Evidencia de Aprendizaje 3. Proceso de transformación de datos y carga en el datamart final

María José Hernández Rosales
Institución Universitaria Digital de Antioquia
Bases de Datos II
Prof. Victor Hugo Mercado
Octubre, 2024

Introducción

En el área del análisis de datos, es fundamental poseer la capacidad de extraer insights significativos de grandes cantidades de data. En el presente documento, se expone brevemente el proceso de construcción de un modelo de esquema estrella para un data mart utilizando como fuente de datos principal la base de datos ***Jardinería***; este serviría para realizar un análisis integral de los insights accionables que son fundamentales para una toma de decisiones informada.

Con el desarrollo de este marco se pretende dar respuesta a tres categorías concretas: la identificación del producto más vendido, de la categoría con más productos y del año con mayor volumen de ventas. Todo esto se logra por medio del uso de un amplio conjunto de datos que incluyen como entidades a oficinas, empleados, clientes, productos, pedidos y pagos.

Es importante no solo almacenar los datos de manera eficiente, sino también transformar y cargar los datos para su análisis, lo cual se logra mediante el proceso ETL (Extract, Transform, Load). Por lo que se detalla el proceso aplicado a la base de datos Jardineria.

Planteamiento y Análisis del Problema

En el caso de estudio se evidencia la necesidad de analizar y responder a ciertas consultas clave relacionadas con las operaciones comerciales y transaccionales de ventas.

El problema radica en la falta de una estructura de datos amigable que permita el análisis eficiente de la información disponible en la base de datos. Se carece de un marco analítico que facilite la identificación de los productos más vendidos, el análisis de las categorías de productos con mayor presencia en el mercado y la determinación de las tendencias de ventas a lo largo del tiempo. Todo esto se detalla brevemente por medio de los siguientes puntos:

- **Complejidad de los datos.** La base de datos contiene una gran cantidad de información sobre oficinas, empleados, clientes, productos, pedidos y pagos; y debido a la estructura inadecuada que poseen puede ser compleja su comprensión.
- **Falta de Estructura Analítica.** No se posee un marco analítico establecido para extraer la información relevante de los datos almacenados, dificultando la identificación de tendencias y/o patrones.
- **Necesidad de Análisis Estratégico.** Se requiere tener la capacidad de identificar los productos más vendidos, las categorías de productos más populares y las tendencias de ventas a lo largo de los años.
- **Desafíos en la Generación de Información Accionable.** Debido a la ausencia de una estructura de datos eficiente es difícil generar información procesable a partir de los datos almacenados.

- **Necesidad de una Arquitectura de Datos Adecuada.** Una estructura de datos, como el modelo de estrella, otorgaría la estructura requerida para realizar análisis de datos de manera eficiente.

Para dar solución a lo anterior, de forma integral, se busca la construcción de un modelo de estrella para un data mart para obtener información procesable que mejore su rendimiento.

Propuesta de Solución

El modelo estrella propuesto posee la estructura clásica, compuesto por una tabla de hechos central que se encuentra rodeada por varias tablas de dimensiones. Tiene como objetivo permitir el análisis de los datos de las transacciones de ventas, para facilitar la identificación de los productos más vendidos, el análisis de las categorías de los productos y la determinación de las tendencias de ventas a lo largo del tiempo.

