ETL y Data Mart (Jardinería)

Jorge Andrés Echavarría Pardo
Yeison Padron Higuita
Juan Carlos Ardila Gómez
Edwin Alberto Ruíz

Docente: Antonio Jesús Valderrama
Grupo: PREICA2502B010064
SEMESTRE IV

Institución Universitaria Digital De Antioquia
Facultad de Ingenierías
Ingeniería en Software y Datos
Medellín, Colombia

Bases de Datos II
Septiembre 29 de 2025

Resumen

Este informe documenta el diseño e implementación de un proceso ETL para construir un Data Mart en esquema estrella a partir de la base de datos Jardinería. Se abordan las etapas de preparación, extracción, transformación (limpieza, normalización, enriquecimiento y SCD Tipo 2), y carga de registros en el Data Mart. Se incluyen las consultas de verificación y la evidencia de calidad de datos (QualityFlag, miembros Unknown, conteos y conciliación de montos).

Palabras clave: ETL, Data Mart, Modelo Estrella, Staging, SQL Server, SCD, Calidad de datos.

Introducción

El objetivo de un Data Mart es optimizar el análisis de información para un dominio específico. En este proyecto, la base Jardinería sirve como fuente para construir un Data Mart de ventas con un modelo dimensional que facilite consultas analíticas y métricas de negocio.

Objetivos

Objetivo general

Diseñar e implementar un proceso ETL de punta a punta para construir un Data Mart en esquema estrella a partir de la base de datos Jardinería, garantizando calidad, consistencia y trazabilidad de los datos

Objetivos específicos

- Analizar el modelo estrella propuesto y definir grano, dimensiones y hechos relevantes.
- Diseñar y poblar la base de datos de Staging, asegurando integridad y consistencia (por lotes).
- Aplicar transformaciones avanzadas: limpieza, normalización, generación de DimTiempo y SCD Tipo 2 en clientes.
- Cargar las dimensiones y hechos (FactVentas, FactPagos) de manera eficiente e idempotente.
- Implementar controles de calidad (QualityFlag, miembros Unknown, checks referenciales).
- Orquestar la ejecución con procedimientos almacenados y, opcionalmente, una tarea programada (SQL Agent).
- Evidenciar resultados con consultas de verificación y documentar hallazgos, limitaciones y mejoras.

