Diseño y evaluación de un pipeline ETL low-code en KNIME para mejorar la eficiencia del proceso de preparación y la calidad de datos en escenarios empresariales tipo PYME

Ivan Ramiro Quiroga Castañeda

Asesor

Mireya García García

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2026

Resumen

En muchas pequeñas y medianas empresas, la información requerida para generar reportes y apoyar procesos de análisis no se encuentra consolidada en una única fuente ni bajo criterios homogéneos de estructura y calidad. Con frecuencia, los datos de ventas, inventario, productos, clientes y abastecimiento se administran en archivos planos, hojas de cálculo o exportaciones parciales de sistemas transaccionales, lo que obliga a ejecutar tareas manuales repetitivas de integración, limpieza, estandarización y validación. Esta situación incrementa los tiempos de preparación, dificulta la trazabilidad del proceso y eleva el riesgo de errores que afectan la calidad del conjunto de datos utilizado para análisis.

En respuesta a esta problemática, el presente proyecto diseñó, implementó y evaluó un pipeline ETL con enfoque low-code utilizando KNIME, orientado a automatizar la preparación de datos comerciales en un escenario empresarial tipo PYME. El trabajo se desarrolló sobre un entorno de datos estructurado con tablas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Sobre estas estructuras se indujeron inconsistencias frecuentes en procesos manuales, tales como valores faltantes, duplicados, formatos inválidos, errores de codificación y llaves inconsistentes, con el fin de simular condiciones realistas de trabajo.

Metodológicamente, el estudio comprendió la caracterización del escenario de datos, la definición del modelo objetivo, el diseño del diccionario de datos y de las reglas de calidad, la construcción del flujo ETL en KNIME y la evaluación comparativa entre un procedimiento manual de preparación y el flujo automatizado. La comparación se apoyó en métricas de eficiencia, particularmente el tiempo de preparación, y en métricas de calidad de datos, como completitud, unicidad, consistencia y validez.

Como resultado, se obtuvo un flujo ETL replicable, documentado y aplicable a contextos similares, capaz de reducir el esfuerzo operativo asociado a la preparación manual de datos y de mejorar la calidad del conjunto de datos resultante para fines analíticos. Entre los productos obtenidos se encuentran el workflow en KNIME, los datasets de entrada y salida, el reporte de métricas y las evidencias técnicas de ejecución del prototipo.

Palabras clave**:** ETL, KNIME, automatización, calidad de datos, PYME.

Abstract

In many small and medium-sized enterprises, the information required for reporting and analytical processes is not consolidated into a single source nor managed under homogeneous quality and structure standards. Sales, inventory, products, customers, and supply data are often distributed across flat files, spreadsheets, or partial exports from transactional systems, which leads to repetitive manual tasks related to integration, cleansing, standardization, and validation. This situation increases preparation time, reduces traceability, and raises the risk of errors that affect the quality of the dataset used for analysis.

In response to this problem, this project designed, implemented, and evaluated a low-code ETL pipeline using KNIME to automate the preparation of commercial data in an SME-type business scenario. The work was developed on a structured data environment composed of customer, product, supplier, sales, purchases, inventory, and inventory movement tables. Controlled inconsistencies frequently found in manual preparation processes, such as missing values, duplicates, invalid formats, coding errors, and inconsistent keys, were induced in these structures in order to simulate realistic working conditions.

Methodologically, the study included the characterization of the data scenario, the definition of the target data model, the design of the data dictionary and quality rules, the construction of the ETL workflow in KNIME, and the comparative evaluation between a manual data preparation procedure and the automated workflow. The comparison was based on efficiency metrics, such as preparation time, as well as data quality metrics, particularly completeness, uniqueness, consistency, and validity.

As a result, a replicable and documented ETL workflow applicable to similar contexts was obtained, capable of reducing the operational effort associated with manual data preparation

and improving the quality of the resulting dataset for analytical purposes. The outputs obtained include the KNIME workflow, the input and output datasets, the metrics report, and technical evidence of the prototype execution.

Keywords**:** ETL, KNIME, automation, data quality, SME.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

En la operación cotidiana de muchas pequeñas y medianas empresas, los datos necesarios para apoyar el análisis del negocio no suelen encontrarse organizados en un único repositorio ni bajo criterios homogéneos de estructura y calidad. Es frecuente que la información de ventas, inventario, productos, clientes y abastecimiento se distribuya en archivos de Excel, reportes en formato CSV y, en algunos casos, tablas almacenadas en bases de datos. Cuando surge la necesidad de generar reportes o realizar análisis, esta dispersión obliga a invertir tiempo en tareas repetitivas de consolidación, corrección de formatos, depuración de duplicados y verificación de reglas básicas del negocio.

Aunque estas actividades pueden parecer operativas, su impacto sobre la analítica es directo. Un proceso manual de preparación de datos no solo consume tiempo que podría destinarse a la interpretación de resultados, sino que también incrementa la probabilidad de inconsistencias y errores difíciles de rastrear. En estas condiciones, la toma de decisiones puede apoyarse en información incompleta, desactualizada o poco confiable, lo que reduce la oportunidad y el valor del análisis. Por esta razón, la automatización de los procesos de extracción, transformación y carga de datos constituye un paso importante para mejorar la trazabilidad, la consistencia y la calidad de la información.

En este contexto, el presente proyecto diseñó, implementó y evaluó una solución ETL low-code en KNIME, orientada a integrar, depurar y validar datos comerciales en un contexto empresarial modelado. El propósito no fue construir una solución empresarial de gran complejidad, sino desarrollar un prototipo funcional, documentado y replicable, capaz de transformar fuentes heterogéneas en un conjunto de datos consolidado y apto para análisis, mediante reglas explícitas de estandarización, validación y control de calidad.

Para el desarrollo del proyecto se trabajó con un entorno de datos estructurado a partir de tablas representativas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Sobre estas estructuras se indujeron inconsistencias frecuentes en procesos manuales, tales como valores faltantes, registros duplicados, formatos no uniformes, errores de codificación y llaves inconsistentes, con el fin de simular condiciones habituales de trabajo y evaluar el comportamiento del flujo en un ambiente controlado.

Además de la construcción del flujo automatizado, el proyecto incorporó un componente de evaluación comparativa. Se definió una referencia del procedimiento manual de preparación de datos y se contrastó con la ejecución automatizada en KNIME, utilizando métricas de eficiencia, particularmente el tiempo de preparación, y métricas de calidad, como completitud, unicidad, consistencia y validez, antes y después del tratamiento. De esta manera, el trabajo busca sustentar con evidencia el aporte del prototipo y ofrecer un proceso documentado que pueda servir como referencia para contextos similares en los que la preparación de datos continúa siendo un cuello de botella para la analítica.

En este sentido, el proyecto se orientó a responder la pregunta de investigación sobre la capacidad de una solución ETL low-code en KNIME para mejorar la eficiencia del proceso de preparación y la calidad de los datos comerciales en un escenario tipo PYME. Para ello, se definieron objetivos enfocados en la caracterización del entorno de datos, la construcción del flujo ETL, la evaluación comparativa frente al procedimiento manual y la documentación técnica de los resultados obtenidos.

El aporte práctico de la solución desarrollada consiste en ofrecer una alternativa funcional para que una PYME pueda transformar datos dispersos y con problemas de calidad en información estructurada, confiable y lista para análisis. Al automatizar actividades como la

integración, limpieza, validación y generación de salidas curadas, el flujo permite reducir el esfuerzo operativo previo al análisis y facilita que la organización disponga de datos más consistentes para construir reportes, indicadores, tableros de control o insumos para modelos analíticos. De esta manera, el proyecto no solo aborda una necesidad técnica de preparación de datos, sino que también contribuye a fortalecer procesos de toma de decisiones basados en información verificable.

Descripción del Problema

En muchas pequeñas y medianas empresas, la información necesaria para el análisis del negocio no se encuentra centralizada ni gestionada bajo criterios homogéneos de estructura, calidad y actualización. Con frecuencia, los datos relacionados con ventas, inventario, productos, clientes y abastecimiento se distribuyen en diferentes archivos de Excel, reportes en formato CSV y consultas puntuales a bases de datos, lo que obliga a realizar procesos manuales de consolidación y depuración antes de cualquier uso analítico.

El problema no se limita a la existencia de múltiples fuentes, sino a la ausencia de un proceso estructurado para integrarlas y transformarlas de manera consistente. Cuando los datos deben prepararse manualmente, es habitual encontrar registros duplicados, valores faltantes, formatos no uniformes, identificadores mal estandarizados e inconsistencias entre tablas o entre campos relacionados. En consecuencia, el esfuerzo operativo no se concentra en analizar la información, sino en corregirla y reorganizarla para que pueda ser utilizada. Esta condición afecta directamente la trazabilidad del proceso, dificulta la repetibilidad de las tareas y reduce la confiabilidad del conjunto de datos final.

Desde una perspectiva analítica, esta problemática repercute en la oportunidad y calidad de los resultados que la organización puede obtener. Si esta labor depende de acciones manuales repetidas, los tiempos de entrega se incrementan, los reprocesos son más frecuentes y se vuelve más difícil garantizar que los datos finales cumplan criterios mínimos de completitud, unicidad, consistencia y validez. En este contexto, esta etapa deja de ser una actividad de apoyo y se convierte en un factor crítico que condiciona la utilidad de los procesos analíticos.

En este trabajo, la problemática se aborda en un contexto PYME, utilizando un entorno de datos estructurado con lógica comercial y operacional, compuesto por tablas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Sobre este entorno se indujeron inconsistencias frecuentes en procesos manuales, tales como duplicados, valores nulos, errores de codificación, formatos inválidos y llaves inconsistentes, con el fin de simular condiciones realistas de trabajo y evaluar de manera controlada el comportamiento del flujo automatizado.