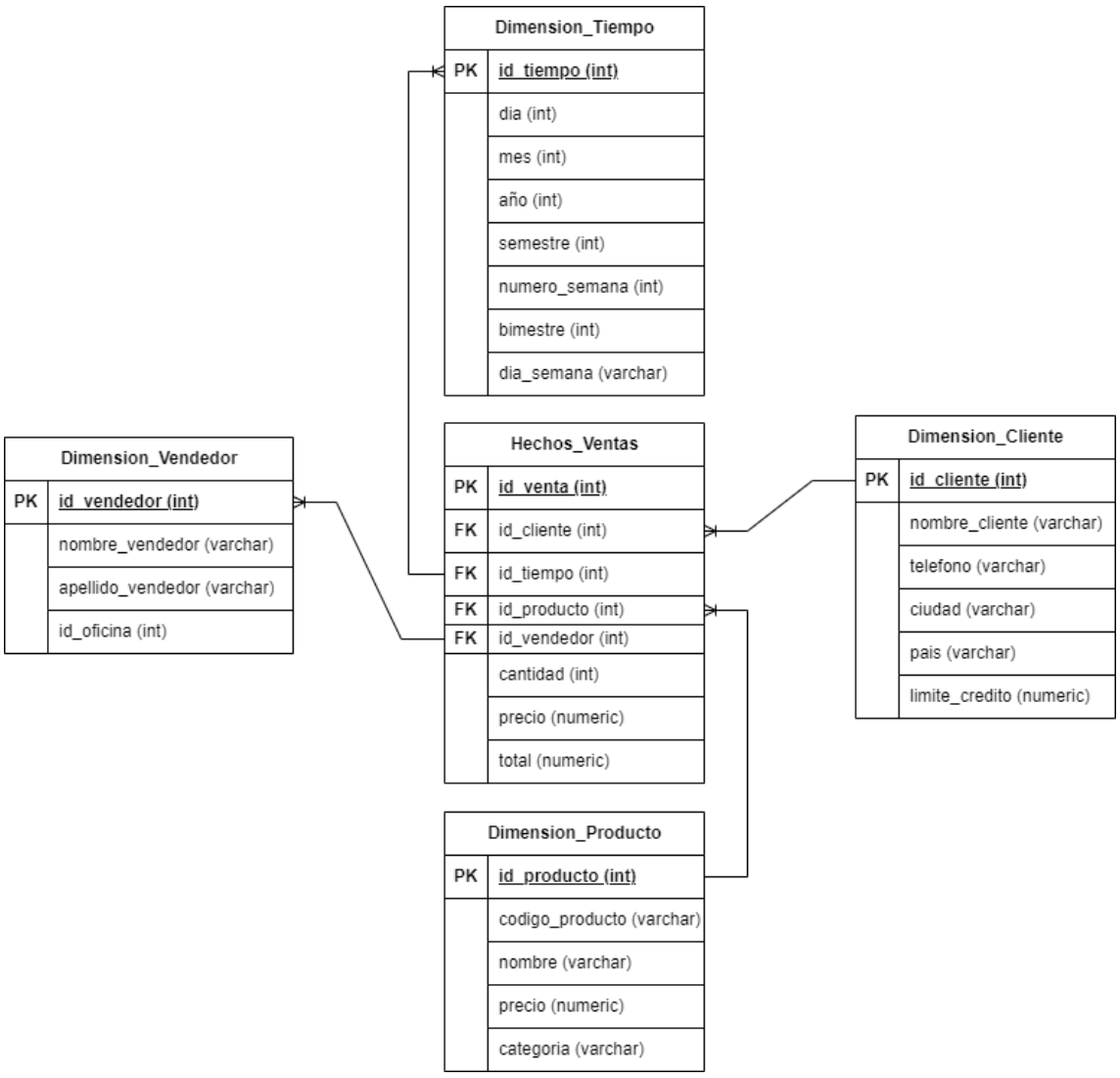

Ilustración 1. Tabla de Hechos

En el diseño del modelo de estrella, se evidencian los siguientes componentes.

- **Tabla de Hechos**

La tabla de hechos es el núcleo del modelo de estrella y contiene las métricas que representan las transacciones de ventas. Los campos y tipos de datos incluidos son los siguientes:

- **Hechos_Ventas**

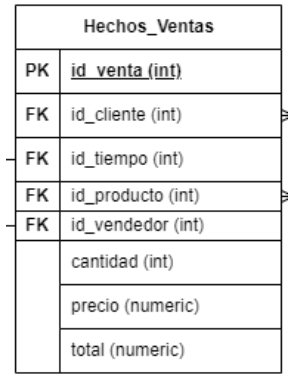

Ilustración 2. Hechos_Ventas

id_venta (int). Clave primaria. Identificador único de cada venta

id_cliente (int). Clave foránea. Identificador del cliente asociado a la venta

id_tiempo (int). Clave foránea. Identificador de las fechas de compra y entrega de la venta

id_producto (int). Clave foránea. Identificador del producto vendido

id_vendedor (int). Clave foránea. Identificador del vendedor

cantidad (int). Cantidad de productos incluidos en la venta

precio (numeric). Precio unitario del producto vendido.

