www.iudigital.edu.co

EVIDENCIA DE APRENDIZAJE 3

PROCESO DE TRANSFORMACIÓN DE DATOS Y CARGA EN EL DATA MART FINAL

INTEGRANTES

JHON JAIRO FUENTES TURIZO

GRUPO PREICA2502B010064

DOCENTE:

ANTONIO JESUS VALDERRAMA

INSTITUCIÓN UNIVERSIDAD IU DIGITAL DE ANTIOQUIA

INGENERIA DE SOFTWARE Y DATOS

2025

1. Objetivo

Documentar el procedimiento de Extracción de datos desde la base transaccional `jardineria` hacia la base de datos de preparación `jardineria_staging`, asegurando una copia íntegra y exacta de los datos antes de su transformación.

El paso de **Extracción de datos** (la "E" en ETL) consiste en recuperar la información relevante de la base de datos de origen (jardineria) y transferirla a la base de datos de preparación o *staging* (jardineria_staging). Este proceso inicial garantiza que los datos en bruto se copien de manera íntegra y exacta antes de aplicar cualquier transformación.

Para lograr esto, se utiliza el método de **Extracción Completa** (Full Extraction), que consiste en copiar todos los datos disponibles de la fuente, ya que es el método empleado en el proyecto. Esto se realiza mediante consultas **SQL INSERT INTO SELECT**.

A continuación, se presentan en el archivo script_extraccion_datos.sql utilizadas para extraer los datos de la base de datos transaccional de origen (jardineria) y cargarlos en las tablas de *staging* (jardineria_staging):

La fase de **Transformación de Datos** (la 'T' en ETL) es fundamental para convertir los datos brutos que residen en el área de *staging* (jardineria_staging) en información estructurada, limpia y analíticamente relevante, lista para ser cargada en el Data Mart (Modelo Estrella).

El objetivo es asegurar la **calidad, coherencia y adecuación dimensional** de los datos para responder a preguntas de negocio específicas, como la **identificación del producto más vendido**.

3.a) Aplicación de Técnicas de Transformación de Datos

La transformación de datos se aplica para adaptar los datos de las tablas transaccionales de *staging* (como stg_detalle_pedido, stg_producto, stg_pedido) a las tablas dimensionales y de hechos requeridas: DimProducto, DimCategoria, DimTiempo y HechosVentas.

Las principales técnicas aplicadas son:

1. Limpieza y Filtrado de Datos (Asegurando Calidad y Relevancia)

La limpieza de datos se realiza para eliminar errores, inconsistencias y duplicados.

- **Filtrado de Transacciones no Válidas:** Se deben **excluir los registros de pedidos que no representan una venta completada**, garantizando que solo los pedidos con estado "Entregado" o "Pagado" se incluyan en la tabla de hechos. Por ejemplo, se deben filtrar los pedidos con el estado **'Rechazado'**.
- **Manejo de Nulos e Inconsistencias:** Se deben tratar los valores faltantes o nulos antes de la carga. Esto asegura que las métricas (como precio_unidad o cantidad_vendida) sean completas y precisas para el análisis.

2. Normalización y Estandarización (Garantizando Coherencia)

Se garantiza que los datos en el Data Mart utilicen formatos y unidades consistentes.

- **Conversión de Tipos de Datos:** Los datos se convierten para que coincidan con los formatos definidos en el Modelo Estrella. Por ejemplo, se asegura que los campos monetarios (total_venta, precio_unidad) tengan la precisión adecuada (decimales 15,2).
- **Creación de Claves Subrogadas (Surrogate Keys):** En lugar de utilizar las claves operacionales originales (natural keys), se crean **claves primarias artificiales auto-incrementales** (id_producto, id_categoria, id_tiempo, id_venta) para cada dimensión y la tabla de hechos. Esto es esencial para el modelado dimensional y permite la **gestión eficiente de los cambios históricos** (aunque esto no se detalla en las fuentes para este modelo, es una función clave de las claves subrogadas).

3. Enriquecimiento y Derivación de Datos (Preparación Analítica)

Esta técnica consiste en crear nueva información y calcular métricas críticas para el análisis que no existían en las tablas de *staging* originales.

