Maestría en Ciencia de Datos (Virtual)

Facultad de Ingeniería

Trabajo Práctico - Introducción a Data Warehousing

Flujo de datos en un DWA

Docentes:

- Esteban Alonso
- Eduardo Poggi

Integrantes del Grupo:

- Brusasca, Lucas
- Durán, Pedro
- Gaddi, Martín
- Lijtmaer, Paul
- Palavecino, Nicolás

Fecha de entrega: 3 de Julio 2025

Repositorio de GitHub: https://github.com/plijtmaer/data-warehousing-tp-grupal

Índice

Resumen Ejecutivo

Este informe presenta, de manera resumida, el desarrollo completo de un Data Warehouse Analítico basado en datos del sistema Northwind, orientado al análisis de ventas, productos, clientes, empleados y desempeño de la organización. El proyecto abarca desde la ingesta inicial y los tratamientos de datos, hasta la generación de productos analíticos de valor y su visualización mediante dashboards (en una herramienta de visualización de elección del equipo: Power BI).

El foco se centró en construir un sistema robusto, confiable y alineado con las mejores prácticas de Data Warehousing, aprendidas a lo largo de la materia, habilitando la detección de insights concretos para la toma de decisiones estratégicas.

Descripción de la Base de Datos Fuente

Northwind simula un entorno real de negocio basado en la industria alimenticia y de bebidas, incluyendo datos de clientes, productos, categorías, empleados, proveedores, transportistas y órdenes de compra.

Esta base de datos fue proveída en forma de 12 csv, con la siguiente composición:

- customers.csv
- products.csv
- orders.csv
- order_details.csv
- employees.csv
- categories.csv
- suppliers.csv
- shippers.csv
- regions.csv
- territories.csv
- employee_territories.csv

La estructura original fue enriquecida con información externa proveniente de 198 países (world-data-2023.csv), la cual incorpora indicadores socioeconómicos como GDP, expectativa de vida, densidad poblacional y otros factores relevantes que aportan contexto al análisis comercial y geográfico. La incorporación de datos externos, agregó una dificultad adicional, al requerir normalización de los datos para su posterior conexión y explotación. Este proceso se explica en los apartados posteriores.

Procesos de Construcción del Data Warehouse

La creación del Data Warehouse se abordó siguiendo un proceso escalonado que garantiza la limpieza, estandarización y transformación progresiva de los datos hasta su disponibilidad final para el análisis:

Área Temporal (TMP_) - Recepción y Validación Inicial de Datos

En esta etapa se realiza la ingesta inicial de los archivos .csv, conservando la estructura relacional original y respetando llaves primarias y foráneas. Esta capa funciona como una zona de aislamiento que permite inspeccionar los datos en su formato bruto, sin alteraciones ni transformaciones. Se validan los registros incompletos, se identifican potenciales inconsistencias y se prepara el terreno para su tratamiento posterior.

Capa Staging (STG_) - Transformación y Homogeneización de Datos

En la capa STG_ se aplican transformaciones intermedias indispensables para asegurar la coherencia y calidad de los datos antes de su carga al modelo dimensional. Las tareas incluyen:

- Estandarización de nombres de países, regiones y categorías.
- Normalización de formatos numéricos, fechas y campos de texto.
- Corrección de errores de codificación (acentos, símbolos, caracteres especiales).
- Validación de relaciones jerárquicas, especialmente en estructuras como empleados y territorios.

Esta capa sirve de filtro de calidad y asegura que los datos cumplen los estándares requeridos.

Modelo Dimensional (DWH_) - Diseño Analítico basado en Esquema Estrella

El corazón del Data Warehouse se construye bajo un esquema copo de nieve, permitiendo una organización eficiente y flexible de los datos para el análisis.

El modelo comprende 10 dimensiones/sub-dimensiones (clientes, productos, tiempo, empleados, proveedores, categorías, transportistas, regiones, países y territorios) y 2 tablas de hechos (órdenes y detalles de productos vendidos) conectadas entre sí en el centro. La estructura favorece consultas rápidas, análisis multidimensionales y la integración de múltiples fuentes de información.

Capas de Memoria e Historización (MEM_) - Conservación de Evolución Histórica

Se implementan mecanismos para preservar el historial de cambios en los datos maestros clave, como clientes, productos y empleados. Esto permite realizar análisis evolutivos, comparar estados en diferentes momentos y comprender la dinámica del negocio a lo largo del tiempo.