Por tanto, la descripción del problema no se centra únicamente en la existencia de información distribuida en diferentes fuentes, sino en la falta de un mecanismo automatizado, documentado y reproducible que permita integrar, limpiar, validar y preparar la información de manera eficiente. En ese vacío operativo y metodológico se sustenta el desarrollo del presente proyecto.

Planteamiento del Problema

A partir de la situación descrita, el problema de investigación se concreta en determinar si una solución ETL con enfoque low-code en KNIME puede mejorar de manera verificable la preparación de datos comerciales en un contexto PYME, frente a un procedimiento manual basado en tareas repetitivas de consolidación, limpieza y validación.

En términos concretos, el problema radica en que muchas actividades de alistamiento de información siguen ejecutándose manualmente, incluso cuando involucran grandes volúmenes de registros, múltiples tablas relacionadas y reglas de validación que deberían aplicarse de forma sistemática. Bajo estas condiciones, esta labor se convierte en un cuello de botella para la analítica, no solo por el tiempo requerido, sino porque la intervención manual incrementa la probabilidad de inconsistencias difíciles de detectar y corregir de manera uniforme.

En consecuencia, el proyecto se orienta a responder la siguiente pregunta central:

¿En qué medida el diseño e implementación de un pipeline ETL low-code en KNIME permite mejorar la eficiencia del proceso de preparación y la calidad de los datos comerciales en un escenario empresarial tipo PYME, en comparación con un procedimiento manual de integración, limpieza y validación?

Esta formulación se ajusta al alcance del trabajo, ya que no pretende resolver de manera integral todos los problemas de gestión de datos de una organización, sino evaluar, dentro de un entorno controlado, el aporte de una solución ETL low-code sobre un conjunto de datos estructurado con errores de calidad inducidos.

De esta manera, el planteamiento del problema delimita claramente el objeto de estudio, centrado en la eficiencia del alistamiento de la información y en la calidad del conjunto de datos resultante. También delimita el contexto, correspondiente a un escenario tipo PYME con

información comercial y operacional distribuida en diferentes estructuras. Finalmente, delimita la estrategia de análisis, basada en la comparación entre una línea base manual y un flujo automatizado en KNIME, utilizando métricas observables y verificables.

Sistematización del Problema

Con el fin de desarrollar de manera articulada el problema de investigación, los objetivos y la metodología del proyecto, la sistematización del problema permite desagregar la pregunta central en interrogantes específicos que orientan tanto el diseño técnico del pipeline ETL como la evaluación de sus resultados.

La pregunta general del proyecto es la siguiente:

¿En qué medida el diseño e implementación de un pipeline ETL low-code en KNIME permite mejorar la eficiencia del proceso de preparación y la calidad de los datos comerciales en un escenario empresarial tipo PYME, en comparación con un procedimiento manual de integración, limpieza y validación?

A partir de esta pregunta central, se derivan las siguientes preguntas específicas:

¿Qué características debe tener el entorno de datos de un escenario empresarial tipo PYME para representar adecuadamente procesos de ventas, inventario, catálogos, clientes, compras y movimientos operativos, y qué inconsistencias inducidas deben incorporarse para simular condiciones comunes de preparación manual?

¿Qué modelo de datos objetivo, diccionario de datos y reglas de calidad y consistencia deben definirse para orientar el proceso ETL y garantizar que el conjunto de datos final cumpla criterios mínimos de completitud, unicidad, consistencia y validez?

¿Cómo debe estructurarse en KNIME un flujo ETL parametrizable que permita integrar las fuentes definidas, perfilar los datos, estandarizar formatos, depurar registros, homologar valores, validar reglas de negocio y generar un dataset curado?

¿Qué diferencias se observan entre el procedimiento manual de referencia y el flujo automatizado en KNIME en términos de tiempo de preparación y niveles de calidad del conjunto de datos antes y después del tratamiento?

¿Qué evidencias técnicas y documentales deben consolidarse para demostrar la reproducibilidad del prototipo y su posible reutilización en contextos empresariales similares?

Estas preguntas específicas guardan correspondencia directa con los objetivos del proyecto y permiten delimitar con claridad las etapas de construcción del entorno de datos, definición de reglas de calidad, implementación del flujo ETL, evaluación comparativa y documentación de resultados. De esta manera, la sistematización del problema no solo organiza el desarrollo metodológico del trabajo, sino que también establece una relación clara entre el problema planteado, las acciones ejecutadas y los resultados esperados del proyecto.

Justificación

La elección de esta propuesta surge de una necesidad frecuente en muchas organizaciones y, de manera particular, en pequeñas y medianas empresas: la preparación de datos continúa realizándose de forma manual y, en numerosos casos, las hojas de cálculo siguen siendo la herramienta principal para consolidar, transformar y reportar información operativa. Aunque este tipo de trabajo permite resolver necesidades inmediatas, suele generar consecuencias recurrentes, entre ellas errores por manipulación manual, pérdida de tiempo en tareas repetitivas y dificultades para disponer de información confiable en el momento oportuno. En la práctica, el problema no es únicamente técnico, sino también organizacional, porque limita la capacidad de responder con agilidad en entornos cambiantes y reduce el aprovechamiento de la analítica como apoyo real para la toma de decisiones.

Si bien actualmente existen herramientas de automatización y plataformas orientadas a la analítica, en el contexto académico y aplicado todavía son necesarios más ejercicios que demuestren, de manera clara y medible, cómo una solución low-code como KNIME puede implementarse para mejorar procesos de integración, limpieza y control de calidad de datos sin depender de desarrollos extensos de programación. Esto representa una oportunidad para plantear un modelo replicable que pueda comprenderse, documentarse y adaptarse en escenarios donde el tiempo, los recursos tecnológicos y el personal especializado son limitados.

Desde una perspectiva operativa, la automatización del alistamiento de información impacta directamente en tres aspectos relevantes para una organización: reducción de errores, disminución de tiempos y mayor confiabilidad de los reportes. Al reemplazar tareas manuales repetitivas por reglas configuradas dentro de un flujo ETL, se reduce la probabilidad de errores asociados a la manipulación directa de archivos, copias incorrectas, omisión de validaciones o

aplicación desigual de criterios de depuración. Además, el proceso automatizado permite ejecutar en menor tiempo actividades que manualmente requieren revisiones sucesivas, cruces entre tablas y controles de consistencia. Como resultado, los reportes empresariales pueden construirse sobre datos más estables, trazables y validados, lo que incrementa la confianza en la información utilizada para el análisis y la toma de decisiones.

Con base en lo anterior, este trabajo se justifica porque busca construir un prototipo ETL en KNIME orientado a un contexto PYME, centrado en procesos de ventas, inventario, catálogos, clientes, compras y movimientos de inventario. Para aproximar condiciones habituales del entorno, se trabajó con una base de datos estructurada para el proyecto y sobre ella se incorporaron fallas controladas, como duplicados, valores nulos, formatos no uniformes, errores de codificación y llaves inconsistentes. Esto permitió documentar el comportamiento del flujo antes y después del tratamiento de los datos en condiciones controladas y metodológicamente coherentes con los objetivos del estudio.

El impacto del prototipo no se planteó como una afirmación general, sino como un resultado verificable mediante métricas de eficiencia y de calidad. En consecuencia, se comparó el procedimiento manual de preparación con la ejecución del flujo automatizado, con el fin de identificar diferencias en el tiempo de preparación y en las condiciones de calidad del conjunto de datos resultante. En el desarrollo del estudio, esta comparación se concretó en un tiempo registrado de 6 horas para el procedimiento manual equivalente frente a 1 minuto de ejecución del flujo en KNIME, lo que permitió sustentar el valor del flujo ETL desde una perspectiva aplicada y no únicamente descriptiva.

Finalmente, la propuesta se encuentra delimitada a un alcance concreto y medible: diseñar el modelo de datos y las reglas de validación, implementar el flujo ETL en KNIME,

incorporar y tratar problemas de calidad en los datos, y evaluar su desempeño con evidencia. La experiencia y los resultados obtenidos constituyen una base para futuras mejoras y, al mismo tiempo, aportan una referencia útil sobre el uso de enfoques low-code para fortalecer procesos de preparación y calidad de datos en contextos empresariales similares.

Objetivos

Objetivo General

Diseñar, implementar y evaluar un pipeline ETL low-code en KNIME para la integración, limpieza y validación de datos comerciales en un escenario empresarial tipo PYME, utilizando un entorno de datos estructurado con inconsistencias inducidas, con el fin de mejorar la eficiencia del proceso de preparación y la calidad del conjunto de datos resultante mediante métricas verificables.

Objetivos Específicos

Caracterizar un escenario empresarial tipo PYME y estructurar el entorno de datos de trabajo a partir de tablas representativas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario, definiendo además un conjunto de inconsistencias inducidas para simular condiciones comunes de preparación manual.

Definir el modelo de datos objetivo, el diccionario de datos y las reglas de calidad y consistencia que orientan el proceso ETL, incluyendo campos críticos, tipos de datos, rangos válidos, integridad de llaves y reglas de negocio aplicables al escenario seleccionado.

Construir en KNIME un flujo ETL parametrizable que integre las fuentes definidas e implemente las etapas de perfilamiento, estandarización, depuración, homologación, validación de reglas y generación del dataset curado.

Evaluar comparativamente el procedimiento de preparación manual de datos frente al flujo automatizado en KNIME, utilizando como criterios de análisis el tiempo de preparación y las métricas de calidad de datos, particularmente completitud, unicidad, consistencia y validez, antes y después del tratamiento.