- **Derivación de la Métrica Clave (total_venta):** Esta es la transformación más importante para la tabla de hechos. Se calcula el monto total de la venta (la métrica principal) a nivel de línea de pedido. $$ \text{total_venta} = \text{cantidad_vendida} \times \text{precio_unidad} $$ Esta métrica es fundamental para el análisis de rendimiento de ventas.
- **Enriquecimiento Temporal (DimTiempo):** La fecha de pedido (fecha_pedido de stg_pedido) se descompone en sus componentes analíticos (año, mes, día, nombre_mes, trimestre, semestre) para crear la dimensión de tiempo. Esta

granularidad temporal permite analizar **tendencias estacionales** y volúmenes de ventas a lo largo del tiempo.

3.b) Ejecución de la Transformación (Consultas SQL)

La transformación de datos se realiza mediante la lógica de negocio aplicada en **consultas SQL** de alta complejidad, las cuales integran los datos de las tablas de *staging* (stg_) y los cargan en las tablas finales del Data Mart. Este proceso se conoce como el desarrollo del **Subsistema de Extracción, Transformación y Carga (ETL)**, que es la base sobre la cual se alimenta el Data Mart.

A continuación, se describen las transformaciones y la lógica de carga para las tablas del Data Mart:

1. Carga de las Dimensiones

Las dimensiones se cargan primero para generar las claves subrogadas necesarias en la tabla de hechos.

Tabla de Destino	Origen (Staging)	Lógica de Transformación/Carga
DimCategoria	stg_categoria_producto	Se seleccionan los atributos (Id_Categoria, Desc_Categoria, descripcion_texto, etc.) y se asignan a las columnas de destino, generando el id_categoria subrogado.
DimProducto	stg_producto	Se seleccionan todos los atributos descriptivos del producto (incluyendo precio_venta, proveedor, etc.), y se genera el id_producto subrogado. La clave id_producto_original mantiene la trazabilidad.
DimTiempo	stg_pedido (tomando la columna fecha_pedido)	**Enriquecimiento:** Se utiliza la columna de fecha de pedido para derivar atributos temporales (año, mes, trimestre, semestre) y crear registros únicos para

cada fecha, asignando el id_tiempo subrogado.

2. Carga de la Tabla de Hechos (HechosVentas)

La tabla de hechos se construye uniendo las tablas de pedidos y detalles de pedidos de *staging* y realizando cálculos y búsquedas de claves dimensionales.

La consulta para cargar HechosVentas involucra los siguientes pasos lógicos (integración, derivación, y búsqueda de claves):

1. **Integración de Datos:** Se unen las tablas stg_detalle_pedido y stg_pedido mediante ID_pedido para combinar las métricas de la línea de pedido (cantidad, precio) con el contexto del pedido (fecha, estado).
2. **Filtrado de Transacciones:** Se filtra la unión para incluir **solo los pedidos que han sido 'Entregado'** (o que han sido procesados como venta, excluyendo 'Rechazado').
3. **Derivación de Métricas:** Se calcula el total_venta multiplicando cantidad por precio_unidad.
4. **Búsqueda de Claves (Key Lookups):** Se buscan las claves subrogadas correspondientes en las dimensiones ya cargadas:
    - Se busca el id_producto en DimProducto utilizando el ID_producto original.
    - Se busca el id_tiempo en DimTiempo utilizando la fecha_pedido.
    - Se busca el id_categoria en DimCategoria a través de la relación del producto.
5. **Inserción Final:** Los resultados de esta consulta integrada y transformada se insertan en la tabla HechosVentas.

Resultado Analítico: Al finalizar esta transformación, la tabla **HechosVentas** contendrá el volumen de ventas (cantidad_vendida) y los ingresos (total_venta) enlazados directamente a las dimensiones **Producto, Categoría y Tiempo** mediante claves subrogadas. Esto permite la consulta rápida y la agregación necesaria para análisis clave, como la **identificación eficiente del producto más vendido**.

La siguiente respuesta se centra en la fase 4, la **Carga de registros en el Data Mart final**, utilizando la estructura del modelo estrella propuesto para la empresa Jardinería (HechosVentas, DimProducto, DimCategoria, DimTiempo) y asumiendo

que las fases previas de Extracción y Transformación (E y T) se han completado en la base de datos de *staging* (jardineria_staging).

---

4) Carga de registros en el Data Mart final:

El objetivo de esta fase es la **Carga (L)**, que implica mover los datos transformados desde el área de *staging* al modelo dimensional de destino. Esta carga debe ser precisa, consistente y escalable.

a) Diseñar consultas SQL o scripts de carga para insertar los registros transformados desde la base de datos de staging en las tablas del data mart final.