Enriquecimiento Analítico (ENR_) - Generación de Indicadores Estratégicos

En esta capa se calculan métricas adicionales, rankings y segmentaciones avanzadas que amplifican el valor analítico de los datos. Algunos ejemplos incluyen la clasificación de clientes según facturación acumulada y frecuencia de compra, el ranking de productos por unidades vendidas y contribución al ingreso total, y métricas de desempeño desagregadas por región, empleado y categoría. Este enriquecimiento facilita la identificación de patrones, tendencias y oportunidades comerciales.

Sistema de Metadata - Documentación y Trazabilidad Integral

Se desarrolla una documentación centralizada y detallada que describe técnicamente la estructura de tablas y campos, las relaciones entre entidades, y las transformaciones aplicadas. La categorización por tipos de tabla (staging, dimension, hecho, etc.) permite gestionar el ciclo de vida de los datos de manera sistemática. Cada campo está documentado con su tipo de dato, descripción funcional y origen, creando un diccionario de datos completo que facilita tanto el desarrollo como el uso analítico del DWH.

Esta documentación asegura trazabilidad completa, facilita auditorías, promueve la gobernanza de datos y simplifica futuras expansiones o mantenimientos. Esta documentación es crucial para el mantenimiento a largo plazo y facilita la incorporación de nuevos usuarios al sistema.

Data Quality Management (DQM) - Monitoreo de Calidad de Datos

El sistema DQM garantiza la calidad continua de los datos a través de métricas de completitud, unicidad y consistencia, reglas de validación configurables para cada entidad y umbrales que definen estándares mínimos de aceptación. A su vez, cuenta con un motor de decisiones que automatiza la aprobación o rechazo de procesos de carga. Este enfoque proactivo minimiza errores y asegura la confiabilidad del Data Warehouse.

Este sistema cuenta con cuatro tablas:

1. DQM_Procesos: registro de procesos ejecutados

2. DQM_Indicadores: métricas de calidad con umbrales
3. DQM_Descriptivos: estadísticas de datos procesados
4. DQM_Reglas: reglas de calidad configurables

Parte de estas se utilizará más adelante para analizar la performance del DWH, a través de una visualización en power BI.

Actualización Incremental (Ingesta2) - Incorporación Segura de Nuevos Datos

Se diseña un proceso robusto que permite incorporar nuevos registros o modificaciones de forma controlada. Se contempla la identificación de cambios (altas, bajas, modificaciones), la preservación del historial de datos relevantes, y se implementan controles de integridad antes y después de la carga. Este mecanismo garantiza que el DWH se mantenga actualizado, sin comprometer la integridad ni la coherencia de la información histórica.

Como primer paso, los datos nuevos se cargaron en el área temporal a través de tablas espejo, las cuales incorporan un campo que identifica el tipo de operación (inserción, modificación o eliminación). De esta manera, se mantiene la trazabilidad de cada registro sin comprometer la estructura productiva. Esta segunda ingestión incorporó 270 órdenes nuevas, 691 líneas de detalle asociadas y actualizaciones sobre dos clientes existentes.

Antes de avanzar, se aplicaron controles específicos para este tipo de actualización incremental. Los umbrales de calidad se adaptaron al contexto, con exigencias elevadas para la integridad referencial y niveles de completitud y validez de los datos acordes a los estándares definidos. Se verificó que el 100% de los clientes fueran válidos y que más del 98% de las órdenes cumplieran con las reglas de negocio, detectándose algunos valores atípicos que fueron marcados para revisión.

Una vez validados los datos, se ejecutó el motor de decisiones automatizado. Este mecanismo analiza los resultados de los controles de calidad y decide si corresponde continuar, procesar parcialmente o cancelar la actualización. En esta instancia, se identificó un error crítico de integridad referencial y una alta cantidad de registros duplicados, lo que activó automáticamente la cancelación del proceso. Gracias a esta decisión, el DWH se mantuvo protegido y sin alteraciones.

Como parte del proceso, se generaron snapshots históricos previos a la actualización, que permiten comparar el estado de los datos antes y después de cada operación. Esta práctica garantiza la trazabilidad y facilita las auditorías futuras. Asimismo, se actualizaron las capas de enriquecimiento y métricas analíticas, pero únicamente en los casos donde se aplicaron cambios, optimizando así los tiempos y recursos.

