



UNIVERSIDAD POLITÉCNICA SALESIANA
SEDE GUAYAQUIL
CARRERA DE COMPUTACIÓN

Diseño de un marco de trabajo para la implementación de procesos ETL

Trabajo de titulación previo a la obtención del
Título de Ingeniero en ciencias de la computación

AUTOR: Diego Andres Encalada Garcia

TUTOR: Raúl Segundo Tingo Soledispa, (MSC)

Guayaquil – Ecuador

2025

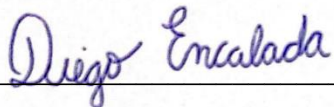
CERTIFICADO DE RESPONSABILIDAD Y AUTORÍA DEL TRABAJO DE TITULACIÓN

Yo, Diego Andres Encalada Garcia con documento de identificación N°0951782838 manifiesto que:

Soy el autor y responsable del presente trabajo; y, autorizo a que sin fines de lucro la Universidad Politécnica Salesiana pueda usar, difundir, reproducir o publicar de manera total o parcial el presente trabajo de titulación.

Guayaquil, 22 de enero del año 2025

Atentamente,



Diego Andres Encalada Garcia

0951782838

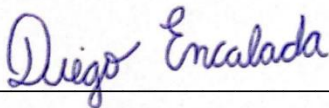
**CERTIFICADO DE CESIÓN DE DERECHOS DE AUTOR DEL TRABAJO DE
TITULACIÓN A LA UNIVERSIDAD POLITÉCNICA SALESIANA**

Yo, Diego Andres Encalada Garcia con documento de identificación No. 0951782838, expreso mi voluntad y por medio del presente documento cedo a la Universidad Politécnica Salesiana la titularidad sobre los derechos patrimoniales en virtud de que soy autor(a) del Artículo Académico: “Diseño de un marco de trabajo para la implementación de procesos ETL”, el cual ha sido desarrollado para optar por el título de: Ingeniero de Sistemas, en la Universidad Politécnica Salesiana, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente.

En concordancia con lo manifestado, suscribo este documento en el momento que hago la entrega del trabajo final en formato digital a la Biblioteca de la Universidad Politécnica Salesiana.

Guayaquil, 22 de enero del año 2025

Atentamente,

A handwritten signature in blue ink that reads "Diego Encalada". The signature is written in a cursive style and is positioned above a horizontal line.

Diego Andres Encalada Garcia

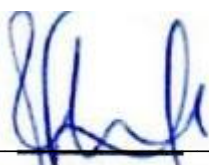
0951782838

CERTIFICADO DE DIRECCIÓN DEL TRABAJO DE TITULACIÓN

Yo, Raúl Segundo Tingo Soledispa con documento de identificación N°0914687165, docente de la Universidad Politécnica Salesiana, declaro que bajo mi tutoría fue desarrollado el trabajo de titulación: Diseño de un marco de trabajo para la implementación de procesos ETL, realizado por Diego Andres Encalada Garcia con documento de identificación N°0951782838, obteniendo como resultado final el trabajo de titulación bajo la opción Artículo Académico que cumple con todos los requisitos determinados por la Universidad Politécnica Salesiana.

Guayaquil, 22 de enero del año 2025

Atentamente,



Raúl Segundo Tingo Soledispa
0914687165

DEDICATORIA

Este trabajo es dedicado a mi familia, quienes, con su amor sin reservas, paciencia y apoyo continuo, han sido mi mayor pilar de fortaleza e inspiración. A ellos, que siempre confiaron en mi capacidad y me incentivaron a vencer cada reto, les dedico este éxito con todo mi afecto y agradecimiento. Este empeño también está dedicado a quienes me instruyeron sobre la relevancia de la constancia y la importancia del aprendizaje, valores que me orientaron en este recorrido.

Diego Andres Encalada Garcia

AGRADECIMIENTO

Estoy profundamente agradecido a mis docentes, quienes con su compromiso y sabiduría me proporcionaron las herramientas requeridas para lograr este éxito. A mis colegas y amigos, quienes con sus recomendaciones, cooperación y estímulo hicieron que este proceso resultara más enriquecedor y sencillo de realizar. Además, expreso mi agradecimiento a todas las personas que, de forma directa o indirecta, me motivaron a continuar con palabras de apoyo en las situaciones más difíciles. Este esfuerzo no habría sido factible sin el respaldo de aquellos que confiaron en mí y me motivaron a entregar lo mejor de mí.

Diego Andres Encalada Garcia

RESUMEN

Dentro de nuestro trabajo se explora el diseño de marco de trabajo, en el cual se utilizará para procesos ETL, dicho trabajo partirá de una revisión detallada de los artículos académicos que engloban en los últimos cinco años, su finalidad fue identificar desafíos en grandes volúmenes de información. Se llegó a un punto el cual busca comprender cuáles son las limitaciones dentro de los métodos tradicionales, por el cual se propone tener un marco más eficiente el cual responda a las demandas actuales. Parte del progreso se cumplirá con datos simulados almacenados en un archivo .csv, los cuales serán procesados en un entorno de prueba, dentro de dicho entorno será configurado como una base de datos, internamente se realizarán las etapas de extracción, transformación y carga, logrando de esta manera simplificar los tiempos de procesamiento y de esta manera afianzar la flexibilidad en la manipulación de los datos.

Para lograr dichos objetivos establecidos, se incorporará herramientas modernas como Python, este mismo concederá automatizar procesos de análisis y procesamiento de datos. El marco que se mencionó previamente, su creación no fue solo para dar eficacia, también fue creado para proporcionar soluciones prácticas para las empresas las cuales poseen un gran flujo de información, el cual garantiza seguridad, dentro de aquello se encarga de identificar datos clave, gracias a ello genera resultados útiles para la toma de decisiones futuras.

Palabras claves: ETL, Optimización, Datos, Marco de trabajo, Estabilidad.

ABSTRACT

The design of this ETL framework is based on a detailed review of the last five years of academic literature, using institutional databases to identify the most common challenges in managing large volumes of information. This analysis aims to understand the limitations of traditional methods and propose a more efficient framework that meets current demands.

As part of the development, simulated data stored in a .csv file will be processed in a test environment configured as a database. Through this environment, the stages of extraction, transformation, and loading will be carried out, aiming to optimize processing times and ensure flexibility in data handling.

To achieve these objectives, modern tools like Python will be integrated, allowing for the automation of data analysis and processing. This framework is not only designed to enhance efficiency but also to provide practical solutions for companies managing large volumes of information by ensuring security, identifying key data, and generating actionable results for decision-making. In summary, this project aims to offer an innovative response to current challenges in data management, providing scalability and efficiency across various business contexts.

Key words: ETL, Optimization, Data, Business Intelligent, Stability.

ÍNDICE DE CONTENIDO

1. INTRODUCCIÓN	10
2. REVISIÓN DE LITERATURA	11
2.1. FLEXIBILIDAD	11
2.2. ESCALABILIDAD	12
2.3. OPTIMIZACION EN EL DISEÑO DEL MARCO ETL.....	12
3. METODOLOGÍA	13
3.1. Métodos y técnicas de Recopilación de datos empleadas	13
3.2. Métodos y técnicas de Análisis de datos	15
4. RESULTADOS.....	16
4.1 Proceso de selección de artículos	16
4.2 Interpretación de resultados.....	18
4.3 Entorno de prueba del marco.....	22
4.3.1 Tiempos Ejecución de Desarrollo.	23
4.3.2 Visualización de datos.....	24
5. DISCUSIÓN	28
6. CONCLUSIÓN.....	29
REFERENCIAS	30

1. INTRODUCCIÓN

En la redacción de este artículo, la aplicación del marco de trabajo especifica cómo se gestionan grandes volúmenes de datos, lo que puede ocasionar varios inconvenientes al exponer los datos debido a su envergadura (Liu et al., 2020). La tecnología progresa de manera constante, y a pesar de que hay diversos enfoques para enfrentar estos retos, el marco de trabajo toma en cuenta herramientas como el Aprendizaje Machine o el Aprendizaje Deep. Sin embargo, también se analizan otras técnicas, como la minería de información, con el objetivo de mejorar el análisis y las soluciones sugeridas (Ren & Ding, 2022).