Documentar los resultados del prototipo mediante evidencias reproducibles, incluyendo el workflow en KNIME, los datasets de entrada y salida, el reporte de métricas y las capturas o registros de ejecución, con el propósito de facilitar su comprensión, validación y posible reutilización en contextos similares.

Marco Conceptual

Para el desarrollo del presente proyecto es necesario precisar los conceptos que orientan tanto la comprensión del problema como la construcción de la solución propuesta. En primer lugar, el proceso ETL, correspondiente a las etapas de extracción, transformación y carga de datos, constituye el núcleo operativo del trabajo. La extracción se refiere a la obtención de datos desde diversas fuentes; la transformación comprende las actividades de limpieza, estandarización, validación, integración y depuración; y la carga corresponde a la generación del conjunto de datos final en una estructura lista para análisis. En escenarios donde la información se encuentra distribuida en múltiples archivos y con diferentes niveles de calidad, el ETL permite convertir datos dispersos y heterogéneos en un activo informacional coherente y utilizable.

En esta misma línea, un pipeline de datos puede entenderse como una secuencia organizada y reproducible de etapas mediante las cuales los datos de entrada son procesados hasta convertirse en salidas con valor analítico. Su importancia radica en que permite definir un flujo claro de operaciones, mantener trazabilidad sobre las transformaciones aplicadas y asegurar que el tratamiento de los datos no dependa de acciones manuales aisladas, sino de reglas explícitas y repetibles. En el contexto del proyecto, el pipeline ETL diseñado en KNIME integra tablas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario, aplicando validaciones estructurales y de negocio para generar datasets curados.

Otro concepto central es el de automatización de datos, entendido como la ejecución sistemática de procesos de preparación mediante flujos configurados previamente, con mínima intervención manual. La automatización no implica únicamente rapidez operativa, sino también uniformidad en la aplicación de reglas, reducción de errores humanos, disminución de reprocesos y mayor capacidad de repetir el proceso bajo las mismas condiciones. En este trabajo, la

automatización se materializa a través de un workflow en KNIME que reemplaza tareas manuales como la identificación de duplicados, la validación de llaves, la homologación de categorías y la exclusión de registros inválidos.

El proyecto se fundamenta también en el enfoque low-code, entendido como un paradigma de desarrollo que reduce la necesidad de programación extensa mediante interfaces visuales, componentes reutilizables y configuraciones declarativas. Este enfoque resulta particularmente pertinente en contextos donde se requiere construir soluciones funcionales en tiempos razonables, con trazabilidad técnica y sin depender completamente de desarrollo tradicional basado en código. A partir de este enfoque, KNIME se adopta como la herramienta principal del proyecto, dado que permite construir pipelines de datos mediante nodos visuales para lectura, perfilamiento, transformación, validación, integración y exportación de resultados. Su uso favorece la comprensión del flujo, la reutilización del proceso y la documentación técnica del prototipo.

La calidad de datos constituye otro concepto transversal del proyecto. Se entiende como el grado en que un conjunto de datos cumple las condiciones necesarias para ser utilizado de manera confiable en procesos analíticos. En este trabajo, la calidad se evalúa a partir de cuatro dimensiones principales. La completitud hace referencia a la presencia de valores en campos críticos; la unicidad alude a la ausencia de duplicados no deseados; la consistencia se relaciona con la coherencia interna entre campos, tablas y reglas definidas; y la validez corresponde al cumplimiento de formatos, dominios, rangos y restricciones del negocio. Estas dimensiones permiten evaluar objetivamente el estado de los datos antes y después del tratamiento realizado por la ETL.

De manera complementaria, la integridad referencial representa la coherencia existente entre tablas relacionadas a través de identificadores comunes. En el proyecto, este concepto es fundamental porque varias de las inconsistencias inducidas se concentran en relaciones incompletas o erróneas entre ventas y clientes, ventas y productos, compras y proveedores, compras y productos, inventario y productos, así como movimientos de inventario y productos. La validación de estas relaciones permite detectar registros huérfanos y asegurar que el dataset final conserve coherencia estructural.

También es necesario considerar el concepto de homologación de datos, entendido como el proceso mediante el cual valores semánticamente equivalentes pero escritos de forma distinta son llevados a una representación estándar. En el flujo desarrollado, esta homologación se aplicó a campos como municipios, tipos de proveedor y temperatura de bodega, con el fin de evitar que errores ortográficos, abreviaturas o variantes de escritura afectaran la interpretación y el análisis posterior; la homologación cumple así una función clave dentro de la calidad semántica del conjunto de datos.

Finalmente, el concepto de inteligencia de negocios se incorpora como marco de uso del resultado final, la inteligencia de negocios comprende el conjunto de procesos y herramientas orientados a transformar datos en información útil para apoyar la toma de decisiones. Sin embargo, su efectividad depende de que los datos estén previamente integrados, depurados y validados. En consecuencia, el pipeline ETL desarrollado en este proyecto se concibe como una etapa habilitadora para procesos posteriores de análisis, visualización o explotación de información, al proveer un conjunto de datos estructurado, confiable y reutilizable.

Estado del Arte

La literatura reciente coincide en que las pequeñas y medianas empresas enfrentan barreras particulares para aprovechar analítica y automatización de datos. Hassani et al. (2020) señalan que la transformación digital y el aprovechamiento analítico dependen de capacidades organizacionales y tecnológicas que no siempre están disponibles en este tipo de contextos. En una línea similar, Mikalef et al. (2018) muestran que el valor de la analítica sobre el desempeño organizacional depende de capacidades internas para transformar datos en decisiones útiles.

En ese contexto, los procesos ETL mantienen un papel central como mecanismo para integrar, transformar y dejar disponibles los datos en condiciones aptas para análisis. Más allá de la función técnica de mover datos entre fuentes y destinos, la literatura resalta que el ETL aporta trazabilidad, estandarización y control sobre la preparación, aspectos especialmente valiosos cuando la información proviene de archivos heterogéneos y presenta inconsistencias estructurales o semánticas. De forma complementaria, la discusión contemporánea sobre ETL y calidad de datos muestra que la mejora analítica no depende solo de almacenar información, sino de garantizar reglas explícitas de validación, consistencia y depuración durante el proceso de transformación.

En este contexto, el enfoque low-code ha ganado relevancia como alternativa para construir soluciones de datos con menor dependencia de programación intensiva. Berthold (2023) destaca que este enfoque acerca capacidades analíticas y de ciencia de datos a perfiles no exclusivamente programadores, mientras que KNIME (2024) presenta su plataforma como un entorno visual orientado al diseño de flujos reproducibles de transformación y análisis de datos.

Dentro del estado del arte, también se identifica una línea de investigación centrada en la calidad de datos como condición previa para el aprovechamiento analítico. Revisiones y marcos

recientes coinciden en que dimensiones como completitud, validez, unicidad y consistencia siguen siendo criterios fundamentales para evaluar si un conjunto de datos es apto para uso analítico. Aunque la literatura reconoce otras dimensiones, estas cuatro mantienen alta relevancia cuando se estudian procesos de preparación y depuración sobre datos estructurados, como los involucrados en escenarios comerciales y operacionales.

Ahora bien, aunque existe abundante literatura sobre calidad de datos, adopción de analítica y ventajas del low-code, se observa una menor cantidad de trabajos aplicados que integren esos tres componentes en un mismo ejercicio metodológico y evaluativo, especialmente en escenarios tipo PYME construidos con lógica empresarial y sometidos a inconsistencias controladas. En otras palabras, buena parte de la literatura describe beneficios potenciales de la automatización o de las plataformas low-code, pero no siempre presenta implementaciones comparativas donde se contraste, con métricas observables, un procedimiento manual de preparación frente a un flujo ETL automatizado sobre tablas relacionadas de clientes, productos, proveedores, ventas, compras, inventario y movimientos. Esa brecha es justamente el espacio en el que se ubica el presente proyecto.

Desde esa perspectiva, este trabajo aporta un caso aplicado y controlado en el que se diseña e implementa un pipeline ETL low-code en KNIME para un escenario empresarial tipo PYME, con énfasis en la detección y tratamiento de duplicados, valores inválidos, errores de codificación, problemas de integridad referencial y homologación semántica. Su aporte no se limita a demostrar que la herramienta puede ejecutar transformaciones, sino a evaluar su efecto sobre la eficiencia del proceso de preparación y sobre la calidad final del conjunto de datos. Así, el proyecto se alinea con las discusiones contemporáneas sobre automatización accesible, calidad

de datos y adopción analítica en organizaciones con restricciones operativas, pero lo hace desde una implementación concreta, documentada y evaluable.

Marco Contextual

El proyecto se desarrolla en un contexto empresarial simulado de tipo PYME, construido con lógica comercial y operativa, orientado a representar condiciones frecuentes de alistamiento de información en organizaciones que gestionan su información mediante archivos planos, hojas de cálculo y exportaciones parciales de sistemas transaccionales. El contexto definido no corresponde a una empresa específica, sino a un entorno controlado y verosímil en el que se modelan procesos habituales de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario, con el propósito de evaluar el comportamiento de una solución ETL low-code en condiciones cercanas a la práctica empresarial.

Desde el punto de vista funcional, el entorno reproduce una operación comercial donde coexisten actividades de abastecimiento, gestión de catálogo, control de inventario y registro de transacciones. En este entorno, la información de clientes permite identificar la base comercial sobre la cual se realizan ventas; la tabla de productos representa el catálogo de referencias; los proveedores soportan los procesos de compra y abastecimiento; las ventas registran la salida comercial de productos; las compras documentan el ingreso por adquisición; el inventario consolida existencias por producto y ubicación; y los movimientos de inventario reflejan eventos operativos asociados a entradas, salidas y ajustes. La articulación de estas siete tablas permite construir un modelo de datos suficientemente completo para simular un proceso real de integración y depuración.

