LAB-3: Dimensional Data Modeling

Brayan Stiven Tigreros

Jose David Santa

Valentina Morales Valencia

Universidad Autónoma de Occidente

Facultad de Ingeniería y Ciencias básicas

ETL (G01)

Breyner Posso Bautista

X de Febrero de 2026

A. Reporte técnico

- Modelo dimensional (Actividad 2):

El diagrama se compone de un fact llamado sales este se compone de atributos tanto propios como llaves foráneas traídas de las dimensiones, sus atributos son: sale_id - time_id - product_id - customer_id - channel_id - quantity - unit_price_sale de los cuales el segundo hasta el quinto son las llaves foráneas el resto son los atributos propios. Las dimensiones que componen el diagrama son: Time, Channel, Product y Customer cada uno guarda información importante que ayudara a cumplir los requerimientos, su distribución fue realizada enfocándose en lo que se necesitaba para resolver las preguntas y sus atributos fueron filtrados para satisfacer únicamente lo que se necesitaba sin tener en cuenta datos que no se iban a implementar en el Datawerhause. Los atributos de Time son: time_id - day - month - year; los de Channel son: channel_id - channel_name - type_channel; los de Product son: product_id - name - category - brand y los de Customer son: customer_id - name - city - country - age. Los cuales proporcionaran la información necesaria para cumplir con los requerimientos y KPIs.

- Definición de los KPI (Actividad 1)

Requerimientos	**KPIs**	**Formula de calculo**	**Tablas de hechos**	**Dimensiones requeridas**	**Tipo de visualización**	**Justificación de su valor comercial**
1.¿Cuál es el volumen de ventas y los ingresos por categoría de producto?	Total de ventas por categoría del producto	Ventas por categoría = SUMA(Total ventas de productos de una categoría)	Sales	Product	Grafico de barras	Te permitira saber cual es la categoría del producto que mas ingresos genera esto ayudara a saber a que categoría invertirle mas o a cual sacarle promociones para que se venda
2. ¿Qué canales de ventas (tienda física frente a en línea) generan los mayores ingresos?	Total de ventas por canal de venta	Ventas por canal = SUMA (Total de ventas de producto por un determinado canal de venta)	Sales	Channel	Grafico de barras	Te permite saber cual es el tipo de canal de ventas que mas esta generando ingresos para saber a cual invertirle mas
3. ¿Cómo evolucionan las ventas a lo largo del tiempo (tendencias mensuales)?	Ventas del mes	Ventas mensuales = SUMA (ventas por mes)	Sales	Time	Grafico de lineas	Te permite saber cuales son las ventas que se realizaron en un mes ademas de poder luego comparar los diferentes meses y tomar acciones para mejorar las ventas o entender por que se relucieron las ventas en determinado mes
4. ¿Qué marcas son las más rentables?	Ventas por marca de producto	Ventas por marca = SUMA (Total de ventas de producto de una marca)	Sales	Product	Grafico de barras	Te permite saber cual es la marca que mas ingresos esta produciendo esto permite saber las elecciones de las personas y a cual marca comprarle mas
5. Cual es la ciudad de la que mas sale clientes?	Ventas por ciudad del cliente	Ventas por ciudad = SUMA (Total de ventas de producto por ciudad del cliente)	Sales	Customer	Grafico de barras	Te permite saber a que ciudad esta llegando mas promociones o cual ciudad pertenece las personas que mas estan comprando, esto permite saber en que ciudades centrar mas las promociones para aumentar las ventas
6. Cual es el rango de edades de los clientes?	Grupo de edad del cliente con mas ventas	Ventas por edad = SUMA (Total de ventas agrupadas por grupo de edad (18 - 24 , 25 - 34, 35 - 44, 45 - 54, 55 - 64, 65 - +))	Sales	Customer	Grafico de barras	Te permite saber cual es el rango de edad que mas esta comprando los productos esto para saber a quien centrar mas las promociones o hacer estrategias para aumentar las ventas en otro largo de edad

- Explicación diseño del ETL

Para llevar a cabo el laboratorio utilizamos datos sintéticos generados por la IA a modo de ejemplo para realizar el flujo de ETL. Estos datos iniciales son un total son 4 archivos csv para canales, clientes, ventas y productos simulando que son provenientes de un sistema OLTP.

