

Propuesta de diseño y evaluación de un pipeline ETL low-code en Python para mejorar la preparación y calidad de datos en una pyme del sector serigráfico.

Trabajo final presentado por:	Oscar Eduardo Tolosa Cesar Augusto Bedoya Gómez
Programa:	Especialización en Inteligencia Artificial Aplicada a la Analítica de Datos
Docente:	Deidy Gutiérrez Niño
Fecha:	28 de julio de 2026

Índice de contenidos

Índice de tablas	3
Índice de ilustraciones	3
Resumen	4
Abstract.....	5
1. Introducción.....	7
1.1. Objetivo General.....	9
1.2. Objetivos Específicos.....	9
1.3. Metodología	9
1.3.1. Identificación y Análisis del Problema	12
1.3.2. Planificación de la Solución.....	13
1.4. Descripción de la empresa a intervenir.....	14
1.4.1. Misión	15
1.4.2. Visión.....	15
1.4.3. Mapa de procesos.....	15
1.4.4. Descripción del proceso a intervenir.....	16
1.4.5. Descripción del problema que presenta el proceso.....	17
2. Establecimiento de planes de acción para la solución del problema	20
2.1. Establecimiento de objetivos	20
2.2. Descripción de la solución	21
2.3. Secuencia de actividades para ejecución de la solución	22
2.4. Descripción de recursos necesarios para la ejecución de la solución	26
3. Definición del alcance de la solución, partes interesadas, cronograma y presupuestos...29	
3.1. Alcance final de la solución.....	29
3.2. Identificación de responsables de las actividades para ejecución.....	31
3.3. Cronograma de ejecución	32
3.4. Presupuesto	34
4. Referencias bibliográficas	37

Índice de tablas

Tabla 1: Matriz DOFA sobre la gestión de datos y analítica.....	19
Tabla 2: Matriz de responsabilidades para la ejecución del proyecto	32
Tabla 3: Cronograma de actividades del proyecto	33
Tabla 4: Presupuesto estimado del proyecto	35

Índice de ilustraciones

Ilustración 1: Mapa de procesos	16
---------------------------------------	----

RESUMEN

El presente trabajo de grado tiene como propósito proponer el diseño y evaluación de un pipeline ETL low-code desarrollado en Python para mejorar la preparación y calidad de datos en una pyme colombiana del sector serigráfico. El problema abordado se origina en la dependencia de procesos manuales para depurar, estandarizar y consolidar archivos CSV exportados desde el sistema ERP, situación que genera tiempos muertos, duplicidad de registros, inconsistencias en inventarios y menor confiabilidad en los reportes usados para la toma de decisiones. Para responder a esta necesidad, la investigación se desarrolla bajo un enfoque de investigación-acción, apoyado en observación directa estructurada del flujo operativo y comercial, cronometraje de micro tareas, análisis de causas raíz y construcción de un ambiente de prueba controlado con anomalías representativas del proceso real.

La solución propuesta consiste en un prototipo automatizado basado en Python y la librería Pandas, estructurado en fases de extracción, transformación, carga y auditoría. Este prototipo permite leer archivos CSV crudos, homogeneizar formatos, depurar campos textuales, detectar duplicados, validar reglas de negocio e identificar registros inconsistentes mediante un archivo de trazabilidad de errores. Los resultados esperados evidencian una mejora significativa frente al procedimiento manual tradicional, al reducir el tiempo de procesamiento de datos de rangos aproximados entre 45 y 90 minutos diarios a una ejecución automatizada de pocos segundos, con mayor integridad referencial, completitud y consistencia de la información. En consecuencia, la propuesta demuestra viabilidad técnica, operativa y financiera para pymes con recursos limitados, al apoyarse en herramientas de código abierto y contribuir a una cultura organizacional orientada a decisiones basadas en datos confiables.

Palabras clave: Analítica de datos; calidad de datos; ETL (Extracción, Transformación, Carga); inteligencia artificial; preparación de datos; Python; PYME.

ABSTRACT

The purpose of this undergraduate thesis is to propose the design and evaluation of a low-code ETL pipeline developed in Python to improve data preparation and data quality in a Colombian SME operating in the screen-printing sector. The problem addressed stems from the dependence on manual processes to clean, standardize, and consolidate CSV files exported from the ERP system, a situation that generates downtime, duplicate records, inventory inconsistencies, and lower reliability in the reports used for decision-making. To respond to this need, the research is developed under an action research approach, supported by structured direct observation of the operational and commercial workflow, microtask time measurement, root cause analysis, and the construction of a controlled testing environment with anomalies representative of the real process.

The proposed solution consists of an automated prototype based on Python and the Pandas library, structured into extraction, transformation, loading, and auditing phases. This prototype reads raw CSV files, standardizes formats, cleans textual fields, detects duplicates, validates business rules, and identifies inconsistent records through an error traceability file. The expected results show a significant improvement over the traditional manual procedure by reducing data processing time from approximately 45 to 90 minutes per day to an automated execution of only a few seconds, with greater referential integrity, completeness, and consistency of information. Consequently, the proposal demonstrates technical, operational,

and financial feasibility for SMEs with limited resources, as it relies on open-source tools and contributes to an organizational culture oriented toward decisions based on reliable data.

Keywords: Data Analytics; Data Quality; ETL (Extract, Transform, Load); ETL (Extract, Transform, Load); Data Preparation; Python; SME

1. INTRODUCCIÓN

En el entorno empresarial contemporáneo, la capacidad de tomar decisiones esenciales en datos, se ha consolidado como un diferenciador competitivo estratégico.

Las pequeñas y medianas empresas (pymes) del sector serigráfico en Colombia, la transición hacia una gestión basada en datos, enfrenta dificultades operativas asociados con la baja calidad, la falta de estandarización y la restringida automatización de la información, generada por los sistemas de programación de recursos empresariales (ERP). De manera periódica, los registros de ventas, consumo de tintas e inventarios de mallas, se exportan en archivos de valores separados por comas (CSV), lo que exige al personal administrativo y operativo la ejecución de los procesos manuales de depuración, validación y consolidación antes de que se realice cualquier análisis.

