**EVIDENCIA DE APRENDIZAJE 3. PROCESO DE TRANSFORMACIÓN DE DATOS**

**OSSMAN MEJÍA GUZMÁN**

**GRUPO PREICA2401B010076**

**OLIMPO DE JESÚS MACEA HERNÁNDEZ**

**GRUPO** [**PREICA2401B010094**](https://iudigital.instructure.com/courses/15609)

**INSTITUCIÓN UNIVERSITARIA DIGITAL DE ANTIOQUIA**

**INGENIERIA DE SOFTWARE Y DATOS**

**BASES DE DATOS II**

**VICTOR HUGO MERCADO**

**MEDELLÍN**

**MAYO DE 2024**

**INTRODUCCIÓN**

Esta evidencia de aprendizaje es la continuación de las evidencias de aprendizaje 1 y 2 que responde a los siguientes objetivos de aprendizaje:

1. Da cuenta del diseño de un modelo estrella para un data mart. El modelo estrella es una técnica de modelado de datos que busca, especialmente, optimizar los datos correspondientes a almacenes de datos y data mart. El esquema lógico de un modelo de estrella consta de una tabla central denominada hechos y diversas tablas de dimensiones. La tabla central de hechos, almacena datos susceptibles de análisis, mientras que las tablas de dimensiones almacenan datos utilizados para agrupar, clasificar o filtrar las medidas de la tabla de hechos. El buen diseño de los data mart, asegura la eficiencia de los procesos de inteligencia de negocios, por lo tanto, esta evidencia de aprendizaje permite desarrollar las competencias necesarias para su análisis y diseño dependiendo de las necesidades de la empresa.
2. Incluye el proceso de diseño y creación de la base de datos staging basada en el modelo relacional de la base de datos jardinería. Esto incorpora las siguientes tareas:
   1. Análisis de los datos almacenados en la base de datos Jardinería.
   2. Revisión de los datos almacenados en el Jardinería para identificar cuáles son relevantes y cuáles se deben trasladar a la base de datos Staging.
   3. Construcción de la base de Datos Staging
   4. Diseño de la estructura tablas que estarán en la base de datos Staging.
   5. Construcción de las consultas que permitan traer los registros de Jardinería a la Base de Datos Staging.
   6. Ejecución de las consultas y validar que los datos queden almacenados correctamente en la base de datos Staging.
   7. Construcción del backup de ambas bases de datos.
3. Desarrollar un proceso de transformación desde la base de datos origen, pasar Staging y luego hasta el data mart final, utilizando la base de datos de staging previamente creada. El objetivo es asegurar que los datos estén preparados para el análisis, incluyendo la identificación del producto más vendido y otros análisis relevantes para la toma de decisiones empresariales.

**OBJETIVO GENERAL**

Desarrollar un proceso de transformación de datos que abarque desde la extracción de datos desde la base de datos origen, su transición a un entorno de staging, y finalmente su carga en el data mart final (pendiente), garantizando que los datos estén adecuadamente preparados para análisis, incluyendo la identificación del producto más vendido y otros análisis críticos para la toma de decisiones.

**OBJETIVOS ESPEFICOS**

* Implementar un flujo de ETL (Extracción, Transformación y Carga) que extraiga datos de la base de datos origen y los traslade al entorno de staging, asegurando la integridad y calidad de los datos durante todo el proceso.
* Diseñar y ejecutar un plan de extracción de datos desde la base de datos origen.
* Aplicar transformaciones necesarias en el entorno de staging para limpiar y normalizar los datos.
* Realizar validaciones y controles de calidad para asegurar que los datos en el entorno de staging sean precisos y consistentes.

**PLANTEAMIENTO DEL PROBLEMA**

En las evidencias de aprendizaje 1 y 2 se tomó como caso de estudio, una empresa de jardinería que carece de un sistema que permita realizar el análisis de sus datos de ventas y entender, por ejemplo, las tendencias clave que impulsan su negocio. En la evidencia de aprendizaje 1 se planteó el análisis y diseño de un data mart que dé respuesta a las necesidades del negocio para comprender procesos propios de las ventas como el producto más vendido, la categoría con más productos vendidos y el año con más ventas. Esta información ayudará al desarrollo de la inteligencia de negocios en la empresa.

Por su parte, en la evidencia de aprendizaje 2 surgió la necesidad de realizar el análisis y migración de los datos almacenados en la base de datos transaccional hacia una base de datos Staging, esta última definida en pocas palabras como un sistema que se encuentra entre las fuentes de datos y el data mart.

En esta evidencia de aprendizaje 3, se continua en proceso de ETL específicamente en la transformación de los datos, se realiza el proceso de enfocándose específicamente en la transformación de los datos. Se realiza el proceso de limpieza de datos, que implica eliminar duplicados y corregir valores inconsistentes y, de normalización y estandarización que consiste en convertir los datos a formatos consistentes y estandarizados.

