Universidad Galileo de Guatemala

Diseño y Construcción de Data Warehouses

Ing. José Rolando Lucero Morales

Proyecto: Construcción de Data Warehouse del Supermercado

Integrantes:

Homero Javier Castañón Morán	24008059
Diego Fernando Álvarez Muñoz	24003874
Oswaldo René López Aquino	18001257

Guatemala, 05 de marzo de 2024

i. Índice

ii. Índice de Figuras

1. Resumen

Se procedió a analizar la información del archivo SuperStoreOutput.csv, con la finalidad de construir el Data Warehouse para el negocio de una “super store”, la granularidad se definió por transacción de venta, dentro de lo dimensión se detallan en Producto, País, Cliente, Modo de Envío y Fecha. Además, permite segmentar las ventas por categoría de productos, marca, precio, entre otros atributos, a nivel de país, cliente o producto.

2. Introducción

El diseño y la implementación de un Data Warehouse es una herramienta fundamental para la toma de decisiones informadas y estratégicas. Este documento presenta el desarrollo de un Data Warehouse específicamente diseñado para un supermercado, utilizando como punto de partida el análisis de la información contenida en el archivo SuperStoreOutput.csv.

El objetivo principal de este proyecto es diseñar un Data Warehouse que permita analizar y comprender en profundidad las transacciones de venta realizadas por el supermercado. Para lograr este propósito, se aplicarán los conocimientos adquiridos en materia de granularidad y dimensiones, siguiendo las directrices propuestas por Kimball y Ross (2013) en su metodología de diseño de Data Warehouses impartidas en clase.

En cuanto a las dimensiones seleccionadas para la construcción de la tabla de hechos, se han identificado cinco elementos clave: Producto, País, Cliente, Modo de Envío y Fecha. Cada una de estas dimensiones desempeña un papel fundamental en el análisis y la comprensión de las transacciones de venta, proporcionando información detallada sobre aspectos como el rendimiento de ventas, la geografía de las ventas, el perfil del cliente, los métodos de envío utilizados y el factor temporal.

En resumen, este proyecto tiene como objetivo finalizar con un Data Warehouse que permita almacenar, gestionar eficientemente los datos de venta del supermercado, además que también facilite su análisis y consulta, proporcionando una base sólida para la toma de decisiones estratégicas y la identificación de oportunidades de mejora en el negocio.

3. Objetivo

Diseñar un DataWarehouse para un supermercado en particular con los conocimientos adquiridos

4. Teoría

4.1. Granularidad

De acuerdo con Kimball y Ross, (2013) la granularidad se refiere al nivel de detalle o especificidad de los datos en un conjunto de datos o sistema. En un contexto de bases de datos y análisis de datos, la granularidad indica qué tan específicos o generales son los datos registrados.

Por ejemplo, en una base de datos de ventas, la granularidad podría ser a nivel de cada transacción individual, a nivel diario, semanal, mensual, etc. Cuanto menor sea la granularidad, más detallados serán los datos, pero también pueden requerir más espacio de almacenamiento y procesamiento Kimball y Ross, (2013).

4.2. Dimensiones

En el contexto de un data warehouse, las dimensiones representan las categorías o aspectos clave a través de los cuales se analizan y se organizan los datos. Son atributos descriptivos que proporcionan contexto y detalles sobre los datos almacenados en la tabla de hechos. Las dimensiones suelen incluir información estática y jerárquica que facilita el análisis y la consulta de los datos Kimball y Ross, (2013).

4.3. Role-playing dimensions

Los atributos de tiempo en la tabla de hechos en un data warehouse son aquellos que representan diferentes aspectos temporales de los datos registrados en la tabla. Estos atributos permiten realizar análisis temporales y evaluar el desempeño, tendencias o patrones a lo largo del tiempo.

El concepto de role-playing dimensions se refiere a la práctica de utilizar una misma dimensión en una tabla de hechos en múltiples "roles", es decir, para representar diferentes aspectos o períodos temporales. Esto permite analizar los datos desde diferentes perspectivas temporales utilizando la misma dimensión, lo que brinda flexibilidad en el análisis de datos temporales.

Por ejemplo, en una tabla de hechos de ventas, podríamos tener dos atributos de tiempo: uno para la fecha de la transacción y otro para la fecha de entrega. Ambos atributos se relacionarían con la misma dimensión de tiempo, pero tendrían roles diferentes, lo que permitiría analizar las ventas desde la perspectiva de la fecha de transacción y la fecha de entrega.

