencia TLC.	Numérico	B00254
pickup_datetime	Fecha y hora de inicio del viaje (activación del taxímetro).	Fecha/hora	2023-10-01 00:24:00
dropoff_datetime	Fecha y hora de final del viaje (desactivación del taxímetro).	Fecha/hora	2023-10-01 00:38:39
PULocationID	ID de la ubicación de recogida.	Numérico	48.0
DOLocationID	ID de la ubicación de entrega.	Numérico	107.0
SR_flag	Indicador (S/N) de si el viaje es compartido.	Texto	
Affiliated_base_number	Número de la base a la que está afiliado el vehículo.	Numérico	B00254

El campo «SR_Flag» indica si el viaje es compartido. Para viajes compartidos, el valor es 1. Para viajes no compartidos, este campo es nulo.

Volumetría

#	Nombre del fichero	Total de registros
1	fhv_tripdata_2023-10.csv	1.628.438
2	fhv_tripdata_2023-11.csv	1.343.846
3	fhv_tripdata_2023-12.csv	1.376.748
4	fhv_tripdata_2024-01.csv	1.290.116
	Total	**5.639.148**

- **yellow_tripdata-001.zip**. Este fichero comprimido con extensión .zip contiene cuatro ficheros planos con extensión .csv (ver volumetría). Cada uno de ellos almacena información de los viajes realizados por los taxis amarillos en la ciudad de Nueva York durante el último trimestre de 2023 y enero de 2024. Son ficheros de texto plano de diecinueve columnas, con la coma (,) como delimitador de campos y el primer registro para las cabeceras. Cada uno de los ficheros .csv de los ficheros comprimidos tiene la siguiente estructura:

Nombre de campo	Descripción	Tipo	Ejemplo
VendorID	Código proveedor de TPEP.	Numérico	1
tpep_pickup_datetime	Fecha y hora de inicio del viaje (activación del taxímetro).	Fecha/hora	2023-10-01 00:16:44
tpep_dropoff_datetime	Fecha y hora de final del viaje (desactivación del taxímetro).	Fecha/hora	2023-10-01 00:16:49
passenger_count	Número de pasajeros.	Numérico	1,0
trip_distance	Distancia del viaje (millas).	Numérico	0,0
RatecodeID	Código de tarifa.	Numérico	1.0
store_and_fwd_flag	Indicador (S/N) de si el viaje almacenado en el taxímetro por falta de Internet.	Texto	N
PULocationID	ID de la ubicación de recogida.	Numérico	168
DOLocationID	ID de la ubicación de entrega.	Numérico	168
payment_type	Tipo de pago.	Numérico	2
fare_amount	Importe del viaje.	Numérico	3,0

Nombre de campo	Descripción	Tipo	Ejemplo
extra	Extras y recargos.	Numérico	1,0
mta_tax	Impuesto MTA de 0,5 $.	Numérico	0,5
tip_amount	Propina.	Numérico	0,0
tolls_amount	Pago por peajes.	Numérico	0,0
improvement_surcharge	Recargo 0,3 $ por mejora.	Numérico	1,0
total_amount	Importe a cobrar.	Numérico	5,5
congestion_surcharge	Recargo por congestión.	Numérico	0,0
airport_fee	Recargo por recogida en aeropuerto.	Numérico	0,0

Notas:

— El campo «VendorID» contiene tres valores posibles: 1 corresponde a Creative Mobile Technologies, LLC; 2 corresponde a VeriFone Inc; 6 corresponde a Others.
— El campo «RatecodeID» cuando viene informado, contiene los siguientes valores: 1 – Standard rate; 2 – JFK; 3 – Newark; 4 – Nassau or Westchester; 5 – Negotiated fare; 6 – Group ride; y 99 – Others.
— El campo «tip_amount» corresponde a las propinas con pago con tarjeta de crédito. Las propinas en efectivo no están incluidas.
— El campo «airport_fee» corresponde a 1,25 $ de recargo por recogida en los aeropuertos LaGuardia y John F. Kennedy.
— El campo «total_amount» se corresponde con el importe total de los viajes. Puede contener valores negativos (posibles reembolsos o ajustes).
— Existen campos «passenger_count», «RatecodeID», «store_and_fwd_flag», «congestion_surcharge» y «airport_fee» que pueden contener valores nulos o faltantes.