Además, es fundamental trabajar con diagramas de flujo que definan los pasos para extraer, transformar y cargar los datos (ETL). Este enfoque permite una gestión óptima y ordenada de los datos, ayudando a superar las adversidades asociadas con los grandes volúmenes de información

Adicionalmente, es crucial manejar esquemas de flujo que establezcan los procedimientos para obtener, transformar y cargar los datos (ETL). Este método posibilita una administración eficiente y organizada de los datos, contribuyendo a vencer los desafíos vinculados a los grandes volúmenes de información (Hor & Sohn, 2021). En la actualidad, otras ocupaciones comprenden la programación como un componente crucial para optimizar los tiempos de procesamiento y la automatización. En esta situación, Python se establece como una herramienta esencial para elaborar diagramas de flujo, trazar rutas de variables, definir vínculos entre campos y efectuar transformaciones en grandes volúmenes de datos (Waghlikar et al., 2022). Esto es fundamental, considerando la cantidad de datos que generan numerosas compañías, y resalta la relevancia de optimizar los periodos de tiempo.

El propósito del marco, reforzado por revisiones técnicas de numerosos artículos publicados en los últimos cinco años, es examinar las técnicas ETL tradicionales, adquirir resultados vinculados a estas y crear un marco optimizado. Este marco se aplicará en un ambiente de prueba para valorar los resultados y cotejar indicadores como la duración del proceso, la calidad del código y el efecto generado por la utilización de tecnologías actuales como Python.

Las siguientes son las preguntas de investigación que orientan el mencionado estudio para alcanzar los objetivos previamente mencionados:

1. ¿De qué forma impactan las deficiencias de los enfoques tradicionales de ETL en la capacidad de gestionar grandes volúmenes de datos de manera eficiente y escalable?
2. ¿Cómo pueden las tecnologías avanzadas, como la integración semántica y el procesamiento distribuido, contribuir a mejorar la flexibilidad y el rendimiento de los procesos ETL?

3. ¿Qué tan efectivo es el marco propuesto para adaptarse a diversos sectores empresariales, mejorando los tiempos de procesamiento y optimizando la toma de decisiones estratégicas?

Estas cuestiones se han propuesto para establecer el contexto laboral y lograr el objetivo establecido. Por lo tanto, este estudio se enfoca en reconocer los aspectos a mejorar de los procesos ETL convencionales, examinar las dificultades presentes, proponer etapas de optimización del tiempo y, finalmente, asegurar una toma de decisiones estratégicas más eficaz para las compañías, fundamentada en la optimización del tiempo al manejar datos.

2. REVISIÓN DE LITERATURA

Dentro del diseño de los procesos ETL, se examina el incremento en la cantidad de información mediante revisiones literarias que resaltan cómo la tecnología sigue progresando mientras los volúmenes de datos se incrementan de manera constante en el sector empresarial. Esta circunstancia presenta retos considerables para profesionales, alumnos y compañías que gestionan grandes volúmenes de datos, necesitando soluciones eficaces para procesar, administrar y exponer la información de forma clara y comprensible (Ramos et al., 2024).

2.1. FLEXIBILIDAD

La adaptabilidad del entorno laboral es esencial para ajustarse a las necesidades particulares de cada compañía. En los procesos de transformación de datos, las compañías a menudo alteran la estructura o el aspecto de los datos, lo cual exige que el marco tenga la capacidad de adaptación suficiente para incluir estos cambios de forma eficaz (Peng et al., 2023). Esto significa que el marco debe facilitar la gestión y cambio de los datos de acuerdo a las demandas del negocio, garantizando de esta manera su utilidad y pertinencia.

Un método esencial para alcanzar esta adaptabilidad es la aplicación de configuraciones parametrizables y variables modificables que permitan la personalización del entorno laboral. Estos atributos no solo hacen más sencillo el reaprovechamiento de elementos y variables, sino que también maximizan los recursos existentes, incrementando la calidad y el manejo de la información. Así, el marco se transforma en un instrumento versátil y eficaz para enfrentar los desafíos de la extracción, transformación y carga de datos (Akram et al., 2020).

2.2. ESCALABILIDAD

La escalabilidad es un elemento crucial en los procesos ETL, puesto que posibilita que el marco gestione cantidades en aumento de datos sin afectar su desempeño, siempre que se destinen los recursos requeridos. Conforme transcurre el tiempo y se incrementan las demandas de las empresas, un marco escalable garantiza que no existan restricciones vinculadas a la capacidad de procesamiento. Además, su vínculo directo con la adaptabilidad y la mejora lo vuelve aún más significativo. Es crucial incorporar variables parametrizables, pues posibilitan a las compañías modificar el marco de acuerdo con sus necesidades particulares, optimizando la eficiencia y la capacidad de adaptación del sistema(- & -, 2023).

Un marco escalable y flexible también ofrece beneficios económicos significativos. La capacidad gestionar flujos de información asegura un beneficio de la inversión a través de decisiones estratégicas más ágiles y precisas. No solo mejora el rendimiento de las compañías, sino que también facilita el trabajo de los especialistas en análisis de datos, que tienen la habilidad de optimizar recursos y generar resultados de alta calidad(Tran et al., 2022)(Galici et al., 2020)

En el contexto empresarial, la escalabilidad permite que las organizaciones integren múltiples fuentes de datos, como sistemas CRM (Customer Relationship Management), ERP (Enterprise Resource Planning) y redes sociales, asegurando una visión integral de sus operaciones (Berti et al., 2022)(Hadhoud & Salameh, 2020)(Deb Nath et al., 2021). Este enfoque garantiza que las empresas puedan tomar decisiones informadas, ya sea a nivel macro o detallado.

El propósito final del marco es procesar datos en tiempo real, permitiendo que áreas clave, como marketing, puedan evaluar la efectividad de sus campañas y optimizar la captación de prospectos (Edhya, 2022). Por otro lado, departamentos como crédito y cobranzas pueden analizar recaudaciones y presupuestos, lo que les facilita ajustar estrategias basadas en los ingresos generados por los clientes (Gonçalves et al., 2023). En resumen, la escalabilidad del marco es esencial para manejar grandes volúmenes de información y asegurar que las empresas puedan tomar decisiones ágiles y acertadas (Ren & Ding, 2022).

2.3. OPTIMIZACION EN EL DISEÑO DEL MARCO ETL

Es esencial perfeccionar el marco ETL para disminuir el tiempo de procesamiento y asegurar la calidad de los datos. La meta de este procedimiento no solo consiste en manejar un volumen

considerable de información, sino también en administrarla de manera eficiente y ajustarla a las demandas presentes (OCHOA-ORNELAS & GUDIÑO-OCHOA, 2022).

