Diseño y Carga de un Área de Staging para la Base de Datos Jardinería

Juan Carlos Ardila Gómez

Jorge Andrés Echavarría Pardo

Edwin Alberto Ruiz Guzmán

Yeison Padron Higuita

Instituto Universitario Digital de Antioquia, Ingeniería de Software y Datos

Base de Datos II

Prof. Antonio Jesús Valderrama Jaramillo

11 de septiembre de 2025

Introducción

Este documento presenta el diseño, la construcción y la validación de un área de Staging para los datos del sistema Jardinería. El Staging es una zona de aterrizaje intermedia en la que los datos se almacenan prácticamente crudos, con normas mínimas de calidad, tipificación y trazabilidad, antes de ser transformados hacia un modelo dimensional (Data Warehouse). La propuesta se implementó en SQL Server e incluye scripts de creación de tablas, procesos de carga (INSERT…SELECT), validaciones y procedimientos de respaldo (backups) para garantizar la recuperabilidad del sistema.

Objetivos

• Diseñar la estructura de tablas Staging (raw + campos técnicos y normalizaciones mínimas).

• Construir consultas que trasladen los datos desde la base fuente Jardinería hacia Staging.

• Ejecutar validaciones para contrastar conteos y reglas básicas de calidad (fechas, claves, duplicados).

• Proveer scripts de backup completos para ambas bases (Jardinería y Jardinería\_Stg).

• Entregar documentación bajo lineamientos APA y anexos con el material técnico.

Planteamiento del problema

La organización requiere un repositorio analítico confiable para explotar información de ventas, clientes y productos. Actualmente, los datos del sistema transaccional “Jardinería” presentan heterogeneidades (mezcla de idiomas en países y regiones, categorías de producto como texto o entero, formatos de dimensiones, comentarios extensos) y posibles inconsistencias (fechas incoherentes, duplicados de clientes, entregas sin fecha). Se necesita una capa Staging que reciba los datos tal como llegan, aplique normalizaciones mínimas, incorpore trazabilidad por lote, y exponga datos listos para su posterior depuración y modelado dimensional.

Análisis del problema

Durante la revisión de las tablas fuente (oficina, empleado, categoría, cliente, pedido, producto, detalle\_pedido y pago) se identificaron:

1) Codificación de caracteres y acentos con mojibake en textos (p. ej., 'EspaÃ±a').

2) País/región/ciudad con nomenclaturas diversas (ESPAÑA, Spain, EEUU, USA, etc.).

3) Posibles duplicados de clientes (mismo nombre + teléfono + dirección).

4) Inconsistencias de fechas en pedidos (fecha\_esperada < fecha\_pedido; entregado sin fecha\_entrega).

5) Incompatibilidad del campo Categoria en producto (texto o id), lo que rompe la FK a Categoría.

6) Formatos heterogéneos en dimensiones de productos (decimales con coma, rangos n/m).

7) Teléfonos y códigos postales sin un estándar único.

La estrategia de Staging propuesta preserva el dato crudo y agrega columnas normalizadas mínimas, junto con metadatos de auditoría (\_batch\_id, \_ingestion\_ts) para permitir reprocesos, controles y trazabilidad completa.

Propuesta de la solución

## Correcciones a la entrega 1

• Se agregó una tabla de bitácora stg.\_etl\_batch\_log para trazar cada lote de carga

• Se diseñaron tablas stg.\* con claves compuestas por la PK de negocio + \_batch\_id, preservando histórico de ingestiones.

• Se añadió mapeo robusto de categorías (Categoria\_raw y Categoria\_id\_norm) para resolver incoherencias texto/entero.

• Se incluyeron normalizaciones mínimas (TRIM/UPPER/TRY\_CONVERT), manejo de dimensiones (dim\_valor\_decimal, dim\_min, dim\_max).

• Se estandarizó estado de pedidos (estado\_norm) y reglas de calidad para fechas y entregas.

• Se creó script de backups con compresión y verificación (RESTORE VERIFYONLY).

## Descripción del análisis y traslado de datos a Staging

Las tablas de origen se trasladan a su homóloga en Staging con el sufijo \_raw. En cada inserción se registra un \_batch\_id común para todas las filas cargadas en el ciclo, lo cual habilita auditorías y comparativas por lote. Se conservan campos de texto largos (comentarios, descripciones) y se agregan columnas normalizadas mínimas para facilitar el consumo por la siguiente capa (DW). Los principales puntos de tratamiento fueron:

• Productos: el campo Categoria se almacena como Categoria\_raw (texto) y se resuelve a Categoria\_id\_norm buscando por id numérico o por coincidencia del nombre en Categoría.

• Dimensiones de producto: se intentan convertir valores con coma a decimal y descomponer rangos n/m en dim\_min y dim\_max.

• Pedidos: se normaliza el estado a ENTREGADO/PENDIENTE/RECHAZADO y se validan reglas de fechas.

• Clientes: se propone estandarizar país a ISO-3 y detectar posibles duplicados.

• Todas las tablas: columnas técnicas (\_src\_system, \_batch\_id, \_ingestion\_ts) para trazabilidad.

Las consultas de carga y validación están incluidas como anexos .sql.

Anexos

• Scripts SQL: 01\_creacion\_staging.sql, 02\_carga\_staging.sql, 03\_validaciones.sql, 04\_backups.sql.

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