Volumetría

#	Nombre del fichero	Total de registros
1	yellow_tripdata_2023-10.csv	3.522.286
2	yellow_tripdata_2023-11.csv	3.339.715
3	yellow_tripdata_2023-12.csv	3.376.567
4	yellow_tripdata_2024-01.csv	2.964.624
	Total	**13.203.191**

- **current_bases.tsv**. Este fichero contiene, en un fichero de formato .tsv, la información correspondiente a las licencias de los vehículos de alquiler que dan servicio en la ciudad de Nueva York. La estructura de este fichero es:

Nombre de campo	Descripción	Tipo	Ejemplo
License Number	Número de licencia.	Numérico	B00056
Entity Name	Nombre base autorizado por TLC.	Texto	Transit Private Car Service INC.
Telephone Number	Número de teléfono.	Texto	7186494100
SHL Endorsed;	Indicador de si es vehículo económico.	Texto	No
Building	Edificio.	Texto	1407
Street	Dirección.	Texto	Rockaway Parkway
City	Ciudad.	Texto	Brooklyn
State	Estado.	Texto	Nueva York
Postcode	Código postal.	Texto	11236
Type of Base	Tipo de base.	Texto	Livery Base
Latitude	Latitud.	Numérico	40.645455
Longitude	Longitud.	Numérico	-73.902663
Date	Fecha de carga.	Texto	03/14/2024
Time	Hora de carga.	Texto	18:00:12
Location	Coordenadas de la ubicación.	Numérico	(40.645455, -73.902663)

Total de registros: 853.

Nota: se ha detectado que la licencia B02920 está **duplicada** en la fuente de datos. Se deberá resolver, pues, en la PR2, ya que los números de licencia son únicos.

- **rate_code.tab .** Contiene la información correspondiente a los diferentes tipos de tarifa que se aplican a los pasajeros al finalizar el trayecto en los viajes en taxi en la ciudad de Nueva York. Es un fichero de texto plano con el tabulador como delimitador de campos.

Nombre de campo	Descripción	Tipo	Ejemplo
code	Código de tarifa.	Numérico	1
rate	Tipo de tarifa.	Texto	Standard rate

Total de registros: seis.

Nota: el fichero no tiene cabeceras.

- **taxi_zone_lookup.csv**. Contiene la información de las ubicaciones de recogida o devolución de pasajeros en los viajes de taxis en la ciudad de Nueva York. Las ubicaciones se refieren a los lugares por los que los taxis realizan los viajes, concretamente a los diferentes distritos, zonas y áreas de la ciudad por los que se mueven. Es un fichero de texto plano, con la coma (,) como delimitador de campos, el primer registro para las cabeceras y con el texto entre comillas dobles para los campos de tipo texto.

Nombre de campo	Descripción	Tipo	Ejemplo
LocationID	Código de zona.	Numérico	1
Borough	Distrito.	Texto	EWR
Zone	Nombre de zona.	Texto	Newark Airport
service_zone	Código de identificación para el estado.	Texto	EWR

Total de registros: 265.

Nota: se incluyen dos registros para el tratamiento de registros con zonas de valores desconocidos.

2.1. Estimación de volumetría

En los proyectos de diseño de la factoría de información corporativa, existe una primera fase en la que se realiza una carga inicial y, *a posteriori*, una segunda fase para realizar las cargas incrementales de los datos nuevos que van llegando.

Una posible estimación del volumen de datos del almacén para la carga inicial de los datos sería la siguiente:

Fichero	Registros	Valores	Datos
Payment_type.xls	6	2	12
fhv_tripdata_YYYY_MM.csv (4 ficheros)	5.639.148	7	39.474.036
yellow_tripdata_YYYY_MM.csv (4 ficheros)	13.203.191	19	250.860.629
Rate_code.tab	7	2	14
taxi_zone_lookup.csv	265	4	1.060
Current_bases.tsv	853	15	12.795
Total			**290.348.546**

3. Análisis funcional

A continuación, se propone el tipo de arquitectura para la factoría de información que mejor se adecua al proyecto. Para ello, se consideran los requisitos funcionales y se establece la prioridad entre exigible (E) o deseable (D). En el contexto de esta actividad, los requisitos exigibles son aquellos que pide el enunciado, mientras que los deseables son los que complementan la actividad.

Además, en términos de la escala de prioridades, se asigna una prioridad de 1 a 3, en la que 1 es completamente prioritario para la actividad y 3 es no prioritario.

A continuación, se describen los requisitos funcionales para el diseño de una factoría de información para la organización, teniendo en cuenta las consideraciones del enunciado:

#	Requisito	Prioridad	Exigible o deseable
1	Se extraerá de manera adecuada la información de las fuentes de datos.	1	E
2	Se creará un almacén de datos.	1	E
3	Se cargará la información para realizar el análisis de registros de viajes de los taxis y vehículos de alquiler en NY.	1	E
4	Se creará un modelo OLAP para consultas multidimensionales de los usuarios.	2	E
5	Se crearán los informes estáticos solicitados.	2	E
6	Se redactará un manual de carga de datos inicial e incremental.	3	D

Cabe comentar que, en un caso genérico real, se pueden encontrar también otros requisitos funcionales, como los que se muestran a continuación:

- Análisis de viabilidad y análisis de riesgos.
- Creación de procesos de calidad de datos.
- Creación de *data marts* (si se analizan otros servicios).
- Creación de procesos de cargas incrementales.
- Creación de un repositorio de metadatos de gestión del almacén de datos, así como de los procesos de ETL, que permitan realizar la trazabilidad a lo largo del ciclo de vida de los datos.

