Extracción de datos desde la base de datos "Jardinería" a una base de datos Staging

Evidencia de aprendizaje 2

Daniela Ceferino Marín

Laura Victoria Navarro Arriola

Alexander Zapata Rico

Institución Universitaria Digital de Antioquia

Facultad de Ingeniería y Ciencias Agropecuarias

Especialización en Analítica y Big Data

Electiva II: Base de datos II

Docente: Victor Hugo Mercado

Semestre 2024-2 Bloque 2

TABLA DE CONTENIDO

Diseño de Modelo Estrella para la base de datos Jardinería

Introducción

Toda empresa y/o organización necesita una estructura de base de datos optimizada para analizar sus transacciones de manera efectiva en sus diversas áreas. Es por ello que resulta necesario plantear el orden y estructura de la información para que al momento de desarrollar sus aplicaciones para los diferentes fines que se creen se pueda obtener la información en corto tiempo, estructurada, y fiel a los procesos que ocurren en tiempo real. De ahí el uso de modelos estructurados de data que parten de la selección de un dominio y unas entidades y relaciones que hagan una adecuada asociación entre ellas para realizar todo tipo de consultas.

En el proceso de aprendizaje académico, se ha tenido la oportunidad de interactuar y mediante la práctica comprender una serie de recursos fundamentales para el procesamiento, almacenamiento y modelado de datos, donde inicialmente se plantea el conocimiento de un modelo relacional que es útil para el almacenamiento pero dificulta la generación de reportes, rendimiento en recursos al hablar de una arquitectura de modelado y eficiencia en la posibilidad de incurrir en la redundancia de datos y repetición de los mismos. Por lo tanto, se hace interesante conocer otras metodologías y aquí es donde se introduce la necesidad de construir un modelo de datos

dimensional a partir de una estructura relacional que permite análisis más eficientes, menor consumo de recursos en hardware y la implementación ligera de consultas.

Existen dos modelos dimensionales muy usados: el modelo estrella y el modelo copo de nieve, cada uno tiene diversas implicaciones, pero el más usado y el que será objeto de este trabajo es el modelo estrella, que permita el análisis de datos y transacciones facilitando la toma de decisiones mediante un esquema de datos centralizado y optimizado para consultas analíticas. (Sánchez, 2023)

En la actualidad, muchas organizaciones manejan múltiples fuentes de datos para recopilar información. Antes de incorporarlos al sistema final, es necesario procesar y depurar los datos, asegurándose de que tengan el formato y la estructura correctos. En este contexto, un área de almacenamiento temporal resulta esencial. Esta capa intermedia permite ajustar, duplicar, enlazar y, cuando es necesario, combinar datos, además de realizar procesos de limpieza para garantizar su calidad antes de la carga definitiva. La base de datos Staging es un entorno de trabajo intermedio dentro de un sistema de procesamiento de datos, diseñado para almacenar temporalmente datos provenientes de diversas fuentes antes de que sean transformados e integrados en un sistema de análisis, como un Data Warehouse o un modelo de datos específico, en este caso caso, un modelo estrella. Allí los datos son extraídos, transformados y validados con el fin de mejorar su calidad, resolver inconsistencias y garantizar que cumplen con los estándares requeridos antes de ser integrados al modelo dimensional. (Hevo Data, 2024).

Como parte siguiente a la elaboración del modelo dimensional de los datos, continúan una serie de fases que son fundamentales en el proceso de gestión de datos donde lo que se busca es recopilar datos de diversas fuentes en un conjunto único de datos para llevarlo al Datawerehouse, DataLake y/o otros. Estas fases se conocen en el ámbito actual de los datos como ETL, que no es más que

una metodología aplicada en gestión de datos de Business Intelligence describiendo un flujo de trabajo en un proceso organizado y coherente que permita visualizar insights, tendencias y patrones en procesos fundamentales para una empresa como las ventas, ganancias, inversiones y que puedan reflejarse en informes de valor para una toma de decisiones más precisas y coherentes con la realidad actual de la empresa. La palabra ETL corresponde a las siglas en inglés de: Extract (Extraer), Transform (Transformar) y Load (cargar), procesos que se utilizan para mezclar datos de múltiples fuentes. (IBM, s. F.).

