Laboratorio #4 – Modelado multidimensional y ETL inicial

Andrés M. Ochoa Toro (201913554)

ISIS 3710 – Inteligencia de Negocios

Universidad de los Andes, Bogotá, Colombia

{am.ochoat[}@uniandes.edu.co](mailto:%7d@uniandes.edu.co" \t "_blank)

Fecha de presentación: Noviembre 5 de 2021

Tabla de contenido

[1 Introducción 1](#_Toc87126339)

[2 Perfilamiento y procesamiento de datos 2](#_Toc87126340)

[3 Herramientas proceso ETL 2](#_Toc87126341)

[3.1 Spoon 2](#_Toc87126342)

[3.2 Herramientas de Google Cloud 8](#_Toc87126343)

[3.3 Talend 12](#_Toc87126344)

[4 Recomendaciones 17](#_Toc87126345)

# Introducción

WWI (World Wide Importers) es una empresa encargada de realizar importaciones y venderlas a diferentes clientes en diferentes ciudades de Estados Unidos. En esta ocasión, WWI desea optimizar sus ganancias, pues consideran que algunos de sus productos no están generando las ganancias que deberían. Para esta primera fase, se desea realizar la creación de la base de datos, la carga de datos y la comprobación del correcto funcionamiento del proceso realizado a través de algunas consultas iniciales.

En esta etapa de la consultoría, la empresa desea implementar el proceso ETL que le permita extraer los datos de órdenes desde unos archivos CSV y almacenarlos en un modelo dimensional tal que les permita realizar análisis para mejorar entre otros elementos, su eficiencia operativa. A continuación, se presenta el modelo multidimensional que se desea obtener:

Diagrama

Descripción generada automáticamente con confianza media

Imagen 1. Modelo Multidimensional propuesto

# Perfilamiento y procesamiento de datos

Los archivos CSV que se van a utilizar en este laboratorio provienen de una base de datos SQLServer, por lo que es importante realizar un perfilamiento de los archivos recibidos para limpiar o transformar datos que puedan generar conflictos en el proceso.

El proceso realizado en el procesamiento se decidieron borrar todas aquellas filas que presentan un valor nulo, excepto en la tabla stock item. Para este conjunto de datos con una gran cantidad de valores nulos se tomaron acciones diferentes. En el caso del color, siendo un item relevante para el producto, se decidió reemplazar el valor nulo por una nueva opción que represente que se desconoce el valor. Y, en el caso de la columna *Brand* dado a su gran número de filas con valores nulos se decide borrar toda la columna.

Este proceso se describe en mejor medida en el notebook adjuntado.

# Herramientas proceso ETL

Para el proceso de ETL se crea una base de datos con PosgreSQL. Esta base de datos es el equivalente a crear un DataMart para la empresa.

## Spoon

En el uso del proceso de ETL, se empieza creando un nuevo *Job.* A este nuevo *Job* creado se le asigna una conexión a la base de datos. Esta base de datos es la creada en los procesos pasados y es la que representa el DataMart.

Después, por cada una de las dimensiones en el modelo dimensional se procede a realizar la carga de los datos en cada una de las tablas que representan esta tabla. Este proceso de carga de datos se realiza a partir de una transformación que lee cada uno de los archivos *CSV* y a partir de los nombres de las columnas se asigna a cada una de las columnas de la tabla SQL.

Como resultado final del proceso de carga de los datos del *job* en Spoon, a través de una *transform* para cada una de las dimensiones, se obtiene el siguiente resultado.

Diagrama

Descripción generada automáticamente

Imagen 2. Resultado *job* en Spoon

Imagen que contiene Gráfico

Descripción generada automáticamente

Imagen 3. *transformation en Spoon*

En el proceso de verificación, primero se comprueba que en el *Schema* se creen las tablas.

Interfaz de usuario gráfica

Descripción generada automáticamente con confianza media

Imagen 4. Creación de tablas Spoon

Finalmente, en el proceso de la verificación se comprueba la existencia de los datos en la base de datos a través de sentencias SQL.

Tabla

Descripción generada automáticamente

Imagen 5. Comprobación datos en tabla city

Tabla

Descripción generada automáticamente

Imagen 6. Comprobación datos en tabla customer

Tabla

Descripción generada automáticamente

Imagen 7. Comprobación datos en tabla date\_table

Tabla

Descripción generada automáticamente

Imagen 8. Comprobación datos en tabla employee

Tabla

Descripción generada automáticamente

Imagen 9. Comprobación datos en tabla fact\_order

Tabla

Descripción generada automáticamente

Imagen 10. Comprobación datos en tabla stockitem

Preguntas:

* ¿Por qué se utiliza el comendo “IF NOT EXISTS” en la sentencia, en el contexto del proceso de ETL?

El comando “IF NOT EXISTS” revisa si la tabla ya existe, y de ser así ignora todo el contenido de la sentencia y no crea una nueva tabla.

* ¿Por qué para la columna de día se utiliza el nombre “day\_val” y no “day”?

Se usa el nombre “day\_val” para representar el valor numérico del día. Es decir, si es el primer día del mes, el segundo, etc. En lugar de si se usa el día se podría referir es al día de la semana, lunes, martes, etc.