El contexto del proyecto también está determinado por la naturaleza de las inconsistencias incorporadas en las fuentes. Con el fin de aproximar condiciones comunes de trabajo en procesos manuales de alistamiento de información, se indujeron problemas como duplicados, valores fuera de rango, errores de codificación, registros huérfanos, dominios no

homogéneos y fallas de integridad referencial. Estos problemas se distribuyeron tanto en tablas maestras como en tablas transaccionales, lo que permitió evaluar el desempeño del proceso automatizado sobre distintos tipos de errores y sobre diferentes niveles de dependencia entre estructuras. En consecuencia, el proyecto no se limita a limpiar datos aislados, sino que aborda un contexto donde la calidad depende de la relación coherente entre múltiples tablas.

En términos de volumen, el entorno construido resulta adecuado para un ejercicio aplicado de especialización, ya que combina un tamaño de datos suficiente para evidenciar problemas de calidad con una escala manejable para el diseño, ejecución y evaluación del prototipo en KNIME. Esto permitió implementar un flujo completo de lectura, perfilamiento, limpieza, homologación, validación y exportación, sin perder trazabilidad sobre las reglas aplicadas ni sobre los efectos observados en cada etapa. Al mismo tiempo, el contexto conserva potencial de escalabilidad conceptual, puesto que la misma lógica del flujo puede extenderse a volúmenes mayores y a escenarios empresariales con más fuentes o más complejidad operacional.

Desde una perspectiva organizacional, el contexto modelado refleja una realidad frecuente en pequeñas y medianas empresas: el alistamiento de información suele depender de actividades manuales fragmentadas, ejecutadas por personal que debe revisar registros, corregir formatos, buscar inconsistencias entre tablas y consolidar archivos antes de poder generar reportes o análisis. Esta situación convierte esta etapa en un cuello de botella, porque consume tiempo operativo, dificulta la repetibilidad del proceso y expone el resultado a errores de manipulación. Por ello, el entorno definido es pertinente para analizar el valor de una solución ETL low-code que automatice tareas de integración, normalización y validación sin exigir un desarrollo complejo de software.

En este marco, la elección de KNIME como plataforma de implementación también adquiere sentido contextual. Al tratarse de un entorno visual basado en nodos, permite representar de forma clara y auditable cada etapa del proceso, desde la entrada de datos hasta la obtención de archivos curados. Esto resulta especialmente útil en contextos donde no se busca desarrollar una solución empresarial de producción a gran escala, sino un prototipo funcional, explicable y reutilizable que demuestre el efecto de la automatización sobre la eficiencia del proceso y la calidad del dato resultante.

En síntesis, el contexto del proyecto corresponde a un entorno PYME simulado, construido con lógica comercial realista, múltiples tablas relacionadas, inconsistencias inducidas y un proceso de preparación susceptible de automatización. Este contexto permite evaluar de forma aplicada y controlada la contribución de un flujo ETL en KNIME, no solo como herramienta técnica de transformación, sino como mecanismo para mejorar la trazabilidad, reducir esfuerzo manual y entregar datos con mejores condiciones para su aprovechamiento analítico.

Marco Teórico

La preparación de datos constituye una etapa crítica dentro de cualquier proceso analítico, debido a que la utilidad de los resultados depende en gran medida de la calidad, consistencia y disponibilidad de la información de entrada. En entornos empresariales tipo PYME, esta fase suele verse afectada por la dispersión de datos en múltiples archivos, la ausencia de estándares homogéneos y la dependencia de tareas manuales de revisión, consolidación y depuración. En consecuencia, antes de que los datos puedan utilizarse para reportes, seguimiento operativo o análisis, es necesario ejecutar procesos que permitan integrarlos, estandarizarlos y validarlos de forma sistemática.

En este contexto, los procesos ETL, correspondientes a extracción, transformación y carga, se consolidan como una estrategia técnica para convertir datos heterogéneos en un conjunto estructurado y utilizable. La etapa de extracción permite reunir información proveniente de diferentes fuentes; la transformación incorpora actividades de limpieza, homologación, validación, eliminación de duplicados, tratamiento de valores inválidos e integración por llaves; y la carga entrega un dataset final listo para consumo analítico. Más que un simple movimiento de datos, el ETL representa un mecanismo de control y trazabilidad sobre la preparación de la información.

El proyecto se inscribe además en el enfoque low-code, el cual propone la construcción de soluciones mediante componentes visuales y configurables, reduciendo la necesidad de programación extensa. Este enfoque resulta pertinente cuando se busca desarrollar prototipos funcionales, reproducibles y comprensibles en tiempos razonables. Dentro de este marco, KNIME se presenta como una plataforma adecuada para el diseño de workflows orientados a la

integración y transformación de datos, ya que permite representar de manera visual cada etapa del proceso, documentar reglas aplicadas y generar salidas verificables.

Otro eje teórico central es la calidad de datos. En términos generales, la calidad puede entenderse como el grado en que un conjunto de datos es apto para su uso. En este trabajo, dicha calidad se aborda a partir de cuatro dimensiones principales. La completitud se refiere a la presencia de valores en campos críticos; la unicidad alude a la ausencia de registros duplicados no deseados; la consistencia evalúa la coherencia entre campos, tablas y reglas del negocio; y la validez verifica el cumplimiento de dominios, formatos y rangos aceptables. Estas dimensiones permiten establecer un marco objetivo para comparar el estado de los datos antes y después del tratamiento realizado por la ETL.

De forma complementaria, la integridad referencial adquiere relevancia cuando el entorno de datos está compuesto por múltiples tablas relacionadas. En escenarios como el modelado en este proyecto, la calidad no depende únicamente de que cada tabla esté limpia de manera individual, sino de que existan correspondencias válidas entre clientes y ventas, productos y ventas, proveedores y compras, así como entre inventario, movimientos y catálogos. La detección de registros huérfanos y la validación de llaves se convierten así en una parte esencial del proceso de depuración.

Este marco teórico del proyecto se articula con la automatización de la preparación de datos como mecanismo para reducir tiempos operativos, minimizar errores humanos y aumentar la repetibilidad del proceso. Bajo esta perspectiva, el valor de una solución ETL no radica solo en producir archivos limpios, sino en ofrecer un flujo estructurado, documentado y reutilizable que permita transformar datos de entrada con criterios explícitos y verificables. Esa es precisamente la base sobre la cual se sustenta la propuesta desarrollada en KNIME.

Marco Normativo

El desarrollo del presente proyecto se relaciona con un marco normativo que orienta el tratamiento responsable de la información, el acceso a los datos y la calidad de los procesos asociados a su gestión. En primer lugar, en Colombia la Ley 1581 de 2012 establece las disposiciones generales para la protección de datos personales, definiendo principios y obligaciones aplicables al tratamiento de información que identifique o pueda asociarse a personas. Esta norma resulta pertinente para el proyecto porque, aunque el escenario utilizado es de carácter académico y controlado, incorpora tablas con información de clientes y proveedores, por lo que el diseño del pipeline debe considerar criterios de manejo responsable, finalidad y protección de los datos.

De manera complementaria, el Decreto 1377 de 2013 reglamenta parcialmente la Ley 1581 y desarrolla aspectos operativos del tratamiento de datos personales, mientras que el Decreto 1074 de 2015 compila disposiciones del sector Comercio, Industria y Turismo e incorpora referencias al cumplimiento normativo en materia de protección de datos personales. Estas disposiciones refuerzan la importancia de estructurar procesos de tratamiento de información bajo criterios formales y verificables.

En materia de acceso y gestión de la información, la Ley 1712 de 2014, conocida como Ley de Transparencia y del Derecho de Acceso a la Información Pública Nacional, regula el derecho de acceso a la información pública y promueve principios de disponibilidad, organización y consulta de la información. Aunque el proyecto no se orienta al cumplimiento institucional de transparencia, esta norma aporta un referente importante sobre la necesidad de contar con información estructurada, accesible y gestionada de manera adecuada, lo cual se relaciona directamente con los procesos de integración y preparación de datos.

Desde el punto de vista técnico, también es pertinente considerar normas y lineamientos asociados a la calidad y gobernanza de datos. En el contexto colombiano, ICONTEC reporta la adopción de referencias como la NTC-ISO 8000-51, relacionada con calidad y gobernanza de datos, la cual sirve como marco general para resaltar la importancia de políticas, estandarización y control sobre los datos dentro de procesos organizacionales. Aunque el presente trabajo no pretende certificar el flujo construido bajo una norma específica, sí se alinea con esa lógica al definir reglas de validación, homologación y consistencia para el tratamiento de la información.

En conjunto, este marco normativo respalda la pertinencia del proyecto desde tres frentes: la protección de datos personales, la organización y acceso a la información, y la necesidad de gestionar los datos bajo criterios de calidad. En consecuencia, el pipeline ETL desarrollado en KNIME no solo se justifica como una solución técnica de automatización, sino también como un mecanismo coherente con principios normativos de tratamiento responsable, estructuración y mejora de la información.

Metodología

La investigación se desarrolló bajo un enfoque aplicado, orientado al diseño, implementación y evaluación de un prototipo ETL en KNIME para un escenario empresarial tipo PYME. El trabajo no se centró en la formulación teórica de un modelo abstracto, sino en la construcción de una solución funcional que permitiera integrar, depurar y validar datos comerciales, así como medir su efecto sobre la eficiencia del proceso de preparación y la calidad del conjunto de datos resultante.

El desarrollo metodológico se organizó en cuatro momentos principales. En primer lugar, se estructuró el entorno de datos de trabajo a partir de tablas representativas de clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Sobre estas tablas se indujeron inconsistencias controladas, tales como duplicados, valores inválidos, errores de codificación, formatos no homogéneos y problemas de integridad referencial, con el fin de simular condiciones comunes de preparación manual.

