Iván Yosef Maldonado Arriaga
14003689

Universidad Galileo

Data Warehouse – Proyecto 1

Introducción

La creación de un almacén de datos (Data Warehouse, DW) representa un paso fundamental en la transformación de datos crudos en información valiosa y accionable para una organización. Este proyecto específico se ha centrado en la construcción de un DW diseñado para capturar, almacenar y analizar las operaciones de ventas de una empresa, aprovechando datos provenientes de diversas fuentes y procesados inicialmente a través de Tableau Prep.

El esquema adoptado sigue la metodología dimensional, una práctica estándar en el diseño de almacenes de datos, que facilita la eficiencia en consultas y la escalabilidad. Dentro de este esquema, se han establecido cinco tablas de dimensiones: **dim_date**, **dim_customer**, **dim_product**, **dim_ship_mode**, y **dim_geography**. Estas tablas de dimensiones ofrecen un granulado detalle sobre el tiempo (fechas y períodos), los clientes, los productos, los modos de envío, y la geografía de las ventas respectivamente, abarcando desde el nivel de detalle más fino hasta agregaciones más amplias que permiten un análisis por diversas perspectivas.

La tabla de hechos, **fac_sales**, actúa como el núcleo del DW, donde se registran las transacciones de ventas con referencias a las dimensiones mencionadas. Incluye métricas claves como las ventas, cantidad, descuento, y beneficio, vinculadas a través de claves foráneas a cada una de las dimensiones, lo que permite realizar análisis complejos y multidimensionales de forma eficiente. Esta estructura no solo refleja el estado actual de las operaciones comerciales de la empresa, sino que también es capaz de adaptarse a cambios y crecimiento futuro.

Adicionalmente, se ha creado una tabla de etapa, **stg_super_store,** diseñada para facilitar la limpieza, validación, y preparación de los datos antes de su inserción en las tablas de dimensiones y hechos. Este enfoque asegura que solo los datos verificados y de alta calidad sean introducidos en el DW, preservando así su integridad y fiabilidad.

Tabla de contenido

Granularidad

Tabla de hechos: **fact_sales**

Una fila por cada producto en cada orden de venta: Cada registro en la tabla representa la venta de un producto

Fecha de orden y fecha de envío: Permite analizar las ventas por cuando se realizó la venta, así como también por cuándo se enviaron los productos

Análisis por cliente, producto, modo de envío y ubicación geográfica: Cada registro permite analizar quien compró (cliente), qué se compró (producto), como se envió (modo de envío) y donde se enviaron o se compraron los productos (geografía)

Dimensiones

dim_date

Dimensión que almacena información detallada sobre las fechas, esta dimensión es una dimensión que hemos utilizado por defecto en el curso, por lo que no se agregara descripción para cada uno de los atributos de esta

dim_customer

Dimensión que almacena información detallada sobre los clientes. Cada fila representa un cliente único en el sistema

sk_customer: Clave surrogada autoincremental que sirve como identificador único para cada cliente dentro del data warehouse
customer_id: Identificador único asignado a cada cliente
customer_name: Nombre completo del cliente
segment: Identifica el tipo de cliente (Corporate, Consumer, Home Office)
timestamp: Marcara el tiempo cuando se creó el registro

Slowly Changing Dimensions (SCD):

customer_name: Puede cambiar debido a errores de tipografías, cambios de nombre o fusiones de cuentas
segment: Los clientes pueden cambiar de segmento debido a la evolución de sus necesidades o comportamientos de compra

Estrategia: Utilizar SCD tipo 2, para mantener un histórico, creando una nueva fila en la dimensión con cada cambio significativo, junto con marcas de tiempo para rastrear vigencia de cada versión

dim_product

Dimensión que almacena información sobre los productos que se venden

sk_product: Clave surrogada autoincremental que actúa como un identificador único para cada producto en el Data Warehouse
product_id: Código único asignado a cada producto, utilizado internamente para identificar productos en diferentes sistemas y bases de datos.
product_name: Nombre del producto, utilizado en reportes y análisis para identificar productos de manera legible
category: Categoría general a la que pertenece el producto, permitiendo análisis agregados por categorías de productos
sub_category: Subcategoría más específica del producto, para análisis más granulares y detallados dentro de cada categoría
timestamp: Marcara el tiempo cuando se creó el registro

Slowly Changing Dimensions (SCD):

product_name: Puede cambiar por renovación de marca o errores iniciales
category: El producto puede cambiar de categoría según estrategia de marketing
sub_category: El producto puede cambiar de subcategoría según estrategia de marketing