Asimismo, dado que estos sistemas frecuentemente forman parte de la implementación de un sistema de inteligencia de negocio, la lista de requisitos funcionales sería mucho mayor (como puede ser la administración de seguridad, en cuanto a datos y usuarios).

En términos de arquitectura funcional, existen los siguientes elementos:

- Las fuentes de datos de las que se dispone son las siguientes:
  — un fichero en formato XLS;
  — un fichero en formato TAB;
  — un fichero en formato TSV;
  — nueve (9) ficheros en formato CSV.

- La arquitectura de la factoría de información puede estar formada por varios elementos alojados en la misma máquina:
  — *Staging area* (opcional): en el caso de tener múltiples fuentes (ficheros, bases de datos, servicios RSS, etc.), es conveniente cargarlas para consolidar la información en una estructura de carga intermedia, que puede ser creada en la misma base de datos. Esta área del DW también puede servir para entender, simplificar y consolidar los procesos de ETL.
  — *Data mart* para el análisis de los registros de transportes de la New York City Taxi and Limousine Commission. Al centrarnos en una única área temática, es más correcto considerar que se está creando un *data mart* en lugar de un almacén de datos corporativo.
  — MOLAP: a partir de la información del *data mart,* se creará un cubo multidimensional.

Según lo comentado anteriormente, se podría elegir entre dos diseños para la arquitectura funcional. Por un lado, tenemos una arquitectura funcional que usa un área intermedia (*staging area*) y que se crearía dentro de la misma base de datos, cuyos objetos se identificarán con un prefijo en los nombres. La siguiente figura resume los elementos de la arquitectura necesarios para esta actividad:

Por otro lado, también sería correcto utilizar una arquitectura sin área intermedia (*staging area*) que identifique las tablas intermedias en el *data mart* con un prefijo en el nombre, como, por ejemplo, «IN_*nombre_tabla_intermedia*».

En esta solución, se propone un diseño que utiliza un área intermedia (*staging area*) que, al tener una única base de datos, simularemos utilizando prefijos en los nombres. Concretamente, las tablas intermedias se identificarán respecto al resto de tablas con el prefijo «IN_» en sus nombres. Por ejemplo, «IN_fhv_data».

4. Diseño del modelo conceptual, lógico y físico del almacén de datos

4.1. Diseño conceptual

Para el correcto desarrollo del *data warehouse* (DW), es preciso definir los hechos (*facts*), las dimensiones de análisis (*dimensions*), las métricas y los atributos que permitan tener la granularidad suficiente para la presentación de los resultados. Estos se han definido en el análisis de requisitos y de las fuentes de datos.

El análisis de requisitos, junto al análisis de fuentes de los viajes en taxi y vehículos de alquiler de la New York City Taxi and Limousine Commission, nos permite determinar que los hechos que queremos analizar son:

- **Transporte en vehículos de alquiler en Nueva York**.
- **Transporte de taxis en Nueva York**.

El análisis de estos hechos permite dar respuesta a las necesidades principales de los usuarios potenciales, permitiendo comprender y mejorar el sistema de transporte de taxis y vehículos de alquiler en Nueva York desde la perspectiva de la demanda del servicio, con el fin de promover prácticas de transporte más sostenibles, buscando reducir la congestión en la ciudad y lograr una mayor ocupación de los taxis mediante una comprensión mejor de los patrones de uso.

El **análisis del transporte en vehículos de alquiler en Nueva York** determina el diseño de la primera tabla de hechos:

Tabla de hechos	Descripción
FACT_FHV_TRIP	Análisis de los viajes en vehículos de alquiler en NY.

