Proyecto 1

Introducción a la Ciencia de Datos

Elaborado por:

Gómez Agudelo, JUAN SEBASTIÁN – 2259474

Henao Aricapa, STIVEN – 2259603

Hernández Ortíz, VÍCTOR MANUEL – 2259520

Docente:

Ocampo Arbeláez, HÉCTOR FABIO

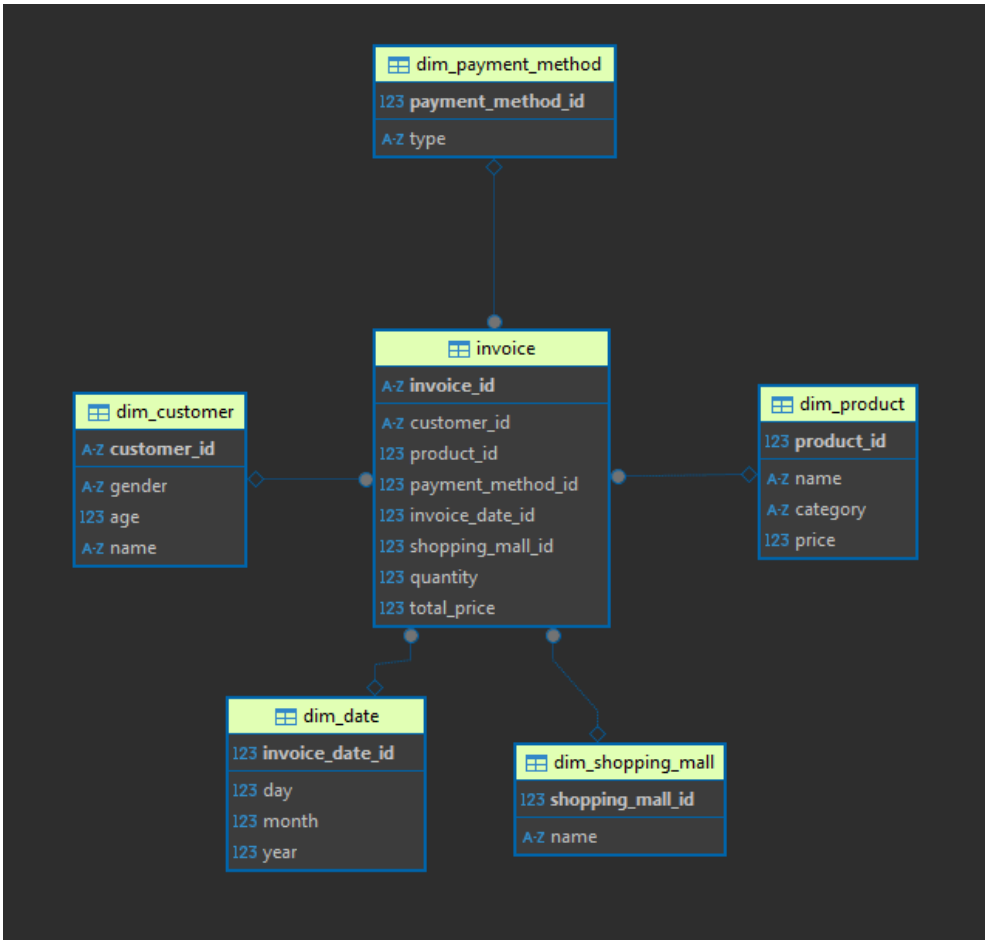
Sede Tuluá

Febrero 2025

Índice

1. Diseño del Modelo de la Bodega de Datos:

Diagrama bodega de datos



Justificación del uso del modelo estrella

El modelo estrella fue elegido para este esquema de bodega de datos debido a su eficiencia y simplicidad para consultas analíticas. Al organizar los datos con una tabla de hechos central conectada a tablas de dimensiones desnormalizadas, se reduce el número de uniones (JOINs) necesarias, lo que mejora significativamente el rendimiento de las consultas. Esta estructura permite un acceso eficiente a la información clave. Además, el modelo estrella facilita los procesos ETL, asegurando una carga de datos más sencilla y un mantenimiento menos complejo, lo que apoya la toma de decisiones oportuna y precisa.

