TEC | Tecnológico de Costa Rica

Escuela de Administración de Tecnologías de Información

Propuesta de mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros para una empresa del sector financiero en Costa Rica durante el I Semestre del 2025

Trabajo Final de Graduación para optar al grado de Licenciatura en Administración de Tecnología de Información

Modalidad Proyecto de Graduación

Elaborado por: Kemuel Abiel Chavarría Moreno

Prof. Tutor: Lic. Michael Lizandro Sánchez Soto

Cartago, Costa Rica

I Semestre

Junio, 2025

Hoja de Aprobación

Posterior a la aprobación de su defensa Usted deberá recopilar las firmas y adjuntar la hoja de aprobación en esta sección del documento final.

INSTITUTO TECNOLÓGICO DE COSTA RICA

ESCUELA DE ADMINISTRACIÓN DE TECNOLOGÍAS DE INFORMACIÓN

GRADO ACADÉMICO: LICENCIATURA

Los miembros del Tribunal Examinador de la Escuela de Administración de Tecnologías de Información, recomendamos que el siguiente informe del Trabajo Final de Graduación del estudiante Kemuel Abiel Chavarría Moreno sea aceptado como requisito parcial para obtener el grado académico de Licenciatura de Tecnología de Información.

MICHAEL LIZANDRO SANCHEZ SOTO (FIRMA)

Digitally signed by MICHAEL LIZANDRO SANCHEZ SOTO (FIRMA)
Date: 2025.07.03 08:03:49 -06'00'

Lic. Michael Lizandro Sánchez Soto

Mgtr. Amed Espinoza Calderón

TEC | Tecnológico de Costa Rica

Firmado digitalmente por MARIA JOSE ARTAVIA JIMENEZ (FIRMA)
Fecha: 2025.07.03 11:01:05 -06'00'

Ing. MAE. María José Artavia Jiménez

YARIMA TATIANA SANDOVAL SANCHEZ (FIRMA)

Firmado digitalmente por YARIMA TATIANA SANDOVAL SANCHEZ (FIRMA)
Fecha: 2025.07.03 11:53:35 -06'00'

Mgtr. Yarima Sandoval Sánchez

Dedicatoria

A Dios, por ser mi refugio constante, mi guía silenciosa y mi fuerza en los momentos más inciertos. Por nunca soltarme la mano, incluso cuando me sentía más débil. Todo esto fue posible gracias a Tu presencia.

A mis padres, pilares inquebrantables de amor y entrega. Gracias por creer en mí siempre, por brindarme todo lo necesario para alcanzar esta meta y por estar siempre, en cada paso, con paciencia y amor. Los amo profundamente. Anhelo, con todo mi corazón, poder devolverles cada esfuerzo con gratitud y orgullo.

A mi hermano, por ser mi compañero leal en este camino. Gracias por compartir conmigo estos cinco años de carrera, por tu apoyo constante, por estar presente en los momentos importantes y también en los cotidianos. Tu compañía ha sido un regalo invaluable.

A mis amigos y amigas de carrera, por las incontables madrugadas entre trabajos, risas y cansancio. Por estar ahí sin dudar cuando se necesitaba una mano, una palabra o simplemente compañía. Los llevo conmigo con sincero aprecio y cariño.

Resumen

Chavarría Moreno, Kemuel Abiel (2025). *Propuesta de mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros para una empresa del sector financiero en Costa Rica durante el I Semestre del 2025*. [Trabajo Final de Graduación para optar por el grado académico de Licenciatura]. Escuela de Administración de Tecnologías de Información. Instituto Tecnológico de Costa Rica.

Este Trabajo Final de Graduación presenta una propuesta orientada a mejorar el proceso de carga de datos en una plataforma de Gestión de Datos Maestros, dentro de una empresa del sector financiero en Costa Rica. El proceso actual, altamente fragmentado, presenta limitaciones en eficiencia operativa, trazabilidad técnica y consistencia de los datos. Como respuesta, se diseñó una solución automatizada utilizando servicios de Amazon Web Services (AWS), con el objetivo de reducir la intervención humana, estandarizar la validación de registros y fortalecer el control del proceso.

El proyecto se desarrolló en cuatro fases: análisis del proceso actual, diseño de la solución automatizada, desarrollo del prototipo de solución y evaluación funcional del prototipo. La propuesta fue validada mediante pruebas controladas que permitieron comparar su desempeño frente al proceso actual. Los resultados evidenciaron una reducción significativa en el esfuerzo requerido, además de mejoras en la precisión de los datos y la estabilidad del flujo automatizado.

La automatización desarrollada representa una alternativa viable y eficaz para mejorar la gestión de datos maestros en el entorno organizacional. Además, se identificaron posibilidades de expansión hacia otras áreas críticas del negocio, así como futuras integraciones con herramientas especializadas en calidad de datos.

Palabras clave: Automatización, Datos maestros, Gestión de datos, AWS, Prototipo.

Abstract

Chavarría Moreno, Kemuel Abiel (2025). *Improvement and Automation Proposal for the Data Load Process in a Master Data Management Platform for a Financial Sector Company in Costa Rica during the First Semester of 2025*. [Final Graduation Project submitted in partial fulfillment of the requirements for the Bachelor's Degree] School of Information Technology Administration. Costa Rica Institute of Technology.

This Final Graduation Project presents a proposal aimed at improving the data load process in a Master Data Management platform within a financial sector company in Costa Rica. The current process, highly fragmented, shows limitations in operational efficiency, technical traceability, and data consistency. In response, an automated solution was designed using Amazon Web Services (AWS), with the objective of reducing human intervention, standardizing record validation, and strengthening process control.

The project was developed in four phases: analysis of the current process, design of the automated solution, prototype development, and functional evaluation. The proposal was validated through controlled testing that allowed a performance comparison with the current process. The results revealed a significant reduction in required effort, along with improvements in data accuracy and the stability of the automated flow.

The implemented automation stands as a viable and effective alternative to improve master data management within the organizational environment. Additionally, opportunities were identified to expand the solution to other critical business areas and to integrate specialized data quality tools in future implementations.

Keywords: Automation, Master data, Data management, AWS, Prototype.

Tabla de Contenidos

Índice de Figuras

Índice de Tablas

1 Introducción

1.1 Descripción General

En el sector financiero, donde la gestión de datos juega un papel fundamental en la toma de decisiones estratégicas, la eficiencia en los procesos de carga y validación de datos es un factor crítico. Sin embargo, en la empresa de estudio, el equipo *Data Operations* enfrenta importantes desafíos en la gestión de datos maestros, ya que el proceso actual de carga de datos es altamente manual, fragmentado y propenso a errores, lo que impacta directamente la calidad y disponibilidad de los datos.

Actualmente, la empresa depende de herramientas como Excel, SharePoint y PostgreSQL para la validación, transformación y carga de datos provenientes de múltiples fuentes. Estos procedimientos carecen de estandarización y automatización, lo que genera retrasos operativos, errores humanos y una sobrecarga de trabajo para el equipo de *Data Operations*.

A lo largo de la industria, diversas organizaciones han implementado soluciones de automatización para la carga y validación de datos, mejorando significativamente la eficiencia operativa y reduciendo la dependencia de procesos manuales. En este contexto, el presente proyecto propone el diseño y desarrollo de un prototipo de solución automatizada, orientado a mejorar la carga de datos en la plataforma de Gestión de Datos Maestros de la empresa. Se busca reducir la intervención humana en tareas repetitivas, disminuir los errores en la manipulación de datos y mejorar la disponibilidad de información clave para las áreas de negocio.

El presente Trabajo Final de Graduación se encuentra estructurado en seis capítulos. El **Capítulo I** introduce el proyecto a través de una descripción general, antecedentes, planteamiento del problema, objetivos, alcance, entregables, supuestos, limitaciones y exclusiones. El **Capítulo II** desarrolla el marco conceptual, donde se abordan los fundamentos teóricos y referencias clave que sustentan la investigación. El **Capítulo III** expone el marco metodológico aplicado, detallando las variables, técnicas, instrumentos y fases del proyecto. El **Capítulo IV** presenta el análisis de resultados, abordando el diagnóstico del proceso actual, así cómo el diseño del nuevo proceso automatizado, la formulación de requerimientos funcionales y la validación de hallazgos. El **Capítulo V** detalla la propuesta de solución, estructurada en dos fases: el desarrollo funcional del prototipo, incluyendo la evidencia técnica de su implementación, las pruebas realizadas y la evaluación para determinar la efectividad del prototipo. Finalmente, el **Capítulo VI** recoge las conclusiones derivadas del estudio, seguido por el **Capítulo VII**, que expone recomendaciones orientadas a la implementación y mejora continua de la solución planteada. En los apartados finales se incorporan las **referencias bibliográficas**, los **apéndices**, los **anexos** y un **glosario** técnico que respalda la comprensión integral del documento.

1.2 Antecedentes

1.2.1 Descripción de la organización

La organización en la que se desarrollará este proyecto es una empresa del sector de servicios financieros, fundada en el año 1900 y con sede en Nueva York, EE. UU. Desde su

fundación, esta empresa se ha dedicado a proporcionar datos, inteligencia y herramientas analíticas que permiten a los líderes empresariales y financieros tomar decisiones informadas con confianza. La organización combina el conocimiento de analistas altamente capacitados, acceso a datos de gran valor, herramientas avanzadas respaldadas por tecnologías innovadoras y una visión del futuro fundamentada en más de 115 años de experiencia.

Esta empresa ayuda a sus clientes a acelerar la creación de valor en un contexto de riesgos crecientes, abordando cuatro áreas clave:

1. *Ratings* (Calificaciones): La organización busca ser la agencia de calificación global preferida por emisores de deuda e inversionistas, proporcionando evaluaciones fiables y reconocidas a nivel mundial.
2. *Research & Insights* (Investigación y Perspectivas): A través de su negocio de investigación, la empresa ofrece análisis de renta fija de alta calidad, siendo uno de los líderes en este campo.
3. *Data & Information* (Datos e Información): La organización posee una de las bases de datos más grandes del mundo sobre compañías y crédito, con más de 450 millones de registros y en constante crecimiento, lo cual sustenta su negocio de datos.
4. *Decision Solutions* (Soluciones de Decisión): La compañía también cuenta con tres plataformas en la nube (SaaS) que apoyan flujos de trabajo críticos en sectores como la Banca, Seguros y Conozca a su Cliente (KYC), permitiendo a estas industrias operar con mayor eficiencia y precisión en sus procesos.

