Universidad CENFOTEC

Maestría en Tecnología de Bases de Datos

Documento final de Proyecto de Investigación Aplicada 2

Tema: Esquema de integración y análisis de datos basado en metodologías ágiles para el fortalecimiento de la gestión en una empresa de insumos industriales SumIndustrial S.A.

Céspedes Torres, Juan Pablo.

Fecha: Julio, 2026.

Esquema de inteligencia de negocios para una distribuidora B2B de insumos industriales

Juan Pablo Céspedes Torres
Maestría en Bases de Datos, Universidad CENFOTEC, Costa Rica

Resumen Ejecutivo—La fragmentación de los datos y la baja madurez analítica limitan la toma de decisiones en las distribuidoras de insumos industriales, que acumulan grandes volúmenes transaccionales sin capacidad de explotarlos estratégicamente. Este estudio tuvo como objetivo diseñar y validar un esquema de integración y análisis de datos para una distribuidora B2B de insumos industriales con cinco sedes en Costa Rica. Se adoptó un enfoque cuantitativo de alcance descriptivo-analítico, bajo un diseño no experimental de corte transversal estructurado como estudio de caso. El procedimiento articuló un diagnóstico de la gestión de datos en cuatro dimensiones, el diseño de una arquitectura por capas con un modelo dimensional en esquema estrella (tres tablas de hechos y seis dimensiones compartidas) y la implementación de cinco análisis sobre datos sintéticos: exploración OLAP, clasificación ABC/XYZ de inventario, segmentación de clientes con K-means, embudo de conversión de cotizaciones y pronóstico de demanda con ARIMA. El almacén integró 32.383 líneas de orden, 72.048 de cotización y 450.000 registros de inventario. La segmentación distinguió dos arquetipos de cliente (silueta = 0,34); el embudo reveló una conversión de cotización a orden del 45 %; la clasificación ABC concentró el 80 % de las ventas en 1.335 SKU; y el modelo ARIMA(0,1,2) generó pronósticos mensuales con un MAPE del 13,8 %. Un asistente conversacional, además, tradujo preguntas en lenguaje natural a consultas ejecutadas sobre el almacén, con respuestas verificadas contra los datos. Se concluye que la separación entre los entornos transaccional y analítico, junto con el modelado dimensional, constituye una condición habilitante para una analítica accionable, y que la accesibilidad de la interfaz determina el valor real de la solución tanto como su sofisticación técnica.

Abstract—*Data fragmentation and low analytical maturity limit decision-making in industrial-supply distributors, which accumulate large transactional volumes without the ability to exploit them strategically. This study aimed to design and validate a data integration and analysis scheme for a B2B industrial-supply distributor with five branches in Costa Rica. A quantitative, descriptive-analytical approach was adopted under a non-experimental, cross-sectional design structured as a case study. The procedure combined a four-dimension assessment of data management, the design of a layered architecture with a star-schema dimensional model (three fact tables and six shared dimensions), and the implementation of five analyses on synthetic data: OLAP exploration, ABC/XYZ inventory classification, K-means customer segmentation, a quotation-to-order conversion funnel, and ARIMA demand forecasting. The warehouse integrated 32,383 order lines, 72,048 quotation lines, and 450,000 inventory records. Segmentation distinguished two customer archetypes (silhouette = 0.34); the funnel revealed a 45% quotation-to-order conversion; the ABC classification concentrated 80% of sales in 1,335 SKUs; and the ARIMA(0,1,2) model produced monthly forecasts with a MAPE of 13.8%. A conversational assistant additionally translated natural-language questions into queries executed against the warehouse, with answers verified against the data. It is concluded that separating the transactional and analytical environments, together with dimensional modeling, is an enabling condition for actionable analytics, and that interface accessibility determines the real value of the solution as much as its technical sophistication.*

Palabras clave—inteligencia de negocios, modelado dimensional, OLAP, segmentación de clientes, pronóstico de demanda, consulta en lenguaje natural, insumos industriales.

I. INTRODUCCIÓN

La distribución de insumos industriales opera en un mercado donde la eficiencia de la cadena de suministro, la optimización de inventarios y la anticipación de la demanda determinan la competitividad. Estas empresas generan volúmenes crecientes de datos de ventas, facturación, inventario y relación con clientes, pero el crecimiento del dato rara vez viene acompañado de una estrategia para explotarlo. Chen, Chiang y Storey [1] describieron precisamente ese tránsito pendiente: pasar del big data al impacto real sobre la organización. La promesa de la inteligencia de negocios consiste en convertir registros operativos en conocimiento para decidir, y su base técnica está bien establecida. Kimball y Ross [2] e Inmon [3] coinciden en que un almacén de datos (integrado, orientado a temas, no volátil y variante en el tiempo) es la pieza que sostiene el análisis histórico que los sistemas operativos no pueden ofrecer.

En la distribución de insumos industriales este desafío se agudiza. Estas empresas operan con catálogos de miles de referencias, carteras de clientes corporativos de comportamiento heterogéneo y una cadena de suministro sensible a la variación de la demanda. Cada sede genera transacciones, cotizaciones y movimientos de inventario que, sumados, conforman un volumen de datos considerable; sin embargo, ese volumen rara vez se traduce en conocimiento, porque la información

permanece dispersa entre sistemas que no dialogan entre sí.

La razón de esa limitación es arquitectónica y explica por qué la solución no puede ser un reporte más. Los sistemas transaccionales están optimizados para registrar operaciones, no para analizarlas: consultar millones de registros agregados exige un entorno separado del operativo [4], [2]. Ese entorno, un almacén alimentado por procesos ETL confiables [5] y organizado bajo un modelo dimensional, es justamente lo que estas empresas no poseen, y es lo que justifica la propuesta: sobre él, las operaciones OLAP habilitan explorar el negocio desde múltiples perspectivas [6] y las técnicas de minería revelan patrones (como la segmentación de clientes) que la intuición no anticipa [7], [8]. Los conceptos no se invocan, entonces, por sí mismos, sino como las piezas que la organización necesita para pasar del dato a la decisión.

