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

PR3 – Explotación de datos

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

Índice

1. Presentación

La PR3 consiste en la explotación de datos de un almacén de datos de los viajes de la New York City Taxi and Limousine Commission (NY taxis).

A partir de las tablas que cada estudiante encontrará en su base de datos, debe implementar los cubos, uno por cada tabla de hechos. Estos cubos son necesarios para la explotación de los datos, facilitarán el análisis de datos y, por consiguiente, la toma de decisiones.

El objetivo de esta actividad es implementar un modelo online *analytical processing* (OLAP) para el análisis multidimensional de la información con el fin de responder a las preguntas definidas en el análisis de requerimientos.

Para que el punto de partida de la PR3 sea el mismo para todo el estudiantado, se facilitarán los scripts necesarios para la generación y carga de las tablas necesarias, o bien se facilitarán las tablas ya pobladas en la base de datos de cada estudiante.

2. Descripción

Más concretamente, esta tercera parte del caso práctico consiste en diseñar un modelo OLAP para el análisis multidimensional de la información disponible en el almacén de datos, que permita dar respuesta a las siguientes cuestiones:

1) ¿Cuántos viajes en vehículo de alquiler se han iniciado en la zona de Times Square durante el año 2023? Se desea conocer la evolución mes a mes.

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

3) ¿Cuáles son las diez zonas de Nueva York (top 10) con mayor cantidad de viajes iniciados por vehículos de alquiler durante todo el periodo del que se disponen datos?

4) 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.

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

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

7) ¿Cuáles son las cinco zonas *service zone* y *zone* de recogida con menor duración total (top 5) de los viajes en vehículos de alquiler durante el año 2023?

Para un desarrollo y diseño correctos del modelo OLAP, el documento debe incluir como mínimo la descripción detallada de los siguientes pasos:

1) Diseño del modelo OLAP:
   a) Creación del proyecto en Visual Studio. Se debe importar el proyecto desde el servidor asignado y revisar el origen de los datos, la conexión a las bases de datos y las propiedades del proyecto.
   b) Creación de la vista de origen de datos. Se deben mostrar capturas de pantalla del proceso de creación del DSV y el diagrama final.
   c) Creación e implementación de los cubos. Se deben mostrar capturas de pantalla del proceso de creación del modelo, el diagrama final y la configuración del «uso de dimensiones».
   d) Configuración de las dimensiones. Se deben mostrar atributos, jerarquías, relaciones de atributos, etc.
   e) Procesado y resolución de errores. Se deben incluir los posibles errores de procesado y explicar las soluciones aportadas en cada caso, así como una captura de pantalla con el proceso completo del modelo.

2) Explotación del modelo OLAP: diseñar seis consultas para dar respuesta a las cuestiones planteadas en este mismo enunciado. Todas las respuestas a las preguntas formuladas en la explotación de datos deben realizarse a partir del cubo OLAP implementado, independientemente de la herramienta utilizada para su análisis y explotación, bien se utilice el browser de Visual Studio, MDX o Power BI. No se darán por válidas las consultas efectuadas directamente sobre el modelo relacional subyacente (*data mart*) en SQL, ni en ningún otro lenguaje de consulta y análisis de datos.

Además de la entrega de la solución de la PR3, en la evaluación de la actividad se podrá considerar también la implementación de la máquina virtual proporcionada en el curso.

En resumen, el documento de la solución de la PR3 debe incluir:

- Las capturas que muestren la correcta realización de los puntos enumerados anteriormente.
- La descripción de todas las acciones realizadas. Es necesario explicar todas las capturas de pantalla facilitadas. La nota será penalizada si únicamente se aportan capturas de pantalla sin ningún comentario.
- Las capturas que muestren la correcta definición de las consultas, la visualización del resultado de las explotaciones y su correcta interpretación.

Además, de manera explícita, se pide incluir en el documento las siguientes capturas de pantalla:

- Test de la conexión, donde se vea el servidor, el usuario o usuaria y la base de datos.
- Captura de la implementación completa del proyecto desde Visual Studio, donde se vea el nombre de la base de datos destino («DEST_loginuoc»), y las fechas de inicio y fin del proceso.

- En Visual Studio, captura de pantalla de los objetos creados tras la implementación del proyecto: «DataSources», «DataSourceViews», «Cubes» y «Dimensions».
- En SQL Management Studio, captura de los objetos creados tras la implementación del proyecto: «DataSources», «DataSourceViews», «Cubes» y «Dimensions», donde también se vea el nombre de la base de datos correspondiente al estudiante («DEST_loginuoc»).

Para cada una de las consultas solicitadas, captura completa donde se vea el entorno de trabajo (Visual Studio, SSMS o Power BI) y la salida de la ejecución.

3. Criterios de evaluación

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

- Diseño del modelo OLAP (50 %).
- Explotación de datos (50 %):
  - Consulta 1 (0 %). Guía de muestra.
  - Consulta 2 (5 %).
  - Consulta 3 (5 %).
  - Consulta 4 (10 %).
  - Consulta 5 (10 %).
  - Consulta 6 (10 %).
  - Consulta 7 (10 %).

4. Guía de muestra

4.1. Diseño del modelo OLAP

Esta guía de muestra se desarrolla con el fin de ayudar a alcanzar los objetivos planteados en la PR3. La guía servirá de ejemplo para saber cómo se deben realizar algunas de las tareas anteriormente descritas, es decir, el diseño de explotaciones de un modelo OLAP para el análisis de los registros de viajes de la New York City Taxi and Limousine Commission (NY taxis).