**ANÁLISIS DEL PROBLEMA**

Luego de diseñar el modelo de estrella para el data mart de la empresa de jardinería, en el cual se realizó un análisis detallado de los requisitos y las necesidades del negocio y del modelo relacional, fue necesario construir la base de datos staging que permitiera un correcto proceso de ETL hacia el data mart. En este sentido, se llevó a cabo el análisis de los datos almacenados en la base de datos Jardinería con el fin de identificar cuáles eran relevantes y cuáles debían ser trasladados a la base de datos Staging, según el modelo de estrella planteado.

Posteriormente, se realizó la construcción de la base de datos, teniendo en cuenta la estructura de sus tablas, el diseño de consultas SQL para trasladar los registros desde el origen y la ejecución de consultas y validaciones para certificar que los datos quedaran almacenados correctamente en la base de datos staging.

En esta evidencia de aprendizaje se da cuenta del proceso de transformación de los datos disponibles en la base de datos staging.

**TRANSFORMACIÓN DE LOS DATOS**

El proceso de transformación de la base de datos staging inició con la creación de la dimensión de Tiempo.

Para este proceso se obtuvieron los datos de la tabla tiempo previamente creada en la base de datos staging y se realizó un proceso de transformación en el cual se derivaron columnas usando las funciones de fecha para día, mes, año, día de la semana y número de semana contable.

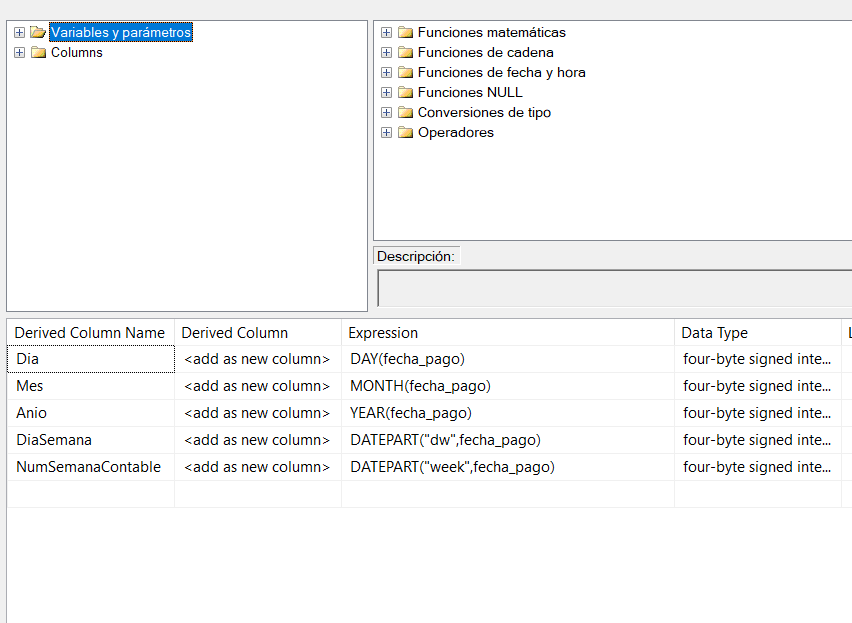


Ilustración 1. Transformación de la dimensión de tiempo

Posteriormente se realizó un sort a la tabla de origen para que ordenara las fechas por el ID\_fecha. Este proceso se acompañó de la un Merge Join en el cual se validó el ingreso de datos a través de un left join a la tabla de destino, teniendo en cuenta además una condición para evitar la duplicación de datos.

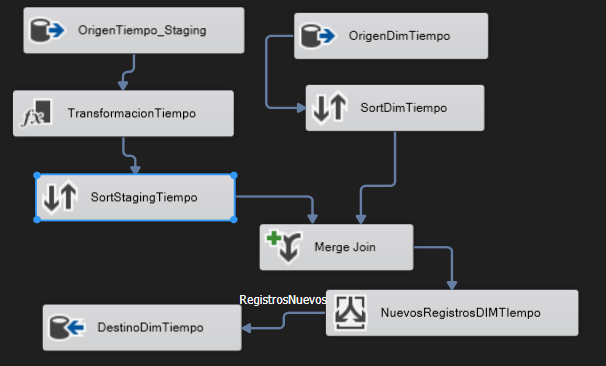


Ilustración 2. Data flow de la creación de la dimensión de tiempo

Para la transformación de las dimensiones de cliente, producto, empleado y la fact de hechos de ventas se siguió el siguiendo proceso:

Se obtuvo un origen de datos de la respectiva tabla en la base de datos de staging a través de una consulta SQL que uniera las demás tablas según el caso. Por ejemplo, la tabla de productos se unió con la tabla de categorías.