Pese a que estas tecnologías son hoy accesibles, buena parte de las distribuidoras industriales presenta una madurez analítica baja. Operan con reportes estáticos, consolidan información a mano en hojas de cálculo y mantienen los datos dispersos entre sistemas que no se comunican, sin un almacén ni procesos ETL formales. Esa carencia tiene un costo concreto: impide segmentar clientes, evaluar el desempeño de productos por región o anticipar la demanda con evidencia. Existe, además, un vacío en la literatura aplicada. La evidencia sobre diseño dimensional y capacidades analíticas se ha concentrado en el comercio minorista y en la administración pública en la que Salvador y Ramió [9] abordan la gobernanza de datos como capacidad institucional previa a la inteligencia artificial, mientras que la distribución industrial B2B, con su lógica de cotización previa a la orden y su gestión multisede de inventarios, ha recibido mucha menos atención.

Atender esa brecha es relevante tanto para la organización, que ganaría capacidad de decisión, como para el campo, que dispondría de evidencia de diseño en un dominio poco documentado. En consecuencia, este trabajo se propuso diseñar y validar un esquema de integración y análisis de datos para una distribuidora B2B de insumos industriales, orientado a sustentar decisiones comerciales y operativas. La pregunta que guía el estudio es en qué medida un esquema dimensional de inteligencia de negocios habilita análisis accionables (segmentación de clientes, clasificación de inventario, conversión de cotizaciones y pronóstico de demanda) para la gestión de una empresa de este tipo.

II. TRABAJOS RELACIONADOS

La inteligencia de negocios cuenta con fundamentos consolidados. Kimball y Ross [2] e Inmon [3] establecieron los principios del almacén de datos y del modelado dimensional, y Chaudhuri y Dayal [6] formalizaron las operaciones OLAP que permiten explorar la información desde múltiples niveles de agregación. Sobre esa base, la minería de datos aportó técnicas para descubrir estructura en grandes volúmenes: Han, Kamber y Pei [7] y Provost y Fawcett [8] documentan el uso del agrupamiento para segmentar clientes según su comportamiento de compra.

La evidencia aplicada, sin embargo, se ha concentrado en dominios específicos. Buena parte de los casos publicados corresponde al comercio minorista, donde el análisis de la cesta de compra y la recomendación de productos dominan la agenda, o a la administración pública, donde Salvador y Ramió [9] sitúan la gobernanza de datos como una capacidad institucional previa a cualquier iniciativa de inteligencia artificial. Sharda, Delen y Turban [11] subrayan, por su parte, el papel de los tableros interactivos en la democratización del acceso a la información dentro de la organización.

La distribución industrial B2B ha recibido mucha menos atención, pese a tener particularidades que la distinguen del comercio minorista. La venta no se origina en una transacción directa sino en una cotización previa, lo que introduce un embudo comercial cuyo análisis rara vez aparece en la literatura centrada en el retail. A ello se suma la gestión multisede de inventarios, con miles de referencias de rotación heterogénea. Este trabajo se ubica en ese vacío: aplica el modelado dimensional y las técnicas analíticas ya consolidadas a un dominio poco documentado, e incorpora de forma explícita el análisis del pipeline de cotizaciones y la clasificación ABC/XYZ del inventario.

III. METODOLOGÍA

Para este trabajo opté por una investigación aplicada, de enfoque cuantitativo y alcance descriptivo-analítico, bajo un diseño no experimental de corte transversal que estructuré como estudio de caso [10]. El caso es una distribuidora de insumos industriales con cinco sedes en Costa Rica, unas 2.500 referencias de producto y cerca de 3.000 clientes corporativos de los sectores de construcción, manufactura, agroindustria y mantenimiento industrial. Como el proyecto es una propuesta de diseño y no un despliegue en producción, generé los datos de forma sintética con la librería Faker de Python: construí un catálogo industrial propio (veinte categorías, desde herramientas manuales hasta lubricantes e instrumentos de medición) y ajusté volúmenes y márgenes (entre 18 % y 32 % sobre el precio de venta) para que se parecieran a los de una distribuidora real del rubro.

El uso de datos sintéticos merece una justificación, porque condiciona el alcance de lo que aquí se afirma. No fue un recurso improvisado, sino una decisión coherente con un estudio orientado a diseñar y validar un esquema, no a desplegarlo en producción. Trabajar

con datos generados me permitió cubrir de forma controlada las veinte categorías del catálogo, las cinco sedes y una variedad de escenarios de demanda que un extracto parcial de datos reales difícilmente habría abarcado, además de garantizar la reproducibilidad del experimento y evitar el manejo de información comercial sensible. Es una práctica habitual en la validación de arquitecturas de datos cuando el objetivo es demostrar el funcionamiento del diseño antes de exponerlo a un entorno productivo; su contraparte, que reconozco de antemano, es que las cifras obtenidas ilustran el comportamiento del esquema y no el de un mercado real.

Conviene aclarar el origen de los datos, porque el proyecto pasó por una migración deliberada. Mi punto de partida fue una plataforma analítica que yo mismo había construido antes en un dominio de comercio electrónico; la reutilicé como prueba de concepto técnica, pero la reescribí por completo para el dominio industrial. Esa migración no fue trivial: además de renombrar el esquema, tuve que regenerar el catálogo y los datos (un primer intento conservó categorías de retail y hubo que rehacerlo) y corregir cerca de una decena de errores en el ETL, entre ellos generar la dimensión de tiempo a partir de las fechas reales de las órdenes cuando la tabla de calendario llegaba vacía. El almacén final reunió 32.383 líneas de orden, 72.048 de cotización y 450.000 registros de inventario mensual del período 2022–2025. Para cada análisis incluí solo los registros con trazabilidad temporal completa y excluí las órdenes canceladas del cálculo de ventas y margen.

La generación de los datos siguió un procedimiento controlado. Con la librería Faker de Python construí un catálogo de 2.500 referencias distribuidas en veinte categorías industriales (desde herramientas manuales hasta instrumentos de medición y lubricantes) y una cartera de 3.000 clientes corporativos etiquetados por sector y por sucursal. Sobre esa base generé las transacciones del período 2022–2025, asignando a cada línea un margen de entre el 18 % y el 32 % sobre el precio de venta y modelando la relación cotización–orden propia del negocio B2B. El uso de una semilla fija garantiza que el conjunto sea reproducible, condición necesaria para que los resultados puedan verificarse.