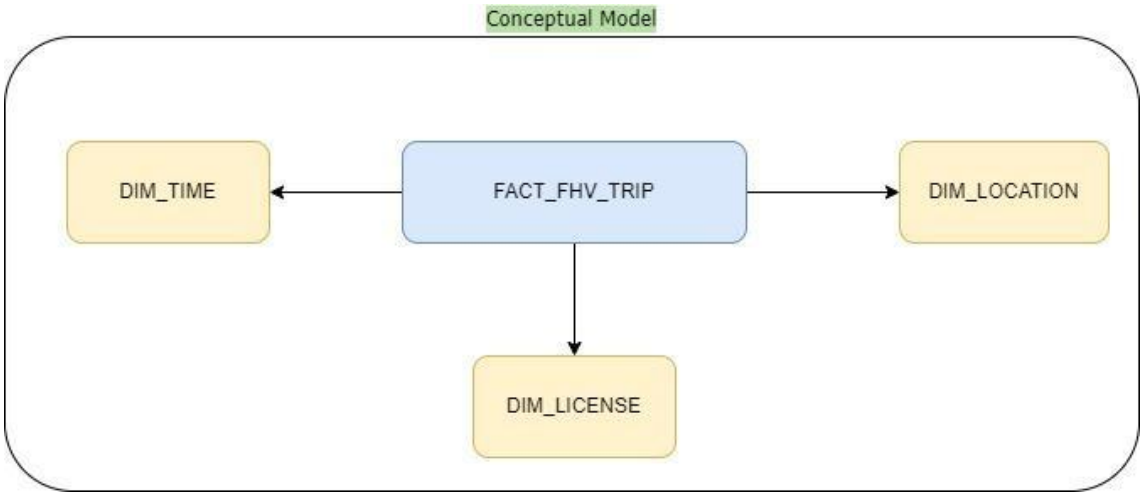
En la siguiente tabla, se indican las métricas de la tabla de hechos FACT_FHV_TRIP.

Métricas	Descripción
fhv_trip	Viaje en vehículo de alquiler.
duration	Duración del viaje.

La métrica de esta tabla de hechos podrá analizarse desde las diferentes perspectivas, a partir de las siguientes dimensiones:

Dimensiones	Descripción
DIM_TIME	Fecha y hora del viaje en taxi.
DIM_LOCATION	Zonas de NY.
DIM_LICENSE	Licencias TLC para vehículos de alquiler.

El diseño conceptual para esta tabla de hechos (FACT_FHV_TRIP) y sus dimensiones con un **diseño en estrella** es el siguiente:



Este diseño considera las siguientes fuentes de datos:

- current_base.tsv.
- taxi_zone_lookup.csv.
- fhv_tripdata-001.zip.

El **análisis del transporte de taxis en Nueva York** determina el diseño de la segunda tabla de hechos, como se puede observar a continuación:

Tabla de hechos	Descripción
FACT_NYTAXI_TRIP	Análisis del transporte de taxi en NY.

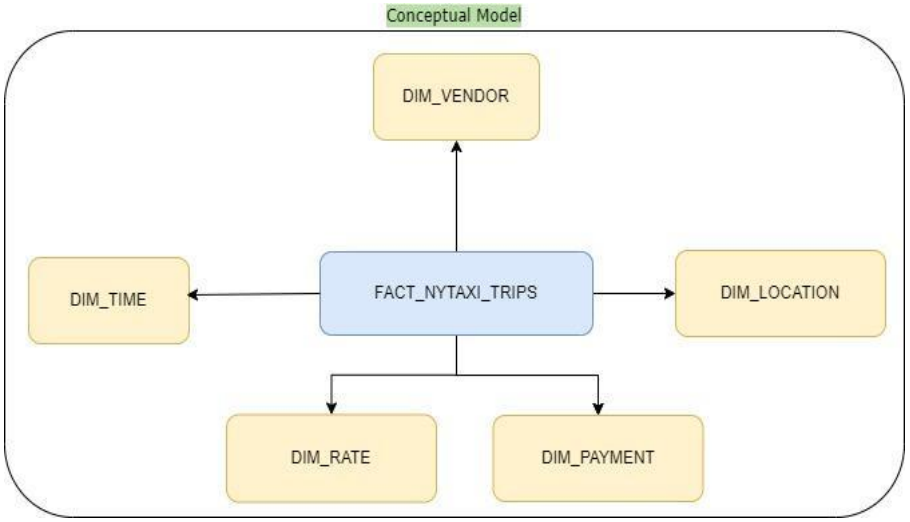
En la siguiente tabla se indican las métricas de la tabla de hechos FACT_NYTAXI_TRIP.

Métricas	Descripción
taxi_trip	Viaje en taxi.
duration	Duración del viaje.
passenger	Número de pasajeros.
distance	Distancia recorrida.
fare_amount	Importe del viaje.
total_amount	Importe total a cobrar por el viaje.

Las métricas de esta tabla de hechos podrán analizarse desde las diferentes perspectivas, a partir de las siguientes dimensiones:

Dimensiones	Descripción
DIM_TIME	Fecha y hora del servicio.
DIM_LOCATION	Zonas de NY.
DIM_RATE	Tipos de tarifa.
DIM_PAYMENT	Formas de pago.
DIM_VENDOR	Proveedor.