1.2.2 Misión

De acuerdo con la página oficial de la empresa, se tiene que su misión corresponde a:

"*Our mission is to be the leading source of relevant insights on exponential risk. Navigating risk is more complex than ever. The company, dedicated to the financial sector, provides rich data, expert analysis, robust tools supported by groundbreaking technologies, and a view of the future to enable our customers to unlock opportunity, advance their business, and act decisively.*"[Nuestra misión es ser la fuente líder de perspectivas relevantes sobre riesgos exponenciales. Navegar el riesgo es más complejo que nunca. La empresa, dedicada al sector financiero, proporciona datos valiosos, análisis de expertos, herramientas sólidas respaldadas por tecnologías innovadoras y una visión del futuro para permitir que nuestros clientes aprovechen oportunidades, hagan avanzar sus negocios y actúen con determinación.] (Empresa dedicada al sector financiero, 2024)

1.2.3 Visión

De acuerdo con la página oficial de la empresa, se tiene que su visión corresponde a:

"*In a world shaped by increasingly interconnected risks, it is more difficult than ever to act with certainty. Our customers need to go beyond data into context, to go beyond context into meaning. The company, dedicated to the financial sector, provides a compass for understanding. With our rich history, innovative technologies, and diverse expertise, we help customers develop a holistic view of their world. We decode complexity, uncovering opportunity amid exponential risk and informing the way forward. Partnering with the company gives customers a comprehensive, global perspective and the confidence to act, and empowers individuals and organizations to thrive.*"[En un mundo moldeado por riesgos cada vez más interconectados, es más difícil que nunca actuar con certeza. Nuestros clientes necesitan ir más allá de los datos para alcanzar el contexto, y más allá del contexto para encontrar el significado. La empresa, dedicada al sector financiero, proporciona una brújula para la comprensión. Con nuestra rica historia, tecnologías innovadoras y experiencia diversa, ayudamos a los clientes a desarrollar una visión holística de su mundo. Desciframos la complejidad, descubriendo oportunidades en medio de riesgos exponenciales y orientando el camino a seguir. Asociarse con la empresa brinda a los clientes una perspectiva global y completa, la confianza para actuar, y empodera a individuos y organizaciones para prosperar.] (Empresa dedicada al sector financiero, 2024)

1.2.4 Valores

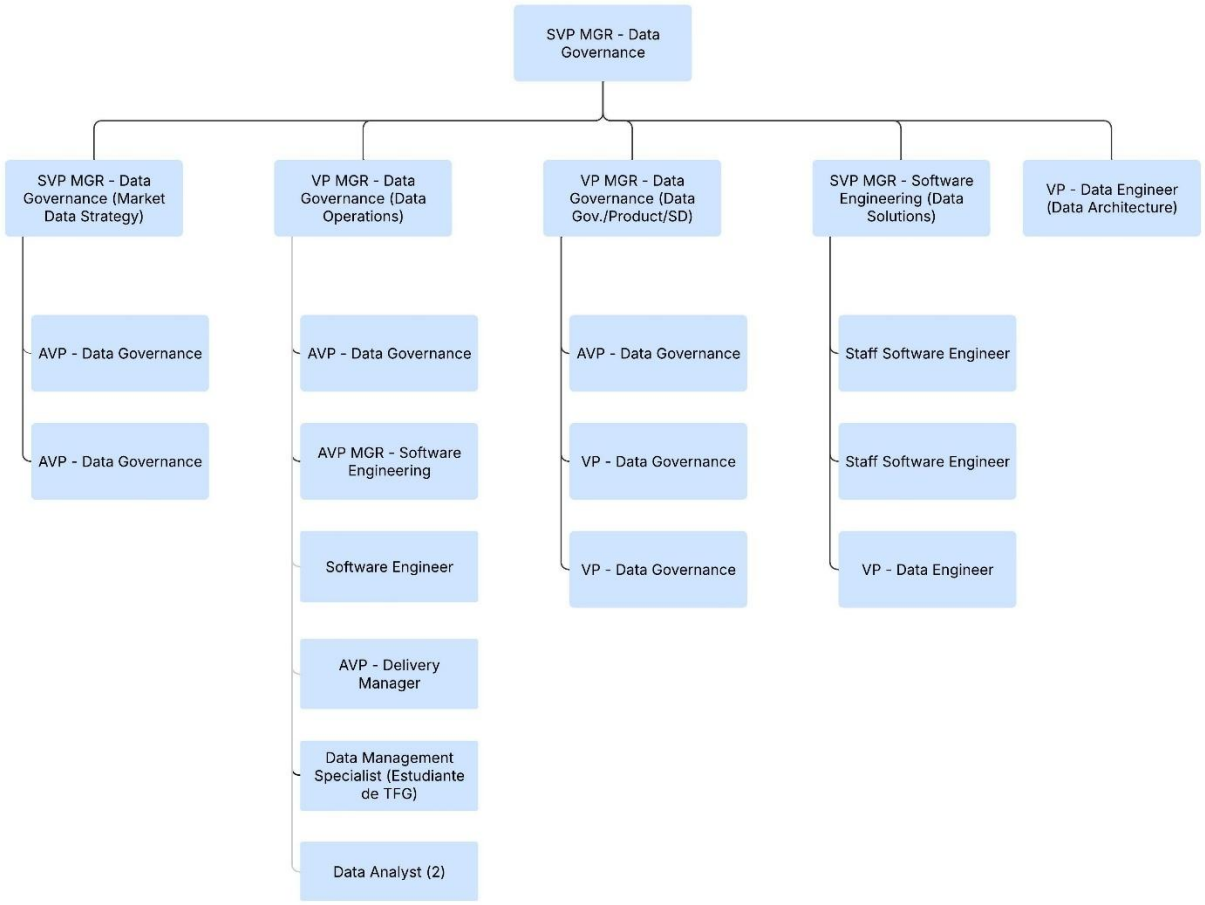
De acuerdo con la página oficial de la empresa dedicada al sector financiero (2024), se tiene que sus valores corresponden a:

- **Invertir en cada relación**: Creamos una experiencia de clase mundial para nuestra gente y nuestros clientes. Priorizamos el valor a largo plazo sobre las ganancias a corto plazo. Nuestro equipo aborda cada interacción con inclusividad, cuidado y compromiso para ayudar a todos a alcanzar su máximo potencial.
- **Liderar con curiosidad**: La curiosidad nos mantiene a la vanguardia. Vamos más allá de lo ya probado. Nuestro equipo combina rigor con una mente abierta, trabajando creativamente en métodos y tecnologías para ampliar nuestro conocimiento y empujar los límites de lo que podemos lograr.
- **Defender perspectivas diversas**: Incluimos voces diversas para tomar decisiones más inteligentes. Trascendemos los silos y estructuras tradicionales. Nuestro equipo une experiencia y conocimientos en toda nuestra organización para formar decisiones holísticas e inteligentes que nos preparan para el éxito.
- **Convertir entradas en acciones**: Entregamos resultados claros. Perseveramos y prosperamos ante la complejidad. Nuestro equipo aprovecha el juicio experto y la tecnología para identificar las entradas correctas y traducirlas en soluciones confiables y prácticas que cumplen con los objetivos.
- **Mantener la confianza a través de la integridad**: La confianza es esencial para quienes somos. Nunca tomamos atajos ni nos conformamos con respuestas fáciles. Juntos, construimos sobre nuestro legado de 115 años haciendo lo correcto, siempre, porque eso nos lleva al mejor resultado para todos.

1.2.5 Equipo de Trabajo

El proyecto se desarrollará en el área de negocio *Enterprise Data Management & Governance*, dentro del equipo de *Data Operations* de la empresa del sector financiero. Este equipo se especializa en soporte de datos maestros y de referencia, gestión horizontal de procesos de *Enterprise Data Management* (EDM), incluyendo la administración del proceso de *intake*, gestión de proyectos, gestión de liberaciones y pruebas, documentación de procesos, y estandarización interna de procesos. A continuación, en la Figura 1 se presenta el organigrama del área de negocio *Enterprise Data Management & Governance*.

Figura 1. Organigrama Enterprise Data Management & Governance



Nota. Adaptado de reunión con manager de Data Operations (2025)

En la Tabla 1 se detallan los roles y responsabilidades del equipo.

Tabla 1. Roles y responsabilidades del equipo Data Operations

Rol	Responsabilidades
VP Mgr - Data Governance	Lidera estratégicamente el equipo, gestionando proyectos clave y coordinando el soporte de producción para mejorar la calidad y eficiencia de los procesos de gobernanza de datos.
AVP - Delivery Manager	Responsable de la planificación, supervisión y coordinación de las iniciativas del equipo, garantizando entregas dentro de los tiempos, costos y alcances establecidos.
AVP - Data Governance	Encargado del soporte de producción relacionado con la gobernanza de datos y la gestión de liberaciones, asegurando la calidad y estabilidad de los cambios implementados en las aplicaciones críticas.
AVP Mgr - Software Engineering	Líder técnico en ingeniería de software, especializado en la generación de reportes estratégicos y la mejora continua de aplicaciones clave dentro del equipo.
Software Engineer	Responsable de garantizar la estabilidad operativa de las aplicaciones del equipo, ejecutando pruebas de regresión y resolviendo problemas críticos en el soporte técnico.
Data Management Specialist	Estudiante responsable del desarrollo del presente Trabajo Final de Graduación. Apoya la gestión documental, contribuye a la generación de reportes y colabora en tareas de análisis de negocio para mejorar los procesos y facilitar la toma de decisiones. Asimismo, cumple un papel estratégico en la gestión de datos, garantizando su trazabilidad, integridad y disponibilidad dentro del entorno organizacional.
Data Analyst (2)	Son responsables de realizar tareas de análisis, validación y seguimiento que contribuyen a la calidad y consistencia de la información manejada por el equipo. Participan activamente en la revisión de datos, identificación de inconsistencias, elaboración de reportes y documentación de hallazgos relevantes.

Nota. Adaptado de reunión con manager de Data Operations (2025)

1.3 Proyectos similares

1.3.1 Proyectos internos en la organización

Entre los proyectos que se han realizado internos en la empresa y resultan ser un valioso insumo para el proyecto se encuentran los siguientes:

1.3.1.1 Implementación de una plataforma de gobernanza de datos

Uno de los proyectos más relevantes realizados dentro de la organización es la implementación de una plataforma de gobernanza de datos, diseñada para ayudar a gestionar y

controlar la información empresarial de manera efectiva. Este sistema ya se encuentra operando, ha sido clave para centralizar la vista de los datos de la organización, facilitando la colaboración entre equipos y departamentos. La plataforma permite a los usuarios descubrir y comprender los datos disponibles en la organización, incluyendo detalles sobre su origen, contexto, calidad y uso. Estas capacidades no solo automatizan procesos y mejoran la calidad de los datos, sino que también aseguran el cumplimiento de regulaciones mediante políticas y estándares de gobernanza centralizados.