La inducción de inconsistencias se realizó de manera controlada sobre las tablas del entorno de datos, procurando representar problemas frecuentes en procesos manuales de preparación de información. Para ello se incorporaron registros duplicados en tablas maestras y transaccionales, valores numéricos fuera de rangos lógicos, errores en dominios categóricos, variaciones textuales en campos sujetos a homologación y llaves inexistentes para simular registros huérfanos. En las tablas maestras, las inconsistencias se orientaron principalmente a duplicidad, errores de codificación, valores inválidos y falta de estandarización semántica. En las tablas transaccionales, se priorizaron errores asociados a cantidades, precios, stocks, tipos de movimiento e integridad referencial con clientes, productos y proveedores.

Las reglas de validación se definieron a partir de criterios técnicos y de negocio asociados a la estructura del modelo de datos. En primer lugar, se identificaron campos críticos para cada tabla, como identificadores, cantidades, precios, plazos, stocks y categorías. En segundo lugar, se establecieron condiciones mínimas de calidad relacionadas con completitud, unicidad, consistencia y validez. Finalmente, estas condiciones se tradujeron en reglas operativas dentro de KNIME, tales como eliminación de duplicados, validación de rangos numéricos, control de dominios permitidos, homologación de valores textuales y verificación de relaciones entre tablas mediante llaves primarias y foráneas. De esta forma, las reglas aplicadas no fueron arbitrarias, sino que respondieron a la lógica funcional del contexto comercial modelado y a las dimensiones de calidad definidas en el proyecto.

Con base en los criterios anteriores, se diseñó el flujo en KNIME, organizado por etapas de entrada de datos, perfilamiento inicial, limpieza de tablas maestras, limpieza de tablas transaccionales, validaciones y generación de salidas limpias. Esta organización permitió mantener una secuencia lógica entre el diagnóstico inicial de las fuentes, la aplicación de reglas de calidad y la obtención de archivos depurados.

Posteriormente, se implementó el proceso ETL utilizando nodos orientados a lectura, perfilamiento, filtros basados en reglas, homologación, eliminación de duplicados, cruces relacionales y exportación de resultados. El flujo permitió producir datasets limpios a partir de las fuentes con problemas de calidad, conservando trazabilidad sobre las transformaciones aplicadas y sobre los registros descartados por incumplimiento de reglas.

Finalmente, se realizó la evaluación del prototipo mediante una comparación entre el escenario de datos inconsistente y el conjunto de datos resultante tras la ejecución de la ETL. Esta evaluación se apoyó en métricas de calidad de datos y en una comparación de eficiencia

entre el tiempo registrado para el procedimiento manual equivalente, medido a partir de la ejecución secuencial de actividades de preparación, y el tiempo de ejecución observado del workflow automatizado en KNIME.

Método

El método empleado fue de carácter deductivo-aplicado. Se partió de criterios generales sobre integración, validación y calidad de datos para definir reglas concretas de tratamiento dentro del flujo ETL. A partir de esas reglas se diseñó un procedimiento sistemático que permitió revisar la estructura de las tablas, detectar inconsistencias, aplicar transformaciones y obtener un conjunto de datos curado. Este método fue adecuado porque permitió traducir principios generales de preparación y calidad de datos en una implementación técnica verificable dentro del entorno de KNIME.

Tipo de Estudio

El estudio es de tipo aplicado, con alcance descriptivo y evaluativo. Es aplicado porque busca resolver un problema concreto de preparación de datos mediante la construcción de un prototipo ETL funcional. Es descriptivo porque caracteriza el entorno de datos, las inconsistencias presentes y las reglas utilizadas para su tratamiento. Es evaluativo porque compara el estado inicial y final de los datos, así como la eficiencia del procedimiento automatizado frente al tiempo registrado para el procedimiento manual equivalente.

Recolección de Datos

La recolección de datos se realizó a partir de un entorno de datos estructurado para el proyecto, compuesto por siete tablas con lógica comercial y operativa: clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Estas tablas constituyeron

la base del experimento y fueron utilizadas tanto en su versión inconsistente como en su versión limpia resultante del proceso ETL.

Sobre las fuentes de entrada se aplicó una estrategia de observación estructurada de los datos, apoyada en el perfilamiento inicial realizado en KNIME. Esta fase permitió identificar tipos de inconsistencias, dominios irregulares, duplicados, valores inválidos y registros huérfanos. Adicionalmente, durante la ejecución del flujo se recolectaron conteos de salida de nodos clave, los cuales sirvieron como evidencia para el análisis de resultados.

La información recolectada se organizó en dos niveles. El primero correspondió a los datos de entrada y salida del proceso ETL, expresados en archivos inconsistentes y archivos limpios. El segundo correspondió a los indicadores observados durante la ejecución del workflow, tales como número de registros por tabla, registros afectados por reglas de validación y registros finales conservados tras el tratamiento. Esta información constituyó la base empírica para la evaluación del proyecto.

Resultados

El desarrollo del proyecto permitió diseñar, implementar y evaluar una solución ETL low-code en KNIME orientada a la integración, depuración y validación de datos comerciales en un contexto PYME. Los resultados obtenidos evidencian, por una parte, la viabilidad técnica del flujo construido y, por otra, su aporte en términos de calidad de datos y eficiencia del proceso de preparación. Para efectos del análisis, los resultados se presentan en dos bloques: el primero asociado al diseño e implementación del pipeline, y el segundo enfocado en la evaluación cuantitativa del tratamiento realizado sobre las fuentes inconsistentes.

Primer Resultado

El primer resultado del proyecto corresponde al diseño e implementación de un pipeline ETL funcional en KNIME, estructurado para procesar siete tablas relacionadas: clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. El flujo fue organizado en bloques de entrada de datos, perfilamiento inicial, limpieza de tablas maestras, limpieza de tablas transaccionales, validaciones de integridad y generación de salidas limpias, lo que permitió mantener trazabilidad sobre cada etapa del tratamiento.

En las tablas maestras se aplicaron procesos de limpieza de texto, normalización, homologación, validación de reglas de negocio y eliminación de duplicados. En clientes se controlaron valores inválidos de cupo de crédito y se homologaron municipios; en productos se validaron relaciones entre costo base y precio de lista; y en proveedores se estandarizó el tipo de proveedor y se filtraron plazos de pago inválidos. Estas transformaciones permitieron consolidar tablas maestras limpias y consistentes, necesarias para soportar las validaciones posteriores sobre las tablas transaccionales.

En las tablas transaccionales se aplicaron validaciones más exigentes, debido a su mayor volumen y a su dependencia de llaves relacionadas. En ventas y compras se verificaron cantidades y precios válidos, y se aplicaron cruces con tablas maestras limpias para detectar registros huérfanos. En inventario se validaron rangos de stock y se homologó la temperatura de bodega. En movimientos de inventario se controló el tipo de movimiento y se verificó la correspondencia con productos existentes. De esta manera, el pipeline no solo automatizó tareas de depuración aisladas, sino que integró validaciones de consistencia estructural y semántica entre múltiples tablas.

Como salida del proceso se obtuvieron siete archivos limpios, correspondientes a clientes, productos, proveedores, ventas, compras, inventario y movimientos de inventario. Estos datasets constituyen el resultado curado del flujo ETL y representan un insumo apto para análisis posterior, generación de reportes y posible reutilización del prototipo en contextos similares. La Figura 1 presenta la vista general del flujo implementado; la Figura 2 muestra el bloque de limpieza de tablas maestras; y la Figura 3 presenta la limpieza y validación de tablas transaccionales.

Un uso posterior de los datos curados podría darse en la construcción de un tablero de inteligencia de negocios para seguimiento comercial y operativo. Por ejemplo, a partir de las tablas limpias de ventas, productos, clientes e inventario, sería posible elaborar indicadores como ventas por período, productos con mayor rotación, clientes con mayor participación, niveles de inventario por bodega y alertas de stock crítico. De igual forma, estos datos podrían alimentar modelos analíticos sencillos orientados a identificar patrones de compra, estimar demanda o priorizar productos con riesgo de desabastecimiento. Esto evidencia que el pipeline ETL no

representa únicamente una etapa de limpieza, sino una base habilitadora para procesos posteriores de análisis, visualización y apoyo a la toma de decisiones.