El diseño conceptual para esta tabla de hechos (FACT_NYTAXI_TRIP) y sus dimensiones con un **diseño en estrella** es el siguiente:



Este modelo considera todas las fuentes de datos proporcionadas:

- Payment_type.xls.txt.
- Rate_code.tab.
- yellow_tripdata-001.zip, yellow_tripdata-002.zip, yellow_tripdata-003.zip.
- taxi_zone_lookup.csv.

4.2. Diseño lógico

Una vez obtenido el modelo conceptual del almacén de datos del análisis de los registros de viajes de la New York City Taxi and Limousine Commission, pasamos a realizar su diseño lógico.

Teniendo en cuenta que vamos a utilizar tecnología relacional y que el modelo de datos va a ser el multidimensional, pasamos a describir el modelo lógico en términos de tablas, atributos, y claves primarias y foráneas.

El primer paso para realizar el diseño lógico implica identificar las métricas de cada tabla de hechos definida a título conceptual.

En el caso de la tabla de hechos FACT_FHV_TRIP, hemos identificado sus métricas:

Tabla de hechos	Métricas
FACT_FHV_TRIP	taxi_trip, duration

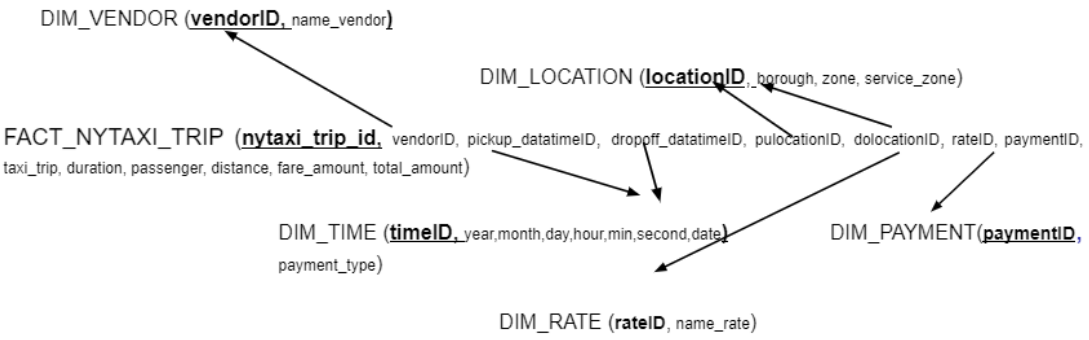
El siguiente paso para realizar el diseño lógico es crear una tabla donde se detallan los atributos de las dimensiones que nos permitan la evaluación del hecho. Específicamente, los atributos de las dimensiones de la tabla de hechos FACT_FHV_TRIP se muestran en la siguiente tabla:

Dimensiones	Atributos
DIM_TIME	timeID, year, month, day, hour, min, second, date
DIM_LOCATION	locationID, borough, zone, service_zone
DIM_LICENSE	licenceID, license_num, entity_name, telephone, SHL_endorsed, building, street, city, state, postcode, type_base, date, time, location

Nota: los atributos «latitude» y «lenght» de la dimensión DIM_LICENSE son coordenadas geoespaciales, y necesitan un tratamiento específico para su carga y visualización. Con

el fin de simplificar la realización del caso práctico, se utilizará solo el atributo «location» que contenga ambas coordenadas.

La representación visual del diagrama de estrella del modelo lógico para el análisis de los viajes de vehículos de alquiler sería la siguiente:



En el caso de la tabla de hechos FACT_NYTAXI_TRIP, hemos identificado sus métricas:

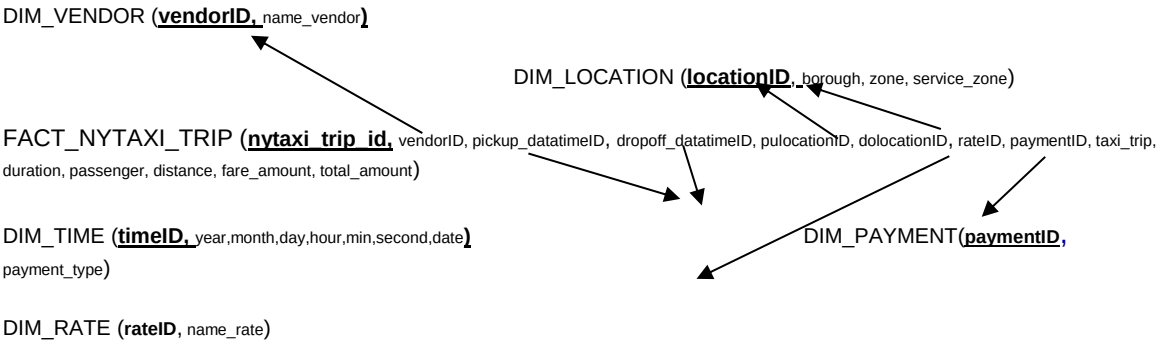
Tabla de hechos	Métricas
FACT_NYTAXI_TRIP	taxi_trip, duration, passenger, distance, fare_amount, total_amount