En cambio, la combinación de herramientas como Power BI con Python ofrece un método de análisis de datos automatizado y eficaz. Estas herramientas le facilitan el manejo de diversos formatos de archivos, tales como CSV, JSON y XML, lo que simplifica la obtención, transformación y carga de datos. Adicionalmente, en este procedimiento, bibliotecas como pandas, NumPy y pyodbc resultan cruciales, dado que agilizan la purificación y organización de los datos, posibilitando su posterior almacenamiento en sistemas como los Data Marts(Foran et al., 2024)(Eyzenakh et al., 2021)(Jensen et al., 2021).

Otro componente de optimización es la elaboración de esquemas de relación entre entidades, los cuales establecen conexiones semánticas evidentes entre la información. No solo estructura los datos, sino que también optimiza el desempeño global de las consultas y las operaciones de datos (Li et al., 2021)(Tanasescu et al., 2022). Este enfoque asegura que las empresas puedan aprovechar al máximo su información, optimizando tanto su gestión como la toma de decisiones estratégicas (Souibgui et al., 2021).

3. METODOLOGÍA

Esta investigación tiene como propósito desarrollar y validar un marco para la implementación de procesos ETL, fundamentado en una revisión sistemática de los artículos publicados en los últimos cinco años. Para garantizar su aplicabilidad, se llevará a cabo una implementación experimental utilizando datos simulados, permitiendo evaluar su desempeño en términos de escalabilidad, flexibilidad y optimización del tiempo de procesamiento. El diseño propuesto se implementará en un entorno de prueba con datos simulados, replicando escenarios empresariales para evaluar la efectividad del marco. Se medirán indicadores clave como la reducción de tiempos de procesamiento, la integración con herramientas como Python y Power BI, y la mejora en la calidad de los datos transformados.

3.1. Métodos y técnicas de Recopilación de datos empleadas

Para esta investigación, se llevará a cabo un análisis detallado de artículos relacionados con métodos ETL convencionales, empleando referencias bibliográficas de bases de datos fiables. Esta recolección facilitará el análisis de las carencias presentes en los métodos anteriores y

actuará como fundamento para sugerir un marco optimizado. Además, se han propuesto tres cuestiones de investigación fundamentales, que orientarán el análisis y ofrecerán una perspectiva completa para crear un marco eficaz, ajustado a las demandas actuales de procesamiento de datos.

1. Preguntas de investigación:

- a. ¿De qué forma impactan las deficiencias de los enfoques tradicionales de ETL en la capacidad de gestionar grandes volúmenes de datos de manera eficiente y escalable?
- b. ¿Cómo pueden las tecnologías avanzadas, como la integración semántica y el procesamiento distribuido, contribuir a mejorar la flexibilidad y el rendimiento de los procesos ETL?
- c. ¿Qué tan efectivo es el marco propuesto para adaptarse a diversos sectores empresariales, mejorando los tiempos de procesamiento y optimizando la toma de decisiones estratégicas?

2. Selección de Fuentes: Se priorizo bases de datos académicas confiables y recomendables como Scopus, Web of Science, ScienceDirect y IEEE, con el fin de tener un buen análisis y revisiones literarias adecuadas.

3. Método de consultas: El método de consultas de diseño de consultas empleados para la extracción de informacion son las siguientes:

- "ETL process optimization" AND "scalability" AND "flexibility"
- "Python" AND "ETL frameworks" AND "data transformation"
- "Power BI" AND "data visualization" AND "ETL integration"
- "real-time data processing" AND "ETL pipelines"

Como ayuda de palabras claves en el método de consultas, utilice como “ETL”, “ETL Worflow”, “real time data processing” y “Business Intelligent”

4. Criterios de Inclusión y Exclusión: En esta investigación, se definieron criterios concretos de inclusión y exclusión, con el propósito de obtener filtros apropiados de la información, eligiendo solo los estudios pertinentes.

Tabla 1.

A continuación, se presentan los criterios empleados

Criterios de inclusión	Criterios de exclusión
Investigaciones que traten asuntos vinculados a los procesos ETL y la mejora de datos.	Investigaciones que no tienen relación con ETL
Trabajos o artículos en los últimos 5 años	artículos publicados hace más de 5 años
Disponibilidad del texto completo	Disponibilidad solo del resumen / Abstracto
Estudios que incluyan herramientas como Python, Power BI u otras tecnologías relacionadas con ETL.	Artículos que no incluyan aplicaciones prácticas de herramientas tecnológicas.

Además, el análisis tiene como objetivo entender cómo optimizar el proceso ETL tradicional que muchas empresas utilizan para gestionar informes diarios, aunque a menudo de forma ineficiente. En este escenario, se utilizan recursos como Python y Power BI para ayudar a que estos procesos sean más rápidos, más versátiles y beneficiosos.

La meta es que este estudio no solo se limite a la teoría, sino que además se emplee en el ámbito práctico. Aspiramos a que los hallazgos resulten ventajosos para las empresas reales, asistiéndoles en la administración de grandes cantidades de información de manera más eficiente y en la toma de decisiones más acertadas. Finalmente, de acuerdo a dicho trabajo proporcione algo tangible y valioso, que pueda aplicarse en diferentes áreas y que realmente produzca una transformación en cómo las organizaciones administran y entienden su contenido.

3.2. Métodos y técnicas de Análisis de datos

Para buscar y examinar los datos necesarios para este análisis, llevamos a cabo una investigación rigurosa en archivos académicos de confianza. La meta es localizar artículos que expliquen cómo la analítica puede potenciar los procesos de extracción de datos, mejorar la

administración de datos y emplear herramientas como Python y Power BI en un entorno de negocios.

Se creó una búsqueda precisa, empleando combinaciones de términos clave como "optimización de procesos ETL", "Python en transformación de datos" y "Power BI y análisis de datos". Adicionalmente, se emplearon operadores como "AND" y "OR" para mejorar los resultados y garantizar que las investigaciones escogidas estuvieran en concordancia con los propósitos de este estudio.

Una vez recopilados los artículos, se revisaron a fondo para entender sus enfoques y aportes. Se prestó especial atención a identificar situaciones en las que las técnicas propuestas realmente afectan la calidad de los datos y el tiempo de procesamiento. También se analizan las dificultades encontradas durante la implementación y las soluciones adoptadas para superarlas.

Este procedimiento no solo nos brindó conocimientos útiles sobre la implementación eficiente de herramientas como Python y Power BI, sino que también nos brindó la oportunidad de adquirir conocimientos de las mejores prácticas y los métodos exitosos empleados en toda la entidad. La meta no solo consiste en perfeccionar el diseño del marco ETL, sino también en asegurar que sea práctico y valioso en la vida real, asistiendo a las compañías en la administración y uso de sus datos de manera más eficaz.