Aunque el formato CSV es muy usado por su interoperabilidad entre plataformas, su uso puede crear dificultades cuando los archivos carecen de esquemas de metadatos y reglas de validación estables. Autores como Fernandes et al. (2023) señalan que la preparación de datos requiere actividades de descubrimiento, limpieza, integración y transformación que suelen reunir una parte significativa del esfuerzo en los flujos analíticos. En efecto, cuando las exportaciones desde el ERP presentan estructuras diversas, codificación variable, campos vacíos o columnas desalineadas, las organizaciones deben invertir tiempo adicional en tareas manuales de perfilado y estandarización antes de disponer de la información confiable para la toma de decisiones.

Esta falta de homogeneidad tecnológica, extiende la dependencia de procesos artesanales de limpieza de datos, en los cuales los analistas destinan una proporción

considerable de su jornada a corregir archivos antes de ejecutar análisis descriptivos o predictivos. Esta situación reduce la eficiencia operativa, retrasa la creación de reportes comerciales y aumenta el costo asociado al procesamiento de información. En el contexto de una pyme serigráfica, el problema adquiere mayor relevancia porque la disponibilidad oportuna de datos limpios incide directamente en el control de inventarios, la programación de la producción, el seguimiento de pedidos y la respuesta al cliente.

Frente a esta problemática, la automatización mediante herramientas de código abierto constituye una alternativa técnica y costo-eficiente para organizaciones con recursos limitados. Python y la librería Pandas ofrecen capacidades para la lectura, depuración, transformación y consolidación de datos tabulares, lo que permite estructurar pipelines locales de extracción, transformación y carga (ETL) sin incurrir en altos costos de licenciamiento (McKinney, 2022). Asimismo, los modelos de procesos de data warehousing y ETL permiten organizar de manera sistemática las fases de preparación de datos, desde la extracción de fuentes heterogéneas hasta la generación de conjuntos depurados y listos para análisis (Dhaouadi et al., 2022).

El presente proyecto de aplicación, desarrollado bajo un enfoque de investigación-acción, se centra en el diagnóstico y tratamiento de esta problemática en una pyme del sector serigráfico. La propuesta integra teoría y práctica, mediante la observación directa del flujo operativo, el análisis de causas raíz y el diseño de un prototipo ETL low-code en Python. De esta manera, el trabajo busca contribuir a una solución posible, reproducible y alineada con las necesidades reales de la organización, orientada a reducir tiempos de procesamiento, mitigar los errores manuales y fortalecer la calidad de los datos utilizados para la toma de decisiones.

1.1. OBJETIVO GENERAL

El objetivo principal del proyecto es proponer el diseño y evaluación de una solución de automatización basada en Python para la depuración, estandarización y consolidación de datos de ventas e inventarios provenientes de sistemas ERP en una pyme del sector serigráfico. La finalidad es reducir los tiempos de procesamiento manual, mitigar errores humanos y habilitar análisis descriptivos y predictivos como soporte a la toma de decisiones estratégicas. La intervención se orienta, además, a formular una propuesta sostenible que pueda integrarse progresivamente a la estructura operativa de la organización.

1.2. OBJETIVOS ESPECÍFICOS

Para cumplir con el objetivo general, se establecieron tres objetivos específicos:

1. Realizar una identificación detallada del problema en su contexto organizacional, determinando causas raíz y evaluando su impacto en el desempeño general del proceso.
2. Diseñar un plan de acción estructurado con objetivos claros, actividades específicas y recursos definidos para la ejecución de la solución.
3. Establecer el alcance de la intervención, definiendo responsables, cronograma y presupuesto, para asegurar que la implementación sea organizada y efectiva.

1.3. METODOLOGÍA

El proyecto se desarrolla mediante la metodología de investigación-acción, entendida como un enfoque participativo que permite analizar una problemática organizacional, intervenir sobre ella y generar aprendizaje a partir de la reflexión sobre la práctica. Esta metodología facilita la colaboración entre el equipo investigador, la docente asesora y los

actores de la empresa vinculados al proceso de gestión de información. Su aplicación resulta pertinente porque permite diagnosticar el problema en el contexto real, diseñar una solución ajustada a las necesidades de la organización y evaluar su viabilidad técnica y operativa.

El desarrollo metodológico se estructura en tres etapas secuenciales. En la primera, se identifica y analiza el problema organizacional mediante observación directa estructurada, revisión del flujo de trabajo y levantamiento de evidencias sobre tiempos, errores e inconsistencias en la manipulación manual de archivos CSV. En la segunda etapa, se planifica y diseña la solución, definiendo las reglas de negocio, los criterios de calidad de datos y la arquitectura del prototipo ETL. Finalmente, en la tercera etapa, se establecen el alcance, los recursos, el cronograma, el presupuesto y los roles necesarios para la ejecución ordenada de la propuesta.

Para comprobar la aplicabilidad del prototipo en un escenario cercano a la operación real, se utiliza una base de datos de prueba conformada por tablas representativas del negocio, entre ellas clientes, catálogo de productos —tintas, mallas y emulsiones—, proveedores de insumos químicos, órdenes de producción y ventas, compras de materia prima, inventario físico y traslados de almacén.

Con el fin de evaluar el alcance y la efectividad operativa del pipeline ETL, la fase de experimentación no se ejecuta sobre la base de datos productiva de la empresa, sino en un entorno de prueba controlado. Para ello, se selecciona una muestra de registros históricos y se incorporan anomalías sintácticas y lógicas representativas del proceso real, tales como espacios en blanco en campos de texto, variaciones en rótulos de identificación, duplicidad de registros y diferencias en la escritura de productos o referencias. Estas alteraciones permiten reproducir

los errores más frecuentes identificados durante la fase de diagnóstico, sin comprometer la integridad de la información original.

La calibración de la base de prueba es fundamental para validar empíricamente las reglas de depuración, estandarización e integridad referencial programadas en Python. La alteración controlada de los datos no afecta las fuentes originales de la empresa; por el contrario, constituye un protocolo técnico que permite comprobar si el motor de automatización identifica, aísla y reporta registros inconsistentes bajo condiciones similares a las observadas en el entorno operativo.