El entorno lo levanté sobre SQL Server 2022, con la base transaccional y el almacén analítico en bases separadas, y trabajé el procesamiento en Python: Pandas para manipular los datos, scikit-learn para la minería, statsmodels para las series temporales y SQLAlchemy para la conexión. La interfaz la desarrollé en Streamlit con Plotly e integré un asistente conversacional sobre la API de Anthropic Claude para permitir consultas en lenguaje natural. El trabajo avanzó en cinco momentos: diagnostiqué la gestión de datos en cuatro dimensiones para fijar la línea base; diseñé una arquitectura por capas que separa fuentes, integración ETL, almacén, análisis y visualización; modelé el almacén como una constelación de hechos en esquema estrella, con tres tablas de hechos y seis dimensiones compartidas; definí cinco análisis estratégicos; y, por último, los ejecuté y medí sobre el almacén ya poblado.

Incorporé, además, un asistente conversacional como capa de acceso más externa del sistema. Su funcionamiento es el siguiente: traduce una pregunta escrita en lenguaje natural a una consulta sobre el modelo dimensional, la ejecuta a través del cubo OLAP sobre el almacén poblado por el ETL y devuelve una respuesta redactada, acompañada de una visualización generada de forma automática. La interpretación de la pregunta y la redacción se apoyan en la API de Claude; las cifras, en cambio, provienen siempre de una consulta ejecutada sobre el almacén, no del modelo de lenguaje. Para comprobar esa distinción, formulé un conjunto de preguntas cuyas respuestas podían contrastarse con los resultados ya tabulados.

Cada técnica siguió su propio procedimiento. La segmentación K-means la apliqué sobre los clientes con dos o más órdenes y elegí el número de grupos contrastando tres criterios (codo, silueta y Davies-Bouldin) con las variables previamente estandarizadas. La clasificación ABC ordenó las referencias por su contribución acumulada a las ventas (A hasta el 80 %, B hasta el 95 %, C el resto) y la XYZ usó el coeficiente de variación de la demanda mensual ($X < 0{,}5$; Y entre 0,5 y 1,0; $Z > 1{,}0$). Para el pronóstico ARIMA partí de la prueba de Dickey-Fuller aumentada, dejé que la búsqueda por criterio de Akaike eligiera el orden del modelo y reporté el error sobre seis meses de prueba con intervalos al 95 %. En la regresión (un bosque aleatorio con partición 80/20 y validación cruzada de cinco particiones) tomé una decisión que vale la pena explicitar: en una primera corrida el margen total dominaba el modelo con una importancia del 96 %, lo que me hizo sospechar de inmediato. Al revisarlo confirmé que el margen es una transformación directa de las ventas, es decir, fuga de información; lo excluí junto con el monto y el descuento totales, y aunque el R^2 cayó, el modelo quedó honesto. En cuanto a lo ético, no usé datos personales reales (toda la información es sintética) y, para un eventual despliegue, dejé previstos controles de cifrado, acceso por roles y auditoría de consultas.

Los criterios de validación de cada técnica se fijaron de antemano. Para la segmentación evalué soluciones de dos a ocho grupos: el coeficiente de silueta alcanzó su valor máximo en dos grupos (0,34) y el índice de Davies-Bouldin su mínimo (1,13), de modo que la solución bimodal quedó respaldada por ambos criterios además del método del codo. En el pronóstico, la prueba de Dickey-Fuller aumentada rechazó la raíz unitaria con un valor p inferior a 0,001, y la búsqueda por criterio de información de Akaike seleccionó un

modelo ARIMA(0,1,2); el error se evaluó sobre un conjunto de prueba de seis meses no visto durante el ajuste. Estos umbrales explícitos permiten juzgar la robustez de cada resultado y replicar el procedimiento.

IV. ARQUITECTURA Y DISEÑO DE LA SOLUCIÓN

La solución se organiza en una arquitectura por capas que separa el entorno operativo del analítico. La Fig. 1 muestra las cinco capas: las fuentes de datos transaccionales, la capa de integración a cargo de los procesos ETL, el almacén de datos dimensional, la capa analítica donde se ejecutan las técnicas y la capa de visualización compuesta por los tableros y el asistente conversacional. Esta separación permite ejecutar consultas agregadas sobre grandes volúmenes sin penalizar el rendimiento de la operación diaria [2], [3].

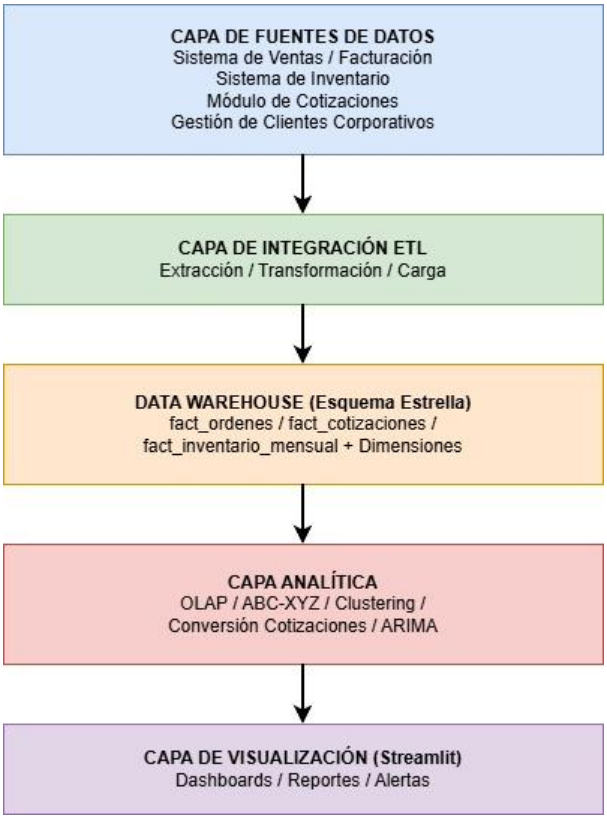

Fig. 1. Arquitectura por capas de la solución

El corazón analítico es un modelo dimensional en constelación de hechos, representado en la Fig. 2. Tres tablas de hechos comparten seis dimensiones. La tabla de órdenes registra cada línea de venta, con métricas de cantidad, precio, descuento, monto, costo y margen; la de cotizaciones captura el estado y la conversión de cada línea cotizada, elemento distintivo del negocio B2B; y la de inventario mensual almacena existencias y rotación por producto, sucursal y mes. Las seis dimensiones (tiempo, cliente, producto, sucursal, proveedor y geografía) proporcionan el contexto común que habilita el análisis cruzado de ventas, pipeline e inventario.

