

Evidencia de aprendizaje #3

Proceso de transformación de datos y carga en el data mart final

Integrantes Grupo #13

JOSE LUIS BENITEZ SALCEDO
GERMAN EDUARDO PAIBA ROCHA
JOSUÉ SANMARTÍN POLO

Institución Universitaria Digital de Antioquia
Ingeniería de Software y Datos

Asignatura:
BASE DE DATOS II

Nombre del tutor:
VICTOR HUGO MERCADO

Marzo 30 de 2025

INTRODUCCIÓN

Este documento describe el proceso de transformación de datos y carga al Data Mart final, basado en el modelo estrella previamente definido para la base de datos Jardinería. El objetivo es garantizar que los datos almacenados en la base de datos Staging sean limpiados, transformados y cargados eficientemente en el Data Mart, asegurando su calidad y coherencia para el análisis.

En el entorno del desarrollo de software y el manejo de información y bases de datos, se manejan las bases de datos Staging, que en nuestros términos es como una “sala de espera” para los datos, ya que los datos se almacenan temporalmente antes de ser procesados y de llevarlos a su destino final, que puede ser un data warehouse o un modelo estrella, manejado en la primera entrega.

OBJETIVOS

- Desarrollar un proceso de transformación y carga de datos desde la base de datos origen, pasar Staging y luego hasta el data mart final, utilizando la base de datos de staging previamente creada.
- Asegurar que los datos estén preparados para el análisis, incluyendo la identificación del producto más vendido y otros análisis relevantes para la toma de decisiones empresariales.
- Revisar el modelo Estrella definido en el proyecto de análisis de la base de datos Jardinería para comprender la estructura y las relaciones entre las tablas de dimensiones y la tabla de hechos.

PLANTEAMIENTO DEL PROBLEMA

Desarrollo de proceso de transformación de datos y carga en el Data Mart final.

1. Preparación:

Revisar el modelo Estrella definido en el proyecto de análisis de la base de datos Jardinería para comprender la estructura y las relaciones entre las tablas de dimensiones y la tabla de hechos.

Verificar la disponibilidad y consistencia de la base de datos de staging previamente creada para el proyecto.

2. Extracción de datos desde la base de datos origen hacia la base de datos de Staging:

Utilizar consultas SQL para extraer datos relevantes de la base de datos origen y cargarlos en las tablas correspondientes de la base de datos de staging.

Verificar la integridad y consistencia de los datos extraídos para asegurar que cumplan con los requisitos del modelo Estrella.

3. Transformación de datos según las necesidades analíticas:

Aplicar técnicas de transformación de datos, como limpieza, normalización y enriquecimiento, para preparar los datos de acuerdo con las necesidades analíticas específicas.

Realizar la transformación de los datos utilizando consultas SQL u herramientas de ETL (Extract, Transform, Load) según sea necesario para garantizar la calidad y coherencia de los datos.

4. Carga de registros en el Data Mart final:

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.

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

5. Documentación y presentación:

Documentar todo el proceso de transformación de datos y carga en un informe detallado que incluya una descripción de las etapas realizadas, las consultas SQL utilizadas.

Presentar el informe de manera clara y concisa, asegurándose de incluir referencias al modelo Estrella y las técnicas de ETL aplicadas.

ANÁLISIS DEL PROBLEMA

La tabla de la base de datos Jardinería nos muestra cómo está organizada y estructurada la información en ese negocio y los tipos de transacciones que se llevan a cabo en él.

Luego de revisar el modelo Estrella definido en el proyecto de análisis de la base de datos Jardinería para comprender la estructura y las relaciones entre las tablas de dimensiones y la tabla de hechos, verificamos la disponibilidad y consistencia de la base de datos de staging previamente creada para el proyecto.

Utilizando consultas SQL para extraer datos relevantes de la base de datos origen y cargarlos en las tablas correspondientes de la base de datos de staging, se verifica la integridad y consistencia de los datos extraídos para asegurar que cumplan con los requisitos del modelo Estrella.