- **En tablas maestras —clientes y productos—:** se incorporan registros duplicados, códigos de referencia escritos de forma heterogénea, campos vacíos y nombres de colores o sustratos sin estandarización. Por ejemplo, una misma tinta puede aparecer como “Plastisol Blanco”, “PLAST-BLANCO” o con variaciones equivalentes.
- **En tablas transaccionales —órdenes de producción y ventas—:** se incluyen cantidades negativas, precios en cero, líneas de producción inconsistentes y registros huérfanos, como órdenes asociadas a clientes inexistentes, con el propósito de evaluar la validación de integridad referencial.

Las variables definidas para el proceso de limpieza se derivan de los requerimientos reales de la planta y de criterios técnicos de calidad de datos, tales como completitud, unicidad, consistencia e integridad referencial.

A partir de estos criterios, se identifican columnas críticas como identificadores de fórmulas, valores por kilo o unidad, inventarios mínimos de seguridad, referencias de

emulsiones y códigos de mallas. Posteriormente, estas reglas operativas se traducen en funciones lógicas dentro del script de Python.

1.3.1. Identificación y Análisis del Problema

En la primera etapa se realiza un análisis de la estructura organizacional y del contexto en el que se desarrolla el proceso afectado. Este diagnóstico permite identificar causas raíz, factores asociados y puntos críticos que inciden en el desempeño del proceso de gestión de información, con el fin de asegurar que la propuesta tecnológica responda de manera pertinente a las necesidades de la pyme serigráfica.

Para comprender con precisión los factores que afectaban la eficiencia del proceso, se aplicó observación directa estructurada en el área administrativa y de producción. Esta técnica permitió registrar de manera sistemática las actividades realizadas por los colaboradores durante la descarga, revisión y preparación manual de los archivos exportados desde el ERP.

La observación se desarrolló en tiempo real, siguiendo el flujo mediante el cual los colaboradores recibían reportes de producción, fórmulas de tintas y pedidos de clientes. Este seguimiento permitió identificar las tareas repetitivas asociadas con la apertura de múltiples archivos CSV, la búsqueda de reglas de cálculo, la validación de códigos de mallas y la corrección manual de inconsistencias.

El acompañamiento permitió evidenciar tiempos muertos, fragmentación del flujo de trabajo y alta dependencia de la revisión celda por celda. Estas condiciones incrementan el riesgo de errores de digitación, retrasan la disponibilidad de información confiable y limitan la capacidad del personal para concentrarse en actividades de análisis y mejora del proceso.

1.3.2. Planificación de la Solución

En la segunda etapa se desarrolla un plan de acción que incluye la definición de objetivos, actividades, recursos y criterios de validación para mitigar el problema identificado. Esta fase contempla el diseño de una estrategia de ejecución organizada, orientada a estructurar el prototipo ETL, documentar sus reglas de negocio y asegurar una implementación coherente con las necesidades del proceso intervenido.

1.3.2.1. Definición de Alcance y Recursos

La tercera etapa se enfoca en delimitar el alcance de la intervención y los recursos necesarios para su ejecución. En esta fase se establecen responsables, cronograma y presupuesto, con el propósito de controlar el avance de la propuesta y optimizar la asignación de recursos. De esta manera, se busca garantizar que las variables críticas del proyecto estén alineadas para una implementación ordenada y sostenible.

Identificación y descripción del problema al interior de la organización

La organización objeto de estudio pertenece al sector de la serigrafía y presenta un problema recurrente en el proceso de gestión de información comercial y de inventarios. Los datos obtenidos desde el ERP se exportan en archivos CSV que, aunque son universales, llegan con estructuras inconsistentes, columnas desalineadas, vacíos de información, duplicados y diferencias en la escritura de campos similares, lo que obliga a realizar depuración manual previa al análisis. Esta problemática se ubica principalmente en el proceso de consolidación y preparación de datos del área administrativa o comercial, antes de la elaboración de reportes gerenciales.

La causa raíz del problema está en la falta de estandarización de los datos de origen, la ausencia de reglas formales de calidad, y la limitada automatización del proceso de transformación de la información. A ello se suma una debilidad en gobernanza de datos, entendida como la carencia de políticas, responsables y procedimientos claros para asegurar que los registros mantengan consistencia, integridad y trazabilidad. Cuando la empresa depende de la depuración manual, aumenta la probabilidad de errores humanos, se prolongan los tiempos de respuesta y se consume una parte importante del esfuerzo del personal en tareas repetitivas en lugar de análisis de valor agregado.

Los efectos del problema impactan directamente la eficiencia del proceso y los objetivos organizacionales. Por un lado, se retrasan los reportes de ventas e inventarios; por otro, disminuye la confiabilidad de los indicadores usados para decidir compras, producción y atención al cliente. En un sector como la serigrafía, donde los tiempos de entrega y la exactitud en los pedidos son críticos, disponer de datos limpios y homogéneos es clave para evitar reprocesos, pérdidas económicas y descoordinación interna. En consecuencia, la intervención propuesta se justifica porque busca pasar de una limpieza artesanal y reactiva a un flujo automatizado, reproducible y escalable.

1.4. DESCRIPCIÓN DE LA EMPRESA A INTERVENIR

Es una compañía colombiana, ubicada en el área metropolitana de Medellín (Itagüí), especializada en la fabricación y comercialización de **insumos y tintas para serigrafía, estampación textil y artes manuales**. Cuentan con más de 34 años de experiencia ofreciendo soporte técnico y productos a nivel nacional e internacional.

1.4.1. Misión

La misión de Tintas Serigráficas, que denominaremos Pinocho, es fabricar y comercializar productos para la estampación (serigrafía solvente, textil) y lavanderías industriales, ofreciendo respaldo técnico, entregas oportunas y precios competitivos para atender tanto el mercado nacional como internacional. Su visión es consolidarse como el principal referente técnico y proveedor de productos para acabados textiles

1.4.2. Visión

Para el año 2030, Tintas Serigráficas Pinocho. busca ser una compañía técnica y organizacionalmente sostenible. Su objetivo es mantener la innovación en las propuestas de valor para sus distribuidores, estampadores y confeccionistas, ampliando su mercado a nivel nacional e internacional.