Marco Conceptual

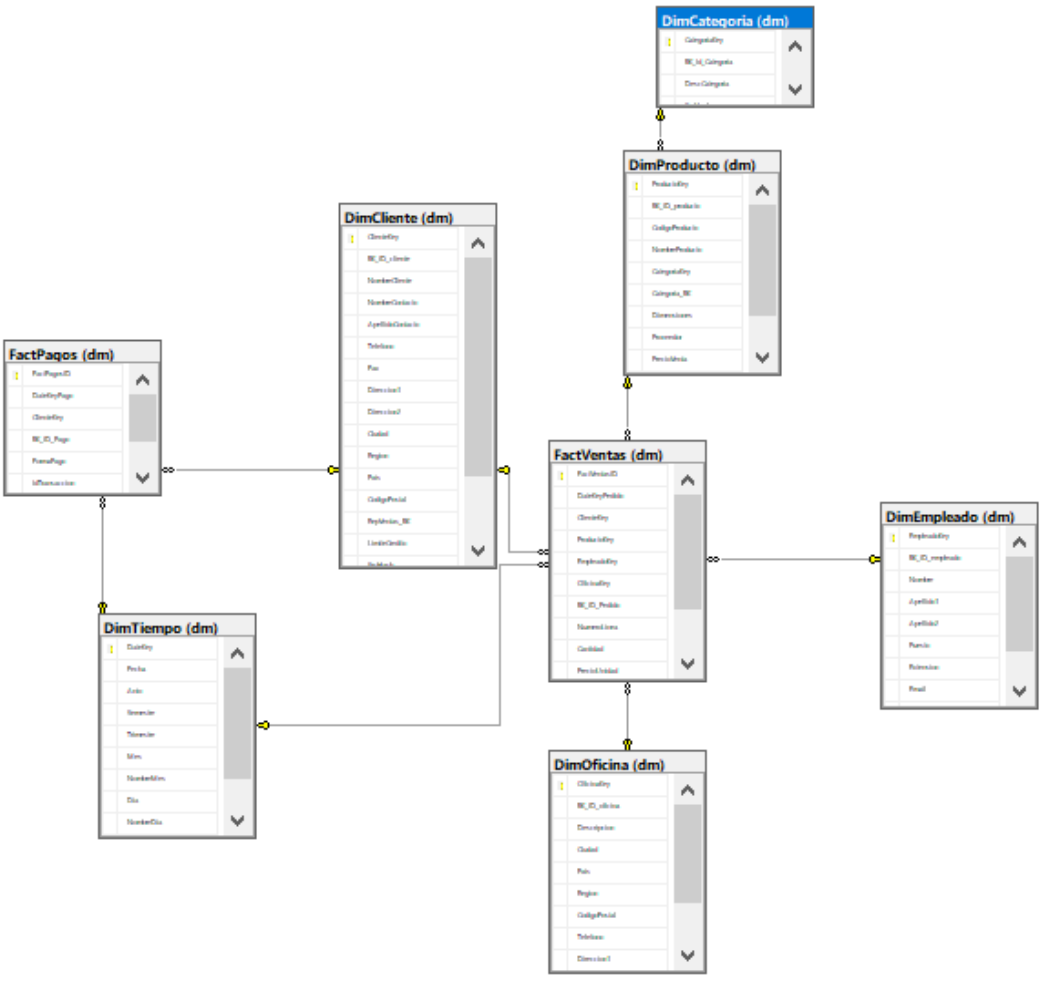
Modelo dimensional (estrella), procesos ETL, staging, Slowly Changing Dimensions (SCD), y métricas de calidad de datos constituyen los pilares técnicos del proyecto. La tabla de hechos se modela al nivel de detalle del negocio (grano); las dimensiones aportan contexto (cliente, producto, tiempo, empleado, oficina, categoría).

Metodología

Se siguió una metodología iterativa: (1) Preparación y análisis del modelo; (2) Extracción desde origen hacia staging; (3) Transformación (limpieza, normalización, enriquecimiento, SCD2); (4) Carga al Data Mart; (5) Verificación de calidad y validación cruzada; (6) Documentación.

Diseño del modelo estrella

Dimensiones: DimTiempo, DimCliente (SCD2), DimEmpleado, DimOficina, DimProducto y, opcionalmente, DimCategoria. Hechos: FactVentas (grano: detalle de pedido) y FactPagos. Se generan claves sustitutas y se resuelven relaciones a través de lookups desde staging.



Transformación de datos

Se implementó la función dm.udf_CleanTrim para limpieza de textos y se aplicó UPPER en campos geográficos; se generó DimTiempo; se gestionó SCD2 en DimCliente mediante cierre/apertura de versiones por hash; se incorporaron miembros Unknown en dimensiones para preservar la coherencia referencial.

```
-- 3.3 DimEmpleado
IF OBJECT_ID('dm.DimEmpleado') IS NULL
BEGIN
    CREATE TABLE dm.DimEmpleado (
        EmpleadoKey        INT IDENTITY(1,1) PRIMARY KEY,
        BK_ID_empleado     INT NOT NULL,
        Nombre             NVARCHAR(50) NULL,
        Apellido1          NVARCHAR(50) NULL,
        Apellido2          NVARCHAR(50) NULL,
        Puesto             NVARCHAR(50) NULL,
        Extension          NVARCHAR(10) NULL,
        Email              NVARCHAR(100) NULL,
        Oficina_BK         INT NULL,
        SrcHash            BINARY(16) NULL
    );
    CREATE UNIQUE INDEX UX_DimEmpleado_BK ON dm.DimEmpleado(BK_ID_empleado);
END;
GO
```