Todo el proceso se encuentra estandarizado mediante scripts secuenciales que aseguran el orden, la validación de prerequisitos y la coherencia general. Al finalizar, se

actualizaron los sistemas de metadata y los indicadores del sistema de calidad, registrando el estado de las tablas, los campos, los históricos de ingestiones y los resultados de los controles.

Es importante destacar que el fallo controlado de Ingesta2 constituye una demostración concreta del valor de la arquitectura implementada. El sistema reaccionó de forma autónoma, protegiendo los datos, bloqueando el avance de registros problemáticos y dejando evidencia documentada de todo el proceso. Este comportamiento refleja la madurez alcanzada, con un DWH que incorpora prácticas empresariales robustas como detección automática de cambios, preservación histórica, controles de calidad avanzados, auditoría integral y mecanismos de recuperación ante errores.

Tratamientos y Controles Aplicados

Durante todo el pipeline de construcción se aplicaron diversas técnicas de limpieza, estandarización y control para asegurar la confiabilidad de los datos. Muchos de ellos mencionados en la etapa previa de ingesta.

En el marco de los tratamientos y controles aplicados, el primer paso consistió en la estandarización de países y regiones, abordando inconsistencias en la denominación de territorios para unificar criterios. Por ejemplo, normalizando los países “United States”, “United Kingdom” y “Republic of Ireland”, a “USA”, “UK” e “Ireland” respectivamente. Esto era especialmente importante para la incorporación de los datos externos al dataset (proveniente de los datos de World_data).

Como segundo paso, se procedió a la corrección de encoding y formatos, solucionando problemas de acentuación, caracteres especiales y símbolos en campos numéricos para mejorar la calidad de los datos.

En una tercera instancia, se implementaron validaciones de integridad referencial y jerárquica, asegurando que las órdenes hicieran referencia a clientes, empleados y transportistas válidos, y que las jerarquías internas no presentaran ciclos o errores. Posteriormente, como cuarto paso, se incorporó enriquecimiento con datos externos (ya estandarizados en pasos previos), especialmente indicadores macroeconómicos como el GDP o la expectativa de vida, fortaleciendo el análisis geográfico y la correlación con el desempeño comercial.

El quinto paso abarcó el cálculo de métricas avanzadas y la segmentación, incluyendo la elaboración de rankings de productos más vendidos, la categorización de clientes según su valor y la identificación de las personas con mejor desempeño, así como el análisis de eficiencia logística mediante tiempos de entrega.

Por último, se llevaron a cabo controles específicos en los procesos de actualización incremental. Esto implicó validar la integración de nuevos datos, detectar duplicados y aplicar chequeos adicionales de integridad y calidad, asegurando la coherencia y robustez del modelo en cada etapa de actualización.

Productos de Datos Generados

Luego del desarrollo del Warehouse y su modelado, se generaron productos analíticos diseñados para responder a preguntas estratégicas del negocio, entre los que se destacan:

- Ventas consolidadas por categoría y país.
- Evolución mensual y estacional de los pedidos.
- Top 10 de productos más vendidos.
- Distribución de pedidos y desempeño por empleado.
- Ranking de clientes por facturación.
- Análisis de tiempos promedio de entrega por país.
- Frecuencia de compra y lealtad de clientes.
- Porcentaje de entregas rápidas y su impacto.

Estos productos están integrados en el DWH y preparados para ser explotados mediante reportes y dashboards.