Nota: al no ser objeto, en este caso práctico, el análisis pormenorizado de los suplementos y recargos aplicados para calcular el importe total a pagar por el viaje en taxi, como impuestos y suplementos, dicha información se transforma en indicadores tipo S/N y serán atributos de las tablas de hechos FACT_NYTAXI_TRIP.

En la siguiente tabla se detallan los atributos de las dimensiones de la tabla de hechos FACT_NYTAXI_TRIP:

Dimensiones	Atributos
DIM_TIME	timeID, year, month, day, hour, min, second, date
DIM_LOCATION	locationID, borough, zone, service_zone
DIM_RATE	rateID, name_rate
DIM_PAYMENT	paymentID, payment_type
DIM_VENDOR	vendorID, name_vendor

En la siguiente imagen se muestra la representación visual del modelo lógico para el análisis de los viajes en taxis amarillos en la ciudad de Nueva York:



4.3. Diseño físico

Tras determinar tablas de hechos, dimensiones, métricas y atributos que existen en el modelo lógico, podremos realizar el diseño físico, lo que implica una implementación del modelo lógico en términos del sistema gestor de bases de datos elegido.

Además, para el correcto diseño físico del almacén, se deben tener en cuenta los siguientes aspectos:

- El **sistema gestor de bases de datos** con el que vamos a trabajar implementará de una manera concreta los distintos elementos del modelo lógico.
- El ajuste del **diseño físico** a las particularidades del sistema gestor de bases de datos, para obtener un buen rendimiento en el procesamiento de consultas.
- La **revisión periódica del diseño físico inicial**, para validar que continúa dando respuesta a las necesidades del cliente.

Puesto que utilizaremos SQL Server, y dado que se trata de un sistema gestor de bases de datos relacional, en esta etapa deberemos tener en cuenta, entre otras cosas, la implementación de las claves primarias y foráneas, tanto en las tablas de hechos como en las de dimensiones.

En este paso, también es necesario tener en cuenta el tamaño adecuado de los atributos (por ejemplo, la longitud de los campos de textos o si los valores numéricos contienen decimales).

Para ello, vamos a detallar los tipos de datos de cada campo que forman parte de las tablas de hechos y dimensiones.

Dado que el modelo de almacén está compuesto por más de una tabla de hechos (*facts*), también se deben revisar las dimensiones que se han definido en el diseño conceptual y en el lógico de cada *fact*, y aplicar una visión conjunta del modelo para determinar si, en el modelo del almacén, existirán dimensiones comunes o conformadas, como DIM_TIME y DIM_LOCATION, y así simplificar el modelo final y lograr un rendimiento óptimo en la ejecución de los análisis.

Como es lógico, primero se crean las tablas de dimensiones y, posteriormente, las tablas de hechos, ya que contienen atributos referenciales a aquellas. De esta manera, se crea cada una de las tablas del almacén de datos.

4.4. Dimensiones

Las dimensiones del modelo podrán estar referenciadas en las tablas de hechos utilizando sus claves primarias o, en inglés, *primary keys* (PK). El modelo físico de las dimensiones es el siguiente:

- **DIM_TIME**. Corresponde a la dimensión temporal para el análisis de la información. Esta dimensión permite analizar los hechos desde un punto de vista temporal, como el análisis de tendencias o los evolutivos. Este tipo de análisis no se puede realizar si el modelo no cuenta con una dimensión de tiempo.

Nombre de campo	Tipo	Tamaño	Ejemplo
timeID (PK)	Numérico	8	1
year	Numérico	4	2022
month	Numérico	2	12
day	Numérico	2	01
hour	Numérico	2	00
min	Numérico	2	30
second	Numérico	2	00
date	Fecha/hora	10	2022-12-01 00:30:00

La dimensión conformada DIM_TIME se utilizará para analizar los momentos de recogida y entrega en los viajes de vehículos de alquiler, y los viajes del servicio de taxi.

- **DIM_VENDOR**. Contiene los datos de tipos de vendedor para los taxis de NY.

Nombre de campo	Tipo	Tamaño	Ejemplo
vendorID (PK)	Numérico	2	1
name_vendor	Texto	100	Creative Mobile Technologies, LLC

- **DIM_LOCATION**. Contiene los datos de los distritos y las zonas de Nueva York.