1.4.3. Mapa de procesos

El mapa de procesos de la pyme del sector serigráfico se diseñó bajo un enfoque de gestión por procesos, siguiendo los lineamientos y buenas prácticas de gestión empresarial para pymes que describe Vargas (2021). Los procesos se clasifican en tres categorías: estratégicos, misionales (u operativos) y de soporte.

Ilustración 1: Mapa de procesos



Nota. Creación propia mediante IA por medio de ChatGPT.

1.4.4. Descripción del proceso a intervenir

El proceso objeto de intervención se ubica dentro de los procesos misionales u operativos del mapa de procesos presentado en el numeral 2.1.3, específicamente en el subproceso de gestión de información comercial y de inventarios, que antecede a la elaboración de los reportes gerenciales de ventas, consumo de insumos y existencias de tintas y mallas.

Tal como se describió en el numeral 1.3.1, la fase de identificación y análisis del problema se apoyó en la observación directa y estructurada del flujo real de trabajo en el taller. Dicha observación evidenció que el proceso se ejecuta actualmente de forma manual y fragmentada, iniciando cuando el sistema ERP genera de manera periódica archivos planos en formato CSV con los registros de clientes, catálogo de productos (tintas, mallas y emulsiones), órdenes de producción y ventas, compras de materia prima e inventario físico. Al carecer de un esquema de metadatos, estos archivos crudos deben ser descargados, abiertos y depurados

manualmente por el coordinador de producción y el asistente de facturación antes de que la información pueda integrarse en cualquier reporte o análisis.

Durante esta etapa manual, los colaboradores ejecutan tareas repetitivas de bajo valor agregado: unificación de archivos sueltos, corrección celda por celda de nombres de productos, fórmulas de tinta o códigos de malla escritos de manera heterogénea, eliminación de registros duplicados y verificación manual de la existencia de clientes, así como de que precios e inventarios no presenten valores incoherentes. Solo una vez depurada, la información puede alimentar los reportes que soportan las decisiones de compras, producción y atención al cliente.

En síntesis, el proceso a intervenir comprende tres momentos secuenciales:

- (i) extracción de datos crudos desde el ERP,
- (ii) depuración y estandarización manual de los archivos CSV, y
- (iii) consolidación de la información para la generación de reportes gerenciales.

Es precisamente en el segundo momento —la depuración manual— donde se concentra el punto crítico identificado en el diagnóstico de campo.

1.4.5. Descripción del problema que presenta el proceso

Planteamiento del problema

En la compañía, el principal desafío asociado con la gestión de información no se concentra únicamente en la infraestructura física del proceso productivo sino en la preparación administrativa de los datos generados por la operación. Hoy en día se gestionan los datos comerciales y operativos mediante exportaciones en CSV del ERP; estos archivos tienen estructuras poco uniformes, no cuentan con metadatos estandarizados, y sus registros necesitan depurarse antes de poder usarse en los reportes de gestión.

A partir de la observación directa permitió confirmar que la dependencia de procesos manuales para depurar, unificar y validar la información genera ineficiencias que afectan el desempeño de la organización. Esto se refleja en tres impactos principales: aumento de tiempos improductivos, reducción de la calidad de los datos y pérdida de capacidad analítica para apoyar decisiones oportunas.

- **Fricción operativa y tiempos improductivos:** los colaboradores pierden varias horas de alto valor que podrían destinar a análisis o propuestas de mejora, porque deben abrir, interpretar y consolidar archivos CSV antes de iniciar la producción. Este formato termina afectando el ritmo del taller y genera retrasos operativos visibles.
- **Baja calidad e integridad de los datos:** al registrar la información manualmente se introducen errores críticos: fórmulas de tinta mal anotadas, códigos de malla inconsistentes, precios en cero, registros duplicados y clientes “fantasma” que distorsionan inventarios y precios reales..
- **Pérdida de competitividad y capacidad de respuesta:** los analistas dedican su tiempo a corregir celdas y buscar fallos en hojas de cálculo en lugar de enfocarse en actividades de mayor impacto, como el control de calidad, la reducción de mermas de tinta y la mejora de la rentabilidad.

En síntesis, la compañía enfrenta un flujo de datos lento, fragmentado y vulnerable debido a la ausencia de un proceso automatizado de integración, depuración, transformación y validación de la información. Esta situación no debe entenderse como una limitación del talento humano, sino como una brecha de proceso y tecnología que dificulta el aprovechamiento oportuno de los datos. Por esta razón, la propuesta de implementar un pipeline ETL se justifica como una alternativa para reducir la carga manual, mejorar la confiabilidad de los registros y fortalecer la toma de decisiones basada en información consistente.

A continuación, se presenta el análisis DOFA de la pyme del sector serigráfico, enfocado en su capacidad de gestión de datos y analítica, con el fin de sustentar la pertinencia y viabilidad de la intervención propuesta.

Tabla 1: Matriz DOFA de la pyme del sector serigráfico

FORTALEZAS	DEBILIDADES
● Conocimiento técnico especializado del proceso productivo de la serigrafía. ● Relación comercial estable con clientes recurrentes del sector textil y publicitario. ● Disposición de la dirección para adoptar mejoras tecnológicas de bajo costo. ● Uso ya establecido de un sistema ERP para el registro de ventas e inventarios.	● Exportación de datos en formatos CSV heterogéneos y sin esquema de metadatos. ● Ausencia de procesos automatizados de depuración y consolidación de información. ● Baja gobernanza y estandarización de los datos entre las áreas comercial y de inventarios. ● Dependencia de procesos manuales que consumen tiempo del personal analítico.
OPORTUNIDADES	AMENAZAS
● Disponibilidad de librerías open source de Python (Pandas, NumPy) de bajo costo de adopción. ● Creciente demanda del mercado por reportes ágiles y decisiones basadas en datos. ● Posibilidad de escalar la solución hacia otras áreas de la organización (compras, producción). ● Acceso a comunidades y documentación técnica que facilitan la curva de aprendizaje.	● Resistencia al cambio por parte del personal acostumbrado a la depuración manual. ● Competencia de empresas del sector con mayor madurez en analítica de datos. ● Riesgo de errores en la toma de decisiones si la automatización no se valida adecuadamente. ● Rotación de personal que concentra el conocimiento técnico del proceso actual.