Por último, se aplican técnicas de transformación de datos, como limpieza, normalización y enriquecimiento, para preparar los datos de acuerdo con las necesidades analíticas específicas.

Al realizar la transformación de los datos utilizando consultas SQL u herramientas de ETL (Extract, Transform, Load) para garantizar la calidad y coherencia de los datos, se diseñan 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 y ejecutar las consultas de carga y verificar que los datos se hayan insertado correctamente en el data mart final.

PROPUESTA DE LA SOLUCIÓN

1. Preparación:

Revisar el modelo Estrella definido en el proyecto de análisis de la base de datos Jardinería para comprender la estructura y las relaciones entre las tablas de dimensiones y la tabla de hechos.

Verificar la disponibilidad y consistencia de la base de datos de staging previamente creada para el proyecto.

Revisión del Modelo Estrella: Se retomó el modelo estrella definido en fases anteriores, compuesto por:

Tabla de hechos: Fact_Ventas (contiene métricas como cantidad_vendida, precio_unidad, total_venta).

Tablas de dimensiones:

Dim_Producto (productos).

Dim_Cliente (clientes).

Dim_Tiempo (fechas).

Dim_Empleado (empleados).

Verificación de la Base de Datos Staging: Se confirmó la disponibilidad de las tablas en la base de datos Staging, creadas en la fase anterior (entrega 2)

2. Extracción de datos desde la base de datos origen hacia la base de datos de Staging:

Utilizar consultas SQL para extraer datos relevantes de la base de datos origen y cargarlos en las tablas correspondientes de la base de datos de staging.

Verificar la integridad y consistencia de los datos extraídos para asegurar que cumplan con los requisitos del modelo Estrella.

Consultas SQL para Extracción: Se utilizaron consultas INSERT INTO ... SELECT para migrar datos desde la base de datos original (Jardinería) a Staging.

```
INSERT INTO staging_pedido (
  ID_pedido, fecha_pedido, ID_cliente, estado
)
SELECT
  ID_pedido, fecha_pedido, ID_cliente, estado
FROM jardineria_pedido;
```

Validación de Integridad: Se ejecutaron consultas para detectar inconsistencias:

```
SELECT *
FROM staging_cliente
WHERE nombre_cliente IS NULL OR ciudad IS NULL;
```

3. Transformación de datos según las necesidades analíticas:

Aplicar técnicas de transformación de datos, como limpieza, normalización y enriquecimiento, para preparar los datos de acuerdo con las necesidades analíticas específicas.

Realizar la transformación de los datos utilizando consultas SQL u herramientas de ETL (Extract, Transform, Load) según sea necesario para garantizar la calidad y coherencia de los datos.

Limpieza y Normalización:

Estandarización de formatos:

```
UPDATE staging_producto
SET nombre = UPPER(nombre);
```

Corrección de duplicados:

```
DELETE FROM staging_cliente
WHERE ID_cliente IN (
  SELECT ID_cliente
  FROM staging_cliente
  GROUP BY ID_cliente
  HAVING COUNT(*) > 1
);
```

Enriquecimiento de Datos: Se agregaron columnas calculadas para análisis:

```
UPDATE staging_detalle_pedido
SET total = cantidad * precio_unidad;
```


4. Carga de registros en el Data Mart final:

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.