4.4. Slow Changing Dimension

Un "slow changing dimension" (SCD), o dimensión de cambio lento en español, es una técnica utilizada en el diseño de bases de datos para manejar cambios en los atributos de una dimensión a lo largo del tiempo. Por ejemplo, supongamos que tienes una tabla de dimensión que representa productos, y uno de los productos cambia su nombre o su categoría a lo largo del tiempo. En este caso, necesitarías manejar este cambio en la dimensión de una manera que sea consistente y eficiente para tus consultas.

Existen varios tipos de SCD, pero los dos tipos más comunes son:

- Tipo 1 (SCD1): En este tipo de SCD, los cambios simplemente sobrescriben los valores existentes en la dimensión. Esto significa que no se mantiene un historial de los cambios; se pierde la información anterior cuando se actualiza.

- Tipo 2 (SCD2): En este tipo de SCD, se mantiene un historial de los cambios creando nuevas filas para cada cambio en la dimensión. Cada fila tiene un identificador único (generalmente una clave principal) y una marca de tiempo que indica cuándo fue válida esa versión de la fila.

Para el caso de nuestro proyecto se utilizo esta tecnica en las tablas de Costumer, Product y Ship Mode, esto debido que son dimensiones que pueden tener cambios en los clientes que puede tener la empresa, así como de los productos que comercializa y los tipos de envíos de mercadería.

5. Datos Prácticos

5.1. Justificaciones de Diseño

Para el desarrollo del proyecto se tomaron algunas decisiones y se identificarón ciertos atributos en las dimensiones que podían utilizarse de maneras distintas y que ayudara a que el Data Warehouse construido fuera mucho mas robusto y sencillo de ejecutar para los usuarios. Entre las decisiones tomadas fueron las siguientes:

5.1.1. Las dimensiones que cambian lentamente son las siguientes:

A. dim_date
B. dim_country
C. dim_ship_mode

5.1.2. Los atributos que cambian lentamente

A. *dim_date*: se realiza una carga inicial con una ventana de tiempo de 10 años y, cuando se acerque la última fecha, se realiza una adición de datos para los próximos 10 años.
B. *dim_country:* se carga inicialmente con estados o regiones predefinidos de un país, por lo que solo habrá cambios según hayan cambios en la geografía de las regiones.
C. *dim_ship_mode:* se cargan los servicios de modo de envío ofrecidos por el proveedor inicialmente, y conforme vayan agregando nuevos servicios se adicionan a la tabla.

5.1.3. Justificar la decisión del modelo a utilizar:

Se decidió diseñar un modelo estrella para mantener una complejidad baja en entre las relaciones de las dimensiones con la tabla de hechos, disminuyendo las necesidades computacionales del Data Warehouse al realizar consultas desnormalizadas.

La creación de las dimensiones para clientes, países, modos de envió, fechas y productos se definieron a partir de los datos proporcionados por el negocio, así como la identificación de campos que contenían información que se duplicaba en

cada registro. Se buscó desnormalizar la información de la tabla de hechos para las órdenes.
Se crea la tabla *snapshot_order* para tener una tabla *snapshot* y el intervalo de tiempo escogido es de una semana laboral, de lunes a viernes. En este caso solo tiene referencia a una tabla dimensional puesto que se va a poblar los datos con un procedimiento almacenado que permite calcular ventas y las ganancias totales correspondientes a la tabla de hechos de f*act_orders.*

5.2. Granularidad

La granularidad que se utilizará para el proyecto es la Granularidad por Transacción. Esta decisión fue tomada debido a que era fundamental para el proyecto debido a su capacidad de para proporcionar un nivel de detalle sin precedente en el análisis de ventas y al ser una base de datos con esas características, era la mejor opción. Al capturar y almacenar cada transacción individual realizada por la empresa, podremos comprender a fondo los patrones de compra, el comportamiento del cliente, entre otros. Esto nos permite identificar de mejor manera, optimizar la carga de datos y el proceso de ETL.

5.3. Dimensiones

Otra parte importante del proyecto se basa en la definición de las dimensiones que se utilizaran para crear la tabla de hechos. Para el proyecto seleccionamos 5 dimensiones: Producto, País, Cliente, Modo de Envío y Fecha. Estas dimensiones se seleccionaron y utilizaron por los siguientes motivos:

5.3.1. Productos:

La utilidad de la dimensión de productos permite analizar el rendimiento de ventas y la popularidad de los productos en diferentes contextos. Permite segmentar las ventas por categoría de productos, marca, precio, entre otros atributos.

5.3.2. País:

La utilidad de la dimensión de país proporciona información sobre la geografía de las ventas, lo que permite analizar las ventas por región, país o ubicación

geográfica. Es útil para comprender las tendencias regionales y las diferencias culturales en el comportamiento del cliente.