El propósito principal de la implementación de esta plataforma fue establecer un catálogo completo de la información empresarial. Por ejemplo, en el caso de bases de datos externas adquiridas que incluyen información sobre mercados específicos, esta herramienta actúa como un catálogo que describe en detalle el contenido de cada columna, su relación con los dominios de datos relevantes y su utilidad para los *stakeholders* internos. Este nivel de organización permite a los equipos internos acceder a información estructurada y comprensible, alineada con los estándares de la organización.

Este proyecto resulta un insumo fundamental para la mejora del proceso de carga de datos en la plataforma de gestión de datos maestros, ya que ambas comparten principios de gobernanza y trabajan con metadatos que debe integrarse bajo estándares comunes. Además, la implementación de esta herramienta estableció lineamientos organizacionales que deben ser respetados en cualquier nueva incorporación tecnológica. Por lo tanto, este proyecto ofrece aprendizajes clave y un marco de referencia para garantizar que el proceso de carga de datos en la plataforma de gestión de datos maestros cumpla con los estándares establecidos y promueva la integración efectiva entre sistemas.

1.3.1.2 Implementación de una plataforma de gestión de licencias

Otro proyecto relevante dentro de la organización fue la implementación de una plataforma orientada al manejo de licencias de acceso a datos, abarcando tanto información financiera como de otros tipos. Esta plataforma desempeña un papel crucial en la administración y procesamiento de metadatos. Este sistema cuenta con una vasta biblioteca de licencias, las cuales contienen información clave como identificadores, nombre de la licencia, proveedor, producto asociado, gasto comprometido total, entre otros. Estos datos son de gran interés para diferentes áreas de negocio dentro de la organización, ya que proporcionan una base estructurada para la toma de decisiones estratégicas.

Una característica destacada de esta plataforma es su integración con otras herramientas de gobernanza de datos. La información procesada en este sistema se carga en otra plataforma centralizada, donde se genera y almacena metadatos adicionales para su posterior uso. Este flujo de trabajo garantiza que los datos estén enriquecidos, organizados y accesibles para los *stakeholders* internos.

Este proyecto constituye un insumo esencial para la mejora del proceso de carga de datos en la plataforma de gestión de datos maestros, ya que ambas herramientas comparten principios similares en cuanto al manejo y carga de metadatos. La experiencia adquirida en la implementación

de esta plataforma ofrece aprendizajes significativos, tanto en términos de estándares como de metodologías aplicadas, que pueden ser replicados y adaptados para mejorar el flujo de carga de datos en el presente proyecto. Además, establece un marco de referencia para garantizar que los procesos sean consistentes y alineados con los objetivos de la organización.

1.3.2 Proyectos externos a la organización

Entre los proyectos que se han realizado externos a la empresa y resultan ser un valioso insumo para el proyecto se encuentran:

1.3.2.1 Propuesta de Estandarización y Automatización para el proceso de Integración de datos en la Empresa Xumtech – Maribel Cordero Pereira

El Trabajo Final de Graduación de Maribel Cordero Pereira (2022) se centra en mejorar la eficiencia y precisión del proceso de integración de datos en una empresa de tecnología. Es un insumo esencial para el desarrollo del presente proyecto, ya que aborda de manera directa los retos de estandarización y automatización en la integración de datos, que son aspectos clave en la mejora de procesos de una plataforma de Master Data Management (MDM). La metodología de análisis, junto con las estrategias de automatización propuestas sirven de base útil para la identificación de herramientas tecnológicas que aseguren la consistencia y calidad de los datos maestros en el proceso de carga.

1.3.2.2 Propuesta para mejoramiento de los procesos de carga de datos sobre el módulo de servicio al cliente que brinda la plataforma Oracle CX, ofrecido por la empresa Xum Technologies - José Carlos Chaves Araya

El Trabajo Final de Graduación de José Carlos Chaves Araya (2023) se enfoca en desarrollar una propuesta para mejorar los procesos de carga de datos en el módulo de servicio al cliente de Oracle CX. Este trabajo es un insumo esencial para el proyecto, ya que aborda la automatización procesos en la carga de datos, aspectos que están directamente relacionados con la eficiencia operativa en una plataforma de MDM. Además, el enfoque en simulación de procesos con herramientas de modelado BPMN es aplicable para el diseño y análisis de mejoras en el flujo de carga de datos del proyecto, proporcionando tanto técnicas como métricas útiles para el objetivo de reducir tiempos y asegurar la calidad de los datos maestros.

1.3.2.3 Propuesta de mejora del proceso de gestión del servicio de Análisis de Datos en la Gerencia de Innovación del Grupo 823 – Jossué Andrey Cascante Molina

El Trabajo Final de Graduación de Jossué Andrey Cascante Molina (2021) se basa en identificar oportunidades de mejora y estandarizar el proceso utilizando prácticas de la industria, como BPMN y marcos de referencia como ITIL y COBIT. Este trabajo es un insumo útil para el proyecto, ya que aborda directamente la mejora y estandarización de procesos, así como la evaluación de eficiencia mediante simulaciones. La metodología aplicada tal cómo el análisis de

brechas constituye una referencia fundamental para estructurar y evaluar el proceso de carga de datos en la plataforma de gestión de datos maestros, asegurando calidad y eficiencia en el manejo de grandes volúmenes de información.

1.4 Planteamiento del problema

En esta sección se describe la situación problemática hallada dentro del entorno de la organización, el cual motiva el desarrollo del proyecto, así como la mención de los beneficios esperados del producto.

1.4.1 Situación problemática

La problemática central que enfrenta el equipo de *Data Operations* de la empresa del sector financiero radica en la ineficiencia operativa en el manejo de grandes volúmenes de datos en una plataforma de Gestión de Datos Maestros. Este problema surge a partir de un proceso altamente manual y fragmentado, donde la carga de datos depende de tareas repetitivas y no estandarizadas, que requieren intervención humana en múltiples etapas. La ausencia de un flujo de trabajo estructurado no solo limita la eficiencia, sino que también incrementa el riesgo de errores humanos y dificulta la implementación de controles efectivos para garantizar la calidad y consistencia de los datos.

Uno de los factores clave en esta problemática es la dependencia de herramientas no especializadas para la carga de datos, lo que ha llevado a una ejecución manual del proceso. Actualmente, Excel, SharePoint y PostgreSQL juegan un rol fundamental en el flujo de trabajo, pero su uso manual y la falta de automatización aumentan significativamente la carga operativa y la probabilidad de errores. Además, la baja escalabilidad del proceso de carga de datos impide manejar grandes volúmenes de información de manera eficiente. A medida que el volumen de datos crece, el tiempo requerido para procesarlos aumenta exponencialmente, generando retrasos y errores en la manipulación de la información.

Otro factor crítico es la definición incompleta de los requerimientos del proceso, lo que ha dificultado la estandarización de la carga de datos. La falta de criterios claros para estructurar los datos ha provocado una gobernanza de datos inconsistente, donde se carece de reglas bien definidas para la validación de datos antes de su integración en la plataforma. Como resultado, los datos ingresados pueden contener errores que afectan su calidad, afectando su confiabilidad para la toma de decisiones estratégicas.

Además, la documentación del proceso es incompleta y desorganizada, lo que ha generado dificultades en la comprensión del flujo de trabajo. La ausencia de lineamientos claros ha limitado el entendimiento de nuevos usuarios y complicado la transferencia de conocimiento entre los miembros del equipo. Esta situación aumenta la dependencia de habilidades individuales y agrava los problemas operativos, ya que la correcta ejecución del proceso recae en el conocimiento empírico de los miembros del equipo, en lugar de estar respaldada por documentación formal y accesible.

El proceso actual de carga de datos refleja múltiples ineficiencias derivadas de la falta de estandarización. Los datos, provenientes de fuentes distintas, se entregan en formato CSV

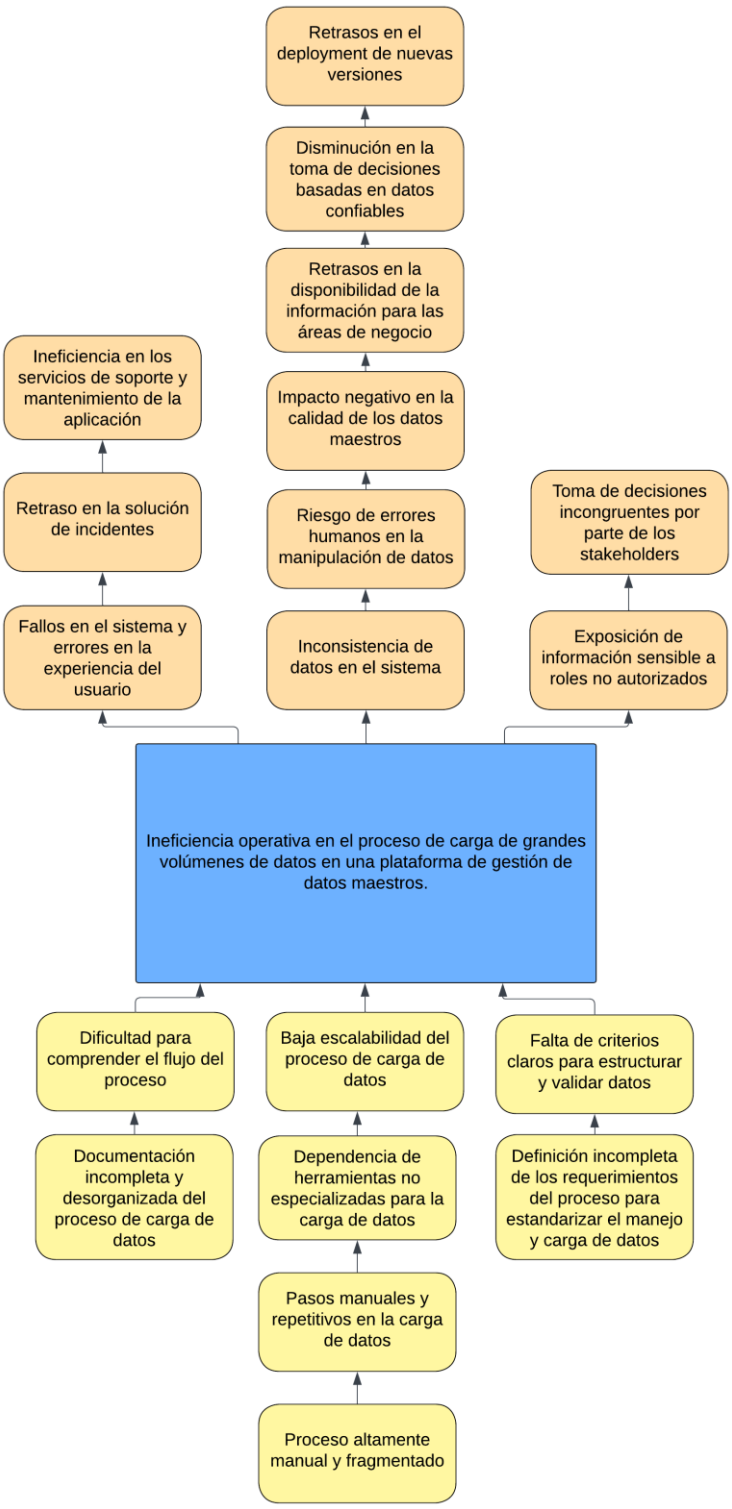
mediante métodos no unificados, como servidores FTP (*File Transfer Protocol*), correo electrónico y APIs. Cada archivo debe ser descargado manualmente y organizado en carpetas específicas dentro de un repositorio en *SharePoint*. Sin un formato estándar ni un flujo automatizado, este procedimiento introduce riesgos significativos de errores humanos y retrasa la disponibilidad de los datos para el equipo.

