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

PR1 – Análisis y diseño del *data warehouse*

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

Índice

1. Presentación

La PR1 consiste en el análisis y el diseño multidimensional de un almacén de datos de los viajes de la New York City Taxi and Limousine Commission (NY taxis).

Para llevar a cabo su desarrollo, se ha publicado la siguiente información:

- **Caso práctico** (22510_Enunciado_PR.pdf). Documento con la información del caso de uso para desarrollar un almacén de datos para el análisis de los registros de los viajes de la New York City Taxi and Limousine Commission (NY taxis). En el documento se describen el contexto, los usuarios potenciales y las fuentes de datos. También se realiza una descripción del enunciado de la práctica con cada una de las fases para construir el almacén de datos, así como los programas que se utilizarán y la bibliografía.
- **Fuentes de datos** (fuentes.zip). Fichero comprimido con las fuentes de datos disponibles para desarrollar el caso práctico.

1.1. Descripción

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiante deberá diseñar y proponer un almacén de datos que ofrezca un apoyo al análisis de los viajes de la New York City Taxi and Limousine Commission (NY taxis).

Para ello, **se pide** que se lleven a cabo los siguientes puntos:

- **Análisis de los requisitos**, que incluya las preguntas a las que el sistema debe dar respuesta. Se debe definir un mínimo de cinco preguntas y un máximo de diez.
- **Análisis de todas las fuentes de datos** proporcionadas. Por cada fuente, se realizará una tabla con la identificación de campos, descripciones, tipos de campo y un ejemplo. Adicionalmente, completará el análisis mediante observaciones que permitan identificar las fuentes y los campos que incluyan listas de valores.
- **Análisis funcional** que, por un lado, incluya los requisitos funcionales para el diseño de una factoría de información, estableciendo la prioridad entre exigible (E) o deseable (D), y que, por otro, proponga el tipo de arquitectura para la factoría de información que mejor se adecue al proyecto.
- **Diseño del modelo conceptual, lógico y físico** del almacén de datos. Se deben identificar y diseñar las tablas de hechos (*facts*), las dimensiones del análisis (*dimensions*) y los atributos que permitan tener la granularidad suficiente para implementar los requisitos.

El documento de la solución de esta actividad (PR1) debe incluir una descripción y una justificación de todas las acciones realizadas en los puntos anteriormente enumerados.

1.2. Criterios de evaluación

La nota final se calculará a partir de la suma de los siguientes apartados:

- Análisis de los requisitos (10 %).
- Análisis de fuentes de datos (15 %).
- Análisis funcional (15 %).
- Diseño conceptual (20 %).
- Diseño lógico (20 %).
- Diseño físico (20 %).

1.3. Guía de muestra

A continuación, encontraréis una guía creada con el fin de ayudar a alcanzar los objetivos planteados de la PR1. Esta guía contiene cada uno de los apartados solicitados en la PR1 sobre uno de los hechos del contexto de nuestro caso de uso.

Importante: la guía no está completa y el estudiantado deberá completarla para cubrir los objetivos marcados para la PR1.

2. 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, se debe ser previsor y pensar más allá de las necesidades actuales para cubrir las futuras.

La necesidad principal de la organización es disponer de la información integrada para su análisis y su 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, del siguiente objetivo: analizar los viajes de los vehículos de alquiler en Nueva York.

A continuación, se indica la información necesaria identificada para analizar este objetivo desde diferentes perspectivas:

- por fechas/horas de recogida;
- por fechas/horas de devolución;
- por distritos y zonas de recogida;
- por distritos y zonas entrega;
- etc.

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

Como ejemplo, se indican de manera específica algunas preguntas que, como mínimo, el sistema debe ser capaz de responder:

- Evolutivo de viajes de vehículos de alquiler iniciados en la zona de Times Square.
- Evolutivo de viajes de taxis amarillos registrados en la zona del aeropuerto LaGuardia durante el primer trimestre de 2021

El estudiantado, en este punto, deberá completar la definición de los requisitos. Para ello, tendrá que completar los objetivos descritos en la guía con otros que permitan cubrir la necesidad principal de la organización encargada del análisis de los viajes de los vehículos de alquiler en Nueva York. Asimismo, deberá proponer otras perspectivas para analizar dichos objetivos y plantear otras preguntas a las que el sistema deberá responder.