Objetivo General:

A partir de un modelo dimensional tipo estrella eficiente crear una base de datos Staging para el ejercicio de extracción de datos de la base de datos "Jardinería"

Objetivos específicos

1. Identificar las tablas y relaciones relevantes en la base de datos "Jardinería" que permitan capturar toda la información pertinente a las ventas y transacciones de la empresa.
2. Construir un modelo estrella que centre los datos de ventas en una tabla de FACTS y proporcione tablas dimensionales adecuadas que enriquezcan la información de cada transacción.
3. Analizar los datos existentes en la base de datos “Jardinería” para diseñar la estructura de las tablas que compondrán la base de datos Staging, asegurando que permitan un almacenamiento temporal eficiente y funcional.
4. Crear consultas que extraigan los registros relevantes desde la base de datos “Jardinería” hacia la base de datos Staging, certificando integridad y precisión en la migración.

5. Ejecutar y validar las consultas para garantizar que los datos se carguen correctamente en la base de datos Staging.

Planteamiento del problema

La empresa enfrenta problemas en la generación de reportes y análisis sobre ventas y transacciones debido a la estructura de su base de datos, actualmente diseñada bajo un modelo relacional ya que es una forma estándar de representar y consultar datos (¿Qué es una Base de Datos Relacional?, s. f.). Este modelo es eficiente para el almacenamiento y gestión transaccional, pero no está optimizado para el análisis de datos, lo que resulta en consultas un poco más complejas debido a la necesidad de implementar una gran cantidad de los joins que van ampliando el recurso de almacenamiento. Ante esta situación, es importante contar con una estructura de datos que permita realizar consultas de forma ágil y obtener insights valiosos de manera oportuna para apoyar la toma de decisiones.

Por lo tanto, se requiere transformar el modelo relacional de la empresa, migrando a un modelo dimensional ya que son estructuras desnormalizadas diseñadas para recuperar datos de un almacén de datos (Datamart) y utilizan tablas de hechos y dimensiones para mantener un registro de datos históricos en almacenes de datos (Datawerehouse) (Ahmed, 2024). Este nuevo enfoque permitirá centralizar y optimizar los datos de las ventas de la empresa de Jardinería en una estructura adecuada mejorando el acceso a la información y facilitando la extracción de indicadores claves por medio de la implementación de un modelo estrella. Para llevar a cabo la construcción del modelo, es esencial crear una base de datos staging que facilite la depuración y preparación de los datos. Este espacio de almacenamiento permitirá filtrar los registros relevantes, asegurando que solo se procesen aquellos que sean necesarios. Además, servirá como una capa intermedia para almacenar los datos de manera temporal, antes de ser transferidos al modelo final. En este contexto,

es fundamental desarrollar consultas que faciliten la transferencia de datos de la base principal a la staging. También es crucial validar que los datos se almacenen correctamente, garantizando la integridad y precisión del proceso (Hevo Data, 2024).

Análisis de la problemática.

La base de datos “Jardinería” de la empresa, está organizada bajo un modelo relacional, lo cual implica que las tablas se encuentran normalizadas para asegurar una consistencia y reducir redundancia. Sin embargo, esta organización resulta poco eficaz al momento de la generación de reportes analíticos en cuanto a las ventas:

- La información sobre las ventas se encuentra distribuida en varias tablas como: pedido, detalle de pedido, producto, categoría de producto, oficina, pago, entre otras y están relacionadas mediante claves lo que implica realizar múltiples uniones o joins en las consultas.
- Estas consultas ralentizan el proceso de generación de reportes y dificultan el acceso de métricas estratégicas para la empresa como son las ventas y sus analíticas.

Para optimizar el análisis de datos y la toma de decisiones, es necesario implementar el modelo dimensional ya que se organiza la tabla de hechos (Facts) que contiene los datos cuantitativos de ventas y varias tablas de dimensiones que contiene los atributos descriptivos relacionados: Dim Producto, Dim Cliente, Dim Tiempo y Dim Empleado. Esto con el fin de centralizar la información de ventas, facilitar el análisis de los datos y la reducir la complejidad en las consultas, así pues, esta

estructura le permitirá a la empresa acceder a información relevante de manera más rápida y fortalecer la toma de decisiones estratégicas.

Descripción del modelo estrella propuesto.

El modelo estrella que se propone se define a partir de la creación de una tabla FACT _Ventas, la cual es la base de todas las consultas cuantitativas hacia las dimensiones (descriptivas-cualitativas) que en consenso grupal se decidieron tuvieran como columnas aquellas que tengan un aporte relevante y sean las variables insumo para comprender los indicadores necesarios para un cierre exitoso de ventas.

Así mismo la absorción o unificación de tablas conectadas por claves foráneas, que estén dentro de una estructura dimensional de jerarquía de datos en la tabla dimensión de mayor granularidad pueda a través de la conexión de dichas claves tener la posibilidad de generar unas respuestas a las consultas que se planteen mejorando su desempeño.

El modelo relacional de Jardinería constaba de 8 tablas relacionadas entre sí a partir de claves foráneas, unas tablas muy íntimamente relacionadas como la tabla clientes que desprendían de ella las tablas unidas entre sí: Pedido, detalle pedido, producto y categoría producto, de aquí pudimos detectar que era una de las tablas principales de dicho modelo relacional, pues de esta misma también salían otras tablas que no estaban relacionadas entre sí como: empleado, oficina y pago.

Al final el modelo dimensional planteado y basándonos en el concepto de jerarquía y granularidad la tabla DIM_PRODUCTO, quedó como aquella tabla que conecta a través de la clave foránea a la tabla detalle_pedido, así mismo la tabla DIM_PRODUCTO es quién tendrá relación con la tabla Categoria_producto que deja de aparecer en el modelo dimensional. En el caso de la tabla oficina, esta queda absorbida por la tabla DIM_EMPLEADO y la tabla pago en DIM_CLIENTE.

Propuesta de la solución: Correcciones Entrega 1

Teniendo en cuenta los comentarios realizados por el profesor en la primera entrega se establece un nuevo modelo dimensional con los ajustes donde se indicaba que la tabla FACT absorbe las tablas Dim_Pedido y Dim_Detalle de pedido, en el nuevo modelo dimensional estas tablas desaparecen del modelo y se crea la tabla de tiempos solicitada por el profesor como la tabla que contiene la información relacionada con las fechas de la base de datos, es importante detallar entonces que en este caso la tabla Fact adquiere una nueva clave foránea para la tabla que denominamos Dim_tiempos

En nuestro análisis consideramos que el modelo estrella planteado es el que tiene mayor coherencia con la definición de las tablas y columnas que pueden describir de manera detallada querys del proceso de venta para Jardinería.

Diseño (Imagen) del modelo estrella donde se puedan observar las dimensiones, la tabla de hechos, sus campos, tipos de datos y relaciones.

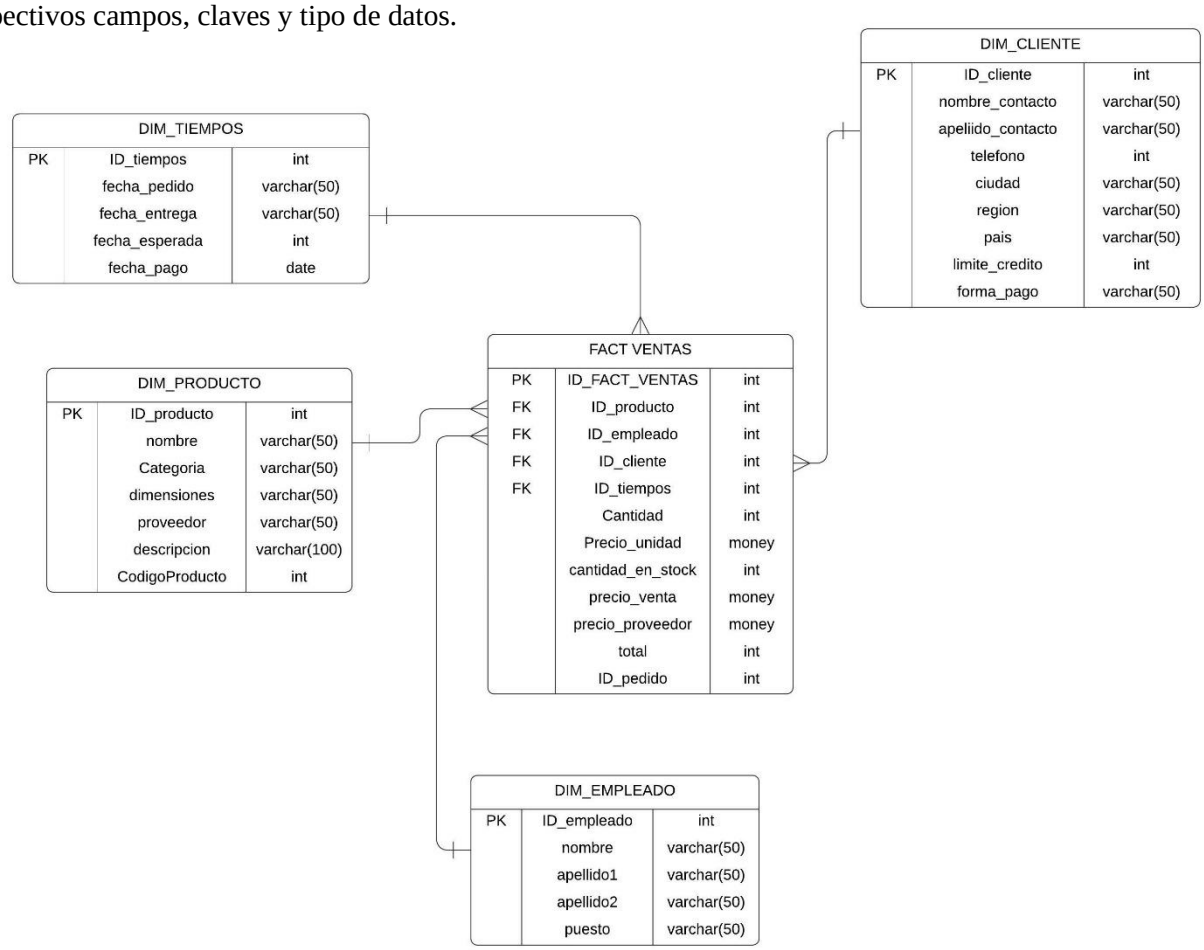
Lista de dimensiones propuestas.

Para la empresa se diseñó el modelo estrella con la base de datos de Jardinería y se tomó como base el modelo relacional para analizar e identificar qué dimensiones son las más pertinentes para análisis de las ventas, de esta manera las tablas dimensionales que se crearon fueron:

Dimensiones	Columnas	Tipo de datos
Producto	ID_producto, nombre, categoría, dimensiones, proveedor, descripción y código producto	Int y Varchar
Tiempo	ID_tiempo, fecha_pedido, fecha_entrega, fecha_esperada y fecha_pago	Date
Empleado	ID_empleado, nombre, apellido1, apellido2 y puesto	Int y Varchar
Cliente	ID_cliente, nombre_contacto, apellido_contacto, telefono, ciudad, region, pais, limite_credito y forma_pago	Int y Varchar

- **Detalla la tabla de hechos, con sus campos y tipos de datos.**

A continuación, se describe el modelo detallado de tablas DIMENSIONALES y de HECHOS con sus respectivos campos, claves y tipo de datos.



Descripción del análisis realizado a los datos Jardinería y cómo estos se trasladaron a la base de datos Staging.

Análisis de los datos realizado a la BD Jardinería

Se analizó la estructura del modelo relacional de la base de datos Jardinería para identificar las tablas relevantes al proceso de ventas, donde este análisis incluyó identificación de las tablas y datos importantes para el modelo y la evaluación de las relaciones entre ellas.

Así que, con el modelo dimensional ya implementado en la primera entrega, donde se incluyeron las tablas Dim_producto, Dim_tiempo, Dim_cliente, Dim_empleado y la tabla de fact_ventas donde esta nos permitió identificar y definir los campos específicos que se trasladan desde el origen (base de datos relacional Jardinería) hacia la base staging, por lo que es importante tener unos criterios para la selección de los campos como la relevancia analítica ya que se seleccionan únicamente los campos que son relevantes para las métricas e indicadores definidos en el modelo dimensional; la integridad y consistencia debido a que se deben priorizar los campos más importantes para el análisis de las ventas y haya una integridad entre las tablas en el origen; y el desempeño ya que esto permite que no tengan campos redundantes o irrelevantes que puedan incrementar la carga del procesamiento, por lo tanto, este proceso nos asegura que solo se transporten y transformen los datos necesarios para una estructura similar al modelo dimensional, optimizando la carga y preparación de datos para el análisis.

Traslado a la base de datos Staging

Este proceso es clave para la integración y preparación de datos para el modelo dimensional de la empresa. Este proceso asegura que los datos extraídos del modelo relacional original sean

transformados, limpiados y organizados para adaptarse a las estructuras requeridas en las tablas del modelo dimensional con las tablas generadas de producto, tiempo, cliente, empleado y ventas. La base Staging se diseñó para reflejar una versión preliminar del modelo dimensional, asegurando que los datos estuvieran organizados de forma consistente antes de la carga definitiva, por lo tanto, la selección de campos desde el origen se basó en un análisis detallado de las tablas y relaciones del modelo relacional, asegurando que únicamente los datos relevantes y necesarios fueran trasladados.

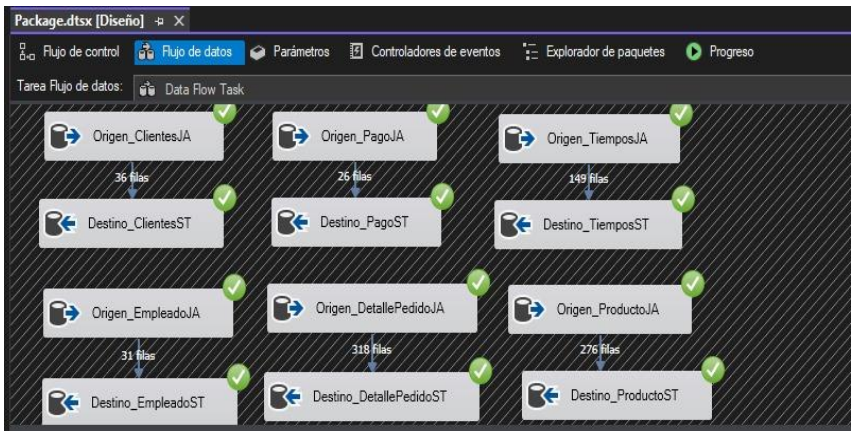
Las tablas generadas en origen y destino son las siguientes:

Tabla	Origen	Destino
Cliente	**ClientesJA**: ID cliente, nombre contacto, apellido contacto, teléfono, ciudad, región, país y limite crédito.	**ClientesST:** ID cliente, nombre contacto, apellido contacto, teléfono, ciudad, región, país y limite crédito.
Pago	**PagoJA:** ID pago, ID cliente, forma pago, fecha pago y total	**PagoST:** ID pago, ID cliente, forma pago, fecha pago y total
Empleado	**EmpleadoJA:** ID empleado, nombre, apellido1, apellido2 y puesto.	**EmpleadoST:** ID empleado, nombre, apellido1, apellido2 y puesto.
Detalle pedido	**DetallePedidoJA:** ID detalle pedido, ID pedido, cantidad, precio unidad, número línea.	**DetallePedidoST:** ID detalle pedido, ID pedido, cantidad, precio unidad, número línea.