4.2. Explotación del modelo OLAP

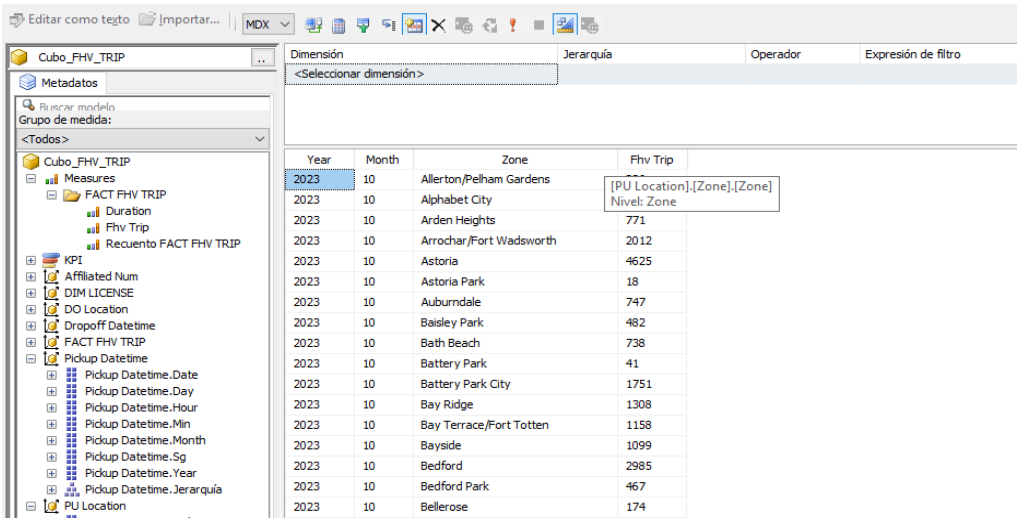
El diseño del cubo se realizará creando un proyecto multidimensional y de minería de datos en Visual Studio. Se deberán definir los orígenes de datos, las vistas de los orígenes de datos, los cubos, las relaciones de atributos, las dimensiones y las jerarquías necesarias para realizar la implementación de la solución y las explotaciones solicitadas en el enunciado de la PR3.

A continuación, se muestra un ejemplo de explotación de datos que se puede realizar tras la implementación de la solución. El documento de solución debe incluir todos los

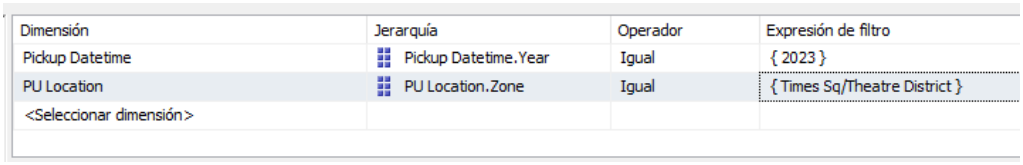
apartados solicitados en este enunciado. Para acceder al visor OLAP, se debe entrar en la pestaña «Browser» del cubo creado.

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

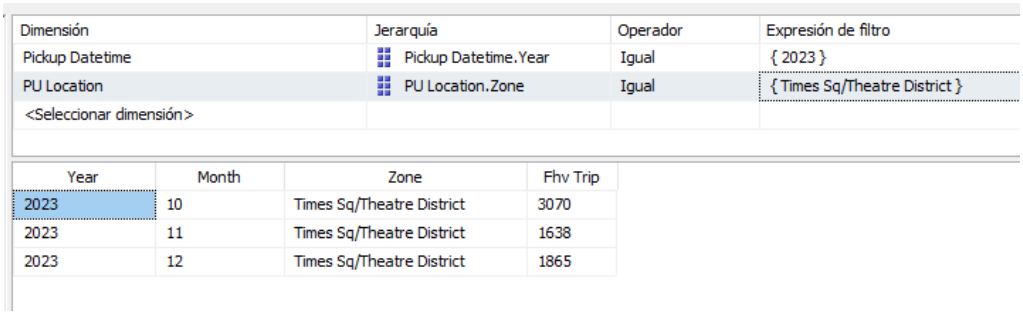
Este análisis se puede empezar utilizando el explorador de cubos de Visual Studio. Se trata de incorporar, desde el cubo «Cubo_FHV_TRIP», los atributos «Year» y «Month» de la dimensión «Pickup Datetime», el atributo «Zone» de «PU Location» y la métrica «Recuento» de «FACT FHV TRIP».



Para filtrar el resultado por año, se debe incluir la dimensión «Year» en la parte superior de los filtros y seleccionar el año 2023 en el desplegable «Expresión de filtro». Para filtrar por la zona de Times Square se incluirá la dimensión «PU Location» en el área de filtros.



Finalmente, se pulsa para ejecutar la consulta y se obtiene el resultado.



Analizando los datos del cuarto trimestre, se puede observar que el mes de octubre fue un año muy bueno, prácticamente doblando los viajes de los meses posteriores en ese mismo año.

5. Formato y fecha de entrega

La entrega de esta actividad debe realizarse a través del enlace de entrega PR3 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_PR3».

Por ejemplo, si el nombre de usuario es «bantich», el nombre del archivo debe ser «bantich_BDA_PR3.pdf».

La fecha máxima de entrega es el xx/xx/xxxx a las 23:59 horas.