Figura 1

Workflow ETL Final Implementado en KNIME

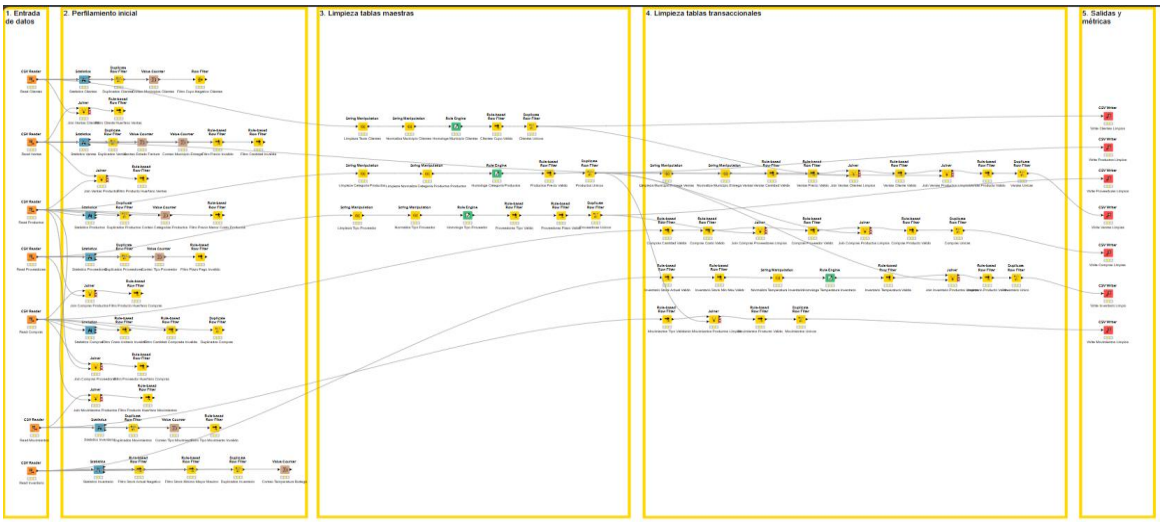


Figura 2

Limpieza y Depuración de Tablas Maestras en el Pipeline ETL

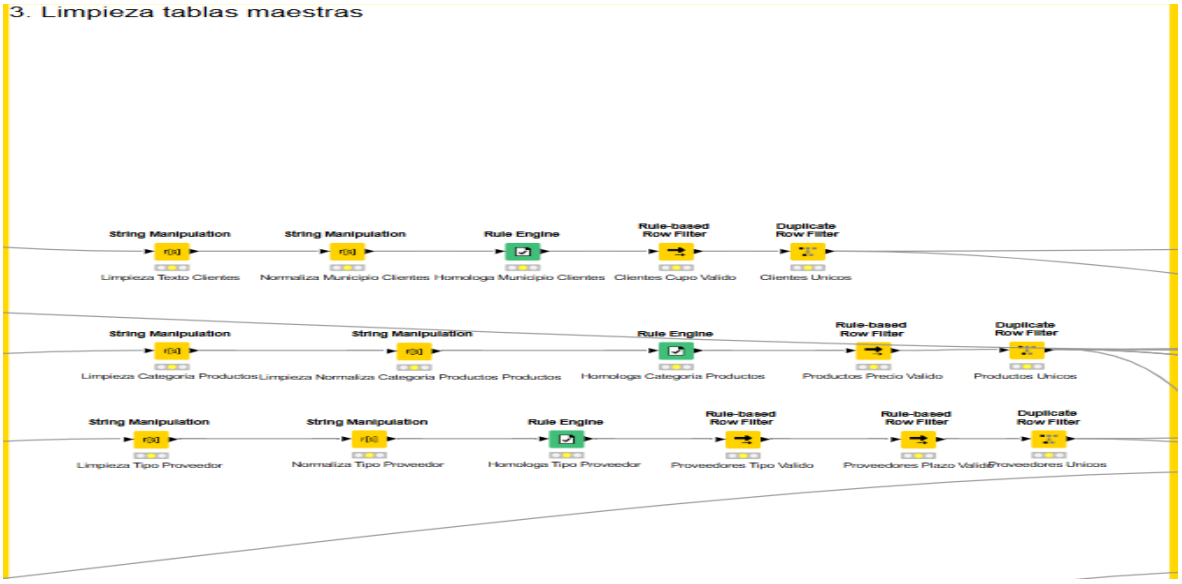
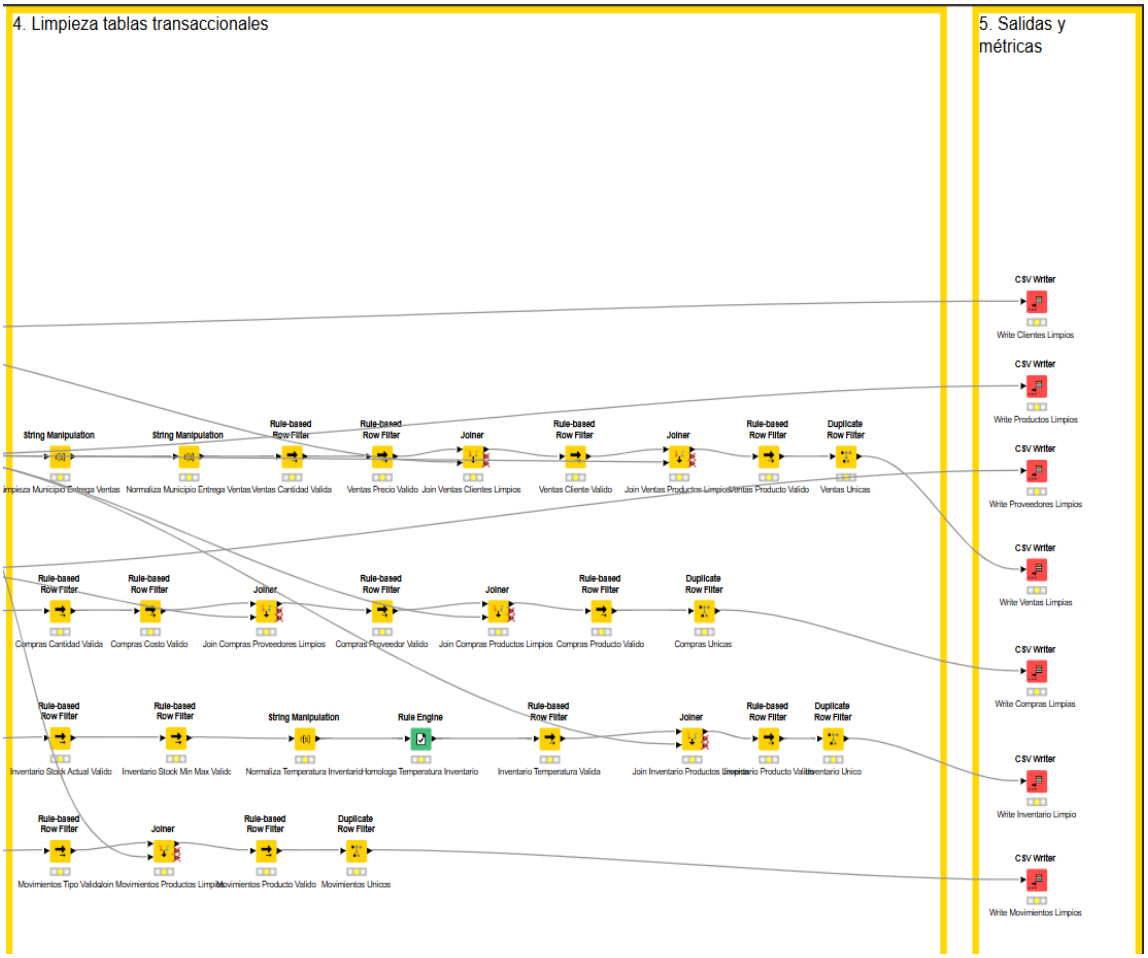


Figura 3

Limpieza y Validación de Tablas Transaccionales en el Pipeline ETL



Segundo Resultado

El segundo resultado corresponde a la evaluación cuantitativa del comportamiento del pipeline sobre los datos inconsistentes. En términos generales, la ETL permitió detectar y tratar registros duplicados, valores inválidos, problemas de integridad referencial y errores de homologación semántica, generando una reducción controlada del volumen inicial y una mejora observable en la calidad del conjunto de datos resultante.

En clientes se partió de 1.520 registros y se obtuvo una salida final de 1.486, con 15 registros detectados con cupo de crédito negativo. En productos se pasó de 252 a 233 registros, identificándose 8 casos en los que el precio de lista era inferior al costo base. En proveedores se inició con 86 registros y se finalizó con 72, luego de aplicar homologación del tipo de proveedor y validación del plazo de pago.

En las tablas transaccionales se observó la mayor afectación por inconsistencias. En ventas se procesaron 60.300 registros inconsistentes y se obtuvo una salida final de 49.880 registros limpios. En esta tabla se detectaron 907 registros con cantidad inválida, 903 con precio inválido, 3.780 con cliente huérfano y 3.932 con producto huérfano. En compras, el volumen inicial fue de 18.120 registros y la salida final de 14.599; en esta tabla se identificaron 273 registros con cantidad inválida, 360 con proveedor huérfano y 760 con producto huérfano. En inventario se pasó de 1.000 a 834 registros, con 29 casos de stock actual negativo y 30 casos de stock mínimo superior al stock máximo. En movimientos de inventario se procesaron 90.250 registros inconsistentes y se obtuvieron 81.999 registros limpios, detectándose 12 tipos de movimiento inválidos y 5.838 registros con producto huérfano.

De manera agregada, los resultados muestran que la mayor afectación de calidad se concentró en las tablas transaccionales, especialmente en ventas, compras y movimientos, donde el volumen de datos y la dependencia de relaciones entre tablas incrementaron la presencia de errores. Esto confirma que el diseño del flujo debía priorizar no solo validaciones de formato y dominio, sino también controles de integridad referencial entre estructuras relacionadas.

En cuanto a la eficiencia del proceso, la ejecución completa del workflow en KNIME tomó 1 minuto para el volumen experimental utilizado. Para efectos comparativos, el procedimiento manual de preparación fue reconstruido y medido a partir de la ejecución

secuencial de actividades equivalentes de carga, revisión estructural, depuración de duplicados, validación de rangos y dominios, verificación de integridad referencial, homologación de valores textuales y consolidación final de salidas. Bajo esta lógica, el tiempo total registrado para el procedimiento manual fue de 6 horas.

Reconstrucción y Medición del Procedimiento Manual de Referencia

Con el fin de complementar la comparación de eficiencia, se documentó el procedimiento manual equivalente al tratamiento realizado por la ETL. Para ello se desagregaron las actividades necesarias para preparar manualmente las siete tablas del escenario, considerando la secuencia operativa que seguiría un analista para organizar archivos, revisar estructuras, detectar duplicados, validar reglas, homologar valores y consolidar resultados. Este ejercicio permitió establecer un tiempo total de 360 minutos, equivalentes a 6 horas de trabajo.