Script SQL con la creación de las tablas en Postgres

Anexado en scripts.sql

2. Exportación, Transformación y Carga de datos

Código del proceso ETL en python

Anexado en ETLProyecto.ipynb

Explicación del proceso y transformaciones realizadas

Cada proceso y transformación especificado en el archivo adjunto denominado *ETLProyecto.ipynb*

Comprobación datos insertados en la base de datos

dim_customer

customer_id [PK] character varying (15)	gender character varying (10)	age integer	name character varying (255)
C241288	Female	28	
C111565	Male	21	
C266599	Male	20	
C988172	Female	66	
C189076	Female	53	
C657758	Female	28	
C151197	Female	49	
C176086	Female	32	
C159642	Male	69	
C283361	Female	60	
C240286	Female	36	
C191708	Female	29	
C225330	Female	67	
C312861	Male	25	
C555402	Female	67	
C362288	Male	24	
C300786	Male	65	
C330667	Female	42	
C218149	Female	46	
C196845	Male	24	
C220180	Male	23	
C125696	Female	27	

dim_date

	invoice_date_id [PK] integer	day integer	month integer	year integer
1	1	5	8	2022
2	2	12	12	2021
3	3	9	11	2021
4	4	16	5	2021
5	5	24	10	2021
6	6	24	5	2022
7	7	13	3	2022
8	8	13	1	2021
9	9	4	11	2021
10	10	22	8	2021
11	11	25	12	2022
12	12	28	10	2022
13	13	31	7	2022
14	14	17	11	2022
15	15	3	6	2022
16	16	7	11	2021
17	17	16	1	2021
18	18	5	1	2022
19	19	26	7	2021
20	20	7	3	2023
21	21	15	2	2023
22	22	1	5	2021

dim_payment_method

	payment_method_id [PK] integer	type character varying (50)
1	1	Credit Card
2	2	Debit Card
3	3	Cash

dim_product

	product_id [PK] integer	name character varying (255)	category character varying (100)	price numeric (10,2)
1	1		Clothing	1500.40
2	2		Shoes	1800.51
3	3		Clothing	300.08
4	4		Shoes	3000.85
5	5		Books	60.60
6	6		Cosmetics	40.66
7	7		Clothing	600.16
8	8		Clothing	900.24
9	9		Food & Beverage	10.46
10	10		Books	15.15
11	11		Toys	143.36
12	12		Books	30.30
13	13		Food & Beverage	15.69
14	14		Food & Beverage	5.23
15	15		Technology	5250.00
16	16		Books	75.75
17	17		Toys	71.68
18	18		Cosmetics	203.30
19	19		Shoes	2400.68
20	20		Cosmetics	121.98
21	21		Toys	107.52
22	22		Clothing	1200.32

dim_shopping_mall

	shopping_mall_id [PK] integer	name character varying (255)
1	1	Kanyon
2	2	Forum Istanbul
3	3	Metrocity
4	4	Metropol AVM
5	5	Istinye Park
6	6	Mall of Istanbul
7	7	Emaar Square Mall
8	8	Cevahir AVM
9	9	Viaport Outlet
10	10	Zorlu Center