Fig. 2. Modelo dimensional en constelación de hechos

El poblamiento del almacén se realiza mediante un flujo ETL en tres fases. En la extracción se obtienen los datos de las fuentes; en la transformación se aplican reglas de limpieza, estandarización de categorías y validación de integridad referencial; y en la carga, los registros se integran de forma incremental en las estructuras dimensionales. La calidad de esta capa condiciona la confiabilidad de todo el análisis posterior [5], por lo que se registran trazas de cada ejecución que permiten auditar el proceso. Sobre este almacén se ejecutan los cinco análisis y el asistente conversacional descritos en las secciones siguientes.

La capa de visualización incluye, junto a los tableros, un asistente que traduce preguntas en lenguaje natural a consultas sobre el modelo dimensional. El componente interpreta la intención del usuario, construye la consulta correspondiente sobre el cubo OLAP, la ejecuta sobre el almacén y devuelve una respuesta redactada acompañada de una visualización generada de forma automática. La interpretación y la redacción se apoyan en un modelo de lenguaje a través de una interfaz de programación; las cifras, en cambio, provienen siempre de una consulta ejecutada, lo que evita que el modelo genere valores por su cuenta. Este diseño acerca el análisis a los usuarios no técnicos sin exigirles conocimiento de SQL.

V. RESULTADOS

Diagnóstico de la gestión de datos

El diagnóstico de la situación reveló una brecha consistente entre el dato disponible y su aprovechamiento. La información se encontraba fragmentada entre sistemas que no se comunicaban, los reportes se generaban de forma manual en hojas de cálculo, no existía un almacén ni procesos ETL formales, y la cultura organizacional dependía de reportes predefinidos, aunque con apertura al cambio. La Tabla 1 sintetiza los hallazgos por dimensión.

Tabla 1. Diagnóstico de madurez en la gestión de datos del caso

Dimensión	Hallazgo principal	Evidencia
Datos	Fragmentados e inconsistentes	Información dispersa sin integración ni históricos consolidados
Tecnología	Sin arquitectura analítica	Ausencia de Data Warehouse, ETL formal y tableros interactivos
Procesos	Manuales y no estandarizados	Consolidación en hojas de cálculo; decisiones reactivas; sin gobernanza
Cultura	Madurez baja, con apertura	Dependencia de reportes predefinidos; disposición a adoptar nuevas herramientas

Fuente: elaboración propia a partir del diagnóstico organizacional.

El diagnóstico se estructuró en cuatro dimensiones. En la de datos, la información aparecía fragmentada entre sistemas que no se comunicaban, sin históricos consolidados. En la tecnológica, no existía un almacén ni procesos ETL formales que separaran el entorno operativo del analítico. En la de procesos, los reportes se producían de forma manual en hojas de cálculo, lo que introducía demoras y riesgo de error. Y en la cultural, la organización dependía de reportes predefinidos, aunque mostraba apertura a adoptar nuevas herramientas. Esta última condición explica en buena medida la viabilidad de la propuesta: la brecha no es solo técnica sino de capacidades, y su cierre requiere tanto la arquitectura como la disposición a usarla.

Sobre el almacén ya descrito, que integró 32.383 líneas de orden, 72.048 de cotización y 450.000 registros de inventario mensual (2.500 referencias, 3.000 clientes corporativos y cinco sucursales), se ejecutaron los cinco análisis estratégicos cuyos resultados se presentan a continuación.

Análisis OLAP multidimensional

Sobre ese almacén, la exploración OLAP confirmó un negocio geográficamente equilibrado. La Fig. 3 muestra las ventas acumuladas y el margen por provincia del cliente: Cartago encabeza con ₡5.080 millones y San José cierra con ₡4.302 millones, dentro de un rango estrecho y con un margen prácticamente uniforme cercano al 30 % en todas las regiones, indicio de una política de precios centralizada.

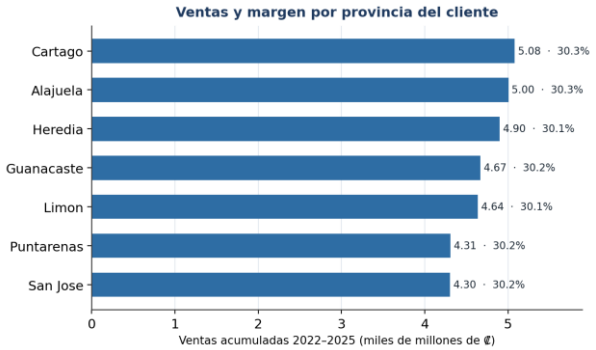

Fig. 3. Ventas y margen por provincia del cliente, 2022–2025

Por categoría de producto, las ventas también se reparten de forma relativamente homogénea entre las veinte líneas industriales (Fig. 4). Instrumentos de medición lidera con ₡1.910 millones, seguido de lubricantes y químicos y de abrasivos; los márgenes oscilan entre el 29,6 % y el 31,0 %. La Tabla 2 resume la evolución anual: las ventas se mantuvieron alrededor de ₡8.500 millones por año entre 2022 y 2024 y descendieron en 2025, con un margen estable en torno al 30,2 %.

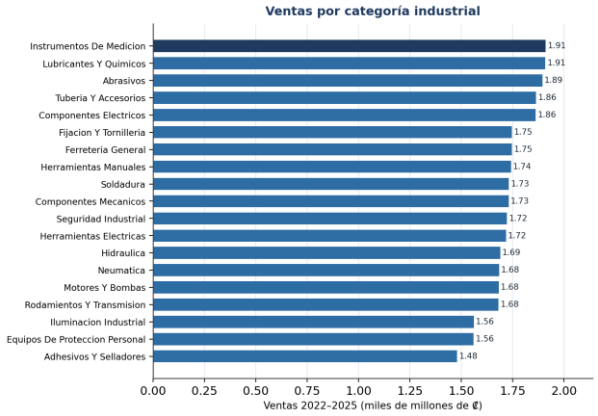

Fig. 4. Ventas por categoría industrial, 2022–2025

Tabla 2. Evolución anual de ventas, órdenes y margen

Año	Ordenes	Ventas (₡)	Margen %
2022	1.932	8.262.793.465	30,1
2023	2.053	8.677.830.349	30,2
2024	2.016	8.708.204.010	30,2
2025	1.699	7.253.158.506	30,2
Total	7.700	32.901.986.330	30,2

Fuente: elaboración propia. Cifras de órdenes no canceladas. El año 2025 cubre hasta noviembre.