3. 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 muestra un ejemplo:

- **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 de la ciudad de Nueva York desde diciembre de 2022 a enero de 2024. Es un fichero de texto plano de siete columnas con la coma (,) como delimitador de campos y 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 base a la que está afiliado el vehículo.	Numérico	B00254

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

Volumetría

#	Nombre del fichero	Total 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

Estimación de volumetría

En los proyectos de diseño de 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
fhv_tripdata-001.zip	5.639.148	7	39.474.0366
….	…	…	…
Total			**98**

En este punto, el estudiantado deberá completar la definición del resto de las fuentes proporcionadas, así como la estimación de volumetría para la carga inicial de todas las fuentes de datos proporcionadas.

4. 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 se piden en el enunciado, mientras que los deseables son los que complementan la actividad.

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

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

#	Requisito	Prioridad	Exigible/deseable
1	Se extraerá de forma adecuada la información de las fuentes de datos.	1	E
2	Se creará un almacén de datos.	1	E
…	…		

El estudiantado, en este punto, deberá:

- Completar la tabla de los requisitos funcionales y asignar las prioridades.
- Elegir la arquitectura funcional que considere más adecuada para el caso de estudio (elegir una arquitectura funcional de todas las estudiadas en los módulos teóricos de la asignatura y describir sus elementos).
- Identificar otros requisitos funcionales.

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

5.1. Diseño conceptual

Para el correcto desarrollo del 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 objetivos. Estos se han definido en el análisis de los requisitos y de las fuentes de datos.

Del análisis de las fuentes de datos y de los requisitos iniciales, se puede determinar que uno de los hechos que se deben considerar es el siguiente:

- Transporte en vehículos de alquiler 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 y con el fin de promover prácticas de transporte más sostenibles, buscando reducir la congestión en la ciudad y una mayor ocupación de los taxis mediante una mejor comprensión 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 Nueva York.

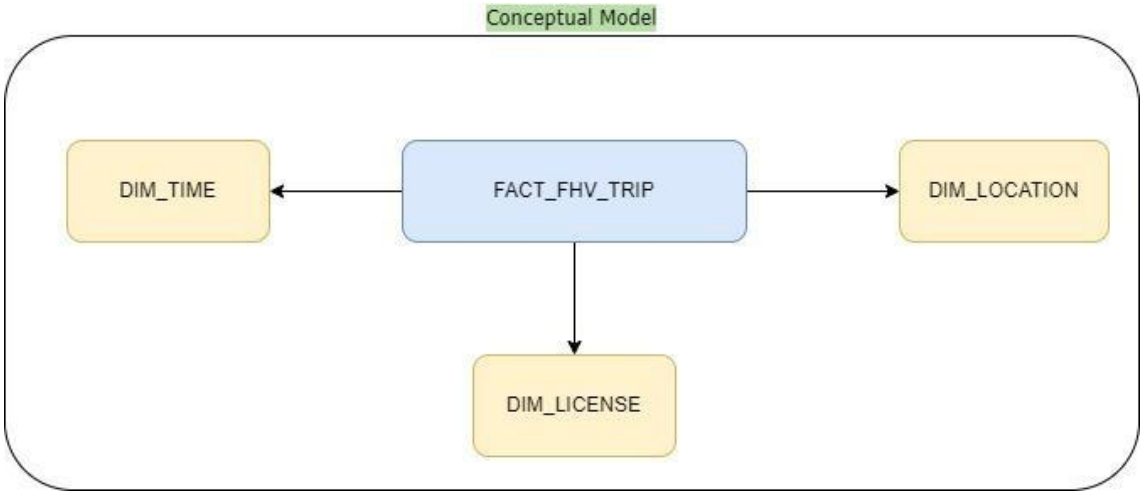
En la siguiente tabla, se indican las métricas de la tabla de hechos FACT_FHV_TRIP.