Nota. Elaboración propia con base en los resultados de la observación directa.

2. ESTABLECIMIENTO DE PLANES DE ACCIÓN PARA LA SOLUCIÓN DEL PROBLEMA

2.1. ESTABLECIMIENTO DE OBJETIVOS

Objetivo general:

Proponer el diseño y la evaluación de un pipeline ETL low-code desarrollado en Python para optimizar la eficiencia del proceso de preparación y mejorar la calidad de los datos en una pyme del sector serigráfico, tomando como caso de estudio la compañía Pinocho.

Objetivos específicos:

- **Diagnosticar** el estado actual del flujo de datos operativos y comerciales en la empresa Pinocho mediante observación directa estructurada, con el fin de identificar causas raíz, tiempos muertos e inconsistencias críticas derivadas de la manipulación manual de archivos CSV.
- **Desarrollar** un prototipo ETL utilizando Python y la librería Pandas, mediante la construcción de un ambiente de prueba controlado que permita validar reglas de negocio, limpieza, verificación e integridad referencial aplicadas a los datos de inventarios de tintas y mallas.
- **Evaluar** el impacto técnico, comercial y operativo del prototipo automatizado, comparando los tiempos de procesamiento y el margen de error del modelo propuesto frente al procedimiento manual tradicional mediante indicadores de calidad de datos y medición de micro tareas.

2.2. DESCRIPCIÓN DE LA SOLUCIÓN

La situación actual de la pyme Pinocho evidencia una brecha tecnológica asociada con la dependencia de tareas repetitivas de depuración manual de datos. Esta práctica prolonga los tiempos de respuesta, incrementa la probabilidad de errores y compromete la integridad de la información utilizada para controlar inventarios de tintas y mallas, así como para apoyar decisiones comerciales y operativas, afectando directamente la competitividad y la toma de decisiones oportunas.

Para mitigar esta problemática, se propone el diseño e implementación de un pipeline ETL (Extracción, Transformación, Carga) de bajo costo y código abierto, desarrollado en Python mediante la librería Pandas. La solución aborda la causa raíz del problema (la falta de estandarización y automatización) mediante un motor de reglas de negocio que actúa como un filtro inteligente entre el ERP y el analista.

La solución funciona a través de los siguientes componentes:

- **Fase de Extracción:** El script ingesta de manera automática los archivos CSV crudos generados por el ERP, sin alterar las fuentes originales de la empresa.
- **Fase de Transformación:** Se aplican funciones vectorizadas para la estandarización semántica (eliminación de espacios en blanco, unificación de mayúsculas/minúsculas), detección algorítmica de duplicados basada en claves únicas (ID de insumos) y validación de integridad referencial (cruce de tablas transaccionales contra tablas maestras de clientes y productos).
- **Fase de Carga y Auditoría:** El motor genera bases de datos purificadas listas para el análisis. Además, se implementa un sistema de gobierno de datos que, al detectar

anomalías lógicas (ej. cantidades negativas de marcos, precios en cero, registros huérfanos), no detiene el proceso, sino que aísla el registro corrupto y lo exporta a un archivo de excepciones (log_errores.csv).

La propuesta permite que una PYME con recursos limitados acceda a herramientas de analítica y preparación de datos de nivel profesional, apoyándose en soluciones de código abierto y reduciendo la carga de trabajo manual. De este modo, el rol del analista se orienta gradualmente hacia la interpretación, el seguimiento y la toma de decisiones, basadas en datos confiables, homogéneos y trazables (McKinney, 2022; Fernandes et al., 2023).

2.3. SECUENCIA DE ACTIVIDADES PARA EJECUCIÓN DE LA SOLUCIÓN

Desarrollo del Objetivo Específico 1: Diagnosticar el estado actual del flujo de datos operativos y comerciales mediante observación directa estructurada.

Actividad 1.

Comprende el diseño y validación de los instrumentos de recolección de campo, la cual se basa en la forma de generar la estructuración técnica de tres herramientas para obtener resultados cualitativos y cuantitativos, esto previo al iniciar el proceso del levantamiento de información en el taller. Se realiza una guía de observación estructurada para mapear la secuencia de clics e interacciones con el ERP; una planilla de transferencia de tiempos para lograr registrar el cronometraje de las micro-tareas en la manualidad de los datos y una bitácora para llevar el control de incidencias críticas. Estos insumos permiten asegurar que las variables medidas respondan directamente a los puntos de fricción del flujo de trabajo sin alterar la rutina del personal.

Actividad 2.

La ejecución del levantamiento en sitio y cronometraje del procesamiento manual consiste en el despliegue del trabajo de campo por un periodo continuo de dos semanas, acumulando 48 horas hombre de observación directa con el líder de producción y el analista de facturación en sus horas de mayor carga operativa. Durante este periodo se registran los tiempos exactos que los colaboradores tardan en delimitar, ajustar y depurar cada celda de los archivos CSV exportados desde el ERP, así como las novedades de digitación más recurrentes en las fórmulas de tintas y códigos de mallas.

Actividad 3.

La consolidación del mapeo y del diagnóstico del flujo operativo actual entrelaza el análisis cualitativo y cuantitativo de los datos recopilados en los documentos de bitácora y plantillas del tiempo, con el objetivo de poder identificar las causas raíz que perjudican la eficiencia del proceso. A partir de estos documentos se llega a saber que la manipulación de los archivos en formato CSV crudos lleva entre 45 y 90 minutos de trabajo al día por analista y que la falta de estandarización disminuye la fiabilidad de los datos de inventario. Dichos apuntes de interés quedan recogidos en la elaboración de un informe técnico de diagnóstico y su representación mediante un diagrama del flujo actual de datos, los cuales son considerados como línea base del diseño del prototipo.

Desarrollo del Objetivo Específico 2: Desarrollar el prototipo ETL utilizando Python y Pandas, estructurando un entorno de prueba controlado.