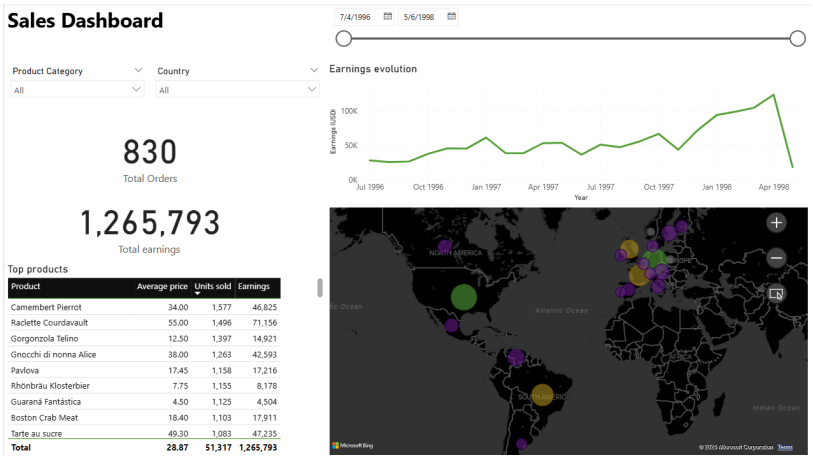
Visualización y Dashboards en Power BI

Se diseñaron dashboards en Power BI que sintetizan los principales indicadores y métricas de negocio. Para probar la funcionalidad de la herramienta, se realizaron distintos reportes.

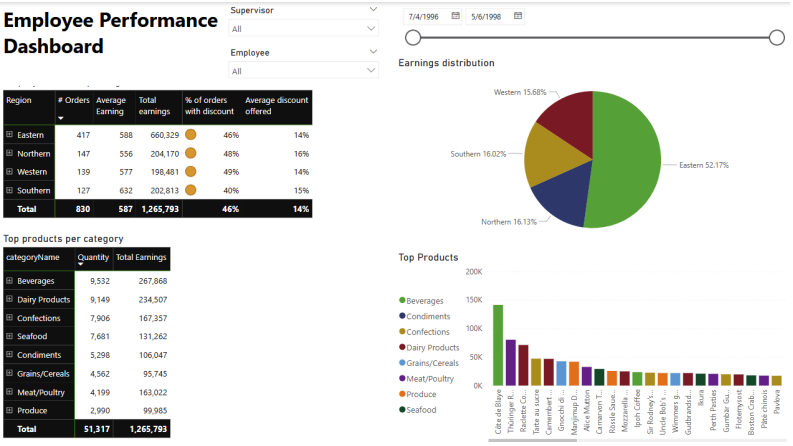
En primer lugar, uno conectado directamente con el Warehouse puro (para validar la integridad de las relaciones, y probar su cohesión). Este se entrega como DB_TP_DW (conectado a DWA). El modelo resultante, una vez conectado, nos muestra que el trabajo realizado en el Warehouse es correcto, y sólido.

El *Dashboard* se encuentra divido en cuatro hojas, cada una referenciando áreas de interés diferentes (Ventas, performance de empleados, *deliveries*, y *customer reviews*).

Ventas:



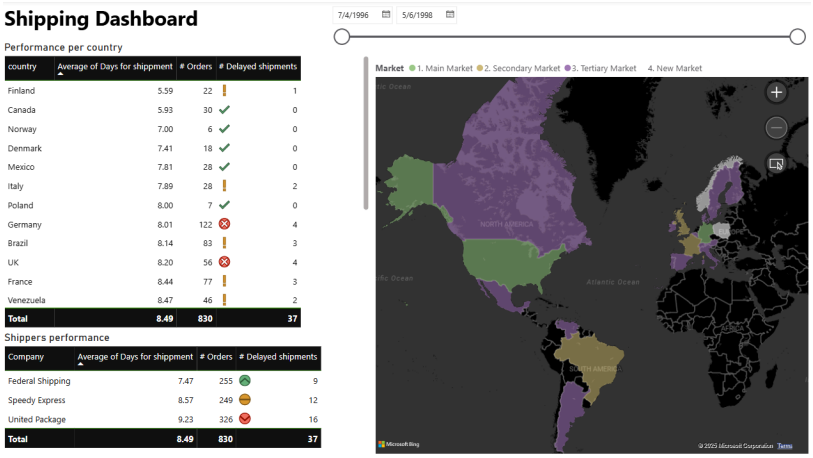
Employees:



Region	# Orders	Average Earning	Total earnings	% of orders with discount	Average discount offered
Eastern	417	588	660,329	46%	14%
Northern	147	556	204,170	48%	16%
Western	139	577	198,481	49%	14%
Southern	127	632	202,813	40%	15%
Total	**830**	**587**	**1,265,793**	**46%**	**14%**

categoryName	Quantity	Total Earnings
Beverages	9,532	267,868
Dairy Products	9,149	234,507
Confections	7,906	167,357
Seafood	7,681	131,262
Condiments	5,298	106,047
Grains/Cereals	4,562	95,745
Meat/Poultry	4,199	163,022
Produce	2,990	99,985
Total	**51,317**	**1,265,793**