4. RESULTADOS

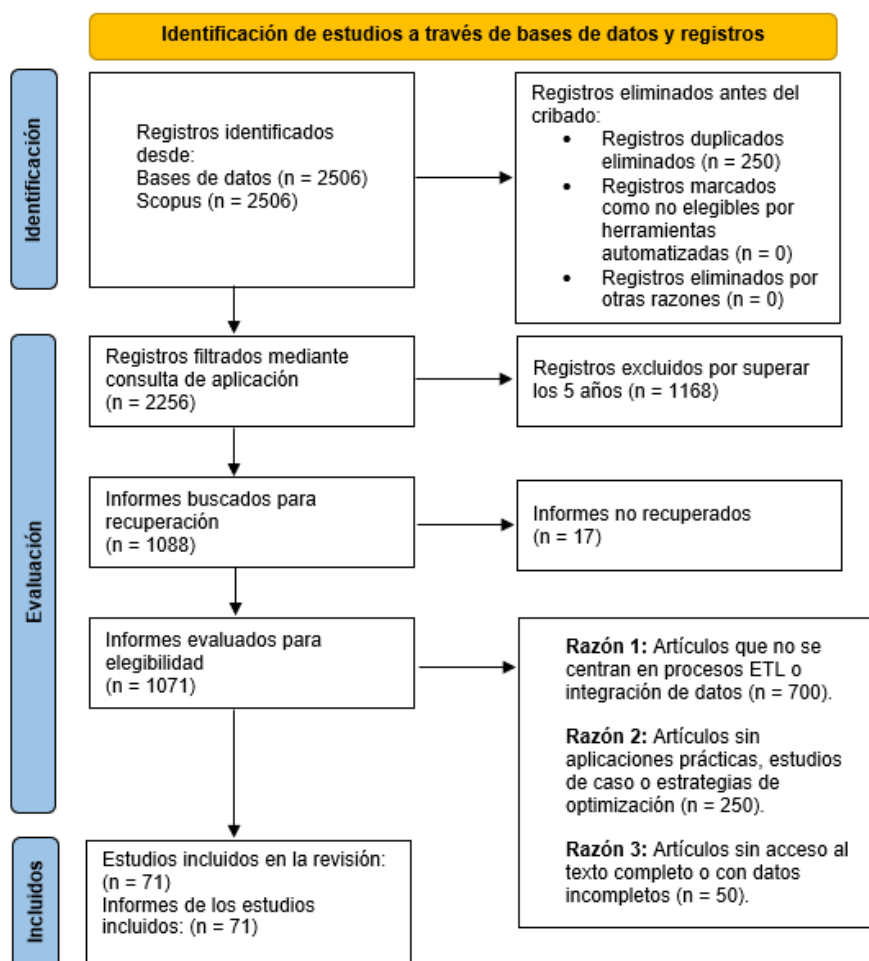
Los resultados obtenidos reflejan el impacto del marco de trabajo en la optimización de procesos ETL. Se observaron mejoras significativas en los tiempos de procesamiento, reducción de errores en la transformación de datos y una mejor integración con plataformas de análisis de datos. Estos hallazgos respaldan la viabilidad del marco propuesto para adaptarse a diversos sectores empresariales.

4.1 Proceso de selección de artículos

Utilizando el método PRISMA, se llevó a cabo un proceso sistemático para filtrar y seleccionar los artículos. A continuación, se describen los pasos principales y los resultados obtenidos

Figura 1.

Declaración Prisma



Se identificaron 2,506 artículos en la consulta inicial en Scopus, se utilizó herramienta como Mendeley ya que los datos fueron descargados y gestionados por el mismo, se identificó y elimino 250 artículos similares, tras este paso, se procedió a aplicar filtros temporales, limitando los resultados a los últimos 5 años, lo que excluyó 1,418 artículos que no cumplían con este criterio.

Al seleccionar sólo artículos relacionados con las disciplinas informáticas y los procesos ETL, el número total se redujo a 365 artículos. Finalmente, considerando sólo los artículos de acceso

abierto en inglés y español, se incluyeron en la revisión 71 artículos. Al examinar estos estudios, notamos cómo herramientas contemporáneas como Python y Power BI pueden ser empleadas para perfeccionar los procesos ETL y solucionar problemas como escalabilidad, adaptabilidad y administración eficaz de datos. Estos hallazgos ofrecen datos relevantes acerca de soluciones vanguardistas para incrementar la calidad y eficacia de los procesos comerciales, conforme a los propósitos de esta investigación.

4.2 Interpretación de resultados

El estudio realizado evidencia cómo la implementación de herramientas tecnológicas actuales, como Python y Power BI, ha transformado los procesos ETL, superando las restricciones de los métodos convencionales. Estas herramientas proporcionan soluciones para administrar grandes volúmenes de datos de forma eficaz, ajustándose a las demandas variables de las entidades. Además, han facilitado la toma de decisiones basadas en datos fiables, al mismo tiempo que potencian tanto la capacidad de adaptación como la reacción de las empresas en un ambiente de negocios cambiante.

¿De qué forma impactan las deficiencias de los enfoques tradicionales de ETL en la capacidad de gestionar grandes volúmenes de datos de manera eficiente y escalable?

Los métodos convencionales de ETL tienen restricciones significativas al manejar grandes cantidades de datos. Estas restricciones comprenden la falta de capacidad para escalar correctamente ante el incremento exponencial de los datos, además de la inflexibilidad en la incorporación de nuevas fuentes de información y la dependencia de procedimientos manuales poco eficaces. Estas limitaciones afectan directamente la rapidez y la excelencia en la toma de decisiones en los contextos corporativos.

Los textos analizados proponen soluciones orientadas a incrementar la flexibilidad, la escalabilidad y la eficiencia mediante arquitecturas sofisticadas y una optimización de recursos. Esto permite desarrollar un marco de trabajo más adaptado a las necesidades actuales de las organizaciones.

En la siguiente tabla se presentan los resultados más relevantes en relación con esta pregunta y sus ventajas

Tabla 2.

Ventajas y puntos destacados de la investigación pregunta 1

AUTOR	TEMA	VENTAJAS O PUNTOS DESTACADOS
(Tong & Tian, 2023)	Soporte de decisiones financieras	Sistemas sofisticados que optimizan la estructura y el estudio de información financiera para la toma de decisiones estratégicas.
(Deb Nath et al., 2021)	ETL en almacenes semánticos	Estructuras de ETL que potencian la capacidad de procesamiento y la capacidad de adaptación a formatos de datos novedosos.
(Mushnoori et al., 2024)	Desarrollo de pipelines de datos	Desarrollo de un proceso end-to-end para mejorar la integración y el manejo eficaz de grandes cantidades de datos.

Según la respuesta a la pregunta, los métodos convencionales de ETL tienen restricciones evidentes en la administración eficaz y escalable de grandes cantidades de datos, impactando en la flexibilidad y la rapidez de los procesos. Mediante el estudio de los artículos escogidos, se descubrieron soluciones como la puesta en marcha de pipelines end-to-end y la utilización de almacenes semánticos, que incrementan notablemente la integración, el procesamiento y la adaptabilidad de los datos, solucionando de esta manera los retos principales. Estos progresos posibilitan a las compañías maximizar sus recursos y tomar decisiones estratégicas de forma más eficaz.

¿Cómo pueden las tecnologías avanzadas, como la integración semántica y el procesamiento distribuido, contribuir a mejorar la flexibilidad y el rendimiento de los procesos ETL?

Los avances tecnológicos, como la integración semántica y el procesamiento distribuido, juegan un papel crucial en la optimización de la flexibilidad y el desempeño de los procesos ETL. La integración semántica facilita una mejor comprensión de la información al establecer vínculos relevantes entre las fuentes, mientras que el procesamiento distribuido mejora la habilidad para gestionar grandes cantidades de datos en tiempo real, utilizando arquitecturas escalables. De acuerdo con los artículos estudiados, estas tecnologías posibilitan a las empresas agilizar los procesos de cambio de datos y ajustarse a ambientes empresariales cambiantes.

En la siguiente tabla mostraremos los resultados relevantes de los puntos destacados de la pregunta y sus ventajas.

Tabla 3.