invoice

	invoice_id [PK] character varying (15)	customer_id character varying (15)	product_id integer	payment_method_id integer	invoice_date_id integer	shopping_mall_id integer	quantity integer	total_price numeric (10,2)
1	I138884	C241288	1	1	1	1	5	7502.00
2	I317333	C111565	2	2	2	2	3	5401.53
3	I127801	C266599	3	3	3	3	1	300.08
4	I173702	C988172	4	1	4	4	5	15004.25
5	I337046	C189076	5	3	5	1	4	242.40
6	I227836	C657758	1	1	6	2	5	7502.00
7	I121056	C151197	6	3	7	5	1	40.66
8	I293112	C176086	7	1	8	6	2	1200.32
9	I293455	C159642	8	1	9	3	3	2700.72
10	I326945	C283361	7	1	10	1	2	1200.32
11	I306368	C240286	9	3	11	3	2	20.92
12	I139207	C191708	10	1	12	7	1	15.15
13	I640508	C225330	11	2	13	3	4	573.44
14	I179802	C312861	7	3	14	8	2	1200.32
15	I336189	C555402	7	1	15	1	2	1200.32
16	I688768	C362288	4	1	16	9	5	15004.25
17	I294687	C300786	12	2	17	3	2	60.60
18	I195744	C330667	13	1	18	10	3	47.07
19	I993048	C218149	7	3	19	4	2	1200.32
20	I992454	C196845	11	3	20	8	4	573.44
21	I183746	C220180	3	1	21	7	1	300.08
22	I412481	C125696	14	3	22	8	1	5.23

3. Consultas analíticas en SQL

Consultas SQL y explicación

1. Total de ventas por categoría de producto

```
SELECT
   p.category AS categoria_producto,
   SUM(i.total_price) AS total_ventas
FROM invoice i
JOIN dim_product p ON i.product_id = p.product_id
GROUP BY p.category
ORDER BY total_ventas DESC;
```

Explicación - Consulta

Calcula el total de ventas (total_price) para cada categoría de producto.

- Une la tabla invoice con dim_product para obtener la categoría de cada producto.
- Agrupa los resultados por categoría (category).
- Suma las ventas (total_price) para cada categoría.
- Ordena los resultados de mayor a menor venta.

Explicación - Resultado

Los resultados muestran que Clothing es la categoría con mayores ventas, seguida de Shoes y Technology, mientras que Souvenir y Books tienen los valores más bajos.

	categoria_producto character varying (100)	total_ventas numeric
1	Clothing	113996791.04
2	Shoes	66553451.47
3	Technology	57862350.00
4	Cosmetics	6792862.90
5	Toys	3980426.24
6	Food & Beverage	849535.05
7	Books	834552.90
8	Souvenir	635824.65

2. Clientes con mayor volumen de compras

```
SELECT
   c.customer_id AS cliente_id,
   c.gender AS genero,
   c.age AS edad,
   SUM(i.total_price) AS total_compras
FROM invoice i
JOIN dim_customer c ON i.customer_id = c.customer_id
GROUP BY c.customer_id, c.gender, c.age
ORDER BY total_compras DESC
LIMIT 10;
```

Explicación - Consulta

Identifica a los 10 clientes que han gastado más dinero en compras.

- Une la tabla invoice con dim_customer para obtener información del cliente.
- Agrupa los resultados por cliente (customer_id), género (gender) y edad (age).
- Suma el total gastado (total_price) por cada cliente.
- Ordena los resultados de mayor a menor gasto y limita a los 10 primeros.

Explicación - Resultado

Se muestra que estos clientes han gastado exactamente la misma cantidad (*26,250.00*), lo que sugiere que hay compras frecuentes de un monto fijo. Además, hay una distribución equitativa entre géneros y un rango de edades variado.

	cliente_id character varying (15)	genero character varying (10)	edad integer	total_compras numeric
1	C169530	Male	34	26250.00
2	C922102	Male	44	26250.00
3	C166881	Female	31	26250.00
4	C254550	Female	28	26250.00
5	C237772	Female	37	26250.00
6	C812985	Male	65	26250.00
7	C259585	Male	30	26250.00
8	C553588	Female	44	26250.00
9	C638391	Female	56	26250.00
10	C160336	Female	20	26250.00

3. Métodos de pago más utilizados

```
SELECT
   pm.type AS metodo_pago,
   COUNT(i.invoice_id) AS total_transacciones,
   SUM(i.total_price) AS total_ventas
FROM invoice i
JOIN dim_payment_method pm ON i.payment_method_id = pm.payment_method_id
GROUP BY pm.type
ORDER BY total_transacciones DESC;
```

Explicación - Consulta

La consulta muestra los métodos de pago más utilizados al comprar y el total de ventas generadas con cada uno.