5.3.3. Cliente:

La utilidad de la dimensión de cliente proporciona información detallada sobre los clientes, como su segmentación demográfica, historial de compras, preferencias y comportamiento de compra. Permite entender quiénes son los clientes m

5.3.4. Modo de Envío:

La utilidad de la dimensión de modo de envío proporciona información sobre cómo se entregan los productos vendidos, como el método de envío, la compañía de envío, el tiempo de entrega, etc. Permite analizar la eficiencia y costos logísticos.

5.3.5. Fecha:

La utilidad de la dimensión de fecha permite analizar las ventas y otros datos relacionados con el tiempo. Permite realizar análisis de tendencias temporales, identificar patrones estacionales, evaluar el impacto de eventos específicos en el tiempo, entre otros.

6. Imágenes de referencia de la elaboración del data warehouse, de Super Store

En la figura 1 se observa el esquema de creacion de tablas desde el archvo “csv” a travez del software

Figura 1.
Esquema de la creación de tablas desde el archivo “.csv” a base de datos

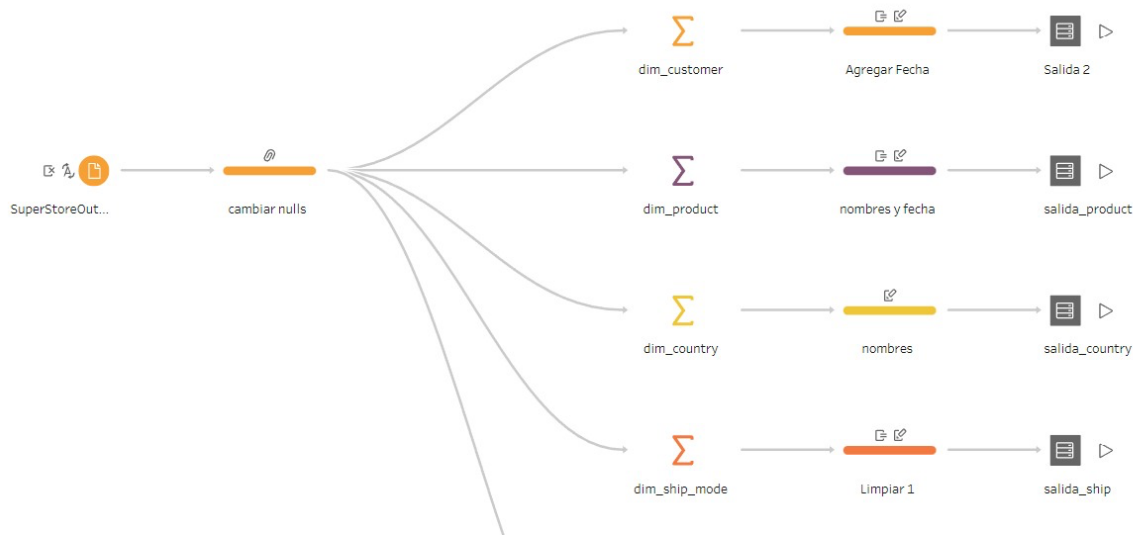

Figura 2.
Esquema de creación de tabla de hechos fact_table, dim_customer y country con limpieza.