Una vez almacenados en SharePoint, los datos son procesados en Excel, donde se asignan reglas de validación a cada columna, como la definición de tipos de datos (*strings, integers o booleans)*. Durante este paso, se corrigen inconsistencias como valores faltantes (*"NAs"*), se aplican filtros manuales y se ajustan errores detectados visualmente. Este trabajo manual, que implica la manipulación de cientos de miles de filas, ralentiza el flujo de trabajo y aumenta la probabilidad de errores. Finalizado este procesamiento, se cargan manualmente en la base de datos de la plataforma de Gestión de Datos Maestros mediante consultas en PostgreSQL. Sin embargo, este último paso hereda inconsistencias generadas en etapas previas, perpetuando problemas de calidad en los datos cargados.

La situación actual no solo genera deficiencias operativas que ralentizan el flujo de trabajo, sino que también compromete la confiabilidad de los datos maestros. La dificultad para manejar grandes volúmenes de información con precisión repercute directamente en la toma de decisiones estratégicas, afectando la capacidad del equipo para garantizar información confiable y oportuna. En un entorno donde la agilidad y precisión son fundamentales, esta problemática representa un riesgo para el cumplimiento de las expectativas de las partes interesadas.

A continuación, se presenta la situación planteada anteriormente de manera más visual en la Figura 2, la cual representa un diagrama tipo árbol del problema, el cual contiene las causas, efectos identificados del problema en cuestión y la problemática a abordar.

Figura 2. Árbol del problema



Nota. Elaboración propia (2025)

1.4.2 Justificación del proyecto

El presente proyecto, se centra en el desarrollo de una propuesta de mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros, abordando una problemática crítica para el equipo de *Data Operations*: la ineficiencia en la gestión y manejo de grandes volúmenes de información. Esta iniciativa resulta pertinente para la Licenciatura en Administración de Tecnología de Información (ATI), dado que integra y aplica conocimientos fundamentales de las áreas de Gestión de Datos e Información y Administración de Procesos de Negocios, como lo estipula el artículo 19 del RETFG-ATI (Instituto Tecnológico de Costa Rica, 2023).

El desarrollo del proyecto utiliza herramientas y metodologías basadas en mejores prácticas de la industria, como la gobernanza de datos y la automatización de procesos, pilares fundamentales en la mejora operativa. Estas tecnologías y enfoques no solo mejoran la calidad y consistencia de los datos, sino que también facilitan la integración de información crítica entre sistemas, habilitando a la organización para tomar decisiones más estratégicas y basadas en datos confiables. Este enfoque está alineado con la formación técnica y administrativa del profesional de ATI, que actúa como un vínculo entre las áreas de TI y de gestión empresarial, permitiendo la creación de soluciones innovadoras y estratégicas (Chavarría Sánchez, 2023).

Además, este proyecto aborda la problemática desde un enfoque integral, combinando los conocimientos técnicos de análisis de datos con la capacidad de gestión para liderar iniciativas que mejoren la infraestructura tecnológica del equipo de *Data Operations*. Según el Instituto Tecnológico de Costa Rica (2023), el egresado de ATI está capacitado para liderar proyectos que incrementen la competitividad empresarial mediante la implementación de tecnologías avanzadas, algo que este proyecto refleja al enfocarse en la mejora de un proceso crítico dentro de la organización. La capacidad de integrar soluciones tecnológicas con procesos administrativos subraya la importancia del proyecto en la formación de un profesional integral.

La implementación de procesos automatizados en la gestión de datos es fundamental para mejorar la eficiencia operativa y la calidad de la información en las organizaciones. Según un estudio de Deloitte (s.f), la transformación digital y la automatización inteligente de procesos permiten a las organizaciones aumentar la transparencia, el control y la eficiencia en sus operaciones, asegurando que las soluciones sean escalables y alineadas con estándares de calidad. Estas prácticas no solo mejoran el flujo de trabajo, sino que también reducen significativamente los riesgos operativos y mejoran la capacidad de respuesta frente a cambios en el entorno organizacional.

Por otra parte, la automatización en el proceso de carga de datos no solo mejora la eficiencia operativa, sino que también reduce significativamente los errores humanos asociados a tareas repetitivas, alineándose con los principios de transformación digital, que según Westerman et al. (2014), permiten a las organizaciones obtener ventajas competitivas sostenibles mediante la implementación de tecnologías disruptivas. La integración de estas tecnologías refuerza la

capacidad de la organización para mejorar sus operaciones, consolidando su posición en un mercado altamente competitivo.

El proyecto también promueve el desarrollo de competencias en Inteligencia de Negocios, al proporcionar datos más precisos y consistentes que pueden ser utilizados para el análisis estratégico y la toma de decisiones informadas. Según Chen et al. (2012), los sistemas de gestión de datos juegan un papel fundamental en la recopilación y análisis de información clave para las empresas, permitiéndoles identificar oportunidades de mejora y responder proactivamente a los desafíos del entorno empresarial.

En términos educativos, el proyecto fomenta la aplicación de habilidades adquiridas durante la carrera, como la planificación, ejecución y gestión de proyectos tecnológicos. Estas competencias son esenciales para la resolución de problemas organizacionales y el diseño de soluciones adaptadas a las necesidades específicas de la organización (Instituto Tecnológico de Costa Rica, 2023). Adicionalmente, la relevancia del proyecto se extiende más allá de la organización, ya que sirve como un ejemplo práctico del impacto positivo que las tecnologías tienen en la mejora de procesos críticos.

Este proyecto se posiciona como una oportunidad para aplicar de manera práctica los conocimientos adquiridos durante la carrera de ATI, proporcionando al equipo, así como a la organización una solución estratégica y alineada con las necesidades actuales. Su temática responde directamente a las áreas de desarrollo del TFG, destacándose como una contribución significativa tanto para el equipo y organización, como para el crecimiento profesional del estudiante.

1.4.3 Beneficios esperados del proyecto

En esta sección, se enlistan los beneficios directos e indirectos que se espera obtener como resultado de resolver la situación problemática.

1.4.3.1 Beneficios directos

- Automatización de la carga de datos: La implementación de herramientas tecnológicas permitirá mejorar las etapas más críticas del proceso de carga de datos, reduciendo significativamente la intervención manual y asegurando un flujo de trabajo más estructurado y eficiente.
- Reducción de errores operativos: La disminución de la intervención manual en tareas repetitivas minimizará la probabilidad de errores humanos, permitiendo un mayor control sobre la calidad de los datos y reduciendo inconsistencias en la información almacenada.
- Mejora en la calidad y consistencia de los datos maestros: Al reducir errores en el proceso y establecer reglas claras para el tratamiento de la información, los datos ingresados en la plataforma de Gestión de Datos Maestros serán más confiables, actualizados y estructurados de manera uniforme.
- Mejora de la eficiencia operativa en el proceso de carga de datos: Como resultado de la automatización, la reducción de errores y la estandarización de los datos, los tiempos de

procesamiento disminuirán significativamente, permitiendo un flujo de datos más ágil y reduciendo la carga operativa del equipo.

1.4.3.2 Beneficios indirectos

- Mejora en la preparación de datos para auditorías y normativas: La automatización del proceso de carga de datos reducirá errores y garantizará que la información esté estructurada de manera uniforme, facilitando el cumplimiento de los requerimientos de auditoría, así como de regulaciones internas y externas.
- Fortalecimiento de la toma de decisiones estratégicas: La disponibilidad de datos más precisos y actualizados en la plataforma de Gestión de Datos Maestros permitirá a los líderes del equipo y a las partes interesadas fundamentar sus decisiones en información confiable, mejorando la planificación y ejecución de estrategias organizacionales.
- Mayor satisfacción de los usuarios internos: Con datos más confiables, las áreas de negocio que dependen de la información en la plataforma de Gestión de Datos Maestros podrán ejecutar sus tareas con mayor eficiencia y seguridad.
- Facilitación de nuevos casos de uso: La centralización de los datos en la plataforma de Gestión de Datos Maestros permitirá su integración con otras herramientas dentro de la organización, favoreciendo el desarrollo de nuevos casos de uso y promoviendo el aprovechamiento estratégico de la información en distintas áreas del negocio.

1.5 Objetivos del Trabajo Final de Graduación

1.5.1 Objetivo general

Diseñar una propuesta de mejora y automatización del proceso de carga de datos en una plataforma de gestión de datos maestros, con el propósito de la eliminación de deficiencias actuales, logrando una gestión más eficiente, dentro de un periodo de 16 semanas.

1.5.2 Objetivos específicos

1. Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros.
2. Diseñar un nuevo proceso de carga de datos integrando herramientas de automatización y alineado con los requerimientos del equipo, con el fin del mejoramiento de la eficiencia del proceso.
3. Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta seleccionada y los requerimientos identificados.
4. Evaluar la efectividad del prototipo de la solución automatizada en términos de precisión, consistencia y reducción de tareas manuales en el proceso de carga de datos, utilizando métricas de desempeño para la determinación de su impacto en la eficiencia del proceso.

1.6 Alcance

El presente proyecto abarca el análisis del proceso actual de carga de datos en una plataforma de gestión de datos maestros, el diseño de un nuevo proceso alineado con herramientas de automatización, el desarrollo de un prototipo de solución y la evaluación de su impacto en la eficiencia operativa. Este alcance se ajusta a los objetivos planteados y busca proporcionar una solución conceptual y técnica que atienda las deficiencias identificadas.

En primer lugar, se llevará a cabo un análisis detallado del estado actual del proceso de carga de datos, identificando deficiencias y tareas manuales que afectan la eficiencia del flujo de información en el proceso. Esta evaluación permitirá documentar el estado actual (*As-Is*) y servirá como base para el diseño de mejoras.

Posteriormente, se procederá con el diseño de un nuevo proceso de carga de datos, alineado con los requerimientos del equipo de *Data Operations* y estructurado para facilitar la automatización. En esta fase, se establecerá un esquema lógico del proceso futuro (*To-Be*), considerando las necesidades operativas y organizacionales.

