ETL Superstore

(Extracción, Transformación y Carga)

Por:
Ysabel Mata

Objetivo!

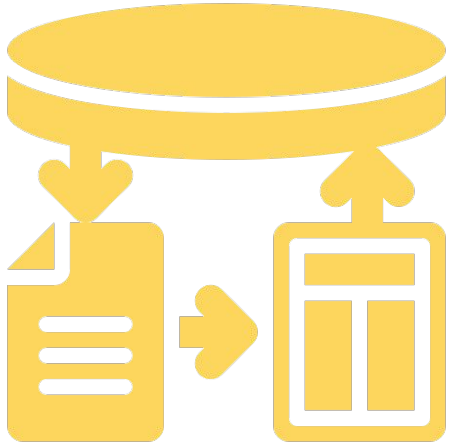
A través del proceso ETL, construir un sistema tabular para la tienda superstore que nos permita almacenar datos de manera eficiente y consultar estos datos más fácilmente.

Metodologia !

ETL (Extracción, Transformación y Carga)

Proceso de tres fases: Extracción (Extraction), Transformación (Transformation) y Carga (Load). Este proceso se utiliza comúnmente en el ámbito de la gestión y análisis de datos, especialmente en el contexto de data warehouses y business intelligence.

- **Extracción:** Los datos se extraen desde una o varias fuentes de datos, que pueden ser bases de datos, archivos planos, servicios web u otras fuentes. La extracción implica recopilar la información necesaria para su posterior procesamiento.
- **Transformación:** En esta etapa, los datos extraídos se transforman según los requisitos del sistema de destino. Las transformaciones pueden incluir limpieza de datos, conversión de formatos, combinación de datos de múltiples fuentes, filtrado y otras operaciones que aseguran que los datos sean coherentes y útiles para el análisis.
- **Carga:** La fase final implica cargar los datos transformados en el sistema de destino, que generalmente es un data warehouse o una base de datos diseñada para el análisis de negocios. Los datos ahora están listos para ser consultados y analizados de manera eficiente.



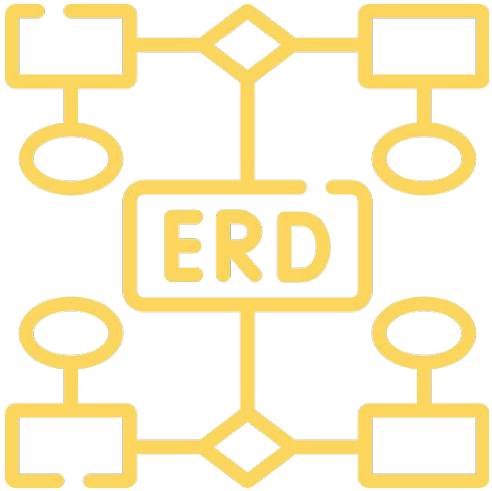
Modelo Entidad Relación (ER)

Un modelo entidad-relación (ER) es una representación gráfica utilizada para diseñar y visualizar la estructura de una base de datos. Se utiliza para describir cómo los datos están organizados y las relaciones entre ellos.