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 Nueva York.
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 estudiantado deberá completar el diseño conceptual para este caso de uso**.**

5.2. Diseño lógico

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

A continuación se muestra, a modo de ejemplo, la tabla con las métricas identificadas en el diseño conceptual de la tabla de hechos FACT_FHV_TRIP:

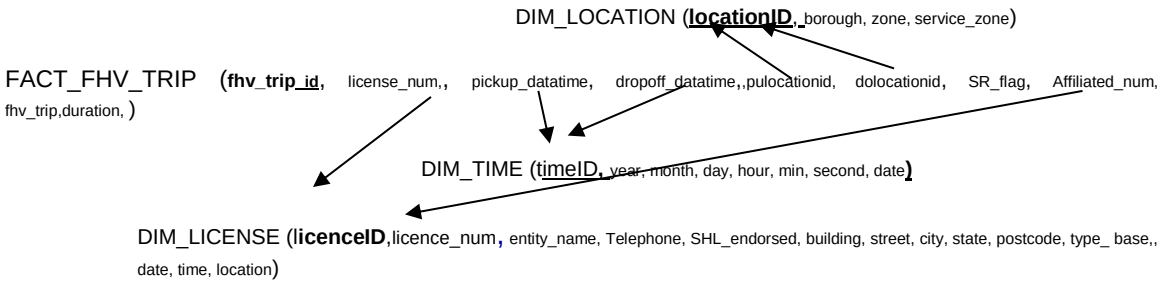
Tabla de hechos	Métricas
FACT_FHV_TRIP	taxi_trip, duration

Después, se detallan los atributos descriptores de las dimensiones de la tabla de hechos FACT_FHV_TRIP:

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:



El estudiante deberá completar el diseño lógico para este caso de uso.

5.3. Diseño físico

Una vez determinadas las tablas de hechos, las dimensiones, las métricas y los atributos que existen en el modelo lógico, podemos pasar a realizar el diseño físico, lo que significa obtener una implementación del modelo lógico en términos del sistema gestor de bases de datos elegido.

En este punto, el estudiantado debería reflexionar acerca de las consideraciones previas que hay que tener en cuenta para el diseño físico.

En esta guía indicamos, como ejemplo, algunos de los aspectos que hay que tener en cuenta.

Sistema gestor de bases de datos

El sistema gestor de bases de datos con el que vamos a trabajar implementará de una manera concreta los distintos elementos del modelo lógico.

Nota: el estudiantado deberá completar el resto de los aspectos que hay que tener en cuenta para el correcto diseño físico para este caso de uso.

Puesto que utilizaremos SQL Server, y este es 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 en las tablas de hechos y 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 y así simplificar el modelo final y conseguir un rendimiento óptimo en la ejecución de los análisis.

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 una de las dimensiones identificadas es el siguiente:

- **DIM_LOCATION**: contiene los datos de los distritos y 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

- **...**

Nota: el estudiantado deberá completar la definición de las dimensiones del diseño físico de la tabla de hechos «FACT_FHV_TRIP», así como del resto de las tablas de hechos identificadas para este caso de uso.

Tablas de hechos

La composición del modelo físico de las tablas de hechos consistirá en la creación de tablas cuyos campos serán las métricas, los atributos y los 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 de estrella definido.

El modelo físico de la tabla de hechos **«FACT_FHV_TRIP»** tendrá los siguientes campos:

Nombre campo	Tipo	Tamaño	Ejemplo
fhv_trip_id (PK)	Numérico	12	1
licenseID(FK)	Numérico	8	254
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	2	No
Affiliated_num	Numérico	8	254
fhv_trip	Numérico	10	1
duration	Numérico	10	18

Nota: el estudiantado deberá completar el diagrama de E/R para el modelo físico de la tabla de hechos «FACT_FHV_TRIP» como la del resto de las tablas de hecho identificadas y sus dimensiones.

Recordad que el objetivo de la guía de muestra es ayudar al estudiante a realizar la entrega completa de la actividad PR1.

6. Formato y fecha de entrega

La entrega de esta actividad debe realizarse a través del enlace de Entrega PR1 del aula, enviando un único archivo en formato Word o PDF. El nombre del archivo debe ser la composición del nombre de usuario y «_BDA_PR1».

Por ejemplo, si el nombre de usuario es «bantich», el nombre del archivo debe ser «bantich_BDA_PR1.pdf».

La fecha máxima de entrega es el 18/11/2024 a las 23:59 horas.