UNIVERSIDAD AUSTRAL | INGENIERÍA

Universidad Austral

Facultad de Ingeniería

Maestría en Ciencia de Datos

Trabajo Práctico Final

Data Warehousing

Autoras:

Alvaro, Giuliana

Guzmán, Mariana

Riera, Natalia

Prof. Eduardo A. Poggi

Prof. Esteban J. Alonso

5 de diciembre de 2025

Contenidos

1. Objetivo del Trabajo

Construir un Data Warehouse (DWA) basado en el modelo Northwind utilizando SQLite, aplicando buenas prácticas de ingeniería de datos:

1. Diseño de capas: TXT → TMP → DWA → DQM → DWM
2. Validación de calidad de datos (PK, FK, tipos, nulos, fechas, outliers)
3. Creación de un Data Quality Mart (DQM)
4. Carga inicial del Data Warehouse
5. Documentación y justificación de decisiones de calidad

2. Arquitectura Implementada

Se utilizó un único esquema físico, diferenciando las capas mediante prefijos, tal como recomiendan las consignas:

Capa	Prefijo	Descripción
Ingesta cruda	TXT_	Tablas espejo de los CSV, todos los campos en TEXT
Validación	TMP_	Tablas tipadas según el DER, con claves primarias
Data Warehouse	DWA_	Dimensiones y hechos finales
Data Quality Mart	DQM_	Proceso, entidades e indicadores
Capa de memoria	DWM_	Snapshot operativo (sólo para Employees, por diseño del TP)

3. Etapa 1 — Adquisición

3.1 Análisis inicial de los CSV

Se revisaron los archivos fuente identificando:

- Tipos inconsistentes
- Fechas con formatos mixtos
- Nulos inesperados
- IDs duplicados
- Campos que no cumplen el DER proporcionado

Hallazgos clave:

- employees.reports_to contiene valores NULL válidos (jefes sin supervisor) → Justificado como “nulo permitido”.
- orders.shipped_date con NULL para pedidos no enviados
- Fechas en formato "YYYY-MM-DD HH:MM:SS" → se corrigieron a “YYYY-MM-DD”.
- Sin duplicados de PK en origen.

3.2 Creación de tablas TXT

Se generaron tablas TXT con todos los campos TEXT, como exige la consigna.

3.3 Validación y carga a TMP

Se validaron tipos, rangos y referenciales antes de tipar.

Validaciones aplicadas en TXT → TMP:

- Tipado correcto de claves numéricas
- Conversión de fechas
- Revisión de rangos en precios, cantidades y descuentos
- PK no duplicada
- FK existente según DER
- Outliers básicos

Resultado: Todas las tablas pudieron cargarse en TMP sin rechazos.

4. Etapa 2 — Ingeniería

4.1 Creación de Metadatos (MET_)

Se documentaron entidades, claves y reglas aplicadas.

4.2 Creación del Modelo Dimensional (DWA)

Dimensiones generadas:

- DWA_DimEmployee
- DWA_DimCustomer
- DWA_DimProduct
- DWA_DimSupplier
- DWA_DimShipper
- DWA_DimCategory
- DWA_DimDate

Hecho generado:

- DWA_FactOrderDetails

Reglas aplicadas:

- Todas las dimensiones usan claves surrogate (INTEGER AUTOINCREMENT).
- La fact utiliza estas claves surrogate para garantizar integridad referencial.
- Se derivó revenue = quantity * unit_price * (1 - discount).

4.3 Diseño y construcción del DQM

Se creó un Data Quality Mart con 3 tablas:

- DQM_Processes: Registra cada corrida de validación.
- DQM_Entities: Registra el resultado por tabla.
- DQM_Indicators: Registra cada indicador medido, su umbral y si pasó o falló.

4.4 Indicadores implementados

Grupo 1 — Integridad estructural

- PK duplicada
- Campo NULL no permitido
- Tipos incorrectos

Grupo 2 — Integridad referencial

- FK inexistente en DimCustomer
- FK inexistente en DimEmployee
- FK inexistente en DimShipper
- FK inexistente en DimProduct
- FK inexistente en DimDate (order y shipped date)

Grupo 3 — Calidad de valores

- quantity <= 0
- unit_price <= 0
- discount fuera de rango (0–1)
- revenue < 0
- Fechas inválidas (formato correcto)

Justificación clave: Los NULL en shipped_date se consideraron válidos porque representan pedidos no enviados.