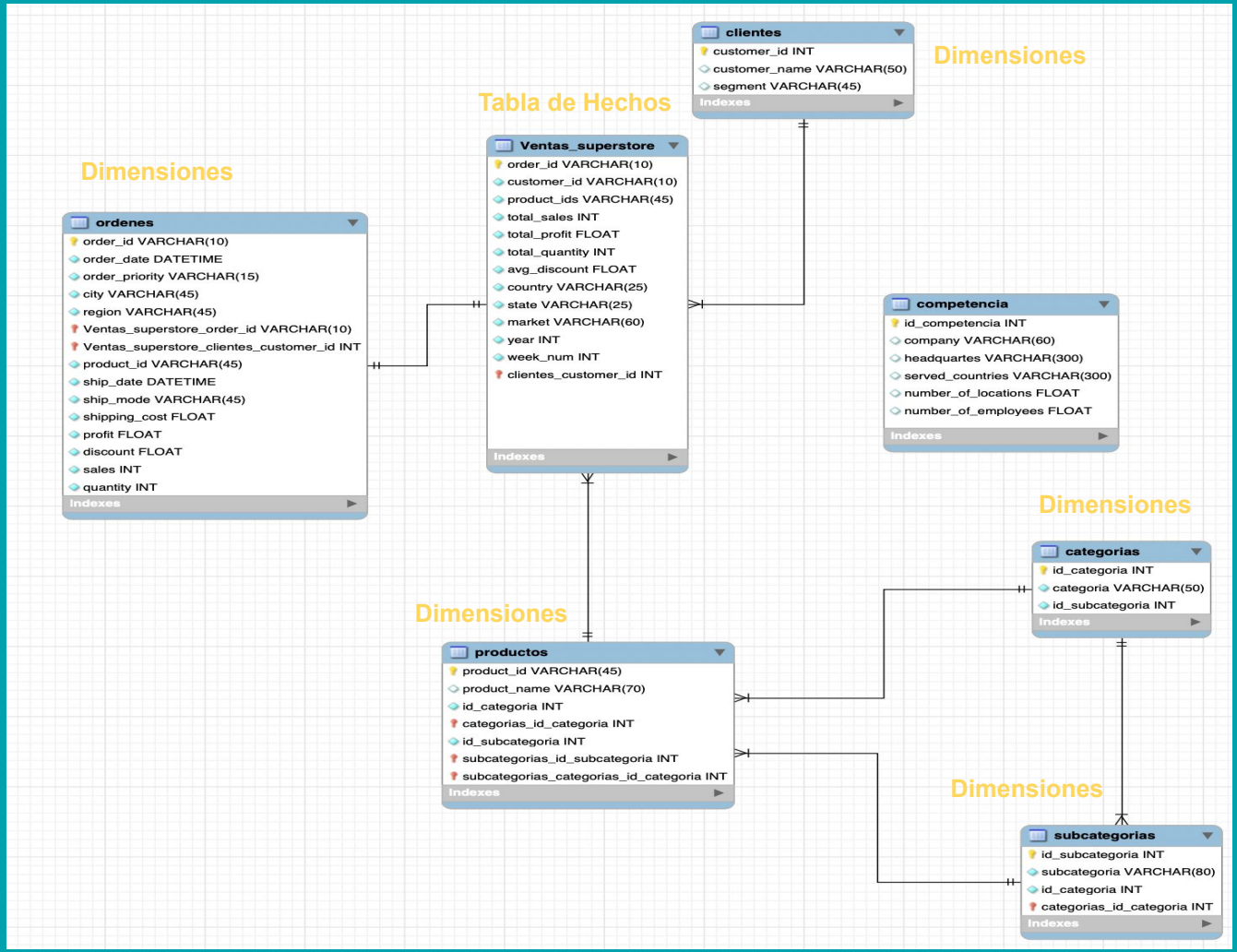
Componentes principales:

- **Entidades:** Representan objetos o elementos que tienen existencia propia dentro del sistema (ej. "Clientes", "Productos", "Órdenes"). Se muestran como rectángulos.
-
- **Atributos:** Son las propiedades o características de las entidades (ej. un "Cliente" tiene atributos como nombre, dirección, etc.). Se muestran como óvalos conectados a las entidades.

- **Relaciones:** Definen cómo las entidades están conectadas entre sí (ej. "Un cliente realiza una orden"). Se representan mediante rombos o líneas que unen las entidades.



Modelo ER Supestore



Flujo de Actualización

1. Clientes → 2. Categorias → 3. Subcategorias → 4. Productos → 5. Órdenes → 6.ventas_superstore

Descripción del Flujo de Actualización

Descripción de las Fuentes de Datos:

Fuente 1: CSV de ventas de superstore

Datos: Ordenes, ventas, clientes, productos.

Frecuencia de actualización: Diaria.

Formato: CSV con columnas: category, city ,country ,customer_ID, customer_name, discount, market, unknown, order_date, order_id, order_priority, product_id, product_name, profit, quantity, region row_id, sales, segment, ship_date, ship_mode, shipping_cost, state, sub_category, year, market2, weeknum.

Fuente 2: CSV de Competencia

Datos: Información de empresas competidoras.

Frecuencia de actualización: Mensual.

Formato: CSV con columnas: company, headquarters, number_of_employees, number_of_locations, served_countries.

Fuente 3: API de envíos

Datos: Información sobre el envío de órdenes.

Frecuencia de actualización: Continua.

Formato: JSON con campos : order_id,order_date, ,order_priority, product_id, quantity, sales,profit, discount,city,region,ship_date, ship_mode, shipping_cost

Proceso de Extracción

Fuente 1: CSV de Ventas

Proceso: El archivo CSV se obtiene desde un sistema de gestión de superstore que exporta datos de ventas a diario.

Herramienta ETL: Se configura una conexión con la carpeta donde se almacenan los archivos CSV diarios.