Carga de registros

Las cargas usan consultas set-based con tablas temporales de mapeo e índices, reemplazo por lote e inserciones en bloque con TABLOCK. FactVentas se enriquece con CostoUnitario y MargenLinea y marca QualityFlag para anomalías. El orquestador dm.usp_Load_DataMart ejecuta el flujo completo.

```
-- 5.1 DimTiempo
IF OBJECT_ID('dm.usp_Load_DimTiempo','P') IS NOT NULL DROP PROCEDURE dm.usp_Load_DimTiempo;
GO
CREATE PROCEDURE dm.usp_Load_DimTiempo
AS
BEGIN
    SET NOCOUNT ON;

    DECLARE @minDate DATE, @maxDate DATE;

    SELECT
        @minDate = (SELECT MIN(fecha_pedido) FROM jardineria_stg.stg.pedido),
        @maxDate = (SELECT MAX(fecha_pedido) FROM jardineria_stg.stg.pedido);

    IF @minDate IS NULL OR @maxDate IS NULL RETURN;

    SET @minDate = DATEADD(DAY, -7, @minDate);
    SET @maxDate = DATEADD(DAY,  7, @maxDate);

    ;WITH d AS (
        SELECT @minDate AS dt
        UNION ALL
```

Pruebas y resultados

Se verificaron: (a) conteos por tabla (v_DM_Counts); (b) checks de referencialidad (v_FactVentas_FKChecks); (c) distribución de QualityFlag; (d) uso de Unknown; (e) conciliación de Monto Fuente vs. Monto en DM. Las capturas de estas consultas se adjuntan en el Anexo A.

	Tabla	Registros
1	DimTiempo	1385
2	DimCliente	36
3	DimEmpleado	31
4	DimOficina	9
5	DimCategoria	5
6	DimProducto	276
7	FactVentas	636
8	FactPagos	52

Distrubucion Quality

	QualityFlag	Registros
	0	318

Sumatoria importe vs fuente

131 %

Results | Messages

	MontoFuente	MontoDM
1	217738.0000	217738.00

Conteos generales

	Tabla	Registros
1	DimTiempo	1385
2	DimCliente	36
3	DimEmpleado	31
4	DimOficina	9

Lista de factventas

	FactVentasID	DateKeyPedido	ClienteKey	ProductoKey	EmpleadoKey	OficinaKey	BK_ID_Pedido	NumeroLinea	Cantidad	PrecioUnidad	ImporteLinea	EstadoPedido	CostoUnitario
1	954	20081028	36	87	31	7	100	1	10	70.00	700.00	Rechazado	56.00
2	953	20090115	36	87	31	7	101	1	10	70.00	700.00	Entregado	56.00
3	952	20081129	36	87	31	7	102	1	10	70.00	700.00	Pendiente	56.00

Conclusiones

El Data Mart resultante soporta análisis de ventas confiable y reproducible. El proceso ETL diseñado ofrece idempotencia, trazabilidad y calidad. Se alcanzaron los objetivos propuestos con evidencia de verificación.

Referencias

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

American Psychological Association. (2020). Publication manual of the American Psychological Association (7th ed.). APA.

Anexo A. Evidencias de verificación

Adjunte aquí capturas de v_DM_Counts, v_FactVentas_FKChecks, QualityFlag, Unknown usage y conciliación de montos.

Bibliografía

Kimball, R., & Ross, M. (2013). The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.). Wiley.

Inmon, W. H. (2005). Building the Data Warehouse (4th ed.). Wiley.

El papel de las tablas de staging en la administración de base datos. Recuperado de https://www.baremon.eu/staging-tables-in-database-administration/

Uso de una base de datos de almacenamiento provisional en Parallel Data Warehouse (PDW). Recuperado de https://learn.microsoft.com/es-es/sql/analytics-platform-system/staging-database?view=aps-pdw-2016-au7