4.5 Resultados del DQM luego de correcciones

- PK sin duplicados
- Tipos correctos
- Descuentos en rango
- Revenue positivo
- Fechas válidas
- Todas las FK corregidas (match con surrogate keys)

El proceso finalizó en estado → OK – Sin errores críticos

4.6 Carga inicial del Data Warehouse (Paso 8.c)

Una vez verificados todos los indicadores:

- Se cargaron las dimensiones completas
- Se cargó la tabla de hechos
- Se marcaron los procesos como OK en DQM

4.7 Capa de Memoria (DWM)

Por diseño del TP y simplicidad, se incluyó solo DWM_Employee. Esta tabla almacena el estado operativo de la dimensión Employee, sin historial.

5. Etapa 3 — Actualización (Ingesta 2)

El objetivo de esta etapa fue incorporar un nuevo conjunto de archivos de origen (*Ingesta2*) que incluían **altas, bajas y modificaciones** a las entidades principales del modelo Northwind, además de integrar una nueva fuente de datos de geolocalización (*world-data-2023.csv*)

5.1 Ingesta a Capas TXT y TMP

La ingesta siguió el diseño de capas definido en la arquitectura: **TXT →TMP →DWA**.

- **Capa TXT_ (Ingesta cruda):** Los nuevos archivos (customers-novedades.csv, products-novedades.csv, orders-novedades.csv, order_details-novedades.csv y world-data-2023.csv) se persistieron "as is" en tablas espejo (e.g., TXT_Customers2, TXT_WorldData), con todos los campos en formato **TEXT**.
- **Capa TMP_ (Validación):** Se aplicaron **validaciones estructurales y de contenido** antes de la tipificación y carga a TMP, registrando todos los resultados en el Data Quality Mart (DQM).

5.2. Controles de Calidad y Correcciones

Grupo	Indicadores de Control (Ejemplos)	Correcciones Aplicadas
Integridad Estructural	Tipos mezclados, fechas mal parseadas.	Normalización de formatos de fecha ("YYYY-MM-DD hh:mm:ss" → "YYYY-MM-DD").
Integridad Referencial	FK inexistentes (e.g., customer_id, product_id).	Descarte de 1 registro con customer_id inexistente, siguiendo el criterio de no asignar orphan facts.
Calidad de Valores	IDs duplicados, valores numéricos inválidos.	Control de PK duplicada con COUNT > 1 y correcciones de rango.
Normalización	Países no reconocidos según TXT_WorldData (e.g., USA, UK)	Normalización de nombres de países en la tabla Customers para lograr consistencia (e.g., USA → United States), asegurando el match con la nueva dimensión DWA_DimCountry.

Se extendieron los controles de calidad para gestionar las inconsistencias de la nueva ingesta.

5.3. Propagación de Cambios al DWA y DWM

Se gestionaron las actualizaciones aplicando la lógica de *Slowly Changing Dimensions (SCD)* y *Merge Incremental*:

5.3.1 Altas: Los nuevos registros se **insertaron** en dimensiones (DWA_DimCustomer, DWA_DimProduct, DWA_DimDate, DWA_DimCountry) y en la tabla de hechos (DWA_FactOrderDetails).

5.3.2 Modificaciones (DWA): Se aplicó la estrategia de **SCD Tipo 0 (Sobreescritura)** para la mayoría de las dimensiones (e.g., cambios en datos de Customers y Products).

5.3.3 Modificaciones (DWM): En la Capa de Memoria (DWM_Employee), se aplicó **SCD Tipo 2**, generando un nuevo registro con los atributos actualizados (e.g., title, reports_to), actualizando los campos valid_to e is_current = 0 en el registro obsoleto.

5.3.4 Bajas: No se aplicaron bajas físicas (eliminaciones) para mantener el historial (criterio de DW). Los registros obsoletos en DWM quedaron marcados con is_current = 0.

6. Etapa 4 — Producto de Datos (Data Product)

El proyecto finalizó con la construcción de un **Data Product** integrado (SPM_FactSales), diseñado específicamente para el consumo de BI y análisis OLAP, cumpliendo con el objetivo de hacer el sistema "listo para análisis OLAP, reporting o dashboards".