Actividad 4.

A partir de este momento se empieza en la construcción y parametrización del ambiente de prueba controlado en un entorno de pruebas aislado a partir de una muestra de registros históricos de la empresa, garantizando la rigurosidad experimental sin comprometer la integridad de la base de datos real. Sobre este conjunto de datos se incluyen alteraciones lógicas y sintácticas que representan los errores humanos detectados en el diagnóstico, como espacios en blanco, variaciones en los nombres de los insumos, códigos de mallas inexistentes y registros de ventas huérfanos, lo que nos permite poner a prueba la capacidad del algoritmo frente a escenarios parecidos a los que se observan en la operación real.

Actividad 5.

Esta tarea consiste en la escritura del código fuente del prototipo que permite automatizar el flujo ETL y lo divide en diferentes módulos —uno para leer, otro para limpiar, otro para transformar— y por lo tanto independientes, con funcionalidad y con posibilidad de reutilización. Mediante el uso de las estructuras de datos de Pandas como los DataFrame se programan funciones de Linux para la lectura automática de fuentes, eliminación algorítmica de duplicados que se hace con base en las claves únicas y la estandarización de los campos de texto mediante operadores lógicos y funciones de higiene de cadenas.

Actividad 6.

Con la implementación de informes de auditoría y reportes de inconsistencias, se puede programar una rutina de control de irregularidades dentro del script con el objetivo de permitir la trazabilidad y la gobernanza de los datos. La lógica se diseña para que, al identificar registros inválidos o inconsistentes —por ejemplo, precios en cero o insumos inexistentes en las tablas

maestras—, el motor no interrumpa la ejecución del flujo, sino que aisle automáticamente la fila observada y la exporte a un archivo alternativo denominado log_errores.csv.

Desarrollo del Objetivo Específico 3: Evaluar el impacto técnico y operativo del prototipo automatizado frente al procedimiento manual tradicional.

Actividad 7.

La ejecución en paralelo y sus respectivas pruebas de estrés del prototipo: consiste en ejecutar el script de Python con el conjunto de datos del ambiente de prueba controlado, que finalmente se toma para verificar que el motor automatizado sea capaz de procesar toda la base de datos, que sea capaz de detectar las anomalías incluidas en el mismo, detecte las anomalías predefinidas y sea capaz de registrarlas en el fichero del log de auditoría y, al mismo tiempo, generar una base de datos de salida limpia, consistente y disponible para posteriores análisis.

Actividad 8.

Esta actividad consiste en analizar y evaluar el resultado de las bases de datos en virtud de los criterios estándar de calidad de la información, como lo son unicidad, completitud, consistencia e integridad referencial. A partir de estos criterios se formula el informe definitivo de impacto operativo a la PYME, el cual contempla la reducción de la carga manual, la disminución de errores asumidos con el inventario y la viabilidad costo-eficiente de la solución basada en tecnologías de código abierto.

Actividad 9.

Esta actividad consiste en analizar y evaluar el resultado de las bases de datos en virtud de los criterios estándar de calidad de la información, como lo son unicidad, completitud,

consistencia e integridad referencial. A partir de estos criterios se formula el informe definitivo de impacto operativo a la PYME, el cual contempla la reducción de la carga manual, la disminución de errores asumidos con el inventario y la viabilidad costo-eficiente de la solución basada en tecnologías de código abierto.

2.4. DESCRIPCIÓN DE RECURSOS NECESARIOS PARA LA EJECUCIÓN DE LA SOLUCIÓN

De acuerdo con los planteamientos de la investigación-acción organizacional, la identificación de recursos constituye una etapa estratégica que determina la viabilidad y sostenibilidad de la intervención. En coherencia con esta perspectiva, los recursos requeridos para la ejecución del proyecto se agrupan en cinco categorías: humanos, tecnológicos, de software, financieros y de capacitación.

Recursos humanos

El conjunto está constituido por los dos estudiantes responsables del desarrollo de la parte de diseño y programación del prototipo ETL, la profesora o la docente asesora que se encargaba de realizar el acompañamiento metodológico, y el personal de la empresa Pinocho que participa como fuente primaria de información: el coordinador de producción y el asistente de facturación, observados en la fase de diagnóstico y que validarán posteriormente la usabilidad del prototipo. Esa combinación de investigadores y colaboradores de la organización va en la línea del enfoque participativo de la investigación-acción, cuya validez de la intervención se sustenta en un vínculo activo de quienes van a vivir la problemática de manera cotidiana

Recursos tecnológicos

Para poder llevar a cabo este proyecto se precisa de un (1) equipo portable que tenga la capacidad suficiente para ejecutar scripts en el lenguaje de Python; además el tener acceso a Internet que permita consultar documentación técnica y repositorios de código; también se debe tener acceso al sistema ERP de la compañía limitado a los ficheros exportados. Por lo tanto, en lo relativo a lo explicado en cuanto a recursos, se puede establecer la infraestructura digital mínima sobre la cual tiene lugar el desarrollo, las pruebas y la monitorización del prototipo.

Recursos de software

El desarrollo se basa plenamente en herramientas de código abierto: el lenguaje de programación Python y la librería Pandas para implementar el motor ETL, un entorno de desarrollo para programar el motor ETL (por ejemplo, Visual Studio Code o Jupyter Notebook), sistemas de control de versiones para asegurar la trazabilidad del código y, por último, se considera la utilización de una suite de oficina para documentar el proyecto y, si lo necesita la organización, de una hoja de cálculo para revisión manual del fichero de auditoría (log_errores.csv) que genera el sistema.

Recursos financieros

Dado que la solución se fundamenta en tecnologías de licenciamiento gratuito, el presupuesto centra muchos de los recursos materiales disponibles, en el tiempo del equipo humano y en unos costes indirectos menores, como conectividad o el almacenamiento local y elementos de papelería para la documentación del proceso. Esta característica de la adopción de baja cuantía hace posible intervenir en una pyme con escasos recursos, dado que se elimina la barrera tecnológica que puede suponer adquirir licencias de software especializado. No obstante, siguiendo este principio de reconocer

una planificación financiera adecuada, se recomienda mantener una reserva para imprevistos de ajustes al script o la eventual necesidad de soporte técnico externo (Kerzner, 2022).