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

Diseño de Consultas de Carga

Se insertaron los datos transformados desde Staging al Data Mart (modelo estrella):

```
-- Cargar datos en Dim_Producto
INSERT INTO Dim_Producto (
  ID_producto, nombre_producto, categoria, precio_venta
)
SELECT
  ID_producto, nombre, Categoria, precio_venta
FROM staging_producto;

-- Cargar datos en Fact_Ventas
INSERT INTO Fact_Ventas (
  ID_producto, ID_cliente, ID_tiempo, cantidad, precio_unidad, total_venta
)
SELECT
  dp.ID_producto, p.ID_cliente, dt.ID_tiempo,
  dp.cantidad, dp.precio_unidad, dp.total
FROM staging_detalle_pedido dp
JOIN staging_pedido p ON dp.ID_pedido = p.ID_pedido
JOIN Dim_Tiempo dt ON p.fecha_pedido = dt.fecha;
```

Verificación de la Carga: Se validó la coherencia de los datos en el Data Mart:

```
-- Comparar totales entre Staging y Data Mart
SELECT
  (SELECT SUM(total) FROM staging_detalle_pedido) AS total_staging,
  (SELECT SUM(total_venta) FROM Fact_Ventas) AS total_datamart;
```



5. Documentación y presentación:

Documentar todo el proceso de transformación de datos y carga en un informe detallado que incluya una descripción de las etapas realizadas, las consultas SQL utilizadas.

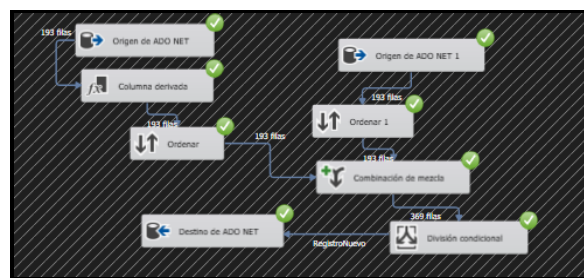
Presentar el informe de manera clara y concisa, asegurándose de incluir referencias al modelo Estrella y las técnicas de ETL aplicadas.

- Para cada punto, se incluyeron las consultas y análisis realizados.
- Se incluye al final del trabajo la carpeta donde se encuentran los documentos adicionales.

Parte 3 del trabajo

Transformacion DIMTiempo

```
CREATE TABLE "DimTiempo" (
  IDDimTiempo int identity(1,1) primary key
  "fecha_pedido" date,
  "fecha_esperada" date,
  "fecha_entrega" date,
  "Dia" int,
  "Mes" int,
  "Año" int,
  "NumeroSemana" int,
  "Trimestre" int,
  "DiaSemana" int
)
```

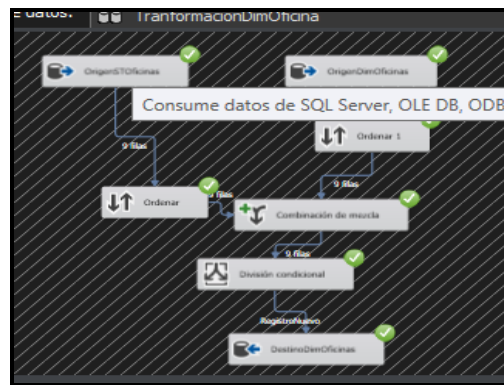
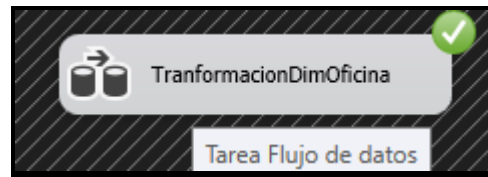


```
1 SELECT TOP (1000) [fecha_pedido]
2     , [Dia]
3     , [Mes]
4     , [Año]
5 FROM [StagingDB].[dbo].[DimTiempo]
6
```

fecha_pedido	Dia	Mes	Año
2006-01-17	17	1	2006
2006-01-19	19	1	2006
2006-05-25	25	5	2006
2006-07-28	28	7	2006
2007-01-07	7	1	2007
2007-01-19	19	1	2007
2007-03-19	19	3	2007
2007-04-24	24	4	2007
2007-05-20	20	5	2007
2007-05-28	28	5	2007
2007-06-20	20	6	2007