Finalmente, el proyecto incluirá el desarrollo de un prototipo de la solución automatizada y la evaluación de su efectividad en términos de precisión, consistencia y reducción de tareas manuales críticas. La evaluación del prototipo permitirá obtener evidencia sobre su impacto en la gestión de datos y su aporte a la mejora de la eficiencia del proceso.

1.6.1 Fuera del alcance

Este proyecto no contempla la implementación final de la solución automatizada, ya que su alcance se limita al análisis del proceso actual, el diseño de una propuesta mejorada y el desarrollo de un prototipo para validar su efectividad. La adopción definitiva de la solución, su despliegue en un entorno productivo y cualquier mejora posterior quedarán bajo la responsabilidad del equipo y la organización.

Asimismo, quedan fuera del alcance actividades relacionadas con la gestión interna de los datos dentro de la plataforma de gestión de datos maestros, dado que el enfoque se centra en la mejora del flujo de carga de datos y no en la manipulación o transformación de la información una vez almacenada.

Tampoco se incluirá el desarrollo de herramientas adicionales para análisis avanzado o visualización de los datos procesados, ya que el propósito del proyecto es mejorar la eficiencia en la carga y estructuración de datos, sin abordar su explotación analítica. Además, no se contempla la integración de nuevas fuentes de datos que no estén dentro del proceso actual, manteniendo el alcance en la mejora del flujo de carga existente.

Por último, la documentación completa de un marco de gobernanza de datos queda fuera del alcance, dado que el objetivo del proyecto es la mejora operativa del proceso de carga y no la definición de estrategias de gobernanza a nivel de equipo y organizacional.

1.7 Supuestos

En la realización del presente proyecto, se asumen los siguientes factores y elementos como ciertos:

- Se asume que el pasante responsable del desarrollo del proyecto contará con total disponibilidad para la empresa durante el I Semestre del 2025, garantizando así el cumplimiento de las actividades planeadas.
- Se supone que el proyecto continuará con aprobación y respaldo por parte de la organización y no será cancelado o pospuesto debido a decisiones de niveles superiores.
- Se asumirá que los recursos técnicos, tecnológicos y humanos requeridos estarán disponibles en cantidad y calidad suficiente para garantizar el éxito del proyecto.
- Se asumirá el apoyo y la disponibilidad de los miembros del equipo para proporcionar información clave sobre el proceso actual, los requerimientos específicos y las fuentes de datos involucradas.
- Se contará con acceso a servicios de AWS, así como a la plataforma de gestión de datos maestros para el diseño, pruebas y validación del prototipo de automatización.
- Se supone que los datos y fuentes requeridas para el proyecto estarán disponibles durante todo el periodo de ejecución, sin interrupciones significativas que puedan afectar el progreso.
- Las fuentes de datos contempladas para el desarrollo del proyecto se mantendrán sin cambios significativos durante el periodo de ejecución del Trabajo Final de Graduación.
- Los entregables propuestos, como el prototipo de automatización y documentación del proceso serán aceptados y utilizados por la organización una vez completados.

1.8 Entregables

1.8.1 Entregables académicos

Los entregables académicos corresponden a los documentos solicitados como parte del desarrollo del Trabajo Final de Graduación y están dirigidos a la Coordinación del Trabajo Final de Graduación y al profesor tutor asignado. Estos entregables incluyen:

- Avances solicitados tanto por el profesor tutor como por la Coordinación del Trabajo Final de Graduación.
- Informe final del Trabajo Final de Graduación
- Presentación y defensa del Trabajo Final de Graduación ante el comité evaluador.

1.8.2 Entregables del producto

Los entregables del producto están directamente asociados a los objetivos específicos planteados y son entregados a la organización como resultado del proyecto. Estos incluyen:

- **Objetivo específico #1**: Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros.

  - Documentación del estado actual del proceso (*As-Is*), identificando deficiencias y tareas manuales que afectan la eficiencia operativa.
  - Diagrama Ishikawa (*Fishbone*) que representa gráficamente las causas raíz que generan ineficiencias en el proceso de carga de datos.
  - Análisis FODA de la situación actual, en el que se sintetizan las fortalezas, oportunidades, debilidades y amenazas que influyen en el desempeño del proceso.
- **Objetivo específico #2**: Diseñar un nuevo proceso de carga de datos integrando herramientas de automatización y alineado con los requerimientos del equipo, con el fin del mejoramiento de la eficiencia del proceso.
  - Documentación del nuevo diseño del proceso (*To-Be*), detallando el flujo lógico, etapas y mejoras en la estructura del proceso de carga de datos.
  - *Checklist* de requerimientos para la automatización, validado con el equipo *Data Operations* para definir criterios funcionales y técnicos del diseño.
  - Matriz de trazabilidad de requerimientos vs diseño, que vincula cada necesidad identificada con los componentes incluidos en el nuevo proceso.
  - Matriz de integración, que demuestra la cohesión técnica entre etapas, herramientas utilizadas y asignación de responsabilidades dentro del diseño propuesto.
- **Objetivo específico #3**: Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta utilizada y los requerimientos identificados.
  - Prototipo funcional de la solución automatizada, incorporando las funcionalidades clave requeridas para la carga eficiente de datos.
  - Documentación del prototipo, describiendo su funcionamiento, roles y los componentes implementados.
  - Arquitectura lógica del flujo automatizado, en la que se ilustran las capas del proceso, los servicios involucrados y la relación secuencial entre las funciones implementadas.
  - Matriz de validación de requerimientos versus prototipo implementado, que demuestra la correspondencia directa entre los criterios funcionales definidos y las funcionalidades desarrolladas.
  - Análisis de riesgos de la solución, en el que se identifican los riesgos técnicos y operativos asociados al prototipo, así como las estrategias propuestas para su mitigación.
  - Análisis costo-beneficio, que evalúa el retorno esperado de la inversión (ROI) con base en la reducción del esfuerzo operativo y el aprovechamiento de servicios de la herramienta seleccionada.
  - Hoja de ruta de implementación de la propuesta de solución, que detalla las fases sugeridas para la adopción progresiva del prototipo en el entorno organizacional real.

- **Objetivo específico #4**: Evaluar la efectividad del prototipo de la solución automatizada en términos de precisión, consistencia y reducción de tareas manuales en el proceso de carga de datos, para la determinación de su impacto en la eficiencia del proceso.
    - Informe de evaluación del prototipo, con métricas de desempeño, resultados de pruebas y análisis del impacto en la eficiencia del proceso.

1.8.3 Gestión del proyecto

En esta sección se describen los artefactos asociados a la gestión del proyecto. Los entregables de gestión tienen como propósito garantizar la adecuada planificación, ejecución, y control del proyecto, asegurando que se cumplan los objetivos planteados dentro de los parámetros establecidos de tiempo, alcance y calidad. A continuación, se describen los entregables de gestión que permitirán supervisar y documentar el progreso del proyecto de forma eficiente y alineada con las expectativas organizacionales y académicas.

1.8.3.1 Minutas

En esta sección se describe la plantilla que se empleará para las minutas de las reuniones llevadas a cabo durante el desarrollo del Trabajo Final de Graduación. El objetivo de esta plantilla es mantener un registro estandarizado y organizado de los temas discutidos, las decisiones tomadas y los acuerdos alcanzados, asegurando así la trazabilidad y claridad en la gestión del proyecto. La plantilla utilizada para las minutas se encuentra adjunta en el Apéndice B.

1.8.3.2 Gestión del cambio

En esta sección se describe la plantilla que se empleará para documentar las solicitudes de gestión de cambio durante el desarrollo del Trabajo Final de Graduación. El objetivo de esta plantilla es garantizar que los cambios solicitados sean registrados y gestionados de manera adecuada, asegurando la trazabilidad y el control de los ajustes realizados en el proyecto. La plantilla para la gestión de cambio se encuentra adjunta en el Apéndice C.

1.8.3.3 Cronograma

El cronograma del proyecto detalla las actividades principales a realizar durante su desarrollo, organizadas en un orden lógico y cronológico. Este esquema permite asegurar una adecuada planificación y gestión del tiempo, garantizando que cada fase del proyecto se complete dentro del período estipulado. En el cronograma se incluyen todas las tareas asociadas a las fases metodológicas previamente descritas, así como los entregables académicos y del producto, asegurando que los objetivos específicos sean alcanzados de manera eficiente y alineados con las metas del proyecto. En el Apéndice A, se especifica el cronograma para la elaboración del proyecto.

1.9 Limitaciones

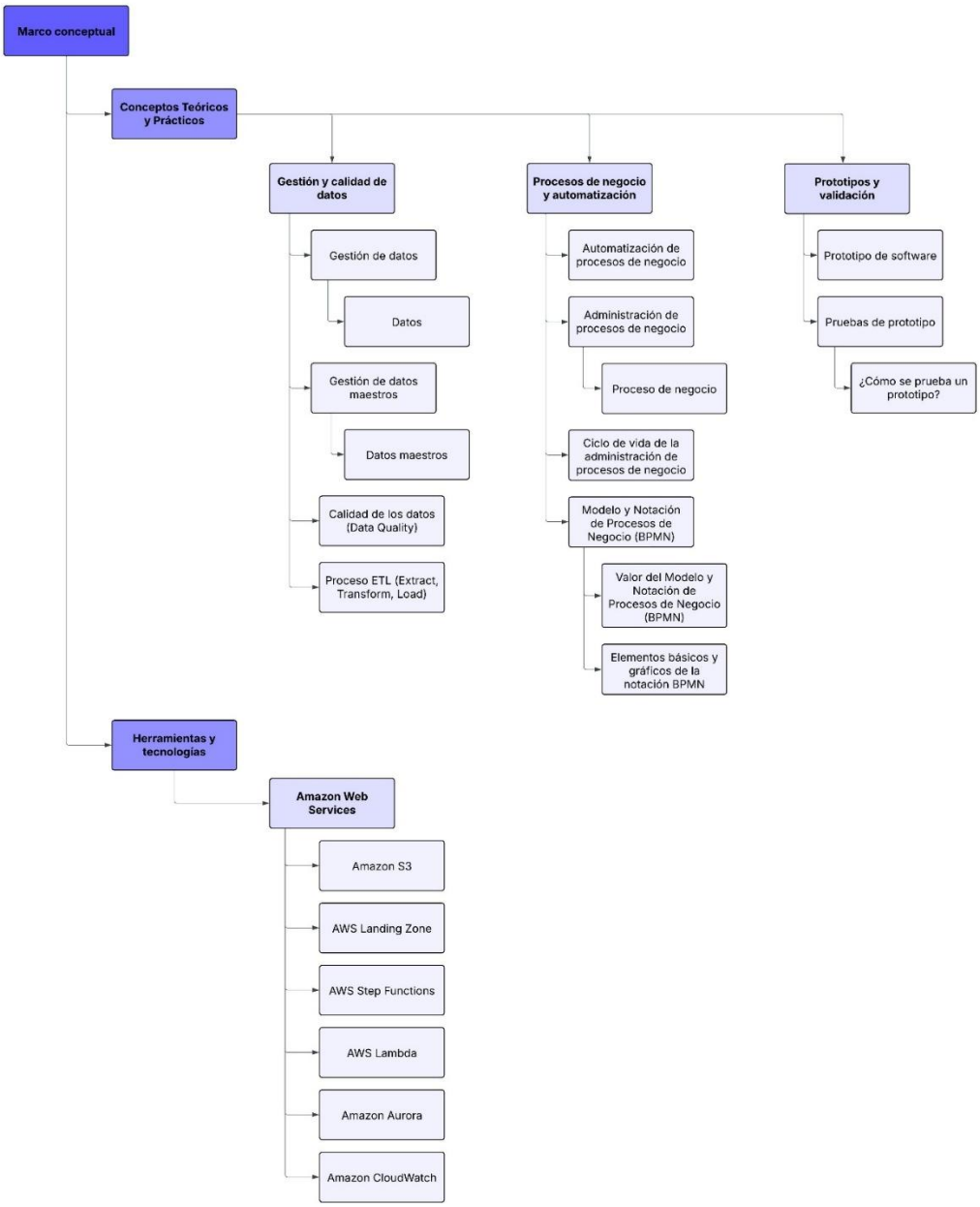
En la realización del presente proyecto, se identifican las siguientes limitaciones que restringen su desarrollo en algún grado.