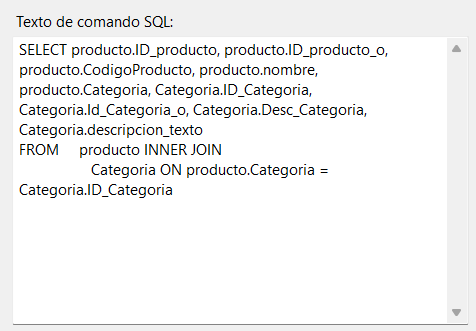


Ilustración 3. Ejemplo de consulta SQL para el origen de datos desde las tablas staging

Posteriormente se realizaron las respectivas validaciones para que las dimensiones de cliente, producto, empleado y la fact de hechos de ventas, recibieran a través de un left join y la condicional de registros nuevos, solamente aquellos registros que no están previamente disponibles en las respectivas tablas de dimensiones y/o hechos.

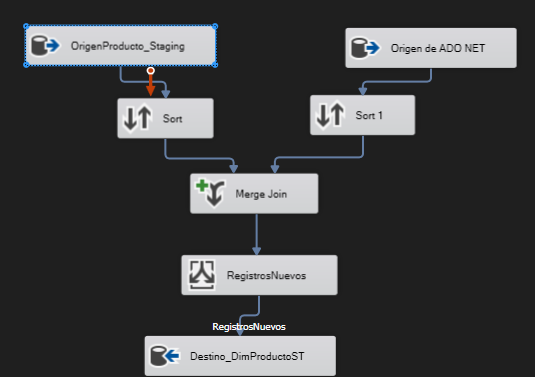


Ilustración 4. Ejemplo de data flow para la transformación de los datos desde la base de datos staging

Para la fact de hechos se creó una consulta SQL que recibiera todos los campos de las dimensiones creadas anteriormente y en la asignación se depuraron los campos para que solo quedara con las llaves a dichas dimensiones y los datos cuantitativos para los cálculos que se realizarán en la carga del data mart.

SELECT \* FROM Pedidos

INNER JOIN Detalle\_Pedidos ON Pedidos.ID\_pedido = Detalle\_Pedidos.ID\_pedido

INNER JOIN DimProducto ON DimProducto.ID\_producto = Detalle\_Pedidos.ID\_producto

INNER JOIN DimCliente ON Pedidos.ID\_cliente = DimCliente.ID\_cliente

INNER JOIN DimEmpleado ON DimCliente.ID\_empleado\_rep\_ventas = DimEmpleado.ID\_empleado

INNER JOIN DimTiempo ON Pedidos.fecha\_pedido = DimTiempo.fecha\_pedido

Los scripts de SQL que se utilizaron para la creación de los orígenes y el back up del proyecto de Microsoft Integration Services se pueden encontrar en el archivo correspondiente de esta evidencia de aprendizaje.

**CONCLUSIONES**

El desarrollo del proceso de transformación de datos abordado en esta evidencia de aprendizaje ha permitido consolidar los conocimientos y habilidades necesarias para el diseño y ejecución de un flujo ETL (Extracción, Transformación y Carga). A través de las actividades realizadas, se han alcanzado los objetivos propuestos y se han puesto en práctica conceptos fundamentales del manejo de bases de datos y la inteligencia de negocios.

En primer lugar, se logró diseñar un modelo estrella para un data mart, lo que permitió optimizar el almacenamiento y acceso a los datos relevantes para el análisis empresarial. Este modelo no solo facilita la organización y estructuración de los datos, sino que también mejora la eficiencia de los procesos analíticos.

Asimismo, se desarrolló la base de datos staging, un paso intermedio crucial que garantiza la correcta preparación y limpieza de los datos antes de su carga en el data mart. Este proceso incluyó un análisis de los datos originales, la identificación de aquellos que eran relevantes para el negocio, y la implementación de técnicas de normalización y estandarización para asegurar su consistencia y precisión.

El proceso de transformación de datos desde la base de datos staging hasta el data mart final fue especialmente significativo, aunque el proceso de carga estará pendiente para la próxima evidencia de aprendizaje, este paso no solo implicó la limpieza y corrección de los datos, sino también su reestructuración para que se ajustaran al esquema del modelo estrella. La creación de dimensiones como Tiempo, Cliente, Producto y Empleado, y la fact hechos de ventas, ejemplifican la capacidad de transformar datos en bruto en información útil para el negocio de jardinería.

**ANEXOS**

Se anexa carpeta comprimida que contiene los siguientes archivos:

* **BK\_Proyecto\_Integration\_Services\_Jardineria:** Carpeta que contiene todo el proyecto de Integration Services de Microsoft Visual Community.
* **Jardineria.bak**: Archivo de respaldo de la base de datos de jardinería.
* **Staging\_jardineria.bak**: Archivo de respaldo de la base de datos staging para el proceso ETL del data mart.
* **QuerysStaging.sql**: Archivo que contiene las consultas SQL que se realizaron para la creación de la base de datos staging y el proceso de transformación de datos
* **Enlace a repositorio GitHub del proyecto**: <https://github.com/ossmanmejia/BI_Jardineria>