total (numeric). Valor total de la venta

- **Dimensiones Propuestas**

1. **Dimension_Producto**

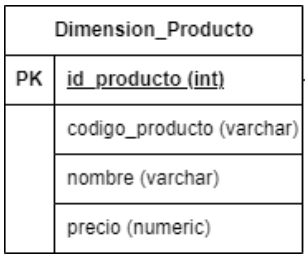

Ilustración 3. Dimension_Producto

id_producto(int). Identificador único del producto

código_producto (varchar). Código del producto

nombre (varchar). Nombre del producto

precio (numeric). Precio unitario del producto

categoría (varchar). Categoría del producto

2. **Dimension_Tiempo**

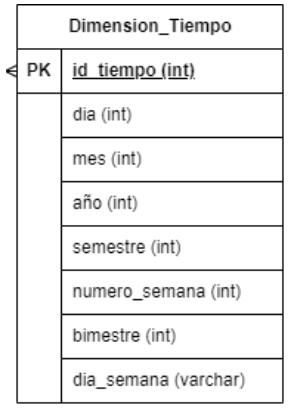

Ilustración 4. Dimension_Tiempo

id_tiempo (int). Identificador único de las fechas de compra

día (int). Día en el que se hizo la compra

mes (int). Mes en el que se hizo la compra

año (int). Año en el que se hizo la compra

semestre (int). Semestre en el que se hizo la compra

numero_semana (int). Número de la semana en el año

bimestre (int). Bimestre en el que se hizo la compra

día_semana (varchar). Día de la semana en el que se hizo la compra

3. **Dimension_Cliente**

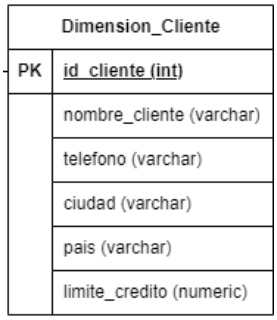

Ilustración 5. Dimension_Cliente

id_cliente (int). Identificador único del cliente

nombre_cliente (varchar). Nombre del cliente

telefono (varchar). Número de teléfono del cliente

ciudad (varchar). Ciudad en la que reside el cliente

pais (varchar). País en el que reside el cliente

limite_credito (numeric). Límite de crédito del cliente

4. **Dimension_Vendedor**

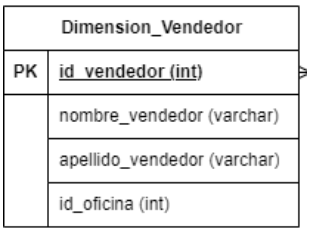

Ilustración 6. Dimension_Vendedor

id_vendedor (int). Identificador único del vendedor

nombre_vendedor (varchar). Nombre del vendedor

apellido_vendedor (varchar). Apellido del vendedor

id_oficina (int). Clave foránea. Identificador de la oficina asociada al vendedor

Con el modelo propuesto se podría realizar un análisis de datos detallado de las transacciones de venta; gracias a la estructura modular del esquema sería posible incorporar nuevas dimensiones en caso de ser necesario.

Para lograr esto, se hace uso de el proceso ETL que consta de las siguientes etapas:

1. **Extracción.** Se extraen los datos de forma “cruda” desde la base de datos Jardineria hacia la base de datos Staging. Esto permite obtener una copia exacta de los datos originales para su posterior transformación.
2. **Transformación.** Los datos obtenidos de la primera fase se limpian, normalizan y transforman para adaptarlos a la estructura requerida por el modelo estrella. Este paso incluye la eliminación de valores nulos, la estandarización de formatos y la conversión de datos.
3. **Carga.** Los datos transformados se cargan en las tablas de dimensiones y hechos del modelo estrella en la base de datos final.