- El proyecto se centra en la automatización y mejora de un subconjunto específico de fuentes de datos. No se abordan todas las posibles fuentes ni las reglas de validación completas, ya que la variedad y complejidad de estas excede los recursos disponibles.
- El acceso a ciertos datos o herramientas podría estar restringido debido a políticas organizacionales de cumplimiento, lo que en cierto grado afecta el nivel de detalle o profundidad alcanzado en algunas partes del proyecto.
- Los proyectos asociados, como el sistema de gobernanza de datos y la plataforma de manejo de licencias, no es permitido detallarlos extensamente debido a su carácter confidencial y a la naturaleza crítica de los datos que manejan.
- El proyecto involucra a un equipo multicultural y multinacional, con miembros ubicados en diferentes partes del mundo, incluyendo Charlotte, New York, Costa Rica e India. Esto podría generar retos de comunicación y coordinación debido a diferencias de zonas horarias y metodologías de trabajo.
- La interacción directa con la plataforma de gestión de datos maestros está limitada, ya que el pasante responsable del proyecto no administra ni manipula los datos internos dentro de esta plataforma.
- Documentar completamente todos los pasos del pipeline podría enfrentar restricciones debido a la falta de detalles previamente establecidos.
- El uso de entornos corporativos como desarrollo, pruebas (QA) y producción está sujeto a restricciones estrictas debido a políticas de confidencialidad y protección de datos sensibles de la empresa.

2 Marco Conceptual

La presente sección tiene como finalidad establecer un sistema articulado y consistente de conceptos y definiciones que fundamentan el abordaje de la problemática planteada. En ella, se describen los elementos teóricos y prácticos que sustentan el desarrollo del estudio, proporcionando el marco referencial necesario para contextualizar y orientar la investigación. En la Figura 3 se visualiza la estructura del marco conceptual que permite cubrir estos aspectos.

Figura 3. Estructura del Marco Conceptual



Nota. Elaboración propia (2025)

2.1 Conceptos teóricos y prácticos

La siguiente sección describe los conceptos teóricos y prácticos que sustentan el desarrollo del presente proyecto.

2.1.1 Gestión de datos

La gestión de datos se entiende como el desarrollo, ejecución y supervisión de planes, políticas, programas y prácticas que permiten entregar, controlar, proteger, así como maximizar el valor de los activos de datos e información a lo largo de su ciclo de vida. Los datos se reconocen como un activo económico, bajo propiedad o control de la organización, con valor estratégico y operativo. Las organizaciones contemporáneas dependen de estos activos para tomar decisiones efectivas, comprender a sus clientes, crear nuevos productos y servicios, mejorar la eficiencia operativa, reducir costos, además de gestionar riesgos. A medida que crece esta dependencia, el valor de los datos se establece con mayor claridad.

Las actividades de gestión de datos abarcan desde la toma de decisiones sobre cómo extraer valor estratégico, hasta la implementación técnica y la operación diaria de bases de datos. Este proceso requiere habilidades tanto técnicas como de negocio. La responsabilidad se comparte entre roles empresariales y tecnológicos, quienes deben colaborar estrechamente para asegurar datos de alta calidad alineados a las necesidades estratégicas de la organización.

Los datos y la información no son únicamente activos en términos de inversión y valor futuro, sino que resultan vitales para las operaciones cotidianas. Han sido denominados como la "moneda", la "savia vital" e incluso el "nuevo petróleo" de la economía de la información. Ninguna organización logra realizar transacciones comerciales sin datos.

Dentro de una organización los objetivos de la gestión de datos incluyen.

- Comprender y respaldar las necesidades de información de la empresa y de sus partes interesadas, incluyendo clientes, empleados y socios comerciales.
- Capturar, almacenar, proteger y garantizar la integridad de los activos de datos.
- Asegurar la calidad de los datos y la información.
- Garantizar la privacidad y confidencialidad de los datos de las partes interesadas.
- Prevenir el acceso, manipulación o uso no autorizado o inapropiado de los datos e información.
- Asegurar que los datos puedan ser utilizados de manera efectiva para agregar valor a la organización

Muchas organizaciones se identifican como "orientadas por los datos" (*data-driven*). Para mantenerse competitivas, deben abandonar decisiones basadas en intuición y utilizar análisis fundamentados en eventos. Esta transformación exige gestionar los datos con eficiencia y disciplina profesional, mediante una colaboración constante entre liderazgo empresarial y experiencia técnica.

(DAMA International, 2017, pp. 5-9)

2.1.1.1 Datos

El término *dato* hace referencia a la representación de hechos, conceptos o instrucciones de manera formalizada, que permite su comunicación, interpretación y procesamiento. De acuerdo con DAMA International (2017, p. 7), los datos representan hechos sobre el mundo real y constituyen la materia prima a partir de la cual se genera información. En el ámbito de las tecnologías de la información, los datos son entendidos como información almacenada en formato digital; sin embargo, este concepto no se limita exclusivamente a los datos digitalizados, sino que también abarca aquellos capturados en medios físicos o documentos impresos.

El valor de los datos radica en su capacidad para ser agregados, analizados y transformados en información útil que facilite la toma de decisiones, la obtención de beneficios económicos, la mejora de procesos o la formulación de políticas. No obstante, los datos no poseen un significado intrínseco; requieren de un contexto que les otorgue sentido y los convierta en información. Este contexto incluye elementos como un vocabulario común y relaciones explícitas entre los componentes de los datos, lo cual es documentado y gestionado mediante metadatos (DAMA International, 2017, pp. 7-8).

Las organizaciones siempre han necesitado gestionar sus datos; sin embargo, los avances tecnológicos han ampliado el alcance de esta necesidad, al transformar la comprensión que las personas tienen sobre qué son los datos. Estos cambios han permitido que las organizaciones utilicen los datos de nuevas maneras para crear productos, compartir información, generar conocimiento y fortalecer su éxito organizacional. No obstante, el rápido crecimiento de la tecnología, junto con la capacidad humana para producir, capturar y extraer significado de los datos, ha intensificado la necesidad de gestionarlos de manera efectiva (DAMA International, 2017, p. 8).

2.1.2 Gestión de datos maestros

La gestión de datos maestros (MDM, por sus siglas en inglés) constituye una disciplina orientada al control de los valores e identificadores de los datos maestros, con el propósito de garantizar un uso coherente y uniforme de los datos más precisos y actualizados sobre las entidades esenciales del negocio, a través de los diferentes sistemas de la organización. Su objetivo principal es asegurar la disponibilidad de datos fiables y vigentes, minimizando los riesgos asociados a identificadores ambiguos o duplicados.

Según Gartner, la gestión de datos maestros es una disciplina habilitada por la tecnología, en la cual las áreas de negocio y tecnología de la información colaboran para garantizar la uniformidad, precisión, administración, consistencia semántica y responsabilidad sobre los activos oficiales de datos maestros de la organización. Estos activos consisten en un conjunto consistente y uniforme de identificadores y atributos extendidos que describen las entidades centrales de la empresa, tales como clientes, proveedores, ubicaciones, jerarquías y estructuras financieras (DAMA International, 2017, p. 70).

DAMA International (2017, p. 71) señala que la evaluación de los requerimientos de gestión de datos maestros (MDM) en una organización implica identificar diversos elementos clave que permiten establecer un control efectivo sobre los datos esenciales del negocio.

- Los roles, organizaciones, ubicaciones y objetos que son referenciados de forma recurrente.
- Los datos utilizados para describir a las personas, organizaciones, ubicaciones y objetos.
- La manera en que los datos son definidos y estructurados, incluyendo el nivel de detalle (granularidad) que poseen.
- El origen, almacenamiento, disponibilidad y acceso a los datos dentro de la organización.
- La forma en que los datos se modifican a medida que se transfieren entre los sistemas organizacionales.
- Los usuarios de los datos y los propósitos para los cuales los utilizan.
- Los criterios empleados para evaluar la calidad, confiabilidad de los datos y sus fuentes.