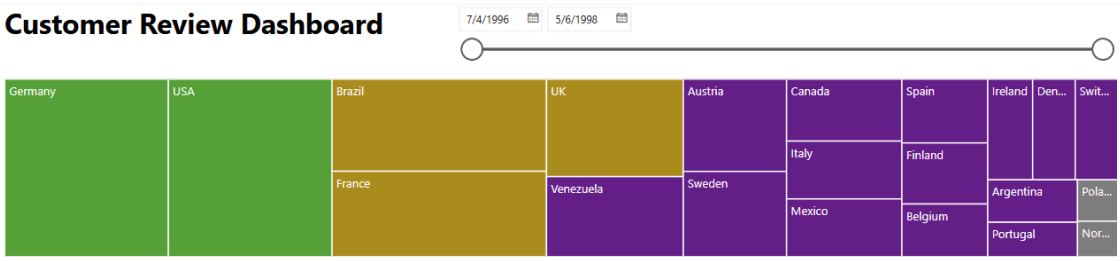
Deliveries:



country	Average of Days for shipment	# Orders	# Delayed shipments
Finland	5.59	22	1
Canada	5.93	30	0
Norway	7.00	6	0
Denmark	7.41	18	0
Mexico	7.81	28	0
Italy	7.89	28	2
Poland	8.00	7	0
Germany	8.01	122	4
Brazil	8.14	83	3
UK	8.20	56	4
France	8.44	77	3
Venezuela	8.47	46	2
Total	**8.49**	**830**	**37**

Company	Average of Days for shipment	# Orders	# Delayed shipments
Federal Shipping	7.47	255	9
Speedy Express	8.57	249	12
United Package	9.23	326	16
Total	**8.49**	**830**	**37**

Customer Review:



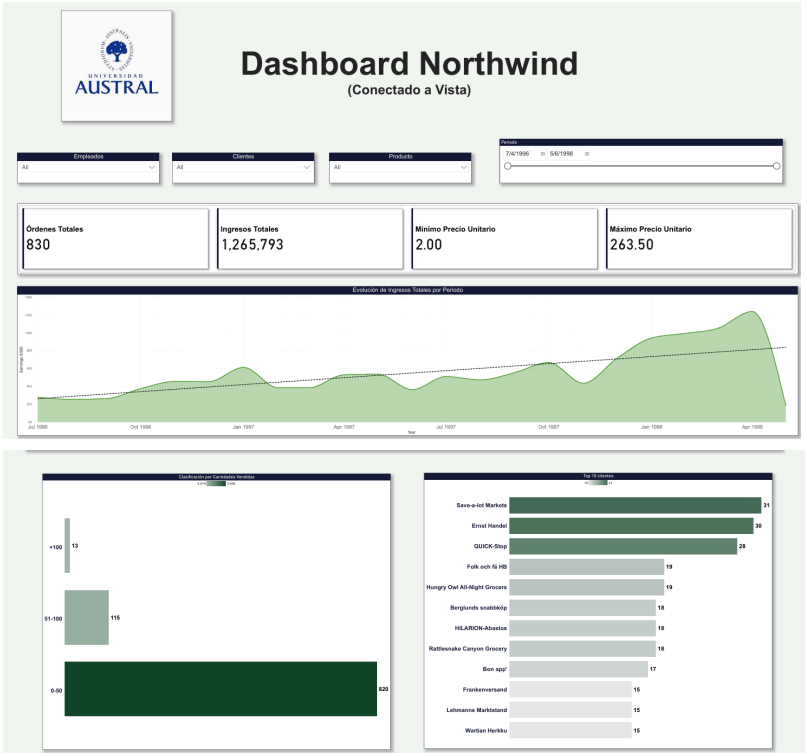
companyName	country	# Orders	Total paid	% of orders with discount	Average discount offered	Average unit price	phone	contactName	address
QUICK-Stop	Germany	28	110,277.32	64%	13%	31.9	0372-035188	Horst Kloss	Taucherstraße 10
Ernst Handel	Austria	30	104,874.99	63%	14%	26.1	7675-3425	Roland Mendel	Kirchgasse 6
Save-a-lot Markets	USA	31	104,361.96	61%	18%	23.1	(208) 555-8097	Jose Pavarotti	187 Suffolk Ln.
Rattlesnake Canyon Grocery	USA	18	51,097.79	50%	10%	30.7	(505) 555-5939	Paula Wilson	2817 Milton Dr.
Hungry Owl All-Night Grocers	Ireland	19	49,979.91	79%	18%	31.3	2967 542	Patricia McKenna	8 Johnstown Road
Hanari Carnes	Brazil	14	32,841.37	43%	17%	43.6	(21) 555-0091	Mario Pontes	Rua do Paço 67
Königlich Essen	Germany	14	30,908.39	50%	12%	31.5	0555-09876	Philip Cramer	Maubelstr. 90
Folk och fä HB	Sweden	19	29,567.57	63%	18%	23.6	0695-34 67 21	Maria Larsson	Åkergatan 24
Mère Paillarde	Canada	13	28,872.19	62%	11%	29.8	(514) 555-8054	Jean Fresnière	43 rue St. Laurent
White Clover Markets	USA	14	27,363.60	50%	15%	32.0	(206) 555-4112	Karl Jablonski	305 - 14th Ave. S. Su...
Frankenversand	Germany	15	26,656.57	67%	14%	20.5	089-0877310	Peter Franken	Berliner Platz 43
Total		**830**	**1,265,793.15**	**46%**	**14%**	**26.2**			