Descripción del Análisis

Para llevar a cabo el análisis de la base de datos Jardineria se examinó la estructura de la data existente para identificar las entidades y atributos más relevantes para el análisis de las ventas. Incluyendo la revisión de cada una de as tablas en la base de datos para poder comprender la data de forma integral.

Extracción (Extract). El primer paso fue la identificación de las entidades y los atributos clave requeridos para el análisis de las ventas. Para eso se extrajeron los datos relacionados con los clientes, productos, fechas de venta, cantidades y precios desde la base de datos Jardineria hacia las tablas crudas en la base de datos Staging. Esto implicó la creación de tablas staging que replican la estructura de las tablas originales de la base de datos Jardineria, tal como se evidencia en el archivo .sql adjunto.

Transformación (Transform). Los datos identificados de la base de datos Jardineria se "limpiaron" y transformaron para que fueran coherentes con la estructura de la nueva base de datos. Esto implicó:

- **Manejo de valores nulos.** Se reemplazaron o eliminaron valores nulos.
- **Estandarización de formatos.** Se aseguraron formatos consistentes para fechas, nombre y otros campos.
- **Conversión de datos.** Se adaptaron los datos para que coincidieran con los tipos de datos requeridos en las tablas de dimensiones y hechos.

Carga (Load). Los datos transformados se cargaron en las tablas correspondientes del modelo estrella. Cada tabla de dimensión se pobló con los datos relevantes y principales, mientras que la tabla de hechos contuvo la información consolidada de las ventas, incluyendo referencias a las dimensiones correspondientes.

Por último, se verificó la integridad de los datos cargados en la base de datos final.

Conclusiones

El modelo propuesto mejoraría significativamente la gestión de los datos y la toma de decisiones, además se identifican varias conclusiones importantes.

Primero, este modelo otorga una estructura analítica sólida que facilita la comprensión y el análisis de las ventas, esto gracias a la separación modular y escalable de las tablas de hechos y dimensiones propuestas.

Segundo, este modelo permite generar información accionable a partir de los datos de las ventas ya que permite identificar los productos más vendidos, analizar las categorías de los productos y ver las tendencias de las ventas a lo largo del tiempo.

Por último, también permite llevar a cabo una toma de decisiones informada al proporcionar información crítica que podría permitir la adaptación de estrategias comerciales y optimización de operaciones.

El proceso ETL descrito fue clave para garantizar la calidad y coherencia de los datos transferidos, asegurando que el modelo estrella sea una herramienta efectiva para el análisis de datos.

Bibliografía

García Mattío, M., & Bernabeu R., D. (s.f.). *Estrella*. Obtenido de Hefesto: https://troyanx.com/Hefesto/estrella.html

How to insert table values from one database to another database? (06 de Febrero de 2019). Obtenido de Stack Overflow: https://stackoverflow.com/questions/3502269/how-to-insert-table-values-from-one-database-to-another-database

IBM Corporation. (08 de marzo de 2021). *Esquemas de estrella*. Obtenido de Documentación de IBM: https://www.ibm.com/docs/es/ida/9.1.2?topic=schemas-star

Myers, P., Patel, M., Sparkman, M., Follis, K., Buck, A., Coulter, D., . . . Saxton, A. (22 de marzo de 2023). *Descripción de un esquema de estrella e importancia para Power BI*. Obtenido de Microsoft Learn: https://learn.microsoft.com/es-es/power-bi/guidance/star-schema

MySQL RESET AUTO INCREMENT. (s.f.). Obtenido de Javatpoint: https://www.javatpoint.com/mysql-reset-auto-increment

SQL INSERT INTO SELECT Statement. (s.f.). Obtenido de w3schools: https://www.w3schools.com/sql/sql_insert_into_select.asp

SQL UPDATE Statement. (s.f.). Obtenido de w3schools: https://www.w3schools.com/sql/sql_update.asp