Fases del Proceso de ETL

Extracción: En esta fase tomamos los datos sintéticos generados por la IA ubicados en la carpeta raw, utilizamos pandas para seleccionar estos archivos csv y convertirlos en data frames de pandas a método de staging area, dejándolos listos para su transformación.

Trasformación: En esta fase lo que hacemos es tomar los dataframes de la extracción, tomamos los dataframes y hacemos una inspeccion de las primeras filas y validamos los tipos de datos. Esto es recomendable para tener claro que columnas vamos a volver atributos y como crearemos las dimensiones.

Aqui el concepto cambia, ahora estos datos no le llamaremos data frames, en vez de ello, despues de transformarlos los llamaremos dimensiones.

Para cada dimension creamos sus atributos y llaves foráneas. En la transformacion usamos llaves foraneas para relacionar la informacion de cada dimension con la tabla de hecho de las ventas

Carga: En esta fase hay 2 funciones, una para cargar los datos ya transformados a una carpeta, guardandolos como csv para visualizarlos luego y tambien definimos una funcion que haga la conexion a mysql workbench y carge los csv en el data warehouse de modelo de estrella definido con anterioridad. Este modelo se crea a partir de un modelo entidad relacion segun las dimensiones y atributos que creamos, para luego usar el script DDL (queries sql) de ese modelo para crear las dimensiones y tabla de hecho en mysql listas para cargar los datos.

Utilizamos un archivo main.py que lo que hace es orquestar todas las fases del proceso ETL permitiendo su ejecución completa

- Querys SQL

1. **Query para la creacion de la dimension channel:**

```
CREATE TABLE `channel_dim` (
  `channel_key` int NOT NULL AUTO_INCREMENT,
  `channel_name` varchar(45) NOT NULL,
  `channel_id` int NOT NULL,
  PRIMARY KEY (`channel_key`),
  UNIQUE KEY `channel_id` (`channel_id`)
```

2. **Query para la creacion de la dimension customer:**

```
CREATE TABLE `customer_dim` (
  `customer_key` int NOT NULL AUTO_INCREMENT,
  `name` varchar(45) NOT NULL,
  `city` varchar(45) NOT NULL,
  `country` varchar(45) NOT NULL,
  `age` int NOT NULL,
  `customer_id` int NOT NULL,
  PRIMARY KEY (`customer_key`),
  UNIQUE KEY `customer_id` (`customer_id`)
```

3. **Query para la creacion de la dimension:**

```
CREATE TABLE `product_dim` (
  `product_key` int NOT NULL AUTO_INCREMENT,
  `name` varchar(45) NOT NULL,
  `category` varchar(45) NOT NULL,
  `brand` varchar(45) DEFAULT NULL,
  `product_id` int NOT NULL,
  PRIMARY KEY (`product_key`),
  UNIQUE KEY `product_id` (`product_id`)
```

4. **Query para la creacion de la dimension time:**

```
CREATE TABLE `time_dim` (
 `time_key` int NOT NULL AUTO_INCREMENT,
 `day` int NOT NULL,
 `month` int NOT NULL,
 `year` int NOT NULL,
 PRIMARY KEY (`time_key`),
 UNIQUE KEY `day` (`day`,`month`,`year`)
```

5. **Query para la creacion del fact sales:**

```
CREATE TABLE `sales_fact` (
 `quantity` bigint DEFAULT NULL,
 `unit_price_sale` double DEFAULT NULL,
 `product_dim_product_key` bigint DEFAULT NULL,
 `customer_dim_customer_key` bigint DEFAULT NULL,
 `channel_dim_channel_key` bigint DEFAULT NULL,
 `time_dim_time_key` bigint DEFAULT NULL
```

- Papel de la IA

La IA nos ayudó en la organización, dejando roles claros y dando directrices para, a la hora de crear código y base de datos, saber cómo y qué se está enviando a la db. Además, fue parte fundamental para saber cómo crear una conexión entre el código y el mysql workbench para poder que este último reciba los datos del proceso de ETL.