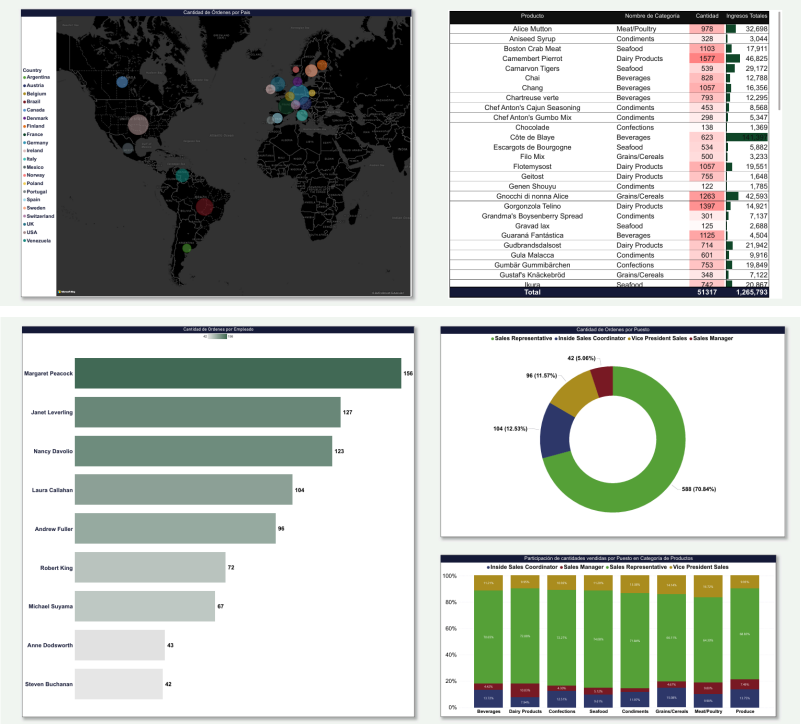
En arquitecturas de Data Warehouse diseñadas para entornos reales de alta demanda y grandes volúmenes de datos, es ampliamente reconocido que una conexión directa desde herramientas de visualización al modelo completo del DWH puede generar múltiples inconvenientes. Fundamentalmente, problemas de performance en las consultas, la saturación de los recursos del servidor y, en ciertos casos, la exposición innecesaria de datos que no son relevantes para los usuarios finales.

Asimismo, permitir el acceso directo al DWH puede derivar en redundancia de cálculos, ya que cada usuario o dashboard podría replicar transformaciones o métricas, aumentando la carga sobre la infraestructura y generando potenciales inconsistencias en los resultados. Por estas razones, se adoptó luego un enfoque más eficiente y seguro, que consiste en la generación de reportes en base a los productos de datos generados previamente, las cuales sintetizan los principales indicadores y métricas clave, prefiltradas y modeladas para los casos de uso específicos. Esto garantiza mayor eficiencia, menor tiempo de respuesta en las consultas y evita la redundancia de cálculos o transformaciones al momento de visualizar la información.