TransformacionDimOficina

```
CREATE TABLE "DimOficina" (
  [ID_oficina]
  ,[Descripcion]
  ,[ciudad]
  ,[pais]
  ,[region]
  ,[codigo_postal]
  ,[telefono]
  ,[linea_direccion1]
)
```

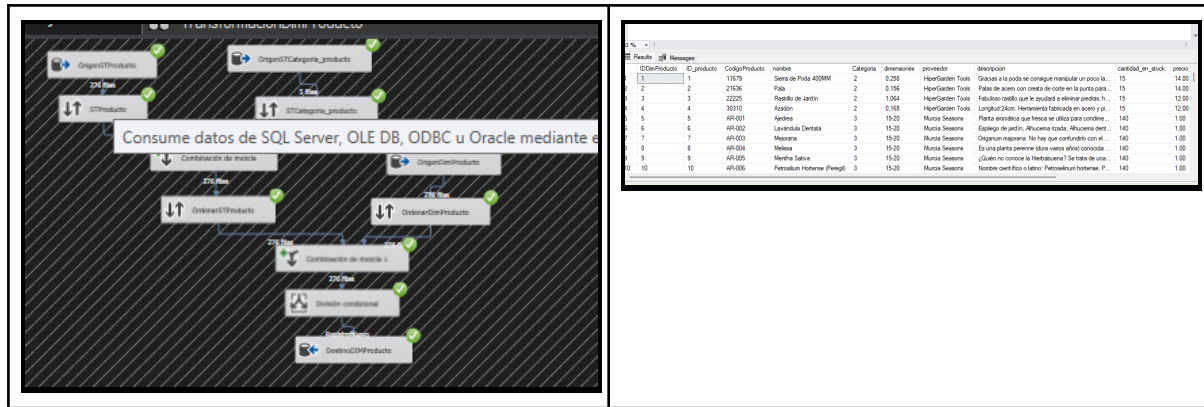


ID_oficina	Descripcion	ciudad	pais	region	codigo_postal	telefono	linea_direccion1
1	BON-ES	Barcelona	España	Barcelona	08019	+34 93 3561182	Avenida Diagonal, 38
2	BOS-USA	Boston	EEUU	MA	02108	+1 215 837 0825	1550 Court Place
3	LON-UK	Londres	Inglaterra	EMEA	EC2N 1HN	+44 20 78772041	52 Old Broad Street
4	MAD-ES	Madrid	España	Madrid	28032	+34 91 7514487	Bulevar Indalecio Prieto, 32
5	PAR-FR	Paris	Francia	EMEA	75017	+33 14 723 4404	29 Rue Joffroy d'abbans
6	SFO-USA	San Francisco	EEUU	CA	94080	+1 650 219 4782	100 Market Street
7	SYD-AU	Sydney	Australia	APAC	NSW 2010	+61 2 9264 2451	5-11 Wentworth Avenue
8	TAL-ES	Talavera de la Reina	España	Castilla-La Mancha	45632	+34 925 967231	Francisco Aguirre, 32
9	TOK-JP	Tokyo	Japón	Chiyoda-Ku	102-8578	+81 33 224 5000	4-1 Kocho