El proceso de carga debe seguir un orden dimensional: primero las tablas de dimensiones, para generar las claves subrogadas (SKs), y luego la tabla de hechos, para referenciar dichas SKs.

Asunción: El Data Mart final se almacena en una base de datos llamada DataMart.

1. Definición de Tablas de Dimensiones y Hechos

El modelo estrella está compuesto por:

- **Dimensiones:** DimProducto, DimCategoria, y DimTiempo.
- **Tabla de Hechos:** HechosVentas.

2. Scripts de Carga de Dimensiones

La carga de dimensiones es el primer paso, ya que proporciona las claves subrogadas (id_producto, id_categoria, id_tiempo) necesarias para la tabla de hechos.

A. Carga de DimCategoria Esta dimensión almacena la clasificación de productos. Se asume que la clave primaria de origen es Id_Categoria de stg_categoria_producto.

Carga DimCategoria (Dimensión de Categorías)

Mapea Id_Categoria de staging a id_categoria_original en el Data Mart.

```
INSERT INTO DataMart.DimCategoria (
  id_categoria_original,
  nombre,
  descripcion_texto,
```

```
  descripcion_html,
  imagen
)
SELECT
  Id_Categoria,
  Desc_Categoria,  Desc_Categoria es el nombre de la categoría en staging.
  descripcion_texto,
  descripcion_html,
  imagen
FROM
  jardineria_staging.stg_categoria_producto;
```

B. Carga de DimProducto Esta dimensión contiene los atributos descriptivos de cada producto. Se asume que la clave primaria de origen es ID_producto de stg_producto.

```
Carga DimProducto (Dimensión de Productos)
Mapea ID_producto de staging a id_producto_original en el Data Mart.
INSERT INTO DataMart.DimProducto (
  id_producto_original,
  codigo_producto,
  nombre,
  dimensiones,
  proveedor,
  precio_venta,
  precio_proveedor
)
SELECT
  ID_producto,
```

```
  CodigoProducto,
  nombre,
  dimensiones,
  proveedor,
  precio_venta,
  precio_proveedor
FROM
  jardineria_staging.stg_producto;
```

C. Carga de DimTiempo Esta dimensión proporciona granularidad temporal (año, mes, día, trimestre, semestre) para analizar las tendencias de ventas. Se extraen las fechas de los pedidos, que son las transacciones en el sistema de *staging* (stg_pedido). Se asume que el proceso de transformación previo ya limpió y estandarizó las fechas.

Carga DimTiempo (Dimensión Temporal)

Requiere la derivación de atributos de tiempo a partir de las fechas de pedido.

```
INSERT INTO DataMart.DimTiempo (
  fecha,
  año,
  mes,
  dia,
  nombre_mes,
  trimestre,
  semestre
)
SELECT DISTINCT
  fecha_pedido AS fecha,
```

```
  YEAR(fecha_pedido) AS año,
  MONTH(fecha_pedido) AS mes,
  DAY(fecha_pedido) AS dia,
   NOTA: Las siguientes columnas (nombre_mes, trimestre, semestre)
   requieren lógica de derivación (Transformación) que debe realizarse
   durante el proceso ETL (ej. usando funciones de fecha).
  (SELECT MONTHNAME(fecha_pedido)) AS nombre_mes,
  (SELECT QUARTER(fecha_pedido)) AS trimestre,
  (SELECT IF(MONTH(fecha_pedido) <= 6, 1, 2)) AS semestre
FROM
  jardineria_staging.stg_pedido;
```

3. Script de Carga de la Tabla de Hechos (HechosVentas)

La tabla de hechos almacena las métricas cuantificables del negocio (medidas) y las claves foráneas de las dimensiones. La consulta principal debe unir (JOIN) los datos transaccionales (stg_detalle_pedido y stg_pedido) con las tablas de dimensiones para resolver las claves subrogadas.

