

LABORATORIO II – INFORME DE PROCESO ETL Y MODELO DE DATOS

1. Proceso ETL (Extracción, Transformación y Carga)

Para este proyecto se trabajó con tres fuentes de datos distintas, cumpliendo con la consigna del laboratorio:

- Los archivos CSV del dataset de Olist (órdenes, clientes, productos e ítems).
- Un archivo Excel externo creado para incorporar la dimensión geográfica Dim_Regiones, con información de estados y regiones de Brasil.
- Una fuente web, agregada mediante conexión URL como tercer origen de datos requerido.

En el Editor de Power Query se realizaron diversas transformaciones:

- Cambio de tipos de datos y renombrado de columnas.
- Ajuste manual de campos de fecha para evitar errores automáticos de conversión.
- Creación de columnas derivadas como AñoMes y AñoMesOrden para análisis temporal.
- Verificación de integridad: se comprobó que las claves order_id, product_id y customer_id no presentaban valores nulos ni inconsistencias; por lo tanto, no fue necesario eliminar registros.
- Normalización y preparación final del conjunto de datos.

Sobre la Tabla Calendario

Durante el proceso, se intentó construir una tabla calendario utilizando DAX (CALENDAR), como práctica estándar de inteligencia temporal. Sin embargo, debido a conflictos regionales (Power BI interpretaba fechas DD/MM/YYYY como MM/DD/YYYY), el modelo arrojaba errores al intentar generar la tabla.

Ante este inconveniente, consulté específicamente a ChatGPT, y se decidió implementar una solución alternativa válida:

→ Generar columnas temporales directamente desde olist_orders_dataset:

- AñoMes (formateado como texto)
- AñoMesOrden (número para ordenar cronológicamente)

Esta solución permitió:

- Mantener el orden cronológico sin errores

- Realizar análisis de tendencia mensual
- Evitar problemas con el formato de fecha
- Cumplir plenamente con los requisitos analíticos del laboratorio

La alternativa fue probada y funcionó correctamente, por lo cual se integró al dashboard final.

2. Modelo de Datos – Esquema Copo de Nieve

El modelo implementado corresponde a un Esquema Copo de Nieve, ya que una de las dimensiones principales se encuentra normalizada mediante una subdimensión adicional.

Tabla de Hechos

- olist_order_items_dataset

Dimensiones

- olist_orders_dataset
- olist_customers_dataset
- olist_products_dataset
- Dim_Regiones (vinculada con customers)

Normalización (Snowflake)

La tabla Dim_Regiones depende de olist_customers_dataset mediante la clave customer_state, lo que convierte el modelo en un esquema Snowflake.

Las columnas AñoMes y AñoMesOrden reemplazan la necesidad de una tabla calendario y permiten ordenar correctamente los análisis temporales.

3. Medidas DAX

Total Ventas

Total Ventas = SUM(olist_order_items_dataset[price])

Cantidad de Pedidos

Cantidad de Pedidos = DISTINCTCOUNT(olist_orders_dataset[order_id])

Ticket Promedio

Ticket Promedio = DIVIDE([Total Ventas], [Cantidad de Pedidos])

Estas medidas alimentan los KPIs principales del dashboard.

4. Dashboard Final

El informe final incluye:

- KPIs: Total Ventas, Ticket Promedio, Cantidad de Pedidos
- Gráfico de líneas (tendencia mensual)
- Gráfico de barras por categoría
- Gráfico horizontal por estado
- Mapa geográfico (Bing Maps)
- Segmentadores por Año y por Mes