6.1. Diseño de la Tabla SPM_FactSales

Se generó la tabla **SPM_FactSales** a partir de una vista desnormalizada (o materializada) de la tabla de hechos **DWA_FactOrderDetails**, uniendo todas las dimensiones clave mediante *LEFT JOINs* para facilitar la exploración y consulta del usuario final.

El Data Product **desnormaliza** la información de ventas para incluir:

- **Métricas Centrales:** revenue, quantity, discount.
- **Contexto de Negocio:** Nombre del producto, categoría, proveedor, y datos del cliente (nombre, región, país).
- **Información Geográfica:** Latitud y Longitud, obtenidas de la nueva dimensión **DWA_DimCountry**.
- **Logística:** order_date, shipped_date y la métrica derivada **ship_delay_days** (calculada como shipped_date_key - order_date_key).

6.2. Actualización de Metadata y DQM

- **DQM:** Se registró el nuevo proceso (Actualización Ingesta2) en DQM_Processes con estado **OK**, documentando la cantidad de registros actualizados, los indicadores validados y las correcciones (e.g., normalización de países).
- **Metadata (MET_):** Se actualizaron los metadatos para reflejar los cambios estructurales, incluyendo la **nueva dimensión DWA_DimCountry** y la documentación de las nuevas reglas de negocio (e.g., normalización de países, descarte de órdenes huérfanas).

6.3.Decisiones tomadas y justificación

- reports_to NULL: Es un nulo válido funcionalmente (directores sin supervisor).

- shipped_date NULL: Las órdenes no enviadas en Northwind tienen fecha nula.
- Uso de surrogate keys: Se aplicó para mantener independencia del sistema transaccional y buena práctica en DW.
- Fechas con hora: Se normalizaron para cumplir el estándar “YYYY-MM-DD”.
- Discount en rango: Los descuentos válidos de Northwind son 0, 0.05, 0.1, 0.15, 0.2, etc. Se validó contra rango 0–1.
- No se implementó capa de enriquecimiento

6.4.Tablero Sales Performance Overview

El dashboard desarrollado permite analizar el rendimiento de ventas completo de Northwind a través de una vista integrada y dinámica de ventas, pedidos, clientes y productos. Su diseño prioriza claridad, exploración interactiva y comprensión rápida de los drivers clave del negocio.

El tablero ofrece una visión ejecutiva del desempeño anual e incluye:

- Vista general del negocio: métricas principales con KPIs de ventas totales, pedidos, ticket promedio, cantidades vendidas y clientes únicos.
- Comportamiento temporal: evolución mensual de ventas para identificar estacionalidades, picos y caídas.
- Distribución geográfica: mapa de calor que muestra los países con mayor volumen de ventas.
- Mix de categorías: visualización de ventas por categoría para entender qué líneas de productos impulsan la mayor parte de la facturación.
- Análisis de concentración: Estas secciones permiten ver cuáles son los productos más relevantes y los clientes que más contribuyen al negocio.
    - Top 10 Most Sold Products
    - Top 10 Customers

Además, el panel lateral permite filtrar dinámicamente por año, mes y categoría, haciendo que el análisis sea flexible y adaptable a distintos escenarios.

Métricas calculadas

- El dashboard se construyó a partir de medidas en DAX que permiten calcular:
- Total Sales = suma de las ventas
- Total Orders = suma de las órdenes
- Average Order Value (AOV) = sales/orders, precio promedio por orden
- Total Quantities Sold = unidades totales vendidas
- Unique Clients = clientes unicos

Estas métricas permiten una lectura integral del negocio, desde la visión global hasta el detalle por categoría, producto o cliente.

7. Conclusiones

El Data Warehouse quedó correctamente construido aplicando:

- Buenas prácticas de ingeniería de datos
- Modelo dimensional coherente
- Validaciones exhaustivas de calidad
- Trazabilidad completa vía DQM
- Correcciones justificadas
- Snapshot en capa de memoria

El sistema está listo para análisis OLAP, reporting o dashboards.

8. Links de referencia

Repositorio: GitHub

Tablero: Power BI