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

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

Índice

Introducción

El caso **«Almacén de datos para el análisis de los registros de viajes de la New York City Taxi and Limousine Commission»** ha sido creado para practicar el diseño y la implementación del almacén de datos como sistema de almacenamiento para el análisis de datos.

El diseño, el desarrollo y la implantación de un sistema de *data warehouse* (DW) en cualquier organización supone llevar a cabo un proyecto que puede durar meses, o incluso años, en función del alcance del proyecto, de la naturaleza y del grado de madurez de la organización, así como de la participación de equipos multidisciplinares que van implementando diferentes proyectos en un proceso de mejora continua del almacén.

El objetivo de este caso no es desarrollar un almacén de datos que dé respuesta a todas las necesidades, sino entender y utilizar las metodologías para desarrollar este tipo de proyectos en un contexto real, pasando por todas las fases que comprenden los proyectos de esta tipología.

La solución del caso práctico se llevará a cabo en tres entregas:

1) **Análisis, diseño e implementación**. Consiste en desarrollar e implementar un almacén de datos que permita la gestión de la información disponible.

2) **Carga**. Implica diseñar e implementar los procesos de carga de datos necesarios para disponer de información en el almacén de datos implementado en la etapa anterior.

3) **Explotación**. El objetivo es llevar a cabo un análisis multidimensional de dicha información que permita su explotación.

Con el fin de poder desarrollar un proyecto lo más específico posible, el estudiantado tendrá que afrontar el reto de desarrollar un almacén de datos que describa parte de los servicios que se pueden ofrecer, basándose en las fuentes de datos del caso y que formarían parte de un sistema real.

A partir del contexto que se describe a continuación, se adquirirá un conocimiento básico del entorno tecnológico, se comprenderán las necesidades existentes y se definirá la propuesta más adecuada que responda a estas necesidades.

Mediante el desarrollo del caso, el estudiantado se va a encontrar con los problemas, las dudas y las dificultades que se plantean en un proyecto de características similares.

Contexto

Nos encontramos en un momento de cambio y transformación, facilitado por una infraestructura digital que impulsa procesos de innovación en materia de movilidades, al mismo tiempo que los actores tradicionales de la movilidad (taxi) muestran una gran resiliencia ante la presión de los nuevos actores (VTC, FHV, etc.), sin olvidar el crecimiento de otras formas de movilidad individual y privada, como bicicletas y patinetes.

Las grandes ciudades, a través de sus organismos competentes en este ámbito, (New York City Taxi and Limousine Commission – TLC, Instituto Metropolitano del Taxi – ImeT, Transport for London – TfL, etc.) se han visto obligadas a revisar sus estrategias y su regulación, para dar cabida a un nuevo modelo de negocio creado por las plataformas de movilidad y basado en el dato como fuente principal de creación de valor.

Para el caso concreto de TLC (https://www.nyc.gov/site/tlc/index.page) se dispone de un equipo de investigadores de políticas propias que utiliza datos generados por los más de 200.000 titulares de licencias (licenciatarios que realizan aproximadamente 1.000.000 de viajes cada día), con el fin de:

- Observar las tendencias cambiantes del sector para informar a la propia comisión.
- Impulsar programas de mejora de la experiencia de los pasajeros con base en información de la fecha y la hora de recogida, la identificación de la ubicación de la zona de taxi, la tarifa aplicada, el tiempo de espera, el número de ocupantes, etc.
- Publicar los datos como *open data* para su libre análisis.

El proceso de planificación y programación de los recursos es parte de su actividad organizativa, orientada a gestionar o asignar los recursos disponibles, considerando preguntas del tipo: ¿cuál es la disponibilidad de cada licenciatario?, o ¿cuántos vehículos se necesitarán para llevar a cabo todos los servicios diarios?

Estas serán algunas de las preguntas que se podrán resolver después del despliegue de este almacén de datos.

Usuarios potenciales

Como fase inicial del diseño del sistema de **análisis de los registros de viajes de la New York City Taxi and Limousine Commission**, se identifican los requisitos de los usuarios potenciales, para que el sistema los pueda tener en cuenta para dar respuesta a sus necesidades y generar información que resulte útil.

Los usuarios finales del sistema serán los siguientes:

- **Licenciatarios**. Con los datos que ofrece el almacén, los titulares de licencias pueden realizar comparativas, tomar decisiones comerciales e iniciativas de ahorro mejor orientadas, incluso optimizar su inversión en los siguientes años.

- **Taxi and Limousine Commission (TLC)**. Para conseguir mayor desarrollo y mejora del servicio de taxi y alquiler en la ciudad de Nueva York, se usan los pronósticos de los viajes para otorgar nuevas licencias y establecer las tarifas de taxi.

- Los **medios de comunicación**, muy relacionados con el perfil de periodista de datos, merecen un punto específico, ya que llevan a cabo un uso diferente de la información, más orientado a la difusión, la denuncia, la reivindicación, etc.

- La **población** en general, puesto que consultando los datos integrados puede disponer de información muy valiosa para valorar qué transporte le conviene más o planificar mejor sus desplazamientos (tiempo de espera, tiempo de trayecto, etc.), entre otros.

Fuentes de datos

Uno de los objetivos de este caso de estudio es integrar las diversas fuentes de datos (y formatos) proporcionadas para poder realizar diferentes tipos de análisis. En concreto, disponemos de información detallada sobre la evolución de los resultados electorales.