Métricas a calcular/incluir: cantidad_vendida, precio_unidad, y el cálculo del **total_venta** (cantidad * precio_unidad).

```
 Carga HechosVentas (Tabla de Hechos Central)
 Resuelve las claves subrogadas (SK) utilizando las claves originales (PK) de staging.


INSERT INTO DataMart.HechosVentas (
  id_producto,
  id_categoria,
  id_tiempo,
  cantidad_vendida,
  precio_unidad,
```

```
    total_venta,
    numero_pedido,
    numero_linea
)
SELECT
    DP.id_producto,  Clave Subrogada de Producto
    DC.id_categoria,  Clave Subrogada de Categoría
    DT.id_tiempo,  Clave Subrogada de Tiempo (fecha_pedido)
    SDP.cantidad AS cantidad_vendida,
    SDP.precio_unidad,
    (SDP.cantidad * SDP.precio_unidad) AS total_venta,  Métrica calculada
    SDP.ID_pedido AS numero_pedido,  Referencia operacional
    SDP.numero_linea
FROM
    jardineria_staging.stg_detalle_pedido SDP
JOIN
    jardineria_staging.stg_pedido SP
    ON SDP.ID_pedido = SP.ID_pedido
JOIN
    jardineria_staging.stg_producto SPR  Necesario para obtener la Categoria (clave
original)
    ON SDP.ID_producto = SPR.ID_producto
 Uniones de Lookup para obtener Claves Subrogadas:
JOIN
    DataMart.DimProducto DP
    ON SDP.ID_producto = DP.id_producto_original  Asumiendo mapeo 1:1 de
ID_producto original a SK
```

```
JOIN
  DataMart.DimTiempo DT
  ON SP.fecha_pedido = DT.fecha
JOIN
  DataMart.DimCategoria DC
  ON SPR.Categoria = DC.id_categoria_original;
```

b) Ejecutar las consultas de carga y verificar que los datos se hayan insertado correctamente en el data mart final.

La ejecución y la verificación son pasos esenciales para garantizar la **calidad de los datos** y la **integridad referencial** del Data Mart.

1. Ejecución de las Consultas de Carga

Este paso implica ejecutar los scripts SQL diseñados en el punto 4a), idealmente a través de un sistema o herramienta de ETL (Extracción, Transformación y Carga). El sistema ETL es la base que alimenta el Data Warehouse (o Data Mart) y asegura que los datos se carguen en un formato acorde para el análisis.

2. Verificación de la Inserción (Validación y Calidad de Datos)

La verificación se centra en asegurar que los datos cargados sean **precisos, completos y consistentes**.

A. Verificación de Conteo de Registros (Integridad y Completitud) Es crucial comparar el número de registros en las tablas de hechos con el número esperado basado en las tablas de origen en *staging* (stg_detalle_pedido en este caso).

1. Verificar conteo total de registros en la Tabla de Hechos

```
SELECT COUNT(*) FROM DataMart.HechosVentas;
```

2. Comparar con el origen (stg_detalle_pedido)

```
SELECT COUNT(*) FROM jardineria_staging.stg_detalle_pedido;
```

Ambos conteos deben coincidir si se cargaron todas las líneas de detalle de pedido correctamente.

B. Verificación de Integridad Referencial (Claves Foráneas) Se debe confirmar que todas las transacciones en HechosVentas estén vinculadas correctamente a las dimensiones mediante claves foráneas (FKs).

3. Verificar registros en HechosVentas con claves subrogadas nulas

Si el mapeo fue exitoso, no debería haber registros con FKs nulas.

```
SELECT COUNT(*)

FROM DataMart.HechosVentas

WHERE

  id_producto IS NULL OR

  id_categoria IS NULL OR

  id_tiempo IS NULL;
```

Si este conteo es mayor que cero, indica fallas en el mapeo entre las claves originales de staging y las claves subrogadas de las dimensiones.

C. Validación de Métricas (Precisión) Se debe verificar que las métricas calculadas, como total_venta, sean correctas comparándolas con el origen. La métrica total_venta es el resultado de la fórmula cantidad_vendida * precio_unidad.

4. Validar agregación total de ventas (ejemplo: suma total)

Suma de ventas en DataMart:

```
SELECT SUM(total_venta) FROM DataMart.HechosVentas;
```

Recálculo de la suma de ventas desde Staging:

```
SELECT SUM(cantidad * precio_unidad)

FROM jardineria_staging.stg_detalle_pedido;
```

Estos totales deben ser idénticos para confirmar la precisión de la transformación.

Al completar estos pasos, se garantiza que los datos están listos y optimizados para el análisis multidimensional en el Data Mart, cumpliendo con la estructura de esquema estrella propuesta.