El registro del tiempo se realizó sumando los tiempos observados en cada actividad manual equivalente, ejecutadas de forma secuencial sobre el entorno de datos del proyecto.

Desde una perspectiva operativa, la diferencia observada no solo refleja una reducción del tiempo de preparación, sino también una disminución del esfuerzo requerido para repetir el proceso bajo condiciones homogéneas. Mientras el procedimiento manual exige múltiples revisiones y cruces ejecutados de forma fragmentada, el pipeline ETL concentra dichas tareas en un flujo reproducible y consistente, fortaleciendo la confiabilidad del tratamiento aplicado sobre los datos.

Con el fin de consolidar la evidencia obtenida durante la implementación y evaluación del prototipo, a continuación se presentan las tablas de soporte relacionadas con la estructura del entorno de datos, las inconsistencias inducidas, las reglas de validación aplicadas, los cambios

observados tras el tratamiento ETL y la comparación de eficiencia entre el procedimiento manual y el flujo automatizado.

Tabla 1

Estructura del Entorno de Datos

Tabla	Tipo	Propósito
clientes	maestra	información comercial de clientes
productos	maestra	catálogo de referencias
proveedores	maestra	fuente de abastecimiento
ventas	transaccional	salida comercial
compras	transaccional	ingreso por adquisición
inventario	transaccional	existencias por producto y bodega
movimientos_inventario	transaccional	trazabilidad operativa

Nota. La tabla presenta las estructuras de datos utilizadas en el entorno experimental del proyecto, clasificadas según su función dentro del modelo de datos.

Tabla 2

Inconsistencias Inducidas por Tabla

Tabla	Inconsistencias principales inducidas
clientes	duplicados, cupo negativo, municipios inconsistentes
productos	duplicados, precio menor al costo
proveedores	tipo de proveedor inconsistente, plazo inválido
ventas	cantidad inválida, precio inválido, cliente huérfano, producto huérfano
compras	cantidad inválida, proveedor huérfano, producto huérfano
inventario	stock negativo, stock mínimo mayor al máximo, temperatura inconsistente
movimientos_inventario	tipo inválido, producto huérfano

Nota. La inducción de inconsistencias se realizó de forma controlada para simular condiciones frecuentes de preparación manual de datos en escenarios empresariales tipo PYME.

Tabla 3

Reglas de Validación Aplicadas

Regla	Aplicación
eliminación de duplicados	llaves principales y combinadas
validación de rangos	cantidades, precios, stock
integridad referencial	ventas, compras, inventario y movimientos
homologación semántica	municipios, tipo de proveedor, temperatura
validación de dominios	estados, tipos, categorías

Nota. Estas reglas orientaron la depuración y validación de los datos durante la ejecución del workflow en KNIME.

Tabla 4

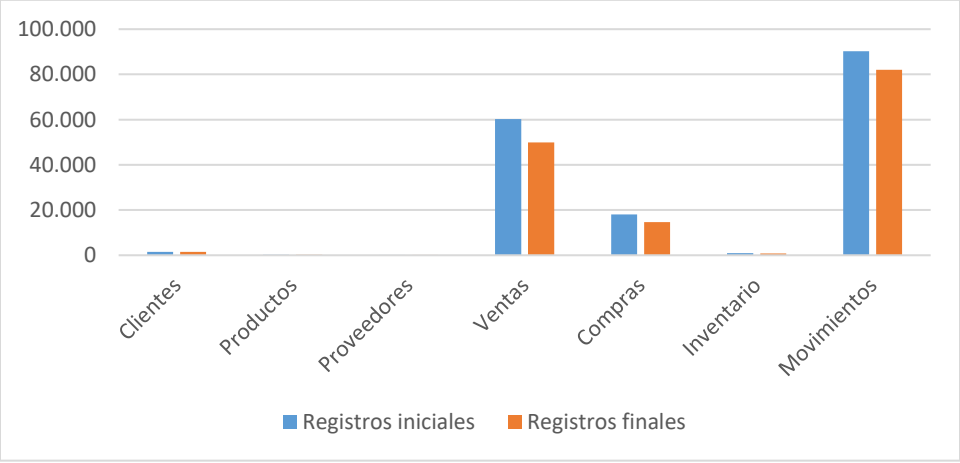
Comparación Antes vs. Después por Tabla

Tabla	Registros iniciales	Registros finales	Reducción absoluta	Reducción %
Clientes	1520	1486	34	2,24 %
Productos	252	233	19	7,54 %
Proveedores	86	72	14	16,28 %
Ventas	60300	49880	10420	17,28 %
Compras	18120	14599	3521	19,43 %
Inventario	1000	834	166	16,60 %
Movimientos	90250	81999	8251	9,14 %

Nota. La reducción porcentual corresponde a la diferencia entre el número de registros iniciales y finales por tabla, luego de aplicar las reglas de validación y depuración del pipeline ETL.

Figura 4

Comparación de Registros Iniciales y Finales Después del Tratamiento ETL



Nota. Registros iniciales y finales obtenidos tras la ejecución del flujo ETL.

Los datos comparativos muestran que la depuración fue más pronunciada en las tablas transaccionales que en las tablas maestras. En particular, compras presentó una reducción del 19,43 %, ventas del 17,28 %, inventario del 16,60 % y movimientos del 9,14 %, lo que sugiere que la mayor carga de inconsistencias se concentró en estructuras con mayor volumen y con dependencia de llaves relacionadas. En contraste, clientes y productos presentaron reducciones más moderadas, coherentes con su función de soporte como tablas maestras del modelo.

Tabla 5

Inconsistencias Detectadas por Tipo

Tabla	Tipo de inconsistencia	Registros detectados
clientes	cupo de crédito negativo	15
productos	precio menor al costo	8
proveedores	plazo de pago inválido	6
ventas	cantidad inválida	907
ventas	precio inválido	903
ventas	cliente huérfano	3780
ventas	producto huérfano	3932
compras	cantidad inválida	273
compras	proveedor huérfano	360
compras	producto huérfano	760
inventario	stock actual negativo	29
inventario	stock mínimo mayor al máximo	30
movimientos_inventario	tipo de movimiento inválido	12
movimientos_inventario	producto huérfano	5838

Nota. Los valores corresponden a registros identificados en nodos de validación del flujo ETL implementado en KNIME.

Tabla 6

Desagregación del Procedimiento Manual de Referencia

Actividad manual equivalente	Tiempo estimado
Carga, organización y revisión inicial de archivos	35 minutos
Revisión de estructura y campos críticos	30 minutos
Identificación y depuración de duplicados	60 minutos
Validación de rangos, formatos y dominios	55 minutos
Verificación de integridad referencial entre tablas	90 minutos
Homologación de valores textuales y categóricos	50 minutos
Consolidación, validación final y exportación	40 minutos
Total estimado	360 minutos (6 horas)

*Nota.*La desagregación presenta las actividades consideradas para medir el procedimiento manual equivalente al tratamiento realizado por la ETL, expresadas en minutos de trabajo secuencial.

Tabla 7

Comparación de Eficiencia entre el Procedimiento Manual de Referencia y la ETL en KNIME

Proceso	Tiempo
Procedimiento manual de referencia	6 horas
Flujo ETL en KNIME	1 minuto

Nota. El tiempo manual corresponde al tiempo registrado a partir de la reconstrucción y ejecución secuencial del procedimiento manual equivalente. El tiempo de la ETL corresponde a la ejecución observada del workflow en KNIME para el volumen experimental del proyecto.

Tabla 8

Métricas de Calidad de Datos Antes y Después del Tratamiento ETL

Dimensión	Situación antes del tratamiento	Situación después del tratamiento	Evidencia de mejora
Unicidad	Presencia de duplicados en todas las tablas	Se conservaron salidas depuradas, reduciendo registros repetidos en el conjunto final.	Mejora alta
Validez	Existencia de valores fuera de rango y dominios inválidos	Se trataron 2.183 registros inválidos asociados a cupo negativo, precios, cantidades, plazos, stocks y tipos de movimiento.	Alta: 100 % de registros inválidos detectados fueron tratados.
Consistencia	Inconsistencias semánticas y relacionales entre tablas	Se trataron 14.670 registros con fallas de consistencia, principalmente clientes, productos y proveedores huérfanos.	Alta: 100 % de registros huérfanos detectados fueron tratados.
Completitud	Afectada por registros inválidos o relaciones incompletas	Las salidas limpias quedaron conformadas por registros con mejores condiciones de uso analítico	Media, con mejora indirecta en la calidad del conjunto final.

Nota. La tabla consolida las dimensiones de calidad evaluadas en el proyecto a partir de los registros identificados en los nodos de validación, depuración e integración del flujo ETL implementado en KNIME.

Desde la perspectiva de calidad de datos, la mejora más evidente se presentó en las dimensiones de validez y consistencia, debido a que estas cuentan con registros cuantificados durante la ejecución del flujo ETL. En validez se trataron 2.183 registros con valores fuera de

rango, dominios inválidos o incumplimiento de reglas numéricas. En consistencia se trataron 14.670 registros con fallas de integridad referencial, principalmente asociados a clientes, productos y proveedores huérfanos. En conjunto, estos resultados evidencian que el proceso implementado no solo redujo el volumen de registros con problemas de calidad, sino que fortaleció las condiciones mínimas requeridas para el uso analítico de los datos.

La comparación evidencia que el mayor aporte del pipeline no se limita a acelerar el tratamiento de los datos, sino a concentrar en un solo flujo actividades que manualmente demandan revisiones repetitivas, validaciones cruzadas y decisiones operativas distribuidas. En este sentido, la mejora observada en tiempo se complementa con una mayor uniformidad en la aplicación de reglas y una mejor trazabilidad del proceso de preparación.