TransformacionDimProducto

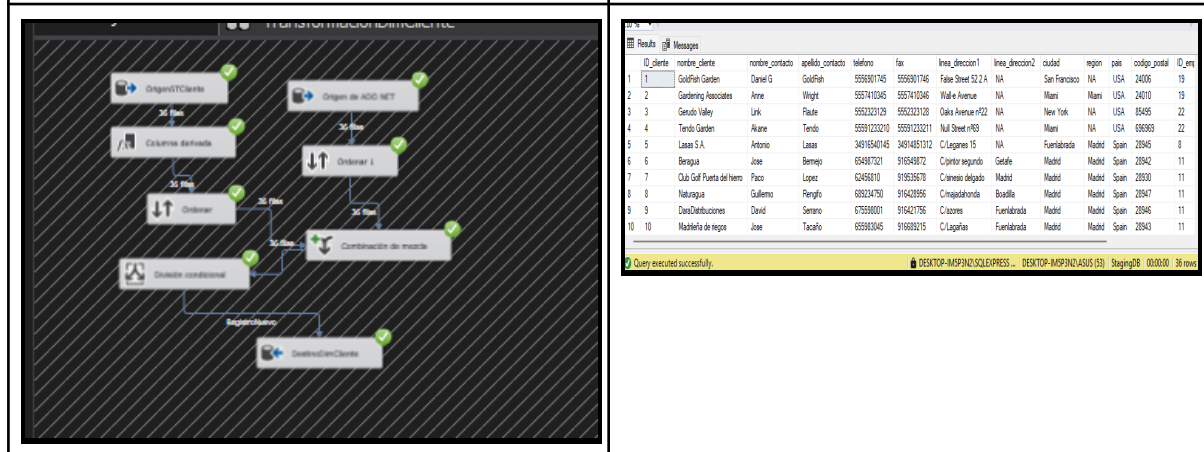
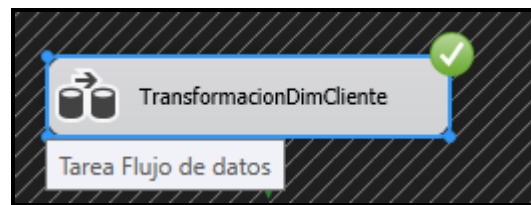
```
CREATE TABLE "DIMProducto" (
  IDDimProducto int identity(1,1) primary key,
  "ID_producto" int,
  "CodigoProducto" nvarchar(15),
  "nombre" nvarchar(70),
  "Categoria" int,
  "dimensiones" nvarchar(25),
  "proveedor" nvarchar(50),
  "descripcion" nvarchar(max),
  "cantidad_en_stock" smallint,
  "precio_venta" numeric(15,2),
  "precio_proveedor" numeric(15,2),
  "Id_Categoria" int,
  "Desc_Categoria" nvarchar(50)
)
```





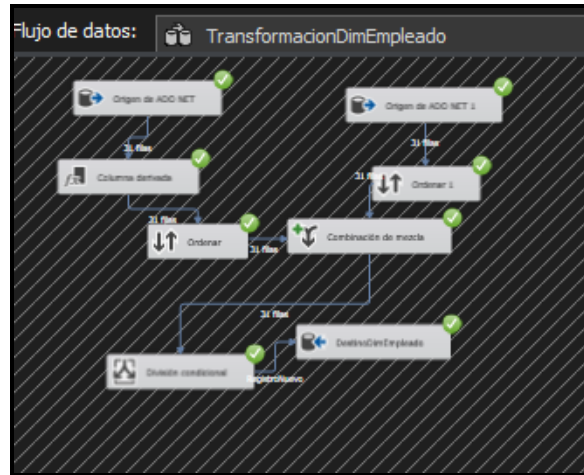
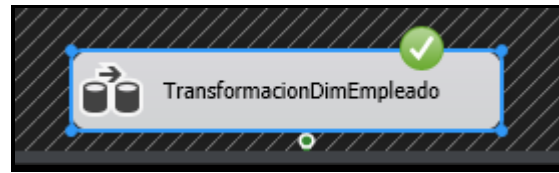
TrasformacionDimCliente

```
CREATE TABLE "DimCliente" (
  "ID_cliente" int,
  "nombre_cliente" nvarchar(50),
  "nombre_contacto" nvarchar(30),
  "apellido_contacto" nvarchar(30),
  "telefono" nvarchar(15),
  "fax" nvarchar(15),
  "linea_direccion1" nvarchar(50),
  "linea_direccion2" nvarchar(50),
  "ciudad" nvarchar(50),
  "region" nvarchar(50),
  "pais" nvarchar(50),
  "codigo_postal" nvarchar(10),
  "ID_empleado_rep_ventas" int,
  "limite_credito" numeric(15,2)
)
```