Ventajas y puntos destacados de la investigación pregunta 2

AUTOR	TEMA	VENTAJAS O PUNTOS DESTACADOS
(Kohler et al., 2023)	Integración semántica en datos de salud	Al incorporar registros de openEHR en el modelo de datos común OMOP, se potencia la interoperabilidad, lo que permite realizar análisis más eficaces.
(Ghosh et al., 2021)	Procesamiento distribuido en entornos empresariales	Estructuras de ETL que potencian la capacidad de procesamiento y la capacidad de adaptación a formatos de datos novedosos.
(Dhaouadi et al., 2022)	Modelado de Procesos de Almacén de Datos: Desde Enfoques Clásicos hasta Nuevas Tendencias - Principales Características y Comparaciones	Evaluación de atributos y métodos para un diseño más eficaz y versátil en la incorporación de datos.

Tecnologías de vanguardia como el procesamiento distribuido, la integración semántica y las corrientes actuales en el modelado de almacenes de datos están revolucionando los procesos de ETL. Estos instrumentos no solo mejoran el desempeño y la adaptabilidad, sino que también posibilitan a las entidades gestionar grandes cantidades de información de forma eficaz. Al implementar estas innovaciones, las compañías pueden ajustarse con rapidez a ambientes en constante cambio y tomar decisiones estratégicas fundamentadas en datos exactos y asequibles.

¿Qué tan efectivo es el marco propuesto para adaptarse a diversos sectores empresariales, mejorando los tiempos de procesamiento y optimizando la toma de decisiones estratégicas?

El sistema sugerido exhibe una eficiencia elevada al fusionar herramientas sofisticadas y técnicas adaptativas, lo que posibilita a las organizaciones administrar grandes volúmenes de datos y efectuar análisis estratégicos en tiempo real. La adaptabilidad del diseño permite su uso en diversas industrias, mejorando notablemente la toma de decisiones fundamentada en datos y disminuyendo el tiempo de operación. Industrias como la salud, las finanzas y la logística pueden aprovechar la puesta en marcha del sistema para alcanzar mejoras globales en los procedimientos comerciales.

Tabla 4.

Ventajas y puntos destacados de la investigación pregunta 3

AUTOR	TEMA	VENTAJAS O PUNTOS DESTACADOS
(Abd Al-Rahman et al., 2023)	ETL basado en procesos web	Partición eficaz entre aparatos y servidores, optimizando la purificación y intercambio de datos.
(Naeem et al., 2020)	Gestión de datos veloces con índices optimizados	Optimización del desempeño al gestionar grandes volúmenes de datos hacia almacenes a través de índices de memoria optimizados.
(Terol et al., 2020)	Reducción de espacio dimensional en grandes datasets	Desarrollo de técnicas híbridas de aprendizaje automático que mejoran el análisis y visualización de grandes datasets, así como la utilización de visualizaciones con Power Bi o python.

Se ha comprobado que el sistema sugerido puede ajustarse eficientemente a diversos departamentos empresariales a través de la incorporación de técnicas avanzadas como la partición eficaz de procesos, la utilización de índices optimizados para grandes cantidades de datos y la disminución del espacio dimensional de grandes volúmenes de datos a través del aprendizaje automático. Estas habilidades no solo incrementan la eficacia en la administración e integración de datos, sino que también potencian la toma de decisiones estratégicas al ofrecer resultados veloces, exactos y muy visuales. La sustitución de entradas con temas de aprendizaje automático en la tabla aporta una perspectiva innovadora al marco, volviéndolo más adaptable y apto para diversas áreas de negocio.

4.3 Entorno de prueba del marco

Para valorar el marco sugerido, se empleó un conjunto de datos simulados adquiridos de la plataforma Kaggle. Este conjunto de datos, creado por Priyam Choksi, una ingeniera de datos experta en aprendizaje automático, se denomina Credit Card Transactions Dataset. Incluye más de un millón de archivos que ofrecen datos exhaustivos acerca de operaciones con tarjetas de crédito. Dentro de la información accesible se encuentran los periodos de las transacciones, las cantidades, datos personales de los usuarios y datos vinculados con los comerciantes. Este conjunto de datos resultó esencial para replicar un ambiente de prueba realista y confirmar el rendimiento del marco en la gestión y estudio de grandes cantidades de información.

El objetivo principal del dataset es proporcionar un recurso valioso para el análisis financiero y la detección de fraudes. Entre sus posibles aplicaciones destacan:

1. **Detección de Fraude:** Permite identificar patrones anómalos en transacciones mediante modelos de aprendizaje automático.
2. **Segmentación de Clientes:** Facilita la agrupación de clientes según sus hábitos de gasto, ubicación y características demográficas.
3. **Análisis Geoespacial:** Ayuda a mapear patrones de gasto y detectar tendencias regionales.
4. **Clasificación de Transacciones:** Agrupa las transacciones en categorías como entretenimiento o alimentos para comprender mejor el comportamiento del gasto.

En esta investigación, se utilizó un subgrupo del dataset, eligiendo columnas pertinentes como la cantidad de transacciones, la localización geográfica, la categoría del comerciante, el sexo del cliente, entre otras. Este método permitió llevar a cabo un estudio exhaustivo de los datos simulados, adaptándolos a las metas del marco diseñado.

4.3.1 Tiempos Ejecución de Desarrollo.

El análisis se realizó en diversas fases, utilizando herramientas como Python y librerías especializadas, como pandas para la gestión de datos y sqlalchemy para la interacción con bases de datos relacionales. Luego se detallaron los procesos y sus respectivos tiempos.

Tabla 5.

Procesos y tiempos de ejecución

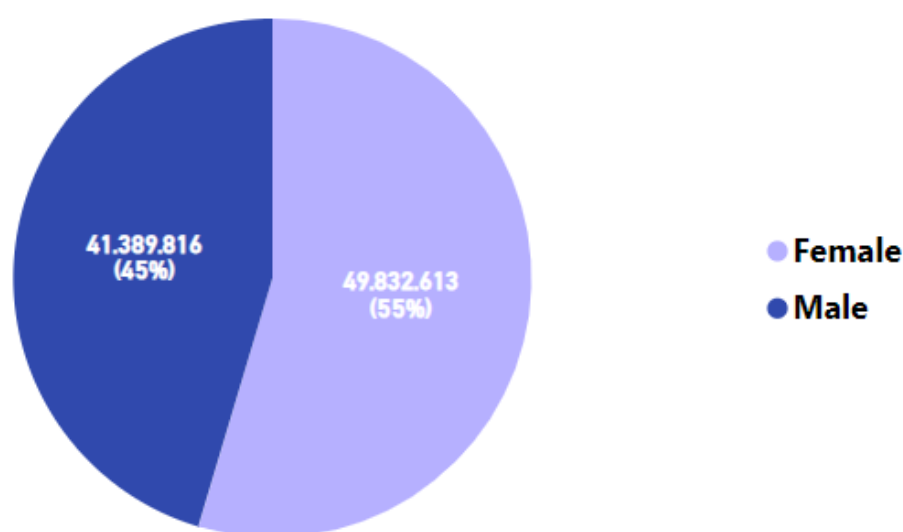
PROCESO	DESCRIPCION	TIEMPO DE EJECUCION
Carga del Dataset	Se cargó el archivo CSV a un ambiente de análisis utilizando pandas, lo que facilitó la exploración de su contenido y la identificación de las columnas más pertinentes para los propósitos del análisis.	7.1 segundos
Transformación de Datos	Elaboración de una columna de nombre íntegro que contenga los campos primero y último. Análisis de género (decrip gender) fundamentado en valores codificados. Descripción del estado de fraude (decrip is_fraud) con el objetivo de mejorar la comprensión de la información. Estas modificaciones garantizaron que la información estuviera organizada y se ajustara a las demandas del análisis.	1.2 segundos
Carga en la Base de Datos	Mediante el uso de sqlalchemy, la información modificada fue almacenada en una base de datos MySQL bajo una tabla llamada Credit_card_tesis. Se definió un modelo de tipos de datos para cada columna, tales como String, Date y DECIMAL, asegurando un correcto almacenamiento de los valores y mejorando las consultas futuras.	1 minuto y 40 segundos

4.3.2 Visualización de datos

Después de convertir y guardar los datos en una base de datos MySQL, se desarrolló un tablero de control interactivo empleando Power BI. El propósito de este tablero es examinar visualmente y de manera fácil los datos de transacciones con tarjeta de crédito, resaltando patrones, categorías y distribuciones pertinentes. A continuación, presentamos la primera ilustración en forma de gráfico.