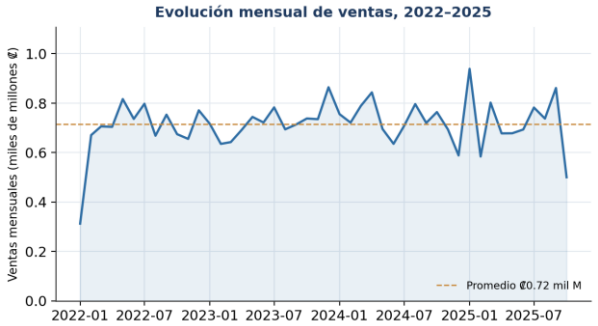

Fig. 5. Evolución mensual de ventas, 2022–2025

El patrón temporal merece una lectura cuidadosa. Las ventas se mantuvieron estables alrededor de ₡8.500 millones anuales entre 2022 y 2024 y descendieron en 2025; sin embargo, ese año solo cubre hasta noviembre, de modo que la caída aparente combina una posible desaceleración real con el efecto de un período incompleto. El margen, en cambio, se sostuvo en torno al 30,2 % en todos los años, lo que sugiere una política de precios estable con independencia del volumen. Separar la tendencia real de la del período incompleto es, precisamente, el tipo de matiz que el análisis multidimensional permite explicitar y que un reporte agregado tendería a ocultar.

Embudo de conversión de cotizaciones

El embudo de cotizaciones cuantifica el pipeline comercial. De las 18.000 cotizaciones emitidas, 8.100

se convirtieron en orden, lo que arroja una tasa de conversión del 45,0 %, con un tiempo medio de respuesta de 16,5 días; el resto se reparte entre cotizaciones rechazadas, vencidas, pendientes y aprobadas sin convertir (Tabla 3). En términos prácticos, más de la mitad de las cotizaciones (y del monto asociado) no llega a facturarse, lo que dimensiona el margen de mejora comercial que ofrece una mejor gestión del pipeline.

Tabla 3. Embudo de conversión de cotizaciones

Estado	Cotizaciones	% del total	Monto cotizado (₡)
Convertida	8.100	45,0	40.446.315.715
Rechazada	3.240	18,0	16.089.483.999
Vencida	2.700	15,0	13.395.486.717
Pendiente	2.160	12,0	10.781.366.453
Aprobada	1.800	10,0	8.962.521.350

Fuente: elaboración propia.

Clasificación ABC/XYZ del inventario

La gestión de inventario se apoyó en dos clasificaciones complementarias. La clasificación ABC, basada en la contribución acumulada a las ventas, asignó 1.335 referencias a la clase A (que concentra el 80 % de las ventas), 595 a la clase B y 570 a la clase C (Fig. 6). La clasificación XYZ, basada en la variabilidad de la demanda mensual, identificó 500 referencias de demanda estable (X), 818 de demanda variable (Y) y 1.182 de demanda irregular (Z), lo que orienta directamente las políticas de reposición y el dimensionamiento del stock de seguridad.

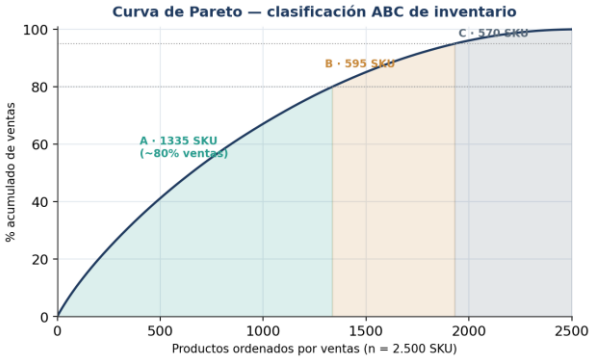

Fig. 6. Curva de Pareto de la clasificación ABC de inventario

Segmentación de clientes

La segmentación de clientes se aplicó sobre los 2.173 clientes con dos o más órdenes. Los tres criterios coincidieron en que dos grupos constituyen la solución más estable (silueta = 0,34; Davies-Bouldin = 1,13). El Segmento 1, de mayor valor, agrupa al 28,9 % de los clientes, con mayor frecuencia de compra, más referencias distintas y una tasa de conversión del 62,6 %; el Segmento 2, mayoritario (71,1 %), reúne clientes menos frecuentes, con mayor recencia y una conversión del 49,6 % (Fig. 7 y Tabla 4). Para la empresa, esto implica que cerca de dos de cada tres clientes activos son cuentas de baja frecuencia con potencial de reactivación, mientras que un núcleo minoritario concentra el valor.

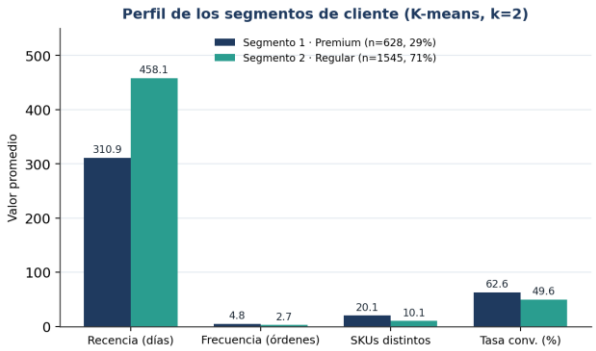

Fig. 7. Perfil comparado de los segmentos de cliente (K-means, k = 2)

Tabla 4. Caracterización de los segmentos de cliente

Métrica	Seg. 1 · Premium	Seg. 2 · Regular
Clientes	628	1.545
% de la base	28,9	71,1
Recencia (días)	311	458
Frecuencia (órdenes)	4,78	2,66
Monto total (₡)	22.596.683	10.471.005
Conversión (%)	62,6	49,6

Fuente: elaboración propia.