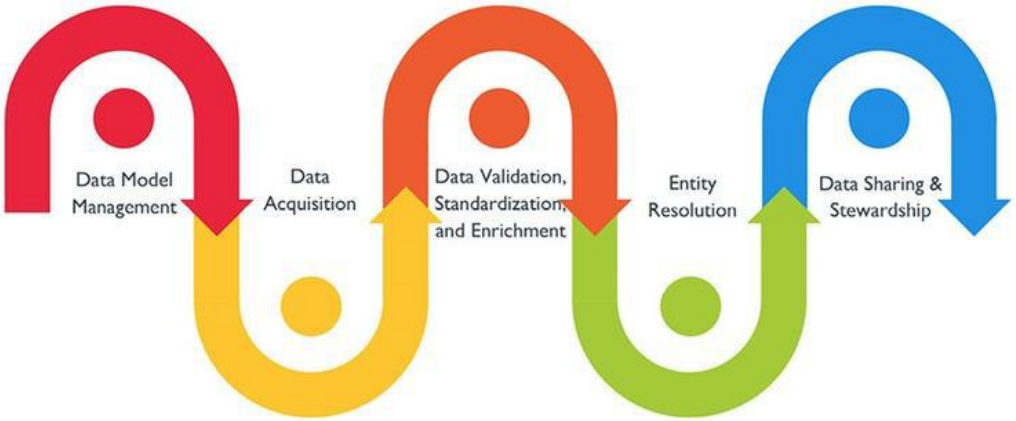
La gestión de datos maestros ayuda a las organizaciones a estandarizar las definiciones y atributos de los elementos de datos (cliente, proveedor, producto, etc.). Facilita el intercambio de datos entre todas las funciones empresariales, departamentos e incluso divisiones, a través de los diferentes sistemas, plataformas y aplicaciones. MDM crea una vista única de los datos del dominio objetivo (Hikmawati, Santosa & Hidayah, 2021, p. 91)

El proceso principal de la gestión de datos maestros consiste en el perfilado de los datos maestros para evaluar la calidad de los datos en diversas fuentes; consolidar los datos maestros en un repositorio y vincularlos con las aplicaciones existentes; limpiar y enriquecer los datos maestros; y sincronizar los datos maestros con los procesos de negocio organizacionales para respaldar la inteligencia empresarial y los sistemas de reporte" (Hikmawati, Santosa & Hidayah, 2021, p. 91)

Uno de los factores que impulsa a las organizaciones a implementar MDM es la necesidad de consistencia y precisión en los datos organizacionales. El núcleo del programa MDM radica en consolidar múltiples conjuntos de datos que representan objetos maestros, como clientes o empleados" (Hikmawati, Santosa & Hidayah, 2021, p. 92)

A continuación en la Figura 4, se representan los pasos clave del procesamiento para la gestión de datos maestros.

Figura 4. Pasos clave del procesamiento para la gestión de datos maestros



Nota. Tomado de Data management body of knowledge (DAMA-DMBOK2) (2nd ed.), por DAMA International, 2017, Technics Publications.

DAMA International (2017, p. 74) explica que los pasos clave en el procesamiento de la gestión de datos maestros comprenden la administración del modelo de datos, la adquisición de información, la validación, estandarización y enriquecimiento de los registros, la resolución de entidades, así como la administración y distribución de los datos. En un entorno integral de MDM, el modelo lógico de datos se implementa en diversas plataformas, orientando la ejecución de la solución y constituyendo la base para los servicios de integración. Este modelo proporciona las directrices necesarias para configurar las aplicaciones, facilitando tanto la reconciliación como la verificación de la calidad de los datos

2.1.2.1 Datos maestros

Según DAMA International (2017, pp. 69-70), los datos maestros hacen referencia a la información sobre las entidades esenciales del negocio, tales como empleados, clientes, productos, estructuras financieras, activos o ubicaciones. Estas entidades corresponden a objetos del mundo real tales como personas, organizaciones, lugares u objetos representados por instancias en forma de datos o registros.

Los datos maestros proporcionan el contexto necesario para las transacciones y los procesos de análisis empresarial. Por esta razón, deben constituir la fuente autorizada y más precisa disponible sobre dichas entidades. Una adecuada gestión de los datos maestros garantiza que estos sean confiables y puedan ser utilizados con certeza dentro de la organización.

DAMA International (2017, p. 70) indica que las reglas de negocio suelen establecer el formato y los rangos permitidos para los valores de los datos maestros. En términos generales, los datos maestros dentro de una organización abarcan información sobre las siguientes categorías.

- **Partes:** Incluye individuos y organizaciones, así como los roles que desempeñan, tales como clientes, ciudadanos, pacientes, proveedores, agentes, socios comerciales, competidores, empleados o estudiantes.

- **Productos y servicios:** Comprende tanto los ofrecidos internamente como los proporcionados externamente.
- **Estructuras financieras:** Datos relacionados con contratos, cuentas del libro mayor, centros de costos o centros de beneficios.
- **Ubicaciones:** Información sobre direcciones y coordenadas geográficas

2.1.3 Calidad de los datos (*Data Quality*)

DAMA International (2017, p. 367) define la calidad de los datos como un concepto que abarca, por un lado, las características asociadas a los datos considerados de alta calidad y, por otro, los procesos implementados para medirla o mejorarla. Esta dualidad implica que la calidad de los datos no solo se refiere a atributos inherentes a la información, sino también a las prácticas destinadas a asegurar que los datos cumplan con los estándares esperados.

La calidad de los datos se determina en función del grado en que estos satisfacen las expectativas y necesidades de los consumidores de datos, es decir, cuando los datos son adecuados para los fines específicos a los que se destinan. En consecuencia, los datos son considerados de baja calidad cuando no cumplen con dichos propósitos. Por lo tanto, la calidad de los datos es contextual y depende de los requisitos definidos por los usuarios de la información.

Uno de los principales desafíos en la gestión de la calidad de los datos radica en que las expectativas de los consumidores no siempre son conocidas o explícitas. En muchos casos, quienes administran los datos no solicitan estos requisitos de manera proactiva. Sin embargo, para garantizar que los datos sean confiables y útiles, los responsables de la gestión de datos deben comprender las necesidades de calidad de sus usuarios y establecer mecanismos que permitan medir y satisfacer dichos requisitos. Este proceso requiere un diálogo constante, dado que las necesidades del negocio y las condiciones del entorno evolucionan con el tiempo.

DAMA International (2017, p. 365) identifica cuatro factores clave que motivan la implementación de un programa formal de gestión de la calidad de los datos. Estos factores, conocidos como *business drivers*, reflejan las razones estratégicas y operativas que justifican la inversión en iniciativas de calidad de datos dentro de las organizaciones.

- **Incrementar el valor de los datos organizacionales:** A través de la mejora continua de la calidad, los datos adquieren mayor valor y utilidad para los procesos de negocio, facilitando su explotación para generar información confiable.
- **Reducir los riesgos y costos asociados a datos de baja calidad:** Los datos inexactos, incompletos o inconsistentes generan costos operativos elevados y derivan en errores de decisión. La gestión adecuada de la calidad de los datos contribuye a mitigar estos riesgos.
- **Mejorar la eficiencia y productividad organizacional:** La disponibilidad de datos confiables permite mejorar los procesos, disminuir retrabajos y fortalecer la capacidad operativa.

- **Proteger y fortalecer la reputación de la organización:** El uso de datos precisos y consistentes respalda la credibilidad de la organización ante sus partes interesadas y reduce la exposición a fallos relacionados con información errónea.

2.1.4 Automatización de procesos de negocio (BPA)

La automatización de procesos de negocio es una estrategia que emplea software para automatizar procesos empresariales complejos y repetitivos, con el propósito de optimizar las operaciones diarias y garantizar la eficiencia organizacional. Esta práctica permite automatizar actividades fundamentales, como el procesamiento de pedidos o la administración de cuentas de clientes, las cuales forman parte de los procesos esenciales para el funcionamiento de la empresa.

Un proceso de negocio consiste en una secuencia de actividades orientadas a alcanzar un objetivo específico, ya sea la producción de bienes, la ejecución de procesos financieros, la incorporación de nuevos empleados o la captación de clientes. Estos procesos, generalmente, abarcan varias áreas organizacionales y admiten automatización total o parcial. Por ejemplo, en la gestión de inventarios, un sistema monitorea los niveles de existencias y genera órdenes de compra de manera automática cuando estos disminuyen por debajo de un umbral predefinido. Asimismo, dicho sistema actualiza la información de productos, elabora informes de tendencias y anticipa la demanda futura.

La automatización de procesos de negocio se caracteriza por su capacidad de integrarse con diversos sistemas de tecnología empresarial, adaptándose a los requerimientos particulares de cada organización. Además, incorpora distintas tecnologías, como la automatización robótica de procesos (RPA), la orquestación de flujos de trabajo, la gestión de procesos de negocio (BPM), la inteligencia artificial y soluciones en la nube. Su principal objetivo radica en mejorar la eficiencia operativa, reducir errores humanos, estandarizar procedimientos y liberar recursos para que los colaboradores se enfoquen en tareas estratégicas.

La automatización de procesos de negocio (BPA) comprende diversas categorías, cada una enfocada en distintos niveles de complejidad, desde la automatización de tareas simples hasta procesos integrales que incorporan tecnologías avanzadas como la inteligencia artificial (IA). A continuación, se describen los principales tipos de BPA:

- **Automatización de tareas:** Consiste en la automatización de actividades manuales individuales dentro de un proceso, con el objetivo de ahorrar tiempo y minimizar errores. Ejemplos comunes incluyen el envío automatizado de correos electrónicos, la generación de documentos, la captura de firmas digitales y la actualización de estados en sistemas.
- **Automatización del flujo de trabajo:** Implica la automatización de una secuencia definida de tareas y actividades, asegurando su ejecución en el orden correcto y facilitando la transición eficiente de una etapa a otra. Algunos flujos de trabajo se ejecutan de manera totalmente automatizada, mientras que otros requieren una combinación de tareas automatizadas e intervención humana, especialmente en actividades que demandan criterio profesional. Por ejemplo, en el procesamiento de pedidos en línea, se automatizan tareas

como el envío de confirmaciones por correo electrónico, la verificación de inventario, la sincronización con pasarelas de pago y la generación de etiquetas de envío.

- **Automatización de procesos:** Consiste en la automatización integral de un proceso de negocio de principio a fin, identificando y automatizando la mayor cantidad posible de componentes, incluyendo tareas discretas y flujos de trabajo que las conectan. El objetivo es optimizar todo el proceso, reducir cuellos de botella y promover la coherencia en toda la organización.
- **Automatización de procesos digitales (DPA):** Amplía el alcance de la BPA tradicional al integrar estrategias de automatización dentro del contexto más amplio de la transformación digital. Busca optimizar procesos de extremo a extremo y mejorar la experiencia del cliente, utilizando tecnología para cerrar la brecha entre iniciativas de automatización individuales y los objetivos digitales generales de la empresa.
- **Automatización inteligente:** Es la forma más avanzada de BPA, combinando la automatización de tareas y procesos con tecnologías sofisticadas como IA, aprendizaje automático (ML), procesamiento de lenguaje natural (PLN) y análisis de datos. Esta modalidad permite ejecutar tareas complejas que requieren capacidad cognitiva y toma de decisiones, como interpretar textos, realizar predicciones basadas en análisis de datos y aprender de decisiones previas para optimizar acciones futuras. Un ejemplo es el uso de *chatbots* impulsados por IA para gestionar consultas rutinarias de clientes, liberando a los agentes humanos para atender asuntos más complejos.