Figura 2.

Diagrama de pastel de cuantos gastos hicieron por genero



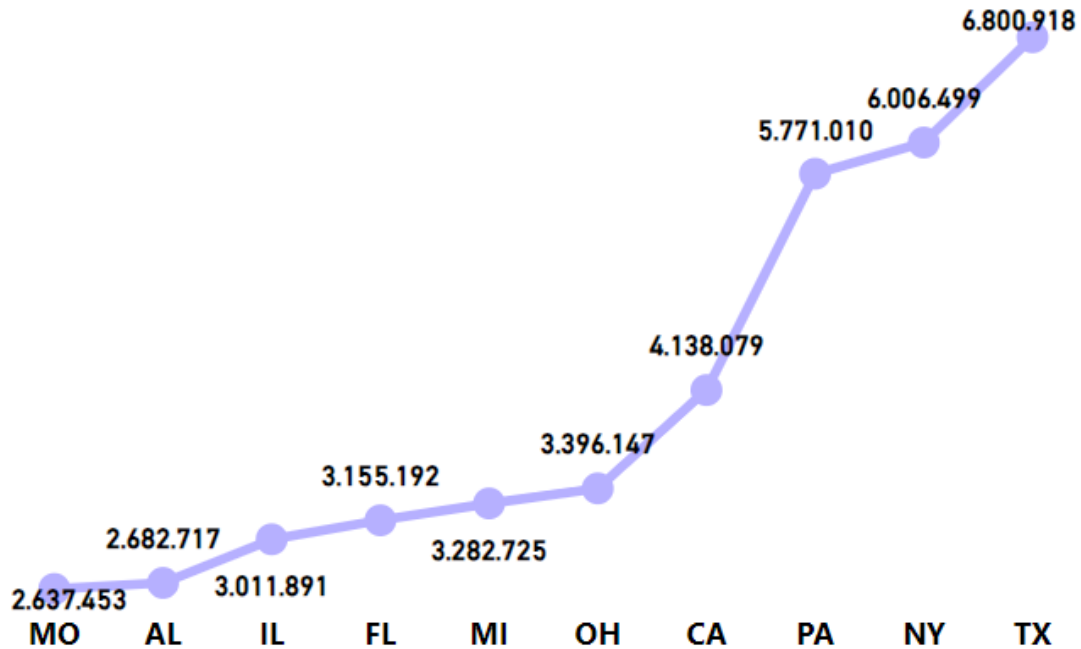
El diagrama circular muestra la repartición del total de dinero (amt) producido por transacciones con tarjetas de crédito, clasificado por sexo. Los descubrimientos indican que:

- Las mujeres lideran en términos de gasto total, con un 55% del monto acumulado (aproximadamente 49.8 millones de dólares).
- Los hombres, aunque representaron un porcentaje ligeramente menor, contribuyeron con el 45% del monto acumulado (alrededor de 41.3 millones de dólares).

Este estudio revela que las mujeres desempeñan un papel preponderante en las transacciones, lo que podría atribuirse a distintos patrones de consumo o a requerimientos específicos. Estas son estas

Figura 3.

Diagrama de línea de cuantos gastos hicieron por estados



La representación gráfica mostrada es un esquema de línea que muestra las sumas totales producidas por operaciones con tarjetas de crédito, repartidas por los estados de Estados Unidos. El eje horizontal (X) presenta las siglas de los estados, mientras que el eje vertical (Y) representa las cantidades acumuladas de dichas transacciones.

Este estudio nos facilita determinar qué estados registran un volumen más elevado de transacciones. Por ejemplo, Texas (TX) sobresale por tener un número notablemente mayor en comparación con otros estados como Missouri (MO) o Alabama (AL). Este tipo de visualización es esencial para analizar tendencias a nivel regional y identificar zonas con una mayor actividad en el ámbito económico.

Figura 4.

Diagrama de barras de cuantos gastos hicieron por categoría



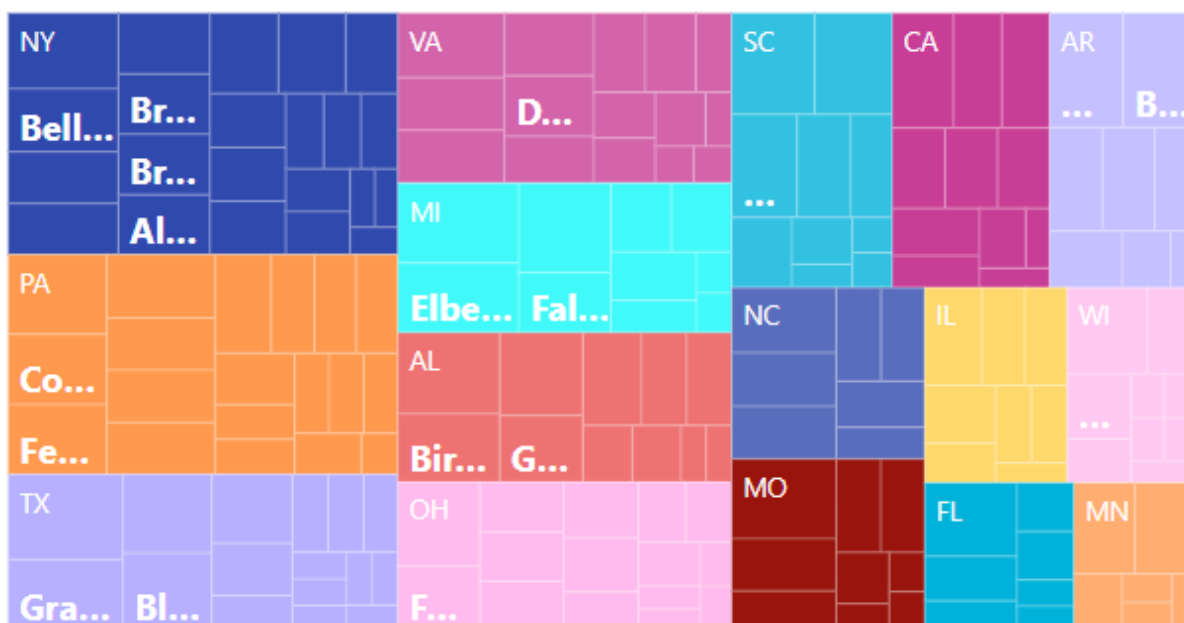
El gráfico de barras ilustra de forma nítida la distribución total de las transacciones efectuadas con tarjetas de crédito, organizadas por dichas categorías de desembolso. Cada barra muestra el acumulado total por categoría, proporcionando una perspectiva global de las zonas donde los usuarios acumulan la mayor parte de sus costos.