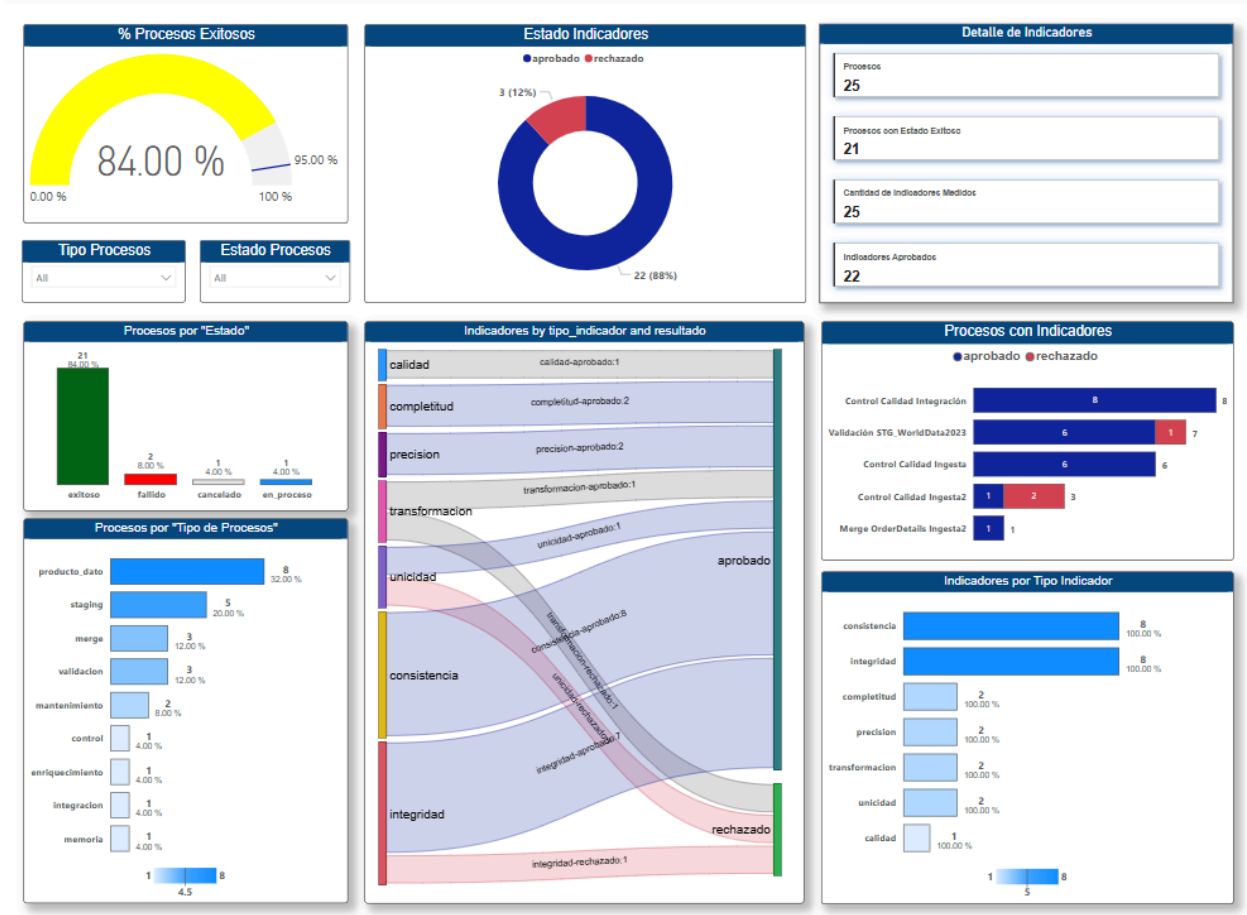
Este esquema representa el producto final de consumo analítico, donde los usuarios acceden únicamente a datos preparados, confiables y diseñados para la toma de decisiones, respetando las mejores prácticas de rendimiento y seguridad en entornos empresariales.

Se adjuntan visualizaciones para mostrar los resultados, y se adjunta en la entrega el “DB_TP_DW (Conectado a Vista)”.





Por último, se realizó un dashboard sobre el DQM, para observar la performance del DWA en relación a los procesos internos. El mismo se adjunta con el nombre “TPDW – DQM”.