Pronóstico de demanda

El pronóstico de demanda se construyó sobre la serie mensual de ventas. La prueba de Dickey-Fuller aumentada rechazó la presencia de raíz unitaria ($p < 0{,}001$) y la búsqueda por criterio de Akaike seleccionó un modelo ARIMA(0,1,2). Sobre el conjunto de prueba de seis meses, el modelo alcanzó un error porcentual absoluto medio del 13,8 % (RMSE = ₡112,7 millones; MAE = ₡87,4 millones). La proyección a doce meses se estabiliza alrededor de ₡689 millones mensuales, con intervalos de confianza que se amplían progresivamente (Fig. 8).

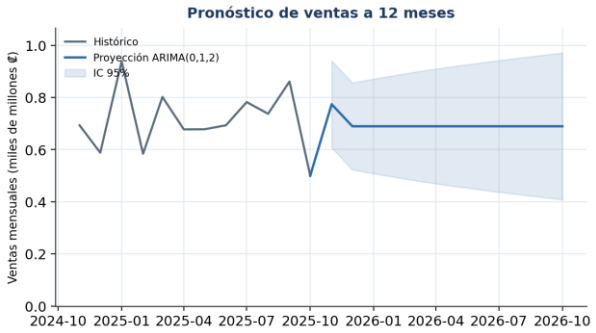

Fig. 8. Pronóstico de ventas a doce meses con intervalos de confianza al 95 %

Modelo de regresión

El modelo de regresión por bosque aleatorio estimó el número de transacciones por grupo de tiempo, categoría, sucursal y vendedor. Tras excluir las variables que producían fuga de información, el modelo explicó el 31,7 % de la varianza en el conjunto de prueba (R^2 de entrenamiento = 0,72; validación cruzada = 0,31 ± 0,01). Las variables de mayor importancia fueron las unidades totales (0,45), el descuento promedio (0,11) y el precio promedio (0,10), seguidas de la categoría y el vendedor (Tabla 5).

Tabla 5. Importancia relativa de variables en el modelo de regresión (sin fuga)

Variable	Importancia	Variable	Importancia
Unidades totales	0,453	Vendedor	0,049
Descuento promedio	0,106	Mes	0,041
Precio promedio	0,101	Día de la semana	0,038
Cantidad promedio	0,090	Sucursal	0,031
Categoría	0,059	Año	0,021

Fuente: elaboración propia. Variable objetivo: número de transacciones. Se excluyeron el margen y el monto totales para evitar fuga de información.

Asistente conversacional y verificación

Por último, el asistente conversacional se sometió a una prueba de correctitud: se le formularon preguntas en lenguaje natural y se contrastaron sus respuestas con los datos ya tabulados. La Fig. 9 muestra una de esas interacciones (la consulta por ventas por provincia), con la respuesta redactada, el gráfico generado de forma automática y el detalle de consumo de la API en el pie. La Tabla 6 resume la verificación: en las tres consultas cruzables con el resto del artículo las cifras coincidieron; una cuarta, sobre órdenes canceladas (un dato no incluido en ninguna tabla), devolvió un total de 400 cancelaciones (tasa del 4,9 %) obtenido en vivo del almacén, y el propio asistente advirtió que el desglose por año era una estimación proporcional y no un valor consultado.

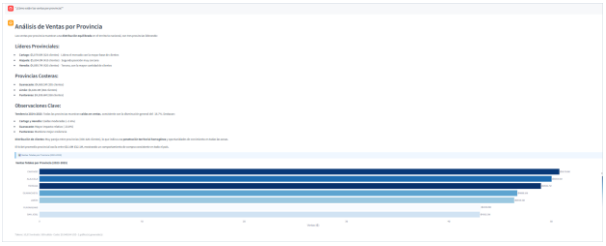

Fig. 9. Interacción con el asistente: consulta en lenguaje natural, respuesta y visualización automática

Tabla 6. Verificación de las respuestas del asistente contra los datos del estudio

Consulta en lenguaje natural	Respuesta del asistente	Contraste	Resultado
¿Cómo están las ventas por provincia?	Cartago lidera con ₡5.079,6 M	Fig. 3	Coincide
Comparación de ventas 2023 vs 2024 vs 2025	₡8.677,8 / 8.708,2 / 7.253,2 M	Tabla 2	Coincide
Análisis de márgenes por categoría	Fijación y Tornillería, 31,0 %	OLAP por categoría	Coincide
¿Cuántas órdenes se cancelaron por año?	400 canceladas; tasa 4,9 %	Dato no tabulado (consulta en vivo)	Ejecución verificada

Fuente: elaboración propia. En la cuarta consulta, el asistente señaló que el desglose anual era una estimación, no un valor consultado.

VI. DISCUSIÓN

Dos ejemplos ilustran la verificación del asistente. Ante la pregunta por las ventas por provincia, respondió que Cartago lidera con ₡5.079,6 millones, cifra que coincide con la Fig. 3. Ante la consulta por las órdenes canceladas, un dato que no figura en ninguna tabla del artículo, devolvió un total de 400 cancelaciones con una tasa del 4,9 % y advirtió de forma explícita que el desglose por año era una estimación proporcional y no un valor consultado. Esa distinción entre lo consultado y lo estimado es la evidencia más clara de que el sistema recupera cifras reales del almacén en lugar de generarlas.

Los resultados respaldan la viabilidad del esquema y, sobre todo, permiten leer qué cambia para la gestión de la distribuidora. El diagnóstico confirmó la brecha que motivó el estudio, y el almacén la cerró al reunir ventas, cotizaciones e inventario bajo un mismo marco: decisiones que antes dependían de reportes dispersos (dónde concentrar inventario, a qué clientes priorizar, cuánta demanda anticipar) pueden ahora apoyarse en una sola fuente consultable. Ese salto, de la hoja de cálculo manual a la exploración multidimensional, es el aporte práctico central; el fundamento técnico de por qué la separación operativo-analítica lo hace posible ya quedó establecido en la introducción y no requiere repetirse.