Recursos de capacitación

Finalmente, se observa una jornada de formación dirigida al equipo del área administrativa y operativo de la empresa, encaminada en explicar el funcionamiento del prototipo, la forma de interpretar los reportes de auditoría y por supuesto las buenas prácticas para el proceso de la exportación de datos desde el ERP. Esta formación resulta precisa para garantizar la sostenibilidad de la solución una vez finalizado el acompañamiento del equipo investigador, dado que la falta de entrenamiento en el uso de una nueva tecnología puede comprometer el éxito de cualquier intervención.

3. DEFINICIÓN DEL ALCANCE DE LA SOLUCIÓN, PARTES INTERESADAS, CRONOGRAMA Y PRESUPUESTOS.

Se desarrolla y se logra implementar una herramienta en código abierto Python, donde se genera un pipeline para un ETL de extracción, transformación y carga que toma los exportables del ERP que arrojan en CSV, que automatiza el proceso de limpieza, estandarización, validación y consolidación de datos de ventas e inventarios. Con ello, se busca disminuir la dependencia de la depuración manual, aumentar la calidad de la información y facilitar el proceso para que se realicen análisis con datos confiables para la toma de decisiones.

3.1. ALCANCE FINAL DE LA SOLUCIÓN

Diseñar y validar una herramienta de manera local de procesamiento automatizado para, depurar y consolidar registros de ventas e inventarios exportados desde el ERP. La herramienta se orienta específicamente a pymes del sector serigráfico que necesitan en robustecer la preparación de datos sin que esto signifique que se traduzca en altos costos de licenciamiento en software y servidores.

Componentes y entregables incluidos en el alcance

Para finalizar el proyecto, se deben entregar tres (3) componentes claves, los cuales están diseñados para garantizar que la solución funcione técnicamente, se pueda tener una trazabilidad del proceso y que los usuarios la adopten con facilidad

- **Pipeline ETL automatizado en Python:** Este componente es el corazón de la solución. Se trata de un script modular, construido siguiendo buenas prácticas de programación, que automatiza la extracción, transformación, validación, carga y auditoría de los datos.
- **Módulo de extracción masiva:** Se encarga de buscar y leer de forma automática los archivos CSV crudos almacenados en una carpeta local, manteniendo el formato original sin alterar el formato y garantizando que el exporte sea seguro.
- **Módulo de homogeneización estructural:** Estandariza los formatos de fecha, transforma valores de moneda y de cantidades números enteros, unifica los nombres de las columnas para el análisis rápido y facilita la lectura.
- **Módulo de depuración sintáctica:** Limpia los textos para eliminar espacios duplicados, caracteres extraños, diferencias entre mayúsculas y minúsculas, unificando cualquier variación de escritura.
- **Módulo de consolidación y carga:** Combina las tablas de ventas e inventarios en un único archivo limpio, depurado, codificado en un formato UTF-8 y exportada en formato CSV o Excel para no alterar los caracteres especiales o tildes y así realizar posterior análisis.
- **Diccionario de datos y reglas de negocio:** Este documento contienen una matriz de metadatos de este proyecto, los tipos de datos que son permitidos, los criterios exactos para la gestión de datos faltantes como campos vacíos o inconsistencias en los registros.
- **Manual de operación y documentación técnica:** Esta guía paso a paso permite a cualquier analista o encargado del sistema, de como se puede operar la herramienta en la pyme, menciona instrucciones del proceso de instalación de Python y así mismo, de la ejecución del script automático, entender el archivo de auditoría y solucionar rápidamente los errores más en la estructura de los archivos de origen.

- **Desarrollo de interfaces gráficas de usuario:** el alcance no contempla el diseño de una aplicación con ventanas gráficas ni una interfaz web o de escritorio. El prototipo se ejecutará mediante consola de comandos, Jupyter Notebook, Google Colab o un entorno equivalente.
- **Conexión directa al ERP:** la herramienta garantiza riesgo cero para no alterar la base de datos ni tampoco genera conexión con los servidores del ERP. Solo se realiza el proceso de forma unidireccional y se ejecutará solamente los archivos previamente descargados del ERP.
- **Infraestructura a la nube:** la ejecución del script, así como el guardado de los archivos será de manera local en los computadores de la empresa. En esta etapa no se incluye la contratación ni uso de servicios en la nube como AWS, Azure o Google Cloud.
- **Modelado predictivo avanzado:** el alcance concluye con la entrega de la base de datos limpia, para lograr ejecutar un análisis descriptivo de forma inmediata. La etapa de creación de dashboard en Power BI o los de modelos analíticos avanzados predictivos quedan abiertos para una fase posterior.

3.2. IDENTIFICACIÓN DE RESPONSABLES DE LAS ACTIVIDADES PARA EJECUCIÓN

Para que la ejecución esté bien coordinada y el proyecto tenga una gobernanza clara, se determinan y definen las partes interesadas principales y responsables a través de una matriz RACI, en la que se diferencian los roles de responsable, aprobador, consultado e informado

Tabla 2: Matriz de responsabilidades para la ejecución del proyecto

Rol en el proyecto	Parte interesada	Responsabilidad clave
Patrocinador / Gerencia	Gerente general de la pyme	Aprobador (A): verifica la viabilidad financiera, autoriza los hitos del cronograma y valora el retorno operativo de la inversión.
Líder del proyecto / Investigador	Analista de datos / Desarrollador Python	Responsable (R): diseña, codifica y despliega el pipeline ETL en Python; documenta el proceso y capacita a los usuarios.
Usuario clave	Analista de operaciones / Inventarios	Consultado (C): aporta reglas de negocio, define campos críticos y participa en las pruebas funcionales del prototipo.
Soporte tecnológico	Administrador de TI / ERP de la pyme	Informado (I): facilita los accesos de descarga del ERP y recibe comunicación previa sobre cambios en la estructura de los archivos de salida.

Nota. Elaboración propia con base en la asignación de roles del proyecto y la matriz RACI.