Fuente 2: CSV de Competencia

Proceso: La información se extrae una vez al mes desde un archivo CSV proporcionado por el departamento de inteligencia de mercado.

Herramienta ETL: La extracción del CSV se realiza automáticamente una vez al mes.

Fuente 3: API de Envíos

Proceso: Se conecta a la API de envíos y extrae en tiempo real los datos de las órdenes enviadas.

Herramienta ETL: La herramienta ETL realiza consultas periódicas a la API para obtener actualizaciones.

Proceso de Transformación

En esta etapa, los datos crudos se normalizan y preparan para la carga en las tablas de hechos y dimensiones. Se aplican las siguientes transformaciones:

Normalización de Fechas

Convertir el formato de fecha de las fuentes de ventas y envíos al formato estándar YYYY-MM-DD para asegurar la consistencia en la base de datos.

order_date, ship_date.

Destino: Tabla de hechos ventas_superstore y tabla ordenes.

Limpieza de Datos Duplicados

Identificar y eliminar registros duplicados de ventas, clientes y productos. Validar que no haya registros repetidos en el CSV de ventas.

Origen: CSV de Ventas.

Destino: Tabla de hechos ventas_superstore, tabla de dimensión clientes , productos, categorías, subcategorías.

Cálculo de Nuevas Métricas

Se agrupan todos los productos de una misma orden de la tabla ordenes (Concatenar productos en una sola columna) para obtener la columna product_ids de la tabla ventas_superstore, se suman las ventas por orden de la columna sale de la tabla ordenes y se obtiene la columna total_sale de la tabla ventas_superstore, se suma el profit de la tabla ordenes para obtener el total_profit de la tabla ventas_superstore, se suma quantity de la tabla ordenes para obtener total_quantity de la tabla ventas_superstore.

Origen: CSV de Ventas.

Destino: Tabla de hechos ventas_superstore.

Mapeo de Relaciones

Asegurar que los campos customer_id, product_id, order_id, id_categoria, id_subcategoria estén correctamente mapeados y no existan valores nulos o no referenciados. Se debe verificar la consistencia de las relaciones entre productos, categorías, subcategorías y clientes.

Origen: CSV de Ventas y CSV de Competencia.

Destino: Tablas de dimensiones (clientes, productos, categorías, subcategorias, competencia).

Actualización de Categorías y Subcategorías

Verificar que cada producto esté correctamente asignado a una subcategoría y que las subcategorías estén asociadas a la categoría correspondiente. Si algún producto no tiene una categoría o subcategoría asignada, se debe generar un log de error para su corrección.

Origen: CSV de Productos.

Destino: Tabla de dimensión productos, categorías, y subcategorías.

3. Proceso de Carga

Carga de la Tabla de Hechos: ventas_superstore, carga: Diaria.

Carga de las Dimensiones:

Clientes: Diaria, tras la limpieza de datos de clientes.

Productos: Diaria.

Categorías y Subcategorías: Diaria.

Competencia: Mensual.

4. Proceso de Automatización (Frecuencia de Actualización)

Ventas y Clientes: Diaria.

Competencia: Mensual.

Ordenes: Continua.

Herramientas de Automatización:

Las tareas ETL diarias se automatizan utilizando herramientas como Talend o Airflow. Estas herramientas permiten ejecutar scripts ETL de forma programada y en función de eventos (como la llegada de un nuevo archivo CSV).

Manejo de Errores:

Se configuran alertas para notificar al equipo en caso de fallos en la actualización o transformación de datos. Los errores se registran y almacenan para facilitar su depuración.

5. Monitoreo y Control de Calidad

Se implementan controles de calidad para asegurarse de que los datos estén correctos y actualizados antes de cargarlos.

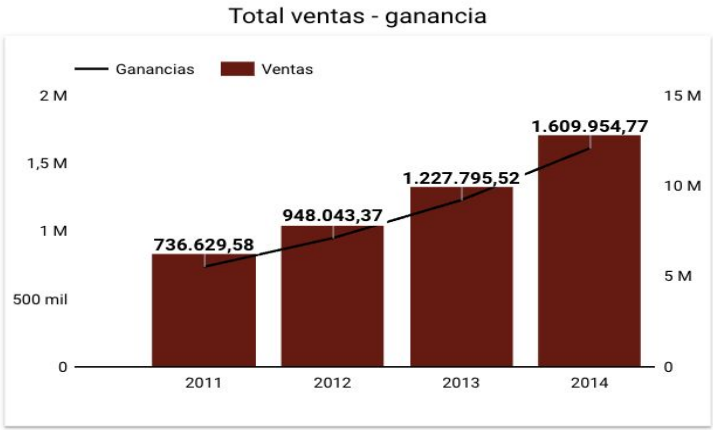
Se mantiene un log de las actualizaciones para auditar el proceso de ETL y corregir cualquier error que ocurra durante el proceso.