La relación de ficheros *open data* que utilizaremos para la carga inicial son los siguientes:

#	Nombre de fichero	Descripción	Fuente
1	yellow_tripdata-001.zip	Viajes realizados por los icónicos taxis amarillos de la ciudad de Nueva York.	Datos de registro de viaje de TLC https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
2	fhv_tripdata-001.zip	Viajes realizados por los vehículos de alquiler (FHV).	Datos de registro de viaje de TLC https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
3	taxi_zone_lookup.csv	Relación de códigos de zona, usados para indicar las zonas de inicio y fin de viajes (las zonas de taxi se basan aproximadamente en los barrios de la ciudad de Nueva York).	Datos de registro de viaje de TLC https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
4	Rate_code.tab	Relación de tarifas vigentes aplicables a los viajes en taxi.	Fuente propia, sobre la base del diccionario de datos de TLC
5	Payment_type.xls	Relación de formas de pago posibles para los pasajeros de un viaje en taxi.	Fuente propia, sobre la base del diccionario de datos de TLC
6	current_bases.tsv	Relación de bases licenciadas por TLC para vehículos de alquiler.	*Current Bases | NYC Open Data* https://data.cityofnewyork.us/Transportation/CURRENT-BASES/eccv-9dzr/data

Las fuentes de datos 1, 2 y 3 que se obtienen de la web TLC vienen en formato «parquet» (específico de sistema *streaming)*. Para facilitar su tratamiento, se entregan convertidos a formato CSV.

Estos datos se recibirán actualizados periódicamente y, por lo tanto, serán necesarias cargas incrementales para su integración en el DW. El desarrollo de estos futuros procesos queda fuera del alcance de esta actividad.

Enunciado

PR1: análisis y diseño del *data warehouse*

A partir del análisis del contexto del caso y de las fuentes de datos disponibles, el estudiantado deberá diseñar y proponer un almacén de datos que permita y facilite el análisis de los registros de viajes de la TLC.

Mediante la metodología de diseño de un DW, propuesta en la asignatura, el estudiantado deberá llevar a cabo las siguientes tareas:

- **Análisis de requerimientos**. Como resultado, se generará un documento que describa las preguntas a las que el sistema dará respuesta para sus usuarios potenciales.

- **Análisis de fuentes de datos**. Se deben revisar las fuentes de datos proporcionadas, qué tipo de información contienen, cuál es su formato y qué cantidad representan para la carga inicial.

- **Análisis funcional**. Se debe proponer el tipo de arquitectura para la factoría de información que mejor se adecue al proyecto (por ejemplo, si es necesario un *data mart* operacional o una estructura de carga intermedia).

- **Diseño del modelo conceptual, lógico y físico del almacén de datos**. Se deben identificar y diseñar las tablas de hechos, las dimensiones y los atributos que describen la información.

En esta parte de la práctica, el estudiantado debe preparar un documento («Solución PR1») en el que se detallen todas y cada una de las secciones anteriores.

Se deberá tener en cuenta que, para el desarrollo del DW, es preciso definir correctamente los hechos (*facts*), las dimensiones de análisis (*dimensions*) y los atributos que permitan tener el nivel de granularidad suficiente para la medida y la presentación de los objetivos que se definan en el análisis de requerimientos.

PR2: carga de los datos

A partir de la solución oficial de la primera práctica (PR1), se deberán diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos (ETL) de las fuentes de datos proporcionadas al DW.

El estudiantado deberá llevar a cabo las tareas siguientes:

- Identificación de los procesos ETL en el almacén de datos.

- Diseño y desarrollo de los procesos de ETL mediante las herramientas de diseño proporcionadas.

- Implementación con los trabajos de los procesos de ETL para su carga efectiva planificada.

PR3: explotación de los datos

Tras la carga efectiva de los datos en el almacén (PR2), se debe implementar un cubo multidimensional para la explotación de la información como apoyo a la toma de decisiones de los usuarios potenciales.

La finalidad del modelo *multidimensional online analytical processing* (MOLAP) será responder a las preguntas definidas en el análisis de requerimientos.

Programas

Para el presente caso, la UOC proporciona un entorno VDI (*virtual desktop infrastructure*) con todo el software preconfigurado con las siguientes características:

- Sistema operativo: Windows 10.
- Base de datos: base de datos remota Microsoft SQL Server 2017 accesible desde clientes mediante SQL Server Management Studio 17.
- Herramienta para la creación de cubos OLAP: Visual Studio 2017.
- Herramienta de diseño de ETL: Spoon – Pentaho Data Integration 9.2.
- Herramienta de creación de informes: PowerBI Desktops.

Bibliografía

Material de la asignatura *Data warehouse* de la UOC

Inmon, W. H. (2005). *Building the Data Warehouse* (4.ª ed.). Nueva York (NY): John Wiley & Sons, Inc.

Inmon, W. H.; Imhoff, C.; Sousa, R. (2001). *Corporate Information Factory*. Nueva York (NY): John Wiley & Sons, Inc.

Inmon, W. H.; Strauss, D.; Neushloss, G. (2010). *DW 2.0: The Architecture for the Next Generation of Data Warehousing*. *The Morgan Kaufman Series in Data Management Systems*. Burlington (MA): Morgan Kaufmann Publishers.

Kimball, R. (2013). *The Data Warehouse Toolkit* (3.ª ed.). Nueva York (NY): John Wiley & Sons, Inc.

Krish, K. (2013). *Data Warehousing in the Age of Big Data*. *The Morgan Kaufmann Series on Business Intelligence*. Burlington (MA): Morgan Kaufmann Publishers.

Enlaces web

MSDN Analysis Services Tutorial

https://docs.microsoft.com/es-es/analysis-services/analysis-services-tutorials-ssas?view=asallproducts-allversions

Tutorial Pentaho Data Integration

https://docs.hitachivantara.com/r/en-us/pentaho-data-integration-and-analytics/9.4.x/mk-95pdia000/getting-started-with-pdi