3.3. CRONOGRAMA DE EJECUCIÓN

Secuencialidad lógica: El cronograma sigue una lógica de trabajo progresiva: lo que se hace al inicio alimenta directamente las fases siguientes de diagnóstico, desarrollo, validación y evaluación. Esta forma de organizar el trabajo permite pasar de manera ordenada de recolectar información, a construir el prototipo, y finalmente a medir los resultados.

Hitos de desarrollo:

Entre los hitos más importantes, se espera obtener:

- Mes 3: Entrega del diagnóstico operativo y del diagrama del flujo actual de datos.
- Mes 6: Entrega del script en Python con la lógica de auditoría implementada.
- Mes 9: Entrega del cuadro comparativo de eficiencia, donde se muestran los tiempos del proceso manual frente a los tiempos estimados con el proceso automatizado.

Tabla 3: Cronograma de actividades del proyecto

Actividad	Mes 1	Mes 2	Mes 3	Mes 4	Mes 5	Mes 6	Mes 7	Mes 8	Mes 9
Actividad 1: Diseño y validación de instrumentos de recolección de campo (manual de observación, planilla de tiempos y bitácora).	X	X							
Actividad 2: Ejecución del levantamiento en sitio, cronometraje del procesamiento manual (48 horas hombre de observación directa).		X	X						
Actividad 3: Consolidación del diagnóstico, análisis estadístico y mapeo del flujo operativo actual (informe técnico y línea base).			X						
Actividad 4: Construcción y parametrización del ambiente de pruebas controlado con inducción de errores humanos típicos.				X	X				
Actividad 5: Programación modular del pipeline ETL utilizando la estructura de datos DataFrame en la librería Pandas.					X	X			

Actividad 6: Implementación del sistema de auditoría interna y generación de archivos de control de mermas de datos (log_errores.csv).						X	X		
Objetivo Específico 3: Evaluación del impacto técnico y operativo									
Actividad 7: Ejecución en paralelo y pruebas de estrés del pipeline usando el dataset de prueba con anomalías controladas.							X	X	
Actividad 8: Medición de tiempos de procesamiento y contraste de eficiencia computacional (análisis comparativo del script vs. el proceso manual).								X	X
Actividad 9: Evaluación de dimensiones de calidad de datos, cuantificación de reducción de fatiga cognitiva y formulación del balance de impacto final.									X

Nota. Elaboración propia con base en las actividades definidas del proyecto.

3.4. PRESUPUESTO

La propuesta es viable financieramente porque parte de un enfoque costo-eficiente: usa herramientas de código abierto y aprovecha la infraestructura tecnológica que ya tiene la pyme. Esto reduce considerablemente la inversión en licencias de software, lo que permite destinar el presupuesto principalmente a desarrollo, capacitación, documentación y aseguramiento operativo.

Tabla 4: Presupuesto estimado del proyecto

Rubro	Detalle	Valor estimado
Recursos humanos	Investigador principal / Desarrollador en Python	\$1.500.000
Recursos humanos	Analista de operaciones / Apoyo técnico	\$0
Recursos humanos	Capacitación operativa	\$500.000
Recursos tecnológicos	Equipo de cómputo existente	\$0
Recursos tecnológicos	Dispositivo de almacenamiento externo para copias de seguridad locales	\$150.000
Recursos operativos	Insumos de oficina y papelería para el levantamiento de campo	\$100.000
Software	Microsoft 365 Business por un año	\$350.000
Software	Servicios en la nube excluidos por ejecución y almacenamiento local	\$0
Software	Python, Visual Studio Code y Pandas	\$0
Total	Presupuesto estimado	\$2.600.000

Nota. Valores en pesos colombianos. Elaboración propia con base en los recursos definidos en la tabla 2.

El presupuesto estimado para llevar a cabo la propuesta es de \$2.600.000 centrándose sobre todo en recursos humanos, capacitación, documentación y elementos de apoyo operativo. Como el proyecto se desarrolla mediante herramientas de código abierto y sobre infraestructura

local, la inversión es asumible por una pyme del sector serigráfico. Por otra parte, el cronograma de 9 meses permite avanzar de forma ordenada las fases de diagnóstico, desarrollo, validación y evaluación, minimizando riesgos de implementación y favoreciendo que la solución sea sostenible en el tiempo.

4. REFERENCIAS BIBLIOGRÁFICAS

- Fernandes, A. A. A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N. W., & Sakellariou, R. (2023). Data preparation: A technological perspective and review. *SN Computer Science*, 4(4), Article 425. <https://doi.org/10.1007/s42979-023-01828-8>
- Dhaouadi, A., Bousselmi, K., Gammoudi, M. M., Monnet, S., & Hammoudi, S. (2022). Data warehousing process modeling from classical approaches to new trends: Main features and comparisons. *Data*, 7(8), Article 113. <https://doi.org/10.3390/data7080113>
- Ghasemaghaei, M., & Turel, O. (2021). Possible negative effects of big data on decision quality in firms: The role of knowledge hiding behaviours. *Information Systems Journal*, 31(2), 268–293.
- Kerzner, H. (2022). *Project management: A systems approach to planning, scheduling, and controlling* (13th ed.). John Wiley & Sons.
- Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK® guide)* (7th ed.). Project Management Institute.
- McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (3rd ed.). O'Reilly Media.
- Mozzillo, A., Zecchini, L., Gagliardelli, L., Aslam, A., Bergamaschi, S., & Simonini, G. (2025). Evaluation of dataframe libraries for data preparation on a single machine. In *Proceedings of the 28th International Conference on Extending Database Technology* (pp. 337–349). <https://doi.org/10.48786/edbt.2025.27>
- Row Sherpa. (2024). *9 smarter data cleaning best practices for analysts in 2024*. <https://rowsherpa.com/guides/data-cleaning-best-practices>
- Segovia-García, N., & Said-Hung, E. M. (2021). Factores de satisfacción de los alumnos en e-learning en Colombia. *Revista Mexicana de Investigación Educativa*, 26(89), 595–621. <https://cutt.ly/LbUPghi>

Vargas, A. (2021). Gestión de proyectos: El marco metodológico con herramientas prácticas (4.^a ed.). Alfaomega.