La categoría "grocery_pos", que simboliza las compras en supermercados de manera presencial, sobresale como la más importante, con un total de 14,460,822.38, lo que señala que una considerable cantidad de uso de tarjetas de crédito se enfoca en este tipo de operaciones. A continuación, se presentan los términos "shopping_pos" (compras en tiendas físicas, 9,307,993.61) y "shopping_net" (compras en plataformas digitales, 8,625,149.68), lo que demuestra que el consumo en establecimientos físicos y en plataformas digitales juega un papel importante en las costumbres de los clientes.

Otras categorías relevantes incluyen "gas_transport" (gastos en transporte y combustible) y "home" (inversiones o compras relacionadas con el hogar). En el otro extremo, vemos categorías como "personal_care" (cuidado personal) y "grocery_net" (compras de supermercado en línea), que muestran montos más reducidos en comparación con las categorías principales.

Figura 5.

Diagrama treemap de cuantos estados en ciudades



A continuación, se muestra un diagrama que ilustra la repartición de las transacciones por estado y ciudad en los Estados Unidos. Cada estado se simboliza por un rectángulo central, mientras que las ciudades que conforman ese estado están separadas en rectángulos de menor tamaño. El tamaño de estos rectángulos se relaciona con el volumen de transacciones registradas en cada sitio, lo que facilita la identificación rápida de las zonas con más o menos actividad.

Por ejemplo, estados como Nueva York (NY), Pensilvania (PA) y Texas (TX) sobresalen por tener territorios más amplios, lo que señala que acumulan un mayor número de operaciones o un gasto más alto. En estos estados, ciertas ciudades sobresalen por su importancia económica, lo que se manifiesta en la magnitud de los rectángulos que las simbolizan. Por otro lado, estados como Arkansas (AR) o Wisconsin (WI) poseen rectángulos de menor tamaño, lo que indica una actividad reducida en cuanto a operaciones con tarjetas crediticias.

Esta representación es útil para comprender patrones de consumo y gasto en diferentes regiones, facilitando el diseño de estrategias comerciales, el análisis de mercados locales y la identificación de áreas prioritarias para futuras acciones.

5. DISCUSIÓN

Esta investigación empleó el método PRISMA para elegir y examinar investigaciones pertinentes acerca de procesos ETL tanto tradicionales como contemporáneos. En un principio, la búsqueda detectó un gran número de artículos y, tras aplicar los criterios establecidos por PRISMA, se escogieron aquellos que se adecuaban más al marco establecido. Este análisis proporciona información sobre los desafíos y las oportunidades de procesar grandes cantidades de datos y utilizar tecnologías avanzadas.

A continuación, se discuten las preguntas planteadas en función de los hallazgos:

1. ¿De qué forma impactan las deficiencias de los enfoques tradicionales de ETL en la capacidad de gestionar grandes volúmenes de datos de manera eficiente y escalable?

El estudio revela que los procesos ETL convencionales se encuentran con significativas restricciones en contextos con gran uso de datos, debido a su diseño sencillo y rigidez. Estos métodos frecuentemente enfrentan problemas para ampliar y administrar la variedad de las fuentes de datos existentes, lo que incrementa los periodos de procesamiento y disminuye la eficiencia global. La bibliografía examinada subraya la importancia de emplear tecnologías actuales con mayor adaptabilidad y escalabilidad para vencer estos desafíos.

2. ¿Cómo pueden las tecnologías avanzadas, como la integración semántica y el procesamiento distribuido, contribuir a mejorar la flexibilidad y el rendimiento de los procesos ETL?

Los textos escogidos de PRISMA subrayan que la integración semántica y la arquitectura distribuida son fundamentales para incrementar la flexibilidad y el desempeño de los procesos ETL. Estas tecnologías facilitan la administración instantánea de datos diversos y distribuidos, disminuyendo la complejidad operacional y potenciando la habilidad para adaptarse a las exigencias variables del mercado. En nuestro sistema, estos procedimientos se modelan a través del proceso de conversión de datos y su respectiva representación visual.

3. ¿Qué tan efectivo es el marco propuesto para adaptarse a diversos sectores empresariales, mejorando los tiempos de procesamiento y optimizando la toma de decisiones estratégicas?

El marco propuesto se puso a prueba con un conjunto de datos analógico, lo que nos facilitó valorar su rendimiento en términos de tiempo de optimización y producción de información visual mediante paneles interactivos de Power BI. Los hallazgos indican que el sistema disminuye eficientemente el tiempo de carga y procesamiento, ofrece una mejor disposición del material y simplifica su interpretación. Además, estas mejoras facilitan la toma de decisiones más ágiles y fundamentadas en datos en todas las unidades de

6. CONCLUSIÓN

Este estudio permitió el diseño y validación de un marco de trabajo para mejorar los procesos ETL en entornos con grandes volúmenes de datos, integrando tecnologías actuales y metodologías innovadoras. La metodología PRISMA proporcionó una base teórica sólida, permitiendo contextualizar los desafíos de los enfoques tradicionales y destacando el potencial de tecnologías como la integración semántica y el procesamiento distribuido.

A través del ambiente de prueba, se evaluó la efectividad del marco con datos simulados, abordando la conversión y transformación de datos, así como su representación en tableros interactivos en Power BI. Los resultados evidenciaron mejoras en la organización de la información, optimización de tiempos de procesamiento y una mayor comprensión de datos estratégicos. Aunque Power BI fue la herramienta principal utilizada en este estudio, es importante destacar que el marco es adaptable a otras soluciones de visualización y análisis de datos, como Tableau o Google Data Studio, lo que amplía su aplicabilidad.

En conclusión, el marco propuesto no solo responde a las preguntas de investigación planteadas, sino que también demuestra su viabilidad para mejorar la flexibilidad, el rendimiento y la escalabilidad de los procesos ETL. Además, su diseño modular permite adaptarlo a diferentes sectores empresariales, facilitando la toma de decisiones basada en datos. Para futuras investigaciones, se recomienda probar el marco con datos reales en entornos empresariales, lo que permitiría validar aún más su aplicabilidad y medir su impacto en la optimización de procesos organizacionales.