La segmentación ilustra el valor del análisis exploratorio. Que los tres criterios converjan en dos grupos no estaba prefijado: el algoritmo reveló una estructura bimodal en la base de clientes, distinguiendo un núcleo de cuentas de alto valor y conversión elevada de una mayoría de clientes ocasionales. Este es el tipo de hallazgo que la minería de datos habilita y que la intuición difícilmente anticipa [7], [8], y que justifica estrategias comerciales diferenciadas: programas de reactivación para el segmento regular y gestión de cuenta para el premium. El embudo de cotizaciones aporta una pieza ausente en la literatura más centrada en el comercio minorista: en el B2B la venta nace de una cotización, y medir su conversión (del 45 % en este caso) ofrece una palanca de mejora que el análisis de ventas por sí solo no capta. En conjunto, los hallazgos coinciden con Salvador y Ramió [9] en que la capacidad analítica es una cuestión institucional, no solo técnica, y con Sharda, Delen y Turban [11] en que los tableros interactivos democratizan el acceso a la información.

Conviene explicitar en qué se diferencia esta propuesta de una implementación convencional de inteligencia de negocios. La mayoría de esas implementaciones se detiene en tableros descriptivos de ventas; aquí, en cambio, el esquema integra en un mismo diseño tres capacidades que el negocio industrial B2B necesita y que rara vez aparecen juntas: la lectura del pipeline de cotizaciones, donde efectivamente se gana o se pierde la venta; la doble clasificación ABC/XYZ, que cruza el valor y la variabilidad de la demanda para decidir la reposición; y una capa conversacional que reduce la barrera de acceso al análisis. El valor diferencial no reside en cada técnica por separado (todas son conocidas), sino en articularlas sobre un modelo dimensional pensado para este negocio y en haberlo validado de forma transparente, incluida la detección y corrección de un sesgo que, de no advertirse, habría inflado los resultados. En esa medida, el trabajo extiende la evidencia de diseño disponible hacia un dominio, la

distribución industrial B2B multisede, poco representado frente al comercio minorista.

La comparación con la literatura de comercio minorista resulta ilustrativa. En el retail, el foco analítico recae sobre la cesta de compra y la recomendación de productos, a partir de transacciones que ya ocurrieron. En la distribución industrial B2B, en cambio, una parte sustancial del valor se decide antes de la venta, en la etapa de cotización, y el inventario debe gestionarse sobre miles de referencias con patrones de demanda muy dispares. El esquema propuesto refleja esa diferencia al tratar la cotización como un hecho de primer orden y al cruzar la clasificación por valor (ABC) con la de variabilidad de la demanda (XYZ), dos decisiones de diseño que rara vez aparecen juntas en las soluciones orientadas al minorista.

El esquema es, además, transferible. Aunque se validó sobre una distribuidora de insumos industriales, su estructura (tres hechos que comparten dimensiones, clasificación ABC/XYZ del inventario y análisis del embudo de cotizaciones) responde a características comunes a otros sectores de distribución B2B, como los repuestos automotrices, los materiales de construcción o los suministros médicos. Adaptar la solución a esos dominios exigiría redefinir el catálogo y las reglas de negocio del ETL, pero conservaría el núcleo dimensional y las técnicas analíticas, lo que sugiere que el aporte de diseño trasciende el caso particular.

La prueba de correctitud del asistente refuerza este punto. Al contrastar sus respuestas con la fuente de verdad se confirmó que ejecuta consultas reales sobre el almacén en lugar de producir cifras verosímiles, y que distingue lo consultado de lo estimado (como cuando aclaró que el desglose anual de cancelaciones era una aproximación); esa honestidad del sistema importa tanto como su exactitud, porque un asistente que devolviera cifras inventadas pero plausibles sería más peligroso que útil. Interrogar el almacén en lenguaje natural acerca la decisión al usuario de negocio sin exigirle conocer SQL, en línea con la democratización del acceso que describen Sharda, Delen y Turban [11]. Como contraparte, la capa conversacional introduce un costo operativo variable, pues se factura por tokens consumidos: en las interacciones registradas, cada consulta con su visualización rondó los USD 0,03 a 0,04, un monto bajo por consulta pero que conviene dimensionar al escalar la solución a muchos usuarios concurrentes.

Ese costo merece una estimación más fina de cara al escalamiento. Cada consulta con su visualización consumió entre 6.000 y 11.000 tokens de entrada y alrededor de 500 de salida. A ese ritmo, un escenario de veinte usuarios que realicen diez consultas diarias supondría del orden de USD 6 a 8 por día, es decir, cerca de USD 150 mensuales. La cifra es modesta frente al valor de las decisiones que habilita, pero crece de forma lineal con el uso, por lo que antes de un despliegue masivo conviene incorporar una caché de las respuestas más frecuentes y límites de consumo por usuario que acoten el gasto.

El asistente aporta, por último, a la dimensión menos técnica del problema. El diagnóstico había señalado una madurez analítica baja, con usuarios de negocio dependientes de reportes predefinidos. Una interfaz que responde en lenguaje natural y devuelve cifras verificadas reduce esa dependencia y, con ella, el cuello de botella que supone que cada consulta deba pasar por un perfil técnico. El valor de la solución no reside únicamente en la sofisticación de sus modelos, sino en acercar el dato a quien toma la decisión.

Algunos resultados merecen una lectura cautelosa, y prefiero señalarlos yo antes que esconderlos. Al revisar la clasificación ABC me llamó la atención lo plana que resultó la curva: la clase A necesitó 1.335 referencias (más de la mitad del catálogo) para reunir el 80 % de las ventas, lejos del principio de Pareto clásico, donde una fracción pequeña de productos explica la mayoría de los ingresos. La causa es que generé la demanda sintética con una dispersión bastante uniforme entre referencias; con datos reales esperaría una concentración mayor. Algo parecido ocurre con la regresión, que explicó apenas el 32 % de la varianza una vez corregida la fuga de información. Antes de corregirla, el margen total (una transformación directa de las ventas) alcanzaba una importancia del 96 % e inflaba el ajuste; al excluirlo, el R^2 cayó, y esa caída es justamente la prueba de que el primer modelo estaba contaminado. Predecir el número de transacciones a partir de variables operativas es difícil de por sí, así que me quedo con un ajuste modesto pero honesto antes que con uno alto pero espurio.