Insights y Findings Principales

En base a los análisis que se pudieron realizar durante el armado, y los resultantes de los productos de datos y visualizaciones generados, se pueden sacar múltiples conclusiones relevantes al negocio.

Para mencionar solo algunos, y evidenciar la potencia de estos desarrollos, podemos observar:

1. Existe una alta concentración de ventas en 5 clientes estratégicos que generan el 33% del total de ingresos de la empresa. En todos ellos, más de la mitad de las órdenes tuvieron descuentos.

2. Liderazgo de la categoría “Beverages”, cercano a la segunda, “Dairy Products”. Entre ambos, representan casi el 40% del total de facturación. Hay un producto de bebida en particular (“Côte de Blaye”) que por sí solo representa el 11% de la facturación total, siendo solamente parte del 1% de las órdenes de venta realizadas. Podríamos decir, el producto de mejor margen.

3. Picos estacionales de ventas en el cuarto trimestre, asociados a campañas navideñas y de fin de año.

4. Correlación positiva entre el GDP de los países y el ticket promedio, validando el enfoque geográfico. Los dos países con mayor facturación, a su vez, son dos de los países mejor posicionados en esa variable (USA, y Alemania).

5. Identificación de Margaret Peacock como la empleada con mayor desempeño en ventas (tanto en cantidad como en facturación). En relación a la performance, Steve Buchanan es el que más proporción de órdenes con descuentos otorgados tiene. Y a su vez, es el de menor cantidad y valor de ventas posee.

6. Detección de un 5% de entregas fuera de los plazos establecidos, con disparidad entra las empresas de transporte en cuánto a la performance (siendo Federal Shipping la que menos tiempo toma en promedio, y la que menos pedidos tardíos posee). Esta tendencia NO es igual por país, ya que por ejemplo, en Brasil, Federal Shipping es la de peor performance. Existe posibilidad de mejorar la logística, delimitando los proveedores a trabajar en las regiones que mejor performen.

7. Se observa que los clientes con menor cantidad de órdenes, prácticamente no tuvieron nunca un descuento. Lo que refleja una política de empresa orientada a "premiar" la lealtad de sus clientes con descuentos, en vez de ofrecer descuentos en pos de obtener nuevos clientes. Esto no es un input para la empresa, pero sí para aquellos que busquen trabajar con la misma o competir.

Estas son solo algunas de las conclusiones que se puede obtener. Para más inputs, pueden utilizarse los dashboards generados.

Conclusión

La implementación integral de este Data Warehouse demostró ser mucho más que un simple ejercicio técnico: se consolidó como una herramienta concreta de apoyo a la toma de decisiones, capaz de transformar datos dispersos y operativos en información estructurada, confiable y enfocada en las verdaderas necesidades del negocio.

El enfoque modular permitió construir una arquitectura ordenada y escalable, que separa de forma clara cada etapa del proceso: desde la ingesta de datos crudos, los controles de calidad, la integración de fuentes externas, hasta la publicación de métricas listas para ser consumidas. Este diseño, combinado con mecanismos de control, auditoría y enriquecimiento, garantiza no solo la precisión de los datos, sino también su contexto y relevancia para quienes deben interpretarlos.

Los análisis realizados sobre el modelo permitieron detectar patrones comerciales relevantes, como la alta concentración de ventas en pocos clientes, tendencias estacionales y oportunidades de mejora en distintos ámbitos de la empresa. A su vez, las métricas de desempeño por empleado, producto y país brindan una visión concreta

que ayuda a entender dónde se generan los mejores resultados y dónde es necesario ajustar las estrategias.

Más allá de lo técnico, este proyecto refleja la importancia de aplicar un enfoque metódico y robusto en la gestión de datos, como requisito indispensable para que los sistemas de información se conviertan en aliados reales del negocio. No se trata solo de almacenar datos, sino de organizarlos, depurarlos y convertirlos en una fuente de valor accesible, confiable y accionable para todos los niveles de la organización.

Este trabajo representa un paso clave en la construcción de una cultura basada en datos, sentando las bases para un ecosistema analítico que, con el tiempo, podrá evolucionar, escalar y seguir generando valor de manera sostenida. Y tener la posibilidad de desarrollarlo en el ámbito académico, permite toparnos con los desafíos de la vida laboral, en un contexto de menor riesgo. Algo que enriquece nuestras capacidades de llevarlo a cabo en el mundo corporativo más adelante.