Estrategia: Similar que en dim_customer, se utilizaría una SCD Tipo 2, para llevar un histórico y validar la evolución de la información del producto

dim_ship_mode

Dimensión que almacena información sobre los modos de envío disponibles

sk_ship_mode: Clave surrogada autoincremental, sirve como identificador único para cada modo de envío.
ship_mode: Nombre o descripción del modo de envío (Standard Class, Second Class, Same Day, First Class)
timestamp: Marcara el tiempo cuando se creó el registro

Slowly Changing Dimensions (SCD):

ship_mode: Los modos de envió pueden cambiar de nombre debido a los proveedores

Estrategia: Similar que en dim_product, se utilizaría una SCD Tipo 2, para llevar un histórico y validar como han cambiado los modos de envío a lo largo del tiempo

dim_geography

Dimensión que almacena información sobre las ubicaciones geográficas relacionadas con las ventas

sk_geography: Clave surrogada autoincremental, actúa como un identificador único para cada entrada geográfica.
postal_code: Código postal de la ubicación, útil para análisis geográficos detallados y para dirigir estrategias de marketing localizadas.
country: País de la ubicación, permite realizar análisis por país y entender mercados internacionales.
region: Región dentro del país, como un estado o provincia, para análisis más específicos dentro de grandes mercados.
state: Estado específico, para análisis detallados y comparaciones entre diferentes estados dentro de un país.
city: Ciudad, útil para análisis urbanos y para entender la distribución de clientes o ventas en áreas metropolitanas.
timestamp: Marcara el tiempo cuando se creó el registro

Slowly Changing Dimensions (SCD):

city: El nombre de las ciudades puede cambiar debido a reorganizaciones territoriales

Estrategia: Similar que en dim_ship_mode, se utilizaría una SCD Tipo 2, para llevar un histórico y validar la evolución de estos cambios

Tabla de hechos

fac_sales

Tabla donde estarán registradas las transacciones de ventas

sk_customer: Clave foránea que referencia a dim_customer. Esta columna enlaza cada registro de venta con un cliente específico, permitiendo analizar las ventas por cliente, entender patrones de compra, y segmentar clientes basado en comportamientos de compra.
sk_product: Clave foránea que referencia a dim_product. Asocia cada venta con un producto específico, lo que es fundamental para analizar el rendimiento de los productos, identificar cuáles son los más vendidos, y entender las tendencias de demanda.
sk_ship_mode: Clave foránea que referencia a dim_ship_mode. Relaciona cada venta con un modo de envío, permitiendo análisis sobre cómo los métodos de envío afectan las ventas, la satisfacción del cliente, y los tiempos de entrega.
sk_geography: Clave foránea que referencia a dim_geography. Vincula las ventas a una ubicación geográfica específica, habilitando análisis geográficos de las ventas para identificar mercados fuertes y áreas con potencial de crecimiento.
sk_date_key_order_date: Clave foránea que referencia a dim_date para la fecha de la orden. Permite realizar análisis temporales de las ventas, identificar tendencias estacionales, y evaluar el impacto de eventos específicos en las ventas.
sk_date_key_ship_date: Clave foránea similar que referencia a dim_date para la fecha de envío. Esto posibilita el análisis del ciclo de cumplimiento de pedidos, desde la orden hasta la entrega, y su impacto en la satisfacción del cliente.
row_id: Identificador único para cada fila en la tabla de hechos. Esencial para mantener la integridad de los datos y para referenciar registros específicos durante el análisis y la resolución de problemas.
order_id: Identificador único de cada orden de venta. Permite rastrear y analizar las ventas a nivel de orden, facilitando la gestión de pedidos y el análisis de transacciones individuales o agrupadas.
sales: Monto total de la venta. Esta métrica es crucial para el análisis de ingresos, permitiendo evaluar el rendimiento financiero de productos, categorías, clientes, y regiones geográficas.
quantity: Cantidad de productos vendidos en la transacción. Esencial para análisis de volumen, ayudando a entender las unidades vendidas más allá del valor monetario de las ventas.
discount: Descuento aplicado a la venta. Permite evaluar el impacto de las estrategias de precios y promociones en el volumen de ventas y en la rentabilidad.
profit: Beneficio generado por la venta. Esta métrica es fundamental para análisis de rentabilidad, permitiendo a las organizaciones identificar qué productos, clientes, o regiones son los más rentables.

Modelo estrella