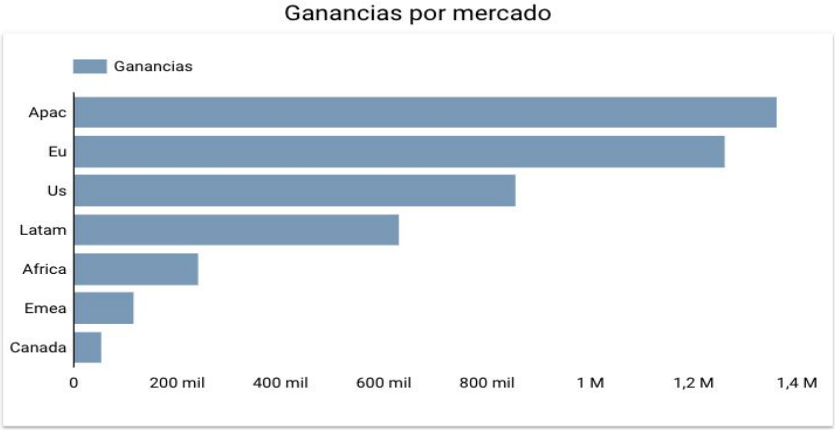
Análisis!

Superstore

Mercados	Total ordenes	Total ventas	Total ganacias	AVG descuentos	Categorias	Subcategorias
7	25.753	39,4 M	4,5 M	3.641	3	17

Año | Market | Segmento | Categoria | Subcategoria





Medidas de Tendencia central y Dispersión

Mln.Ganancia	Max.Ganancia	AVG Ganancia	STD Ganancia	Min.Ventas	Max.Ventas	AVG Ventas	STD Ventas
-22.746,98	43.811,95	165,58	1.078,17	1	165.620	1.440,84	3.708,79

Conclusiones

Usar una herramienta ETL o plataforma de integración de datos simplifica y automatiza el flujo de actualización de tu modelo de base de datos. En lugar de manejar todo manualmente, puedes crear pipelines de datos visuales, programar tareas recurrentes y centralizar el monitoreo de tus flujos de actualización. Esto mejora la eficiencia y reduce el riesgo de errores humanos en procesos repetitivos.

Recomendaciones

- Actualizaciones continuas en tiempo real: Esto es clave para análisis críticos que requieren decisiones inmediatas, como monitoreo de rendimiento o seguimiento en tiempo real. Es importante implementar mecanismos que soporten actualizaciones constantes, como triggers o replicación de datos.

- Asegurar la consistencia: Es fundamental garantizar que no haya inconsistencias entre la base de datos original y las actualizaciones. Mecanismos como control de transacciones y verificación ayudan a mantener la integridad de los datos en sistemas de alta concurrencia.

- Monitoreo de procesos: Implementar auditorías para monitorear el éxito de las cargas de datos es crucial. Esto asegura la identificación de errores, como fallas en la carga o datos incompletos, antes de que afecten el análisis.

Monitoreo de procesos: Implementar procesos de auditoría para revisar el éxito de las cargas de datos. Un buen monitoreo permitirá detectar problemas (como cargas incompletas o fallas en los datos) antes de que afecten el análisis.

Documentar el flujo ETL: Documentar las fuentes de datos, las transformaciones realizadas y el flujo de actualización para asegurar que cualquier cambio en el futuro pueda ser realizado de manera controlada.

Probar el sistema con datos de ejemplo: Antes de implementar el ETL y el flujo de actualización en un entorno de producción, probar con datos de ejemplo para asegurar que el sistema funcione correctamente.

Optimizar las consultas y procesamiento: A medida que los datos aumenten, es crucial optimizar las consultas SQL, el uso de índices y la estructura de la base de datos para mejorar la eficiencia del flujo ETL.

Escalabilidad: Asegúrarse de que el flujo ETL y de actualización sean escalables. Esto es esencial si se espera que el volumen de datos aumente con el tiempo.

Seguridad: Proteger los datos sensibles durante todo el proceso, especialmente si se está trabajando con datos personales o confidenciales. Implementar encriptación y controles de acceso.

GRACIAS !

Link al análisis

https://lookerstudio.google.com/reporting/afa01aad-e961-4175-8125-81fb18df782e