En conjunto, los resultados evidencian que el pipeline ETL implementado en KNIME cumplió el propósito del proyecto: transformar un conjunto de datos inconsistentes en un entorno de datos limpio, estructurado y validado, al tiempo que redujo de forma significativa el esfuerzo operativo asociado a la preparación manual de la información.

Conclusiones

El proyecto permitió comprobar que el diseño e implementación de un proceso ETL low-code en KNIME mejora de manera verificable la preparación de datos comerciales en el contexto PYME modelado. La automatización del flujo hizo posible integrar, depurar y validar siete tablas relacionadas, aplicando reglas explícitas de calidad sobre duplicados, rangos inválidos, dominios inconsistentes e integridad referencial. En consecuencia, se obtuvo un entorno de datos limpio, estructurado y apto para análisis, lo que responde de forma directa a la pregunta de investigación planteada.

Desde la perspectiva de calidad de datos, los resultados evidenciaron mejoras claras en las dimensiones de unicidad, validez y consistencia. La unicidad se fortaleció mediante la eliminación de duplicados en tablas maestras y transaccionales; la validez mejoró por la exclusión de registros con cantidades, precios, plazos y stocks inválidos; y la consistencia se incrementó gracias a la homologación semántica y a la validación de relaciones entre tablas. Esto confirma que la calidad del dato no depende únicamente de limpiar registros aislados, sino de asegurar coherencia estructural y semántica en todo el modelo de datos.

La evaluación comparativa mostró una diferencia significativa en eficiencia entre el procedimiento manual equivalente y el flujo automatizado en KNIME. Mientras el tratamiento manual requirió un tiempo registrado de 6 horas, la ejecución completa del flujo en KNIME tomó 1 minuto para el volumen experimental utilizado. Esta diferencia demuestra que la automatización no solo reduce tiempos operativos, sino que también disminuye el esfuerzo requerido para repetir el proceso bajo criterios homogéneos, con mayor trazabilidad y menor dependencia de intervenciones fragmentadas.

Otro aspecto relevante demostrado por el proyecto es que la solución construida no se limita a una ejecución puntual, sino que constituye un prototipo funcional y reproducible. El flujo desarrollado en KNIME organiza las etapas de lectura, perfilamiento, limpieza, homologación, validación y exportación mediante reglas explícitas, lo que permite repetir el proceso bajo condiciones homogéneas y verificar los resultados obtenidos. Esta característica fortalece su aplicabilidad en contextos empresariales similares, especialmente en organizaciones que gestionan información comercial y operativa desde múltiples fuentes y requieren mejorar la calidad de sus datos sin implementar desarrollos complejos de software.

El proyecto también permitió evidenciar que, en escenarios empresariales tipo PYME, la principal carga de inconsistencias suele concentrarse en las tablas transaccionales, especialmente cuando existen relaciones entre múltiples estructuras y altos volúmenes de registros. Por ello, un aporte importante del trabajo fue demostrar que una solución ETL no debe limitarse a transformaciones superficiales de formato, sino incorporar validaciones de negocio, controles relacionales y homologación semántica para asegurar condiciones mínimas de calidad analítica en el conjunto de datos final.

Recomendaciones

A partir de los resultados obtenidos, se recomienda que futuros trabajos amplíen el prototipo hacia escenarios con mayores volúmenes de información y con fuentes adicionales, de modo que sea posible evaluar el comportamiento del proceso ETL en condiciones más cercanas a un entorno empresarial real. Aunque el flujo implementado demostró ser funcional y eficiente para el contexto experimental desarrollado, su validación sobre conjuntos de datos más extensos permitiría profundizar en aspectos de escalabilidad, rendimiento y robustez operativa. Esta recomendación se sustenta en que el proyecto ya mostró mejoras claras en eficiencia y calidad, por lo que una siguiente fase natural sería probar el mismo diseño bajo una carga de datos mayor.

También se recomienda incorporar en futuras versiones del flujo mecanismos adicionales de trazabilidad y monitoreo, tales como bitácoras de ejecución, indicadores automáticos de calidad por corrida y reportes comparativos generados de forma automática. Esto permitiría fortalecer el uso del proceso ETL no solo como herramienta de transformación, sino como un componente de control continuo sobre la calidad de los datos. Dado que el proyecto ya documentó métricas de unicidad, validez, consistencia y completitud, una evolución lógica sería convertir esas verificaciones en salidas automáticas de seguimiento para facilitar la supervisión periódica del proceso.

Desde una perspectiva aplicada, se recomienda que organizaciones con características similares a las del escenario modelado consideren la adopción de flujos ETL low-code para reducir la dependencia de tareas manuales de consolidación, limpieza y validación de datos. Los resultados del proyecto mostraron que la automatización mediante KNIME no solo disminuye el tiempo operativo de preparación, sino que además mejora la uniformidad en la aplicación de reglas y favorece la trazabilidad del tratamiento. En contextos PYME, donde los recursos

técnicos suelen ser limitados, este tipo de solución puede representar una alternativa viable para fortalecer procesos analíticos sin requerir desarrollos extensos de programación.

Finalmente, se recomienda profundizar en la integración del proceso ETL con actividades posteriores de análisis y visualización, por ejemplo mediante herramientas de inteligencia de negocios o tableros de seguimiento. El prototipo desarrollado dejó como salida datasets limpios y estructurados, por lo que una extensión natural del trabajo consiste en aprovechar esos datos curados para construir reportes, indicadores o modelos analíticos que permitan evidenciar el impacto del ETL no solo en la preparación del dato, sino también en su uso efectivo para la toma de decisiones. Asimismo, futuras implementaciones del prototipo podrían incorporar tableros de control conectados a las salidas limpias generadas por la ETL, con el fin de validar el uso práctico de los datos curados en procesos de inteligencia de negocios. Esta integración permitiría pasar de la preparación automatizada de información a la generación de indicadores visuales para apoyar decisiones relacionadas con ventas, inventario, abastecimiento y comportamiento de clientes.

Referencias Bibliográficas

Al Alamin, M. A., Malakar, S., & Gias Uddin, M. (2021). *Un estudio empírico de las discusiones de desarrolladores sobre los desafíos del desarrollo de software low-code.* https://arxiv.org/abs/2103.11429

Amazon Web Services. (2025). *Explicación sobre la automatización inteligente de procesos*. https://aws.amazon.com/es/what-is/intelligent-automation/

Boada Pacheco, J. D., Bustos Caycedo, P. A., & Gómez Castañeda, J. A. (2025). *Automatización de procesos de facturación de ventas mediante RPA en una empresa del sector gastronómico en Sabana Centro* [Tesis de pregrado, Universidad de Cundinamarca]. https://repositorio.ucundinamarca.edu.co/items/9a6b6bcc-1137-4ba5-b455-8ea944e4d0be

Berthold, M. R. (2023, February 28). *Low-code brings business and data science closer. KNIME Blog*. https://www.knime.com/blog/low-code-brings-business-and-data-science-closer

Cadili, R. (2025, January 15). *How to automate data access and collection*. Medium. https://medium.com/low-code-for-advanced-data-science/how-to-automate-data-access-and-collection-e89f86a1cb14

Hassani, H., Huang, X., & Silva, E. (2020). *Digital transformation and big data analytics: A research agenda*. Journal of Business Research, 120, 328–337. https://doi.org/10.1016/j.jbusres.2020.07.037

IBM. (2025). *ETL moderno: El cerebro de la inteligencia artificial empresarial*. https://www.ibm.com/think/insights/modern-etl

Impacto TIC. (2024, May 5). *Data Science: Herramientas, retos y futuro en Colombia.* https://impactotic.co/tecnologia/data-science-herramientas-retos-y-futuro-en-colombia

Ideal Data. (2024). *Business Intelligence y la Automatización de Procesos*. https://idealdata.com.mx/automatizacion-business-intelligence/

Integrate.io. (2025). *Guía del analista veterano sobre herramientas ETL low-code.* https://www.integrate.io/blog/low-code-etl-tools-a-veteran-analysts-guide/

Jitterbit. (2025). *¿Qué es la integración de sistemas? Tipos, ejemplos y desafíos.* https://www.jitterbit.com/es/blog/what-is-system-integration/

KNIME. (2024). *KNIME Analytics Platform*. https://www.knime.com/knime-analytics-platform

KNIME. (2025). *Guía de instalación de la integración de Python en KNIME.* https://docs.knime.com/latest/python_installation_guide/

Mandala, N. R. (2019). *La evolución de la arquitectura ETL: De los procesos tradicionales por lotes a la integración de datos en tiempo real.* https://doi.org/10.30574/wjarr.2019.3.1.0033

Mikalef, P., Krogstie, J., Pappas, I. O., & Giannakos, M. (2018). *Investigating the effects of big data analytics capabilities on firm performance: The mediating role of dynamic capabilities. Information & Management*, 55(8), 103162. https://doi.org/10.1016/j.im.2018.03.004

Panchwadkar, D. (2023). *How to use the KNIME Business Hub REST API*. https://www.knime.com/blog/how-use-knime-business-hub-rest-api

Salesforce. (2025). *Herramientas de Business Intelligence: Definición y ventajas*. https://www.salesforce.com/es/analytics/business-intelligence-tools/

Silipo, R. (2021, April 12). *Low Code for Data Science: A new journal*. Medium. https://medium.com/low-code-for-advanced-data-science/low-code-for-advanced-data-science-a-new-journal-8bebee7ed619

Uzuntaş, G. (2024). *Integración de KNIME con Power BI*. https://medium.com/low-code-for-advanced-data-science/integrating-knime-with-power-bi-137abf2b685e

Van der Aalst, W. (2018). Process mining and Robotic Process Automation: Partners in process improvement. Cognitive Systems Research, 50, 1–7. https://doi.org/10.1016/j.cogsys.2018.05.004