- Une la tabla invoice con dim_payment_method para obtener el nombre del método de pago.
- Agrupa los resultados por método de pago (name).
- Cuenta el número de transacciones (invoice_id) y suma el total de ventas (total_price) para cada método.
- Ordena los resultados por el número de transacciones de mayor a menor.

Explicación - Resultado

El resultado indica que el efectivo es el método más utilizado, seguido por la tarjeta de crédito y la tarjeta de débito. A pesar de tener menos transacciones, los pagos con tarjeta de crédito generan un alto volumen de ventas, lo que sugiere que los clientes podrían estar realizando compras de mayor valor con este método.

	metodo_pago character varying (100)	total_transacciones bigint	total_ventas numeric
1	Cash	44447	112832243.02
2	Credit Card	34931	88077123.77
3	Debit Card	20079	50596427.46

4. Comparación de ventas por mes

```
SELECT
   d.year AS año,
   d.month AS mes,
   SUM(i.total_price) AS total_ventas
FROM invoice i
JOIN dim_date d ON i.invoice_date_id = d.invoice_date_id
GROUP BY d.year, d.month
ORDER BY d.year, d.month;
```

Explicación - Consulta

La consulta realizada compara las ventas totales por mes y año.

- Une la tabla invoice con dim_date para obtener el año y mes de cada transacción,.
- Agrupa los resultados por año (year) y mes (month).
- Suma las ventas (total_price) para cada mes.
- Ordena los resultados por año y mes.

Explicación - Resultado

Los datos muestran fluctuaciones en las ventas mensuales, con algunos picos significativos en ciertos meses. Es posible identificar patrones estacionales o tendencias de crecimiento o disminución en distintos años.

	año integer	mes integer	total_ventas numeric
1	2021	1	9641614.62
2	2021	2	8772315.22
3	2021	3	9455359.38
4	2021	4	9389541.54
5	2021	5	9771756.97
6	2021	6	9286271.35
7	2021	7	10311119.68
8	2021	8	9630655.70
9	2021	9	9188165.62
10	2021	10	10263015.06
11	2021	11	9265555.29
12	2021	12	9585200.16
13	2022	1	9764311.14
14	2022	2	8344111.92

	año integer	mes integer	total_ventas numeric
15	2022	3	9986685.16
16	2022	4	9326144.44
17	2022	5	9947574.13
18	2022	6	9647503.95
19	2022	7	10067602.95
20	2022	8	9651705.59
21	2022	9	9607629.29
22	2022	10	10282075.37
23	2022	11	8941584.66
24	2022	12	9869885.48
25	2023	1	9485599.83
26	2023	2	9508662.96
27	2023	3	2514146.79

4. Análisis Descriptivo y Visualización de Datos

Código Python y análisis de gráficos obtenidos enlazados en GraficasProyecto.ipynb

5. Conclusiones

Diseño de la Bodega de Datos

Se implementó un modelo estrella para optimizar consultas analíticas y facilitar la integración de datos, mejorando la eficiencia en el análisis de ventas, clientes y métodos de pago.

Proceso ETL

Se desarrolló un proceso ETL en Python para limpiar, transformar y cargar datos en PostgreSQL, asegurando coherencia y calidad en los datos.

Consultas Analíticas en SQL

Se identificaron patrones de ventas, segmentación de clientes por género y edad, y preferencias de pago (efectivo como el más usado, seguido de tarjetas de crédito). También se detectaron variaciones estacionales en las ventas.

Análisis Descriptivo y Visualización

El análisis visual reveló que las mujeres son el segmento predominante, el efectivo es el método de pago más común, y el rango de edad más frecuente es de 30 a 50 años.

Conclusión General

El proyecto integró herramientas como SQL y Python para procesar y analizar datos, proporcionando insights clave para optimizar estrategias comerciales y mejorar la experiencia del cliente. Los resultados ofrecen una base sólida para futuras investigaciones y decisiones empresariales basadas en datos.