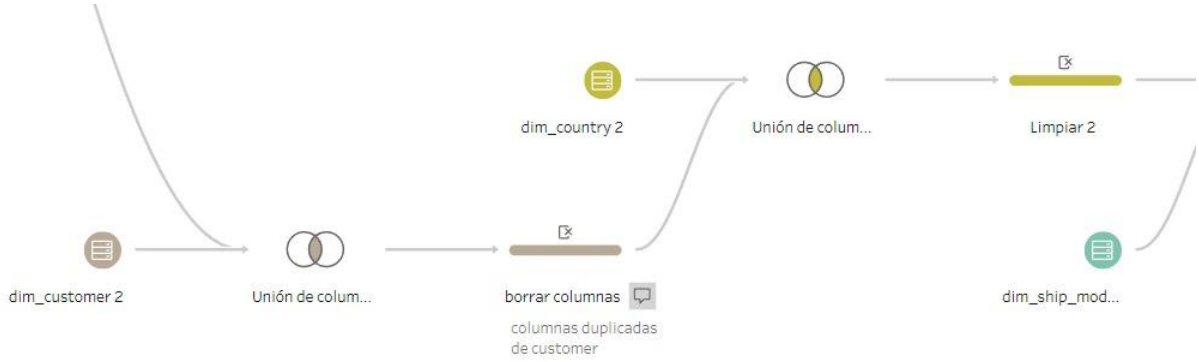

Figura 3.
Diagrama de limpieza de dim_date

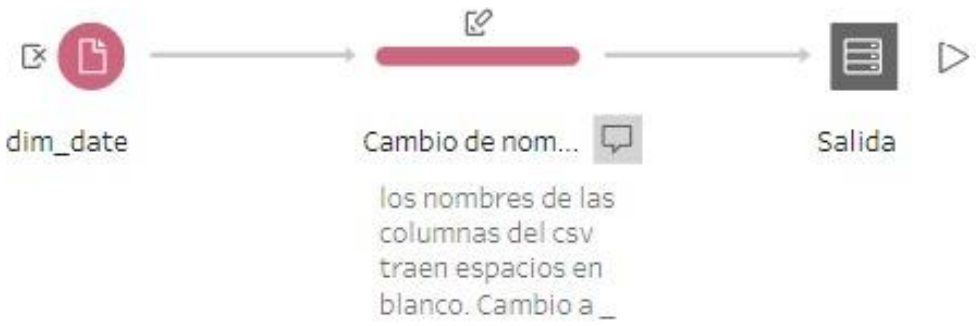

Figura 4.
Limpieza de de ship_mode, dim_date para order

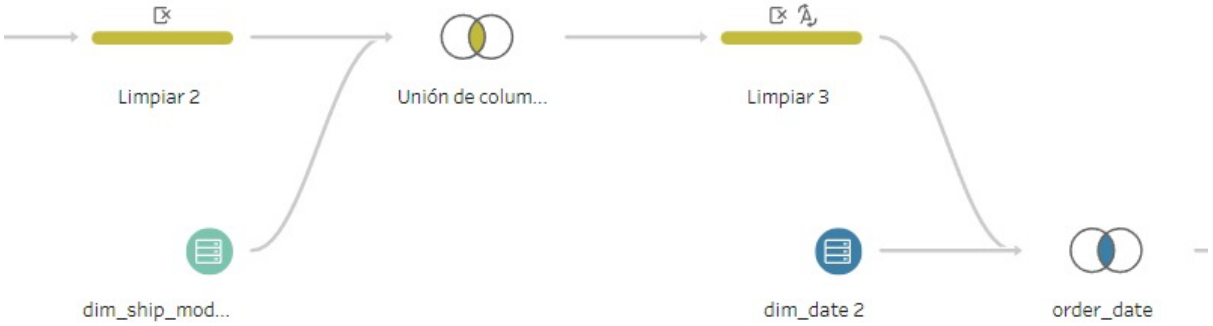

Figura 5.
Elaboración de product_id y ship_date

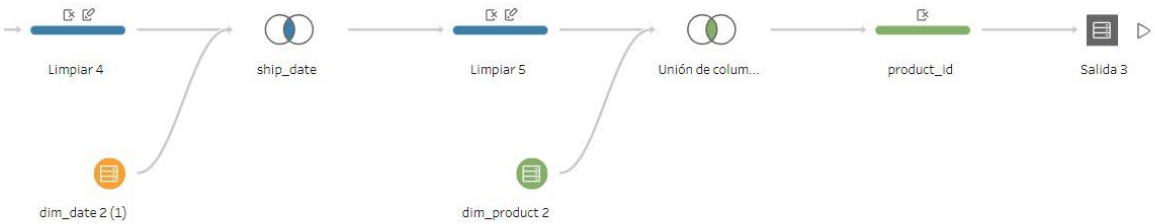

Figura 6.
Bases de datos creada para el trabajo

Base de datos creada

7. Colusiones

1. La implementación de un Data Warehouse en el contexto de un supermercado es fundamental para la comprensión de las transacciones de venta y el análisis de múltiples dimensiones que influyen en el desempeño del negocio. La elección de una granularidad por transacción ha permitido la identificación de patrones de compra, tendencias del mercado y comportamientos del cliente que se observa en el Dashbord de visualización presentado.

2. La selección cuidadosa de las dimensiones adecuadas, incluyendo Producto, País, Cliente, Modo de Envío y Fecha, ha enriquecido significativamente el análisis de datos, proporcionando una perspectiva completa y detallada de las ventas del supermercado. Estas dimensiones han permitido segmentar las ventas, comprender la geografía de las ventas, analizar el comportamiento del cliente indispensable para la toma de decisiones estratégicas.

8. Recomendaciones

1. Continuar mejorando y optimizando el proceso de carga de datos y la limpieza de datos, ya que una calidad de datos óptima es fundamental para garantizar la precisión y fiabilidad de los análisis realizados en el Data Warehouse. Se recomienda implementar técnicas de validación y depuración de datos de manera regular para mantener la integridad y coherencia de la información almacenada.

2. Explorar y aprovechar al máximo las capacidades de análisis y visualización de datos disponibles en el Data Warehouse para obtener insights más profundos y significativos. Además, la creación de paneles de control y reportes automatizados facilitará la monitorización continua del desempeño y la toma de decisiones ágiles basadas en datos en tiempo real.

9. Bibliografía.

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons.
Conclusiones: