Proceso de transformación de datos y carga en el data mart final

Presentado por:

Juan Esteban Casadiego & Evelyn Cerro Acuña

Ingeniería de software y datos, Institución Universitaria Digital de Antioquia

Bases de Datos II

Entregado a:

Antonio Jesús Valderrama

Barranquilla, Atlántico

2025

Tabla de contenido

Introducción

En el contexto actual, donde la calidad y la confiabilidad de la información determinan la efectividad de los procesos analíticos y de inteligencia de negocios, contar con entornos intermedios de preparación de datos resulta esencial. Entre estos, las bases de datos Staging se han consolidado como un componente crítico en los procesos de integración y migración de información, al permitir la depuración, validación y organización de los datos antes de su carga en sistemas de análisis. De acuerdo con Kimball y Caserta (2004), “el área de staging constituye el espacio de trabajo donde los datos son extraídos, transformados y preparados para su posterior explotación analítica, garantizando su consistencia y trazabilidad”.

El presente trabajo tiene como objetivo diseñar y construir una base de datos Staging a partir de la base transaccional Jardinería, que sirva como repositorio temporal para asegurar la integridad y confiabilidad de los datos antes de ser utilizados en procesos de análisis o migración hacia un data warehouse.

Se realizó un análisis exhaustivo de la estructura de la base de datos Jardinería, identificando las tablas relevantes, así como sus relaciones de dependencia. A partir de este estudio se definió la estructura de la base Staging, replicando las claves primarias y foráneas, y añadiendo columnas técnicas para el control de cargas. Posteriormente, se desarrollaron y ejecutaron consultas SQL que trasladaron la información de la base transaccional hacia la base Staging, validando la consistencia de los datos mediante comparaciones de conteo e integridad referencial.

Este trabajo se justifica en la necesidad de contar con un entorno controlado que actúe como filtro de calidad de datos, lo cual facilita posteriores procesos de análisis, evita errores en la información consolidada y mejora la gestión estratégica de la organización. Como señalan Coronel y Morris (2017), “los procesos de preparación de datos en entornos intermedios son indispensables para garantizar que la información final en un almacén de datos sea confiable, precisa y útil para la toma de decisiones”.

Objetivo general

Diseñar y ejecutar un proceso de transformación y carga de datos desde la base de datos Staging hacia un Data Mart bajo el modelo estrella de la base Jardinería, garantizando la calidad, consistencia e integridad de la información para su posterior análisis empresarial.

Objetivos específicos

A. Revisar el modelo estrella propuesto para identificar las dimensiones y la tabla de hechos necesarias en el Data Mart.

B. Verificar la disponibilidad y consistencia de los datos en la base Staging como punto intermedio del proceso ETL.

C. Desarrollar consultas SQL que realicen la transformación, limpieza y normalización de los datos provenientes de Staging.

D. Implementar la carga de registros en las tablas de dimensiones y en la tabla de hechos del Data Mart final.

E. Validar la inserción de datos mediante conteos, chequeos de calidad y consultas de verificación.

F. Elaborar documentación del proceso ETL, detallando las etapas realizadas, las consultas aplicadas y los resultados obtenidos.

Planteamiento del problema

En la actualidad, las organizaciones requieren infraestructuras analíticas que permitan transformar grandes volúmenes de datos en información confiable para la toma de decisiones estratégicas. De acuerdo con Kimball y Ross (2013), “los modelos dimensionales, como el esquema en estrella, proporcionan una estructura simple e intuitiva para organizar los datos, facilitando el análisis y mejorando la eficiencia de las consultas”. Sin embargo, para que estos modelos cumplan su propósito, es indispensable contar con procesos sistemáticos de extracción, transformación y carga (ETL) que aseguren la calidad de la información antes de integrarla en un Data Mart.

En el caso de la base de datos Jardinería, su estructura transaccional fue diseñada para registrar operaciones diarias —como clientes, pedidos, productos, oficinas y empleados—, pero no está orientada a la explotación analítica. Esto genera dificultades al consolidar información, ya que pueden existir redundancias, registros incompletos o dependencias que impiden un análisis eficiente. Además, la ausencia de un proceso definido de transformación de datos limita la posibilidad de obtener indicadores clave, como el producto más vendido o el desempeño de las ventas en diferentes periodos.

Según Coronel y Morris (2017), “la preparación y depuración de los datos en entornos controlados resulta indispensable para garantizar su consistencia y confiabilidad en almacenes de datos”. En este contexto, la falta de un flujo de transformación y carga hacia un Data Mart restringe la capacidad de la organización para aprovechar el potencial de sus datos en la toma de decisiones.

Por lo tanto, surge la necesidad de diseñar e implementar un proceso ETL que traslade los datos desde la base transaccional Jardinería hacia un Data Mart construido bajo un modelo en estrella, utilizando como entorno intermedio la base Staging. Este proceso permitirá validar la integridad, aplicar controles de calidad y cargar las dimensiones y la tabla de hechos

requeridas para el análisis, garantizando así un repositorio confiable y orientado al soporte estratégico de la organización.

Análisis del problema

El análisis de la base de datos Jardinería evidencia que, aunque el uso de la base Staging facilita el almacenamiento intermedio de la información, persisten retos importantes cuando se requiere garantizar la consistencia y calidad de los datos en su traslado al Data Mart final. Dichas dificultades se relacionan con la necesidad de transformar los registros y ajustarlos a la estructura del modelo en estrella, compuesto por dimensiones y una tabla de hechos central.

En primer lugar, la correspondencia entre tablas de staging y las dimensiones del Data Mart exige un control estricto de integridad. Relaciones como la de clientes con sus representantes de ventas o la de pedidos con clientes y fechas deben asegurarse en el proceso de carga, de lo contrario podrían generarse claves nulas o registros huérfanos que comprometan la calidad del análisis posterior.

En segundo lugar, la tabla de empleados presenta una complejidad especial debido a la jerarquía definida por el campo ID_jefe. Este tipo de dependencia autorreferencial requiere un manejo cuidadoso durante la transformación hacia la dimensión de empleados, ya que un error en la carga podría romper la relación entre jefes y subordinados, afectando la trazabilidad de la información organizacional.

En tercer lugar, la integración de las dimensiones de fechas, productos y pedidos con la tabla de hechos demanda la alineación de claves sustitutas (surrogate keys). Si no se controlan estos mapeos entre las claves naturales de staging y las claves artificiales del Data Mart, es posible que los hechos queden sin vinculación a las dimensiones correspondientes, limitando el análisis de indicadores como las ventas totales o el producto más vendido.

Finalmente, la ausencia de un proceso sistemático de validación después de la carga hacia el Data Mart puede ocultar inconsistencias en medidas críticas, como la cantidad y el precio de los pedidos. Cargar estos registros de forma directa sin verificaciones implicaría exponer el esquema analítico a errores que afectarían la confiabilidad de los reportes.

Estos factores evidencian que, aunque la base Staging permite organizar la información de Jardinería, no es suficiente sin un proceso robusto de transformación y carga hacia el Data Mart. Es necesario implementar rutinas ETL que garanticen la depuración de los datos, el respeto de las relaciones entre entidades y la correcta integración de dimensiones y hechos. Solo de esta forma el modelo estrella podrá consolidarse como un entorno confiable para apoyar la toma de decisiones estratégicas de la organización.

Propuesta de la solución

La construcción de un Data Mart bajo el modelo en estrella para la empresa Jardinería tiene como propósito central garantizar que la información disponible para los procesos analíticos sea confiable, consistente y lista para apoyar la toma de decisiones estratégicas. Este repositorio analítico se alimenta a partir de la base Staging, donde previamente los datos fueron depurados, organizados y validados, para luego ser transformados e integrados en dimensiones y hechos.

La estructura del Data Mart se compone de una tabla de hechos (ventas) y múltiples tablas de dimensiones —clientes, productos, empleados, oficinas, pedidos y fechas— que permiten organizar la información de manera intuitiva y orientada al análisis de negocio. Durante el proceso de carga se emplearon claves sustitutas para mantener la integridad referencial y asegurar la trazabilidad de cada registro.

Este modelo simplifica y controla el proceso de migración de datos, ya que permite:

- Relacionar dimensiones clave (clientes, pedidos, productos, empleados y oficinas) con la tabla de hechos, asegurando la consistencia de los análisis.
- Manejar dependencias jerárquicas complejas, como la relación entre empleados y sus jefes, mediante reglas de carga que preservan la integridad.
- Detectar y controlar posibles inconsistencias (valores nulos, claves huérfanas o duplicadas) durante el proceso ETL, evitando que lleguen al esquema analítico.
- Realizar análisis estratégicos, como la identificación del producto más vendido o la evaluación de ventas por categoría y periodo, gracias a la estructura dimensional.

De esta manera, el Data Mart no solo resuelve las limitaciones del modelo transaccional y del entorno Staging, sino que también consolida la información en un formato óptimo para el análisis multidimensional. En consecuencia, la empresa Jardinería podrá disponer de un repositorio que garantice la confiabilidad de sus datos y que brinde soporte efectivo a la gestión estratégica y la competitividad organizacional.

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

El análisis de los datos de la base Jardinería permitió evidenciar que, si bien la información ya se encontraba organizada en la base Staging, era necesario aplicar un proceso de transformación y carga hacia un entorno analítico diseñado bajo el modelo en estrella. Este modelo se compone de una tabla de hechos central (fact_venta) y diversas dimensiones que permiten contextualizar los registros: cliente, empleado, oficina, producto, pedido y fecha.

La revisión inicial de la base Staging se enfocó en identificar las entidades clave para el Data Mart y verificar la calidad de los registros. Para ello, se realizaron consultas de validación orientadas a detectar inconsistencias como clientes sin representante de ventas, pedidos sin cliente asociado, productos sin referencia válida o empleados con relaciones jerárquicas incompletas. Estos chequeos aseguraron que los datos de Staging eran adecuados para ser integrados en un modelo dimensional.

Posteriormente, se definió la estructura del Data Mart jardineria_dm, creando las tablas de dimensiones y la tabla de hechos con sus respectivas claves sustitutas (surrogate keys). Este diseño garantizó la independencia entre las dimensiones y permitió mantener la trazabilidad mediante la relación de claves naturales con las tablas de Staging.

Posteriormente, se diseñó la base jardineria_stg, replicando la estructura de las tablas originales y definiendo en cada caso sus claves primarias y foráneas, con el fin de mantener la integridad referencial. Se añadieron además columnas técnicas como ETL_LoadDate y ETL_Source, destinadas a registrar la fecha de carga y la procedencia de los datos, aportando trazabilidad al proceso.

El proceso de carga se llevó a cabo mediante consultas MERGE y INSERT … SELECT, siguiendo un orden lógico que respetó las dependencias entre entidades:

1. Dimensión fecha, construida a partir de las fechas de pedido, entrega y esperada registradas en Staging.
2. Dimensión oficina, como entidad independiente.
3. Dimensión empleado, cargada con especial cuidado en la relación jerárquica del campo

ID_jefe.

4. Dimensión cliente, vinculada a los representantes de ventas (empleados).
5. Dimensión producto, relacionada con su categoría y con control de descripciones largas.
6. Dimensión pedido, asociada a las fechas de pedido, entrega y esperada.
7. Tabla de hechos (fact_venta), integrada a partir de detalle_pedido, relacionando cada registro con sus claves sustitutas de cliente, producto, empleado, oficina, pedido y fecha.

Finalmente, se realizaron conteos comparativos entre Staging y el Data Mart, así como validaciones de integridad en la tabla de hechos para garantizar que no existieran claves nulas. Adicionalmente, se generó una vista analítica (vw_producto_top) que permitió identificar los productos más vendidos, demostrando la utilidad del modelo en estrella para responder preguntas estratégicas de negocio.

En conclusión, el proceso de análisis y traslado aseguró que los datos de Jardinería fueran transformados y cargados en un Data Mart estructurado, confiable y optimizado para la explotación analítica, constituyéndose en una base sólida para la toma de decisiones empresariales.

Conclusiones

- El diseño e implementación del Data Mart bajo el modelo en estrella para la empresa Jardinería constituye una solución eficaz para garantizar la organización, calidad y consistencia de la información antes de ser utilizada en procesos analíticos y estratégicos.
- Permite integrar de manera coherente las dimensiones de clientes, empleados, oficinas, productos, pedidos y fechas en torno a una tabla de hechos central, lo que asegura la integridad referencial y la trazabilidad de los datos en los análisis.
- Optimiza el proceso de transformación y carga de información al establecer un flujo ETL estructurado que detecta y corrige inconsistencias, evitando que registros incompletos, huérfanos o duplicados afecten los resultados del entorno analítico.
- Fortalece la trazabilidad y confiabilidad del proceso al implementar claves sustitutas en las dimensiones, garantizando la independencia de los datos transaccionales y facilitando la gestión de cambios en el tiempo.
- Contribuye a la eficiencia en la explotación de la información, ya que el modelo en estrella simplifica las consultas analíticas y habilita reportes estratégicos, como la identificación del producto más vendido y la evaluación de las ventas en diferentes periodos y categorías.
- En conclusión, la construcción del Data Mart de Jardinería no solo asegura la confiabilidad de los datos para el análisis estratégico, sino que también fortalece la infraestructura tecnológica de la organización, brindando una base sólida para la toma de decisiones fundamentadas en información precisa, consistente y orientada al negocio.

Bibliografía

- Coronel, C., & Morris, S. (2017). *Database systems: Design, implementation, & management* (12th ed.). Cengage https://www.cengage.com/c/database-systems-design-implementation-management-12e-coronel/9781337627900
- *Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data.* Wiley. https://www.wiley.com/en-us/The+Data+Warehouse+ETL+Toolkit-p-9780764567575
- Microsoft. (2023). *CREATE DATABASE (Transact-SQL)*. Microsoft Learn. https://learn.microsoft.com/es-es/sql/t-sql/statements/create-database-transact-sql
- Microsoft. (2023). *ALTER TABLE (Transact-SQL)*. Microsoft Learn. https://learn.microsoft.com/es-es/sql/t-sql/statements/alter-table-transact-sql
- Microsoft. (2023). *INSERT (Transact-SQL)*. Microsoft Learn. https://learn.microsoft.com/es-es/sql/t-sql/statements/insert-transact-sql
- Microsoft. (2023). *BACKUP DATABASE (Transact-SQL)*. Microsoft Learn. https://learn.microsoft.com/es-es/sql/t-sql/statements/backup-database-transact-sql