TransformacionDimEmpleado

```
CREATE TABLE "DimEmpleado" (
  "ID_empleado" int,
  "nombre" nvarchar(50),
  "apellido1" nvarchar(50),
  "apellido2" nvarchar(50),
  "extension" nvarchar(10),
  "email" nvarchar(100),
  "ID_oficina" int,
  "ID_jefe" int,
  "puesto" nvarchar(50)
)
```



ID_empleado	nombre	apellido1	apellido2	extension	email	ID_oficina	ID_jefe	puesto
1	Marcos	Magaña	Perez	3037	marcos@jardineria.es	8	999	Director General
2	Ruben	López	Martinez	2899	rlopez@jardineria.es	8	1	Subdirector Marketing
3	Alberto	Soria	Carrasco	2837	asoria@jardineria.es	8	2	Subdirector Ventas
4	Maria	Solis	Jerez	2847	msolis@jardineria.es	8	2	Secretaria
5	Felipe	Rosas	Marquez	2844	frosas@jardineria.es	8	3	Representante Ventas
6	Juan Carlos	Oritz	Serrano	2845	cortiz@jardineria.es	8	3	Representante Ventas
7	Carlos	Soria	Jimenez	2444	csoria@jardineria.es	4	3	Director Oficina
8	Mariano	López	Murcia	2442	mlopez@jardineria.es	4	7	Representante Ventas
9	Lucio	Campoamor	Martin	2442	lcampoamor@jardineria.es	4	7	Representante Ventas
10	Hilario	Rodriguez	Huetas	2444	hrodriguez@jardineria.es	4	7	Representante Ventas
11	Emmanuel	Magaña	Perez	2518	manu@jardineria.es	1	3	Director Oficina

CONCLUSIONES

Dentro de las actividades realizadas, uno de los aspectos más importantes del proyecto fue la limpieza y transformación de datos. Durante el proceso, observamos que los datos no siempre están listos para su uso, lo que podría afectar los análisis futuros.

Este proyecto reforzó habilidades en ETL, SQL avanzado y gestión de calidad de datos, críticas para roles como Ingeniero de Datos o Analista BI.

Con este trabajo, estamos adquiriendo herramientas valiosas que nos ayudarán a destacar en el mercado laboral y contribuir al desarrollo tecnológico del país. Las técnicas aprendidas son escalables a empresas colombianas, las cuales en su mayoría, son ajenas a este tipo de análisis y de uso de la tecnología. Un modelo estrella bien diseñado optimiza el rendimiento de consultas.

ANEXOS

Anexo 1. Carpeta comprimida con la información y proyecto.

<https://drive.google.com/drive/folders/1bSpGYKlpGYLXdYIPSCW4p7Km4q8hAtfX?usp=sharing>

Trabajo Entrega No. 2 Y 3:

https://drive.google.com/drive/folders/1QKrserYWclbi2tNSJftfm31_jfWlghGQ?usp=drive_link

BIBLIOGRAFÍA

García, F., & Romero, O. (2015). *Data warehousing y business intelligence: Fundamentos, herramientas y aplicaciones*. Editorial UOC.

Hernández Orallo, J. (2004). *Data mining: Conceptos y herramientas*. Pearson Educación.

Inmon, W. H. (2005). *Construcción del data warehouse (3ª ed.)*. Prentice Hall.

Kimball, R., & Ross, M. (2013). *El data warehouse toolkit: Las técnicas definitivas para el diseño dimensional (3ª ed.)*. Wiley. (Traducción al español)

Marqués, J. (2018). *Business intelligence: Análisis y explotación de datos para la toma de decisiones*. Ediciones Paraninfo.

Pérez López, C. (2016). *Business intelligence: Data mining y análisis predictivo*. RC Libros.