REFERENCIAS

- , A. M. K., & -, S. (2023). The Role of Data Warehousing in Revolutionizing Banking and User Experience. *International Journal For Multidisciplinary Research*, 5(6). <https://doi.org/10.36948/ijfmr.2023.v05i06.8890>
- Abd Al-Rahman, S. Q., Hasan, E. H., & Sagheer, A. M. (2023). Design and implementation of the web (extract, transform, load) process in data warehouse application. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 12(2), 765. <https://doi.org/10.11591/ijai.v12.i2.pp765-775>
- Akram, M., Dar, J. M., & Naz, S. (2020). Pythagorean Dombi fuzzy graphs. *Complex and Intelligent Systems*, 6(1). <https://doi.org/10.1007/s40747-019-0109-0>
- Berti, A., Park, G., Rafiei, M., & van der Aalst, W. M. P. (2022). An Event Data Extraction Approach from SAP ERP for Process Mining. *Lecture Notes in Business Information Processing*, 433 LNBIP. https://doi.org/10.1007/978-3-030-98581-3_19
- Deb Nath, R. P., Romero, O., Pedersen, T. B., & Hose, K. (2021). High-level ETL for semantic data warehouses. *Semantic Web*, 13(1), 85–132. <https://doi.org/10.3233/SW-210429>
- 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). <https://doi.org/10.3390/data7080113>
- Edhya, B. F. P. (2022). BUSINESS INTELLIGENCE DATA MARKETING MENGGUNAKAN METODE KIMBALL DAN ETL DENGAN POWER BI. *Kurawal - Jurnal Teknologi, Informasi Dan Industri*, 5(2). <https://doi.org/10.33479/kurawal.v5i2.642>
- Eyzenakh, D., Rameykov, A., & Nikiforov, I. (2021). High Performance Distributed Web-Scraper. *Proceedings of the Institute for System Programming of the RAS*, 33(3). [https://doi.org/10.15514/ispras-2021-33\(3\)-7](https://doi.org/10.15514/ispras-2021-33(3)-7)
- Foran, D. J., Chen, W., Kurc, T., Gupta, R., Kaczmarzyk, J. R., Torre-Healy, L. A., Bremer, E., Ajjarapu, S., Do, N., Harris, G., Stroup, A., Durbin, E., & Saltz, J. H. (2024). An Intelligent Search & Retrieval System (IRIS) and Clinical and Research Repository for Decision Support Based on Machine Learning and Joint Kernel-based Supervised Hashing. *Cancer Informatics*, 23. <https://doi.org/10.1177/11769351231223806>
- Galici, R., Ordile, L., Marchesi, M., Pinna, A., & Tonelli, R. (2020). Applying the ETL Process to Blockchain Data. Prospect and Findings. *Information*, 11(4), 204. <https://doi.org/10.3390/info11040204>
- Ghosh, P., Sadhu, D., & Sen, S. (2021). A real-time business analysis framework using virtual data warehouse. *International Arab Journal of Information Technology*, 18(4). <https://doi.org/10.34028/18/4/11>
- Gonçalves, C. T., Gonçalves, M. J. A., & Campante, M. I. (2023). Developing Integrated Performance Dashboards Visualisations Using Power BI as a Platform. *Information (Switzerland)*, 14(11). <https://doi.org/10.3390/info14110614>
- Hadhoud, R., & Salameh, W. A. (2020). How business intelligence can help you to better understand your customers. *International Journal of Business Intelligence Research*, 11(1). <https://doi.org/10.4018/IJBIR.2020010104>

- Hor, A. H., & Sohn, G. (2021). DESIGN and EVALUATION of A BIM-GIS INTEGRATED INFORMATION MODEL USING RDF GRAPH DATABASE. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 8(4/W2-2021). <https://doi.org/10.5194/isprs-annals-VIII-4-W2-2021-175-2021>
- Jensen, S. K., Thomsen, C., Pedersen, T. B., & Andersen, O. (2021). pygrametl: A Powerful Programming Framework for Easy Creation and Testing of ETL Flows. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 12670 LNCS. https://doi.org/10.1007/978-3-662-63519-3_3
- Kohler, S., Boscá, D., Kärcher, F., Haarbrandt, B., Prinz, M., Marschollek, M., & Eils, R. (2023). Eos and OMOCL: Towards a seamless integration of openEHR records into the OMOP Common Data Model. *Journal of Biomedical Informatics*, 144. <https://doi.org/10.1016/j.jbi.2023.104437>
- Li, J., Xian, G., Zhao, R., Huang, Y., Kou, Y., Luo, T., & Sun, T. (2021). RDFAdaptor: Efficient ETL Plugins for RDF Data Process. *Journal of Data and Information Science*, 6(3), 123–145. <https://doi.org/10.2478/jdis-2021-0020>
- Liu, J., Cai, D., Zhu, D., & Huang, S. (2020). A regional industry intelligence business platform based on adaptive clustering. *International Journal of Circuits, Systems and Signal Processing*, 14. <https://doi.org/10.46300/9106.2020.14.84>
- Mushnoori, S. C., Zang, E., Banerjee, A., Hooten, M., Merzky, A., Turilli, M., Jha, S., & Dutt, M. (2024). Pipelines for automating compliance-based elimination and extension (PACE2): a systematic framework for high-throughput biomolecular materials simulation workflows. *JPhys Materials*, 7(1). <https://doi.org/10.1088/2515-7639/ad08d0>
- Naeem, M. A., Mirza, F., Khan, H. U., Sundaram, D., Jamil, N., & Weber, G. (2020). Big data velocity management-from stream to warehouse via high performance memory optimized index join. *IEEE Access*, 8. <https://doi.org/10.1109/ACCESS.2020.3033464>
- OCHOA-ORNELAS, R., & GUDIÑO-OCHOA, A. (2022). Implementation of models to predict avocado exports. *Journal Mathematical and Quantitative Methods*. <https://doi.org/10.35429/jmqm.2022.11.6.6.17>
- Peng, Y., Henke, E., Reinecke, I., Zoch, M., Sedlmayr, M., & Bathelt, F. (2023). An ETL-process design for data harmonization to participate in international research with German real-world data based on FHIR and OMOP CDM. *International Journal of Medical Informatics*, 169, 104925. <https://doi.org/10.1016/j.ijmedinf.2022.104925>
- Ramos, R. G. G., De Oliveira, A. A., De Souza, A. F., Gatti, D. C., Lemos, E. P. C., Da Silva, H. M., & Dos Santos, E. Q. (2024). Data analysis as a digital competence for information technology professionals: implementation and use of business intelligence. *OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA*, 22(2). <https://doi.org/10.55905/oelv22n2-051>
- Ren, X., & Ding, X. (2022). Motivation Analysis of Technological Startups Business Models Based on Intelligent Data Mining and Analysis. In *Journal of Electrical and Computer Engineering* (Vol. 2022). <https://doi.org/10.1155/2022/9316534>
- Souibgui, M., Atigui, F., Yahia, S. Ben, & Cherfi, S. S. S. (2021). IRIS-DS: A new approach for identifiers and references discovery in document stores. *Proceedings of the Annual Hawaii International Conference on System Sciences, 2020-January*. <https://doi.org/10.24251/hicss.2021.118>

- Tanasescu, L. G., Vines, A., Bologa, A. R., & Vaida, C. A. (2022). Big Data ETL Process and Its Impact on Text Mining Analysis for Employees' Reviews. *Applied Sciences*, 12(15), 7509. <https://doi.org/10.3390/app12157509>
- Terol, R. M., Reina, A. R., Ziaei, S., & Gil, D. (2020). A Machine Learning Approach to Reduce Dimensional Space in Large Datasets. *IEEE Access*, 8. <https://doi.org/10.1109/ACCESS.2020.3012836>
- Tong, D., & Tian, G. (2023). Intelligent financial decision support system based on big data. *Journal of Intelligent Systems*, 32(1). <https://doi.org/10.1515/jisys-2022-0320>
- Tran, B. H., Aussenac-Gilles, N., Comparot, C., & Trojahn, C. (2022). Semantic Integration of Raster Data for Earth Observation on Territorial Units. *ISPRS International Journal of Geo-Information*, 11(2). <https://doi.org/10.3390/ijgi11020149>
- Waghlikar, K. B., Ainsworth, L., Zelle, D., Chaney, K., Mendis, M., Klann, J., Blood, A. J., Miller, A., Chulyadyo, R., Oates, M., Gordon, W. J., Aronson, S. J., Scirica, B. M., & Murphy, S. N. (2022). I2b2-etl: Python application for importing electronic health data into the informatics for integrating biology and the bedside platform. *Bioinformatics*, 38(20). <https://doi.org/10.1093/bioinformatics/btac595>