* ¿De dónde se obtiene la información sobre las columnas que hay que crear en la tabla?

La información se obtiene a partir de las sentencias SQL que hay que crear. Esta sentencia se hace a partir del nodo SQL.

* ¿Cuál es la diferencia entre un “Job” y una “Transformación” en Spoon?
  + Un trabajo tiene un lugar de inicio y se ejecuta un paso a la vez, con un flujo a través de los pasos.
  + Una transformación tiene muchos lugares de inicio posibles y todos los pasos se ejecutan en paralelo. Si un paso tiene un paso antes, tomará los datos allí y los usará.
* ¿Por qué se hace uso del nodo `Insert/Update` y no del `Table Output`?

En este caso se prefiere hacer uso del ‘Insert/Update’ para actualizar ciertos registros en caso de ya existan, mientras que el uso de table output sólo sirve para la inserción de datos. Esto genera que si hay cambios en una fila se actualicen en lugar de insertarlos.

* ¿Cuál es la finalidad de los campos “The key(s) to lookup the values:” y “Update fields” del nodo `Insert/Update`?

La funcionalidad “The key(s) to lookup the values:” hace una comparación entre los valores de las del archivo CSV y la tabla. A partir de esta columna hace un match entre estos dos archivos.

En el caso del Update fields es representar qué valores de la tabla se llenarán a partir de los valores del CSV.

* ¿Será posible programar la fecha y hora en la cual la organización puede correr de forma automática el proceso de ETL`?

Desde la versión paga de Pentaho sí es posible programar la ejecución automática de Jobs y transformaciones. Esto se hace a partir de la funcionalidad de scheduler.

## Herramientas de Google Cloud

Siguiendo el proceso de creación de la cuenta en Google Cloud, y la creación del respectivo proyecto. Se procede a cargar cada uno de los archivos *CSV.* Este proceso es el equivalente a crear las tablas en la base de datos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Imagen 11. Creación del proyecto y carga tablas

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen 12. Comprobación datos en tabla city

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Imagen 13. Comprobación datos en tabla customer

Tabla

Descripción generada automáticamente

Imagen 14. Comprobación datos en tabla date

Tabla

Descripción generada automáticamente

Imagen 15. Comprobación datos en tabla employee

Tabla

Descripción generada automáticamente

Imagen 16. Comprobación datos en tabla fact\_order

Tabla

Descripción generada automáticamente

Imagen 17. Comprobación datos en tabla stockitem

## Talend

Como primer paso, se procede a crear una base de datos local. Desde Talend se procede a crear la conexión a la base de datos y la carga local de los archivos CSV. Estos archivos se ven reflejados con la representación de los archivos separados por coma.

Texto

Descripción generada automáticamente

Imagen 17. Carga de CSV

A continuación, se procede a realizar el procesamiento de los datos que se leen desde los archivos y seleccionar los datos que se quieren pasar a la base de datos. Este proceso se replica para cada una de las dimensiones del modelo multidimensional esperado.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Imagen 17. Comprobación datos en tabla stockitem

Entre este proceso, se procede a realizar el procesamiento para cada una de las dimensiones.

Calendario

Descripción generada automáticamente

Imagen 18. Proceso de carga de datos

En el costado derecho del diagrama se realiza la carga de los datos a partir de los datos de entrada que provienen del proceso de mapeo del archivo CSV.

Ya en la revisión de los datos, primero se observa la creación de las tablas en la base de datos. Esto se puede observar desde pgAdmin y la base de datos.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Imagen 19. Creación tablas en la base de datos

Después de comprobar la existencia en la tabla de datos se revisa que los datos si se cargaran exitosamente.

Tabla

Descripción generada automáticamente

Imagen 20. Comprobación datos en tabla city

Interfaz de usuario gráfica, Tabla

Descripción generada automáticamente

Imagen 21. Comprobación datos en tabla customer

Tabla

Descripción generada automáticamente

Imagen 21. Comprobación datos en tabla date\_table

Tabla

Descripción generada automáticamente

Imagen 22. Comprobación datos en tabla employee

Tabla

Descripción generada automáticamente

Imagen 23. Comprobación datos en tabla fact\_order

Tabla

Descripción generada automáticamente

Imagen 2. Comprobación datos en tabla stockitem

# Recomendaciones

Entre la comparación de la base de datos se obtienen diferentes resultados respecto a la implementación. Se tiene que la implementación por GoogleQuery es la más rápida. Desde esta herramienta la implementación es a partir de la carga de los datos. En el segundo caso se tiene que la herramienta Spoon de Pentaho provee una implementación bastante robusta, ya que permite directamente la creación de las tablas en la base de datos con sentencias SQL, pero así mismo permite importar los datos desde los archivos CSV de una forma muy práctica. Finalmente, la herramienta de Talend permite la creación de la tabla en la base de datos únicamente a partir de la carga de los archivos CSV.

A partir de la implementación de las 3 herramientas de ETL, se elige Spoon. Esta herramienta es bastante robusta para la implementación de este proceso. Se combina el uso de sentencias SQL con la carga y modificación de los datos por medio de las transformaciones. Estas transformaciones son más fáciles de implementar que en Google Query o Talend.