Producto	**ProductoJA:** ID producto, código producto, nombre, categoría, dimensiones, proveedor, descripción, cantidad en stock, precio venta y precio proveedor	**ProductoST:** ID producto, código producto, nombre, categoría, dimensiones, proveedor, descripción, cantidad en stock, precio venta y precio proveedor
Pedido (solo se toman las fechas, se hace unión y crea en origen tiempo)	**TiemposJA:** fecha pedido, fecha esperada y fecha entrega.	**TiemposST:** fecha pedido.

Nota: La columna fecha de pago no se cargó en la tabla de tiempo porque no contábamos con las herramientas de conocimiento de SQL y Visual Studio que nos permitan extraer de varias tablas datos como las fechas que se puedan almacenar en una sola tabla que en este caso sería Tiempos en Staging, esto lo intentamos hacer en varias opciones, pero siempre nos presentaba error porque nos decía que no podía crear una tabla que ya se había creado.

A continuación se da cuenta de la creación de la base de datos de Statging con sus tablas y la creación de las mimas de forma completa desde Visual Studio.



Este proceso de Staging permitió realizar las validaciones sin afectar la base dimensional final con datos no importantes o duplicados, además todas las correcciones y enriquecimientos aseguraron que los datos cargados al modelo estuvieran listos para consultas analíticas rápidas y precisas, generando eficiencia al consolidar y desnormalizar los datos en esta etapa, y una mayor reducción de la carga de procesamiento en el Data Warehouse.

Conclusiones

- Los modelos dimensionales permiten una mejor visualización y comprensión del contexto con el que se definen las tablas que van a conformar la base de datos.
- Hablar de modelos dimensionales permite tener un mayor enfoque al momento de realizar las consultas disminuyendo joins y procedimientos más extensos que pueden llevar a errores.
- Para definir un modelo relacional es necesario plantear las relaciones entre tablas, y por ello es importante contar con un modelo Entidad Relación.
- La selección de campos desde el origen hacia staging debe seguir un enfoque sistemático que priorice la calidad, relevancia y adecuación de los datos al modelo dimensional.
- Las transformaciones realizadas en staging aseguran que los datos lleguen a su destino final siguiendo el modelo dimensional en un estado óptimo para soportar consultas de interés según la información más analíticas rápidas y confiables.
- La base Staging garantizó que los datos del modelo relacional fueran extraidos y preparados de manera eficiente y confiable, alineándose con las necesidades del modelo

dimensional, lo que asegura la calidad de la información y que el sistema pueda responder adecuadamente a las demandas analíticas y toma de decisiones.

- La optimización de las consultas SQL y la indexación adecuada de las tablas en la base de datos staging contribuyeron a acelerar los procesos de transformación y carga de datos, mejorando el rendimiento general del sistema.
- La construcción de la base de datos staging y la implementación del proceso ETL en su primera fase de extracción ha sido fundamental para que en las fases de transformación se pueda garantizar la calidad, consistencia y disponibilidad de los datos necesarios para soportar las operaciones de análisis y toma de decisiones de la empresa.

Referencias

Sánchez, M. G. (2023, julio 19). ¿Qué es el Modelo Estrella? Tecon. https://www.tecon.es/que-es-el-modelo-estrella/

Hevo Data. (2024, octubre 23). What is staging area in data warehouse?. Hevo Data. https://hevodata.com/

¿Qué es una base de datos relacional? (s.f). Oracle.com. Obtenido de https://www.oracle.com/co/database/what-is-a-relational-database/

IBM. (s. f.). ¿Qué es ETL (extracción, transformación, carga)? IBM. https://www.ibm.com/mx-es/topics/etl

Ahmed, I. (2024, junio 07). What is dimensional data modeling? Examples, benefits & more. Astera. https://www.astera.com/es/knowledge-center/dimensional-modeling-guide/

Datha Learning. (2022, diciembre 12). Introducción al Modelado Dimensional. YouTube. https://youtu.be/Ewh8WsCOFQ4?feature=shared