Nombre de campo	Tipo	Tamaño	Ejemplo
locationID (PK)	Numérico	4	15
borough	Texto	50	Queens
zone	Texto	100	Bay Terrace / Fort Totten
service_zone	Texto	100	Boro zone

- **DIM_RATE**. Contiene los datos de los diferentes tipos de tarifa para pagar los taxis en Nueva York.

Nombre de campo	Tipo	Tamaño	Ejemplo
rateID (PK)	Numérico	2	1
name_rate	Texto	50	Standard rate

- **DIM_PAYMENT**. Contiene los datos de las diferentes formas de pago del transporte de taxi en Nueva York.

Nombre de campo	Tipo	Tamaño	Ejemplo
paymentID (PK)	Numérico	2	1
Payment_type	Texto	100	Credit card

- **DIM_LICENSE**. Contiene los datos relativos a las licencias de los vehículos de alquiler que dan servicio en la ciudad de Nueva York.

Nombre de campo	Tipo	Tamaño	Ejemplo
licenseID (PK)	Numérico	6	1
license_num	Texto	6	B00056
entity_name	Texto	300	Transit Private Car Service INC.
telephone	Texto	12	7186494100
SHL_endorsed	Texto	3	No

Nombre de campo	Tipo	Tamaño	Ejemplo
building	Texto	10	1407
street	Texto	150	Rockaway Parkway
city	Texto	50	Brooklyn
state	Texto	50	NY
postcode	Texto	10	11236
type_base	Texto	50	Livery Base
date	Fecha	10	03/14/2024
time	Hora	8	18:00:12
location	Texto	50	(40.645455, -73.902663)

Nota: los atributos «latitude» y «lenght» de la dimensión DIM_LICENSE son coordenadas geoespaciales, y necesitan un tratamiento específico para su carga y visualización. Con el fin de simplificar la realización del caso práctico, se utilizará solo el atributo «location» de tipo texto que contenga ambas coordenadas.

4.5. Tablas de hechos

La composición del modelo físico de las tablas de hechos consistirá en la creación de tablas, cuyos campos serán métricas, atributos y atributos referenciales definidos en el modelo conceptual y en el modelo lógico. Para crear los atributos referenciales en las tablas de hechos, se definen como *claves foráneas* las primarias de las dimensiones con las que están relacionadas, siguiendo el diagrama en estrella definido.

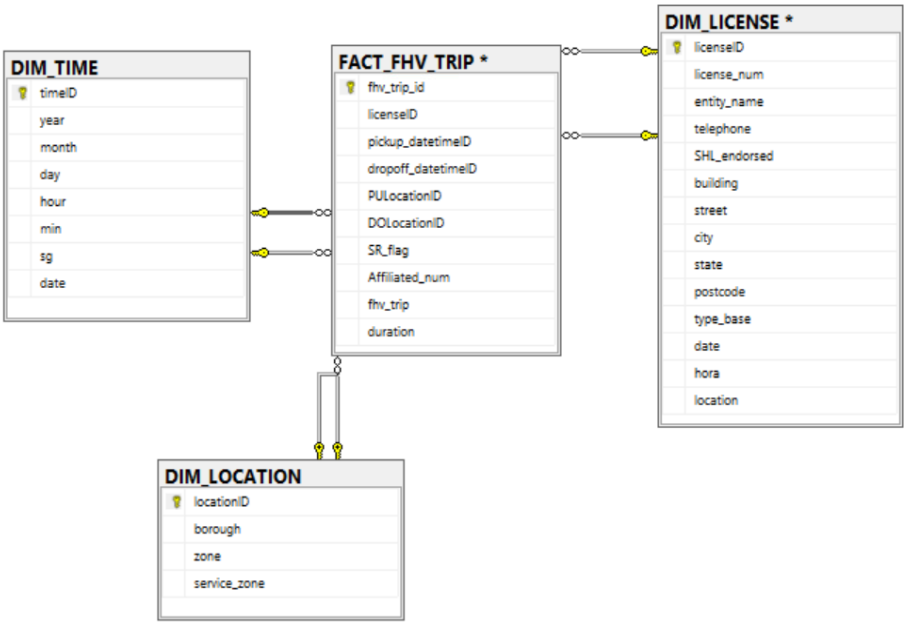
El modelo físico de las tablas de hechos del almacén de datos para el análisis de los registros de viajes de la New York City Taxi and Limousine Commission está compuesto de las siguientes tablas:

- **FACT_FHV_TRIP**. Es la tabla física que contendrá la información que permitirá realizar el análisis de los registros de viajes de la New York City Taxi and Limousine Commission. Tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
fhv_trip_id (PK)	Numérico	12	1
licenseID (FK)	Numérico	8	254