El estudio tiene límites que conviene reconocer. Se trata de una propuesta de diseño validada sobre datos sintéticos y un único caso, por lo que sus cifras ilustran el funcionamiento del esquema, no el comportamiento de un mercado real; la generalización requiere prudencia. El modelo ARIMA, aunque alcanzó un error razonable, asume que la dinámica histórica se mantiene y no incorpora variables externas como la estacionalidad sectorial o los ciclos de la construcción. La clasificación XYZ, por su parte, depende de la calidad de la serie de inventario. Estas limitaciones trazan las líneas futuras: instanciar el esquema sobre datos reales de la empresa en una fase de prototipo, enriquecer el pronóstico con variables económicas y de demanda sectorial, incorporar técnicas de inferencia causal para evaluar el efecto de campañas comerciales y consolidar una función de gobernanza de datos que sostenga la iniciativa en el tiempo, en la línea que sugieren Salvador y Ramió [9].

Estas lecturas tienen implicaciones prácticas directas. La clasificación ABC/XYZ orienta la política de inventario: las referencias de clase A y demanda estable justifican reposición automática, mientras que

las de clase C y demanda irregular obligan a revisar su nivel de stock de seguridad para no inmovilizar capital. La segmentación sugiere una estrategia comercial diferenciada, con gestión de cuenta para el segmento premium y campañas de reactivación para el regular. Y el pronóstico, aun con su margen de error, ofrece a la planeación una banda razonable para dimensionar compras y capacidad. En conjunto, el esquema no se limita a describir el negocio: habilita decisiones concretas sobre inventario, clientes y demanda.

VII. AMENAZAS A LA VALIDEZ

Como todo estudio de caso, este trabajo está sujeto a amenazas a la validez que conviene declarar. En cuanto a la validez interna, el riesgo principal fue la fuga de información en el modelo de regresión: el margen total, una transformación directa de las ventas, inflaba el ajuste hasta un 96 % de importancia. Su detección y corrección, con la consiguiente caída del R^2, es la evidencia de que el control se aplicó; aun así, no puede descartarse que subsistan correlaciones espurias en otras variables.

La validez externa es la limitación de mayor peso. Los datos son sintéticos y corresponden a un único caso, de modo que las cifras ilustran el comportamiento del esquema y no el de un mercado real; la generalización a otras distribuidoras exige instanciar el diseño sobre datos reales. La generación sintética introdujo, además, un sesgo visible: la clasificación ABC resultó poco concentrada, con una clase A que abarca más de la mitad del catálogo, lejos del principio de Pareto que cabría esperar en una operación real.

Respecto de la validez de constructo, el modelo ARIMA asume que la dinámica histórica se mantiene y no incorpora variables externas como la estacionalidad sectorial o los ciclos de la construcción, por lo que su horizonte útil es acotado. En cuanto a la confiabilidad, el uso de una semilla fija en la generación de datos y la ejecución reproducible de los análisis permiten repetir el experimento y obtener los mismos resultados, lo que facilita su verificación por terceros.

VIII. CONCLUSIONES

El estudio diseñó y validó un esquema de integración y análisis de datos para una distribuidora B2B de insumos industriales, y demostró que es técnicamente viable y que responde a la brecha diagnosticada entre la generación de datos y su aprovechamiento estratégico.

En términos de los objetivos planteados, el diagnóstico caracterizó las fuentes de datos y confirmó la brecha de aprovechamiento; la arquitectura por capas centralizó la información al separar el entorno operativo del analítico; el modelo dimensional en constelación habilitó la exploración multidimensional del negocio; el conjunto de cinco análisis, sumado al asistente conversacional, materializó el esquema de análisis con apoyo de inteligencia artificial; y la ejecución sobre un prototipo funcional, con validación de las respuestas contra los datos, evidenció la viabilidad técnica de la propuesta. Cada objetivo específico encontró, así, una respuesta concreta en los resultados.

Los hallazgos confirman que la separación entre los entornos transaccional y analítico, junto con un modelo dimensional en esquema estrella, habilita una exploración multidimensional accesible y un conjunto de análisis accionables. Sobre el almacén poblado, la segmentación distinguió dos arquetipos de cliente, el embudo cuantificó una conversión de cotizaciones del 45 %, las clasificaciones ABC y XYZ ordenaron el catálogo para la gestión de inventario, y el pronóstico de demanda alcanzó un error razonable para la planeación operativa.

La contribución del trabajo es doble: aporta evidencia de diseño en un dominio (la distribución industrial B2B multisede) poco documentado frente al comercio minorista, e ilustra que el valor de una solución de inteligencia de negocios depende tanto de la accesibilidad de su interfaz como de la sofisticación de sus técnicas. Como toda propuesta de diseño validada sobre datos sintéticos, su alcance es ilustrativo. El paso siguiente es instanciar el esquema sobre datos reales en una fase de prototipo funcional, y a partir de ahí se abren líneas concretas de continuidad: enriquecer el pronóstico con variables externas de demanda sectorial, incorporar inferencia causal para medir el efecto de las decisiones comerciales y consolidar una función de gobernanza de datos que dé sostenibilidad a la iniciativa.

Más allá del caso, el trabajo deja una lección de método: en un entorno de madurez analítica baja, el mayor obstáculo no suele ser la técnica sino la distancia entre el dato y quien decide. Una arquitectura sólida es condición necesaria, pero el valor se realiza cuando la información se vuelve accesible y confiable para el usuario de negocio. Esa convicción orienta las líneas futuras y, en particular, el paso hacia un prototipo sobre datos reales que permita contrastar el diseño con la operación de la empresa.

IX. REFERENCIAS

[1] Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188.

[2] Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3.ª ed.). Wiley.

[3] Inmon, W. H. (2005). *Building the data warehouse* (4.ª ed.). Wiley.

[4] Silberschatz, A., Korth, H. F., & Sudarshan, S. (2020). *Database system concepts* (7.ª ed.). McGraw-Hill.

[5] Golfarelli, M., & Rizzi, S. (2009). *Data warehouse design: Modern principles and methodologies*. McGraw-Hill.

[6] Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record, 26*(1), 65–74.

[7] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3.ª ed.). Morgan Kaufmann.

[8] Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

[9] Salvador, M., & Ramió, C. (2020). Capacidades analíticas y gobernanza de datos en la administración pública como paso previo a la introducción de la inteligencia artificial. *Revista del CLAD Reforma y Democracia*, (77), 5–36.

[10] Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2018). *Metodología de la investigación* (6.ª ed.). McGraw-Hill.

[11] Sharda, R., Delen, D., & Turban, E. (2020). *Business intelligence, analytics, and data science: A managerial perspective* (4.ª ed.). Pearson.