Nombre campo	Tipo	Tamaño	Ejemplo
pickup_datetimeID (FK)	Numérico	8	2022-12-01 00:30:00
dropoff_datetimeID (FK)	Numérico	8	2022-12-01 00:48:00
PULocationID (FK)	Numérico	4	170
DOLocationID (FK)	Numérico	4	237
SR_flag	Texto	3	No
Affiliated_num (FK)	Numérico	8	254
fhv_trip	Numérico	10	1
duration	Numérico	10	18

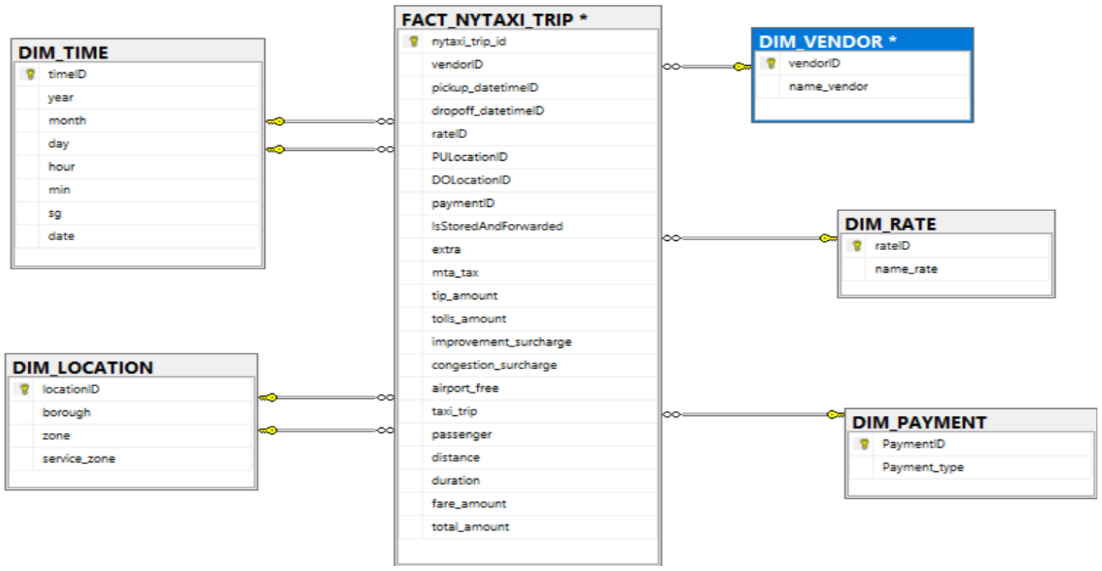
En la siguiente imagen se muestra una posible implementación del diseño del modelo físico para la tabla de hechos FACT_FHV_TRIP:



- **FACT_NYTAXI_TRIP**. Es la tabla física que contendrá la información para realizar el análisis de los registros de viajes de la New York City Taxi and Limousine Commission. Tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
nytaxi_trip_id (PK)	Numérico	15	1
vendorID (FK)	Numérico	2	1
pickup_datetimeID (FK)	Numérico	18	2022-12-01 00:37:35
dropoff_datetimeID (FK)	Numérico	18	2022-12-01 00:47:35
rateID (FK)	Numérico	2	1
PULocationID (FK)	Numérico	4	170
DOLocationID (FK)	Numérico	4	257
paymentID (FK)	Numérico	2	1
IsStoredAndForwarded	Texto	1	S
extra	Texto	1	N
mta_tax	Texto	1	S
tip_amount	Texto	1	S
tolls_amount	Texto	1	N
improvement_surcharge	Texto	1	S
congestion_surcharge	Texto	1	S
airport_free	Texto	1	N
taxi_trip	Numérico	15	1
passenger	Numérico	15	1
distance	Numérico	15	2
duration	Numérico	15	123
fare_amount	Numérico	15	8,4
total_amount	Numérico	15	15,4

En la siguiente imagen se muestra una posible implementación del diseño del modelo físico para la tabla de hechos FACT_NYTAXI_TRIP:


DIM_TIME
timeID
year
month
day
hour
min
sg
date
FACT_NYTAXI_TRIP *
nytaxi_trip_id
vendorID
pickup_datetimeID
dropoff_datetimeID
rateID
PULocationID
DOLocationID
paymentID
IsStoredAndForwarded
extra
mta_tax
tip_amount
tolls_amount
improvement_surcharge
congestion_surcharge
airport_free
taxi_trip
passenger
distance
duration
fare_amount
total_amount
DIM_VENDOR *
vendorID
name_vendor
DIM_RATE
rateID
name_rate
DIM_LOCATION
locationID
borough
zone
service_zone
DIM_PAYMENT
PaymentID
Payment_type