La automatización de procesos de negocio (BPA) ofrece una serie de beneficios clave que fortalecen la eficiencia operativa y competitividad de las organizaciones. Entre los principales beneficios identificados se encuentran:

- **Mejora de la eficiencia y estandarización:** Al disminuir la dependencia de procesos manuales, como el uso intensivo de hojas de cálculo, la BPA permite liberar tiempo para que los colaboradores se concentren en tareas estratégicas. Además, promueve la estandarización de los procesos, facilitando su comprensión, administración y escalabilidad conforme la organización crece.
- **Reducción de costos y aumento de la productividad:** La automatización de procesos contribuye a disminuir los costos operativos y elevar la productividad. Las tareas repetitivas se ejecutan sin errores ni interrupciones, garantizando resultados consistentes. Asimismo, el uso de herramientas de BPA basadas en la nube favorece el acceso centralizado a los datos, habilitando la trazabilidad y la supervisión en tiempo real.
- **Mejor atención al cliente y cumplimiento normativo:** La automatización agiliza los tiempos de respuesta y mejora la precisión en la prestación de servicios. Además, facilita la generación de registros de cumplimiento, lo que permite monitorear el desempeño de los procesos y detectar áreas de mejora.

(Mucci & Stryker, 2024)

2.1.5 Administración de procesos de negocio (BPM)

Dumas et al. (2018, p. 6) definen la Administración de Procesos de Negocio (*Business Process Management*, BPM) como un conjunto estructurado de métodos, técnicas y herramientas destinadas a identificar, descubrir, analizar, rediseñar, ejecutar y monitorear los procesos de negocio, con el propósito de optimizar su desempeño. Esta definición enfatiza que los procesos de negocio constituyen el eje central de la disciplina BPM, abarcando todas las fases y actividades de su ciclo de vida.

Desde esta perspectiva, la administración de procesos no se limita a la mejora aislada de actividades individuales, sino que implica gestionar cadenas completas de eventos, actividades y decisiones que generan valor para la organización y sus clientes. Por consiguiente, BPM integra un enfoque holístico y sistemático para garantizar que los procesos produzcan resultados consistentes, alineados con los objetivos estratégicos de la organización. La disciplina también reconoce la necesidad de identificar oportunidades de mejora continua, ya sea mediante iniciativas incrementales o transformaciones radicales, siempre orientadas a incrementar la eficiencia, reducir costos, minimizar errores y obtener ventajas competitivas.

De acuerdo con IBM (s.f.), la Administración de Procesos de Negocio (BPM) se clasifica en tres categorías principales: centrada en la integración, centrada en el factor humano y centrada en documentos. Estas categorías reflejan la diversidad de enfoques que BPM abarca para mejorar los procesos empresariales.

- **BPM centrado en la integración:** Este enfoque se dirige a procesos que requieren una mínima intervención humana, dependiendo en mayor medida de interfaces de programación de aplicaciones (APIs) y mecanismos que integran datos entre sistemas. Ejemplos comunes incluyen la gestión de recursos humanos (HRM) y la gestión de relaciones con clientes (CRM).
- **BPM centrado en el factor humano:** A diferencia del enfoque anterior, este tipo de BPM prioriza procesos donde la intervención humana es esencial, especialmente en tareas que requieren aprobaciones. Interfaces de usuario intuitivas con funciones de arrastrar y soltar permiten asignar tareas a diferentes roles, facilitando la rendición de cuentas a lo largo del proceso.
- **BPM centrado en documentos:** Este enfoque se focaliza en procesos alrededor de documentos específicos, como contratos. Por ejemplo, al adquirir un producto o servicio, es necesario gestionar diversos formularios y rondas de aprobación para formalizar un acuerdo entre el cliente y el proveedor.

IBM (s.f.) destaca que la adopción de soluciones de Administración de Procesos de Negocio (BPM) genera múltiples beneficios que incrementan el valor organizacional mediante la mejora continua de los procesos. Entre las principales ventajas identificadas se encuentran:

- **Incremento de la eficiencia y reducción de costos:** Los sistemas BPM permiten optimizar los procesos existentes y estructurar adecuadamente el desarrollo de nuevos procesos. La eliminación de redundancias y cuellos de botella favorece la eficiencia operativa y eleva la

productividad. Esta agilidad facilita el cumplimiento de los objetivos empresariales en un menor tiempo, además de posibilitar la reasignación de recursos hacia actividades de mayor prioridad.

- **Mejora en la experiencia de empleados y clientes:** Las herramientas BPM contribuyen a eliminar tareas repetitivas y garantizan un acceso más ágil a la información. Al reducir las distracciones operativas, los colaboradores se concentran en sus funciones y en la atención al cliente, fortaleciendo la satisfacción de los usuarios. Asimismo, la existencia de flujos de trabajo claros acorta el proceso de incorporación de nuevos empleados, incrementando la productividad y el compromiso organizacional.
- **Escalabilidad de los procesos:** La aplicación de BPM favorece la ejecución eficiente de procesos y la automatización de flujos de trabajo, facilitando la expansión de estos procesos a nuevas geografías o contextos. Las herramientas BPM aportan claridad respecto a los roles y responsabilidades, aseguran la coherencia del proceso y permiten incorporar reglas de negocio que promueven la innovación.
- **Transparencia organizacional:** La automatización de procesos de negocio establece de manera explícita los responsables de cada tarea, lo que incrementa la transparencia y la rendición de cuentas a lo largo del proceso. Este enfoque fomenta una comunicación más efectiva entre los equipos de trabajo.
- **Reducción de la dependencia de los equipos de desarrollo:** BPM integra funcionalidades de bajo código que disminuyen la necesidad de intervención de los equipos de desarrollo. Esto permite que los usuarios de negocio adopten rápidamente las herramientas y aceleren la automatización de procesos en toda la organización.

2.1.5.1 Proceso de negocio

Dumas et al. (2018, p. 6), definen un proceso de negocio como un conjunto de eventos, actividades y puntos de decisión interrelacionados, que involucran a diversos actores y objetos, y que en conjunto conducen a un resultado que genera valor para al menos un cliente. Esta definición enfatiza que los procesos de negocio no se limitan a una secuencia de tareas aisladas, sino que comprenden una cadena estructurada de acciones que, de manera coordinada, permiten alcanzar un propósito organizacional.

Los ingredientes esenciales de un proceso de negocio incluyen: eventos que desencadenan o condicionan la ejecución de las actividades, actividades que representan acciones realizadas por personas o sistemas, puntos de decisión que determinan el rumbo del proceso según ciertas condiciones, y actores que participan directa o indirectamente en la ejecución del proceso. Además, cada proceso genera un resultado que entrega valor a uno o varios clientes, lo que justifica su existencia dentro de la organización.

Asimismo, un proceso involucra actores, objetos físicos y objetos informacionales. Los actores incluyen personas, organizaciones o sistemas informáticos que actúan en representación de una organización o individuo. Estos actores se clasifican como internos, conocidos como participantes del proceso, o externos, como proveedores o clientes. Por otra parte, los procesos de negocio integran objetos físicos como equipos o documentos y objetos informacionales, tales como

registros electrónicos o archivos digitales. La ejecución de un proceso genera uno o varios resultados que, idealmente, entregan valor a los actores involucrados. No obstante, en determinadas circunstancias, los resultados no aportan valor alguno para las partes interesadas, dando lugar a un resultado negativo (Dumas et al., 2018, p. 4).

La gestión adecuada de los procesos de negocio es esencial para garantizar que los resultados sean consistentes y alineados con los objetivos estratégicos. Comprender la estructura y dinámica de los procesos permite identificar oportunidades de mejora, aumentar la eficiencia y reforzar la capacidad competitiva de la organización.

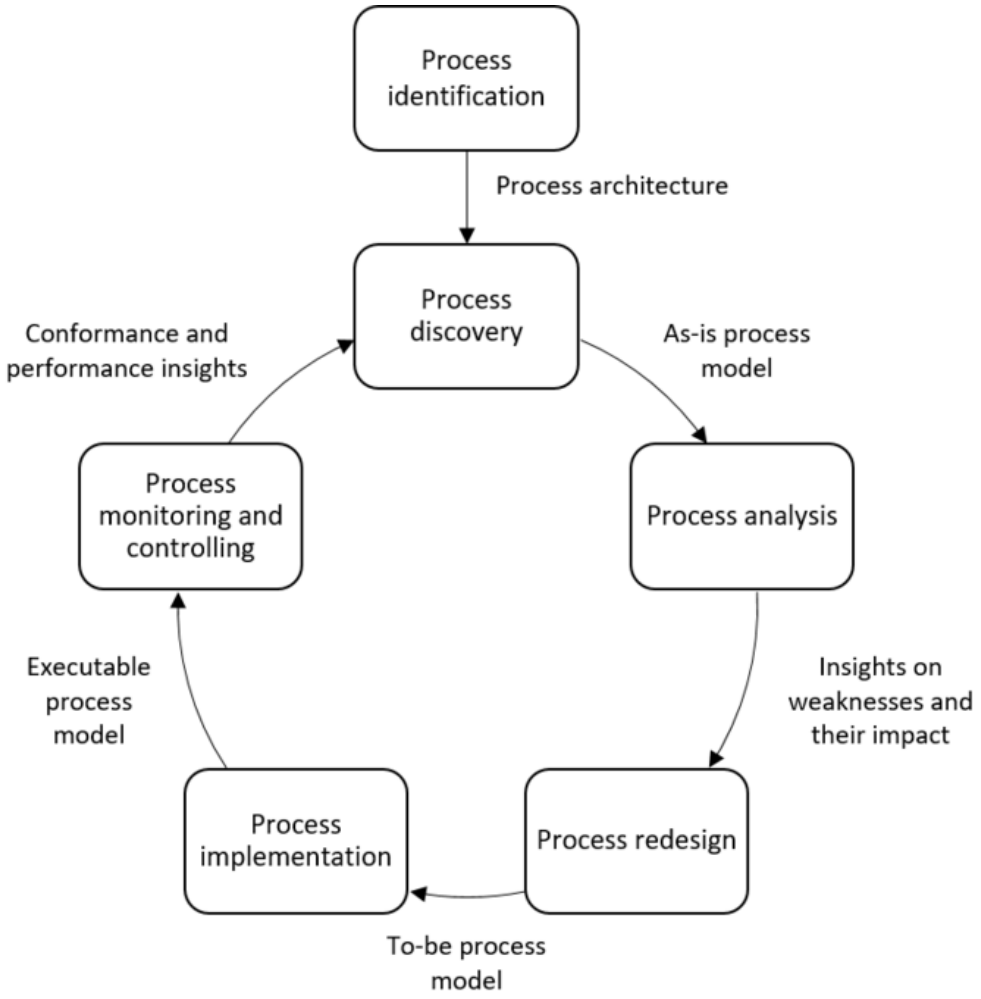
2.1.6 Ciclo de vida de la administración de procesos de negocio

Dumas et al. (2018, pp. 16-24) explican que la Administración de Procesos de Negocio (BPM) se estructura en torno a un ciclo de vida compuesto por seis fases esenciales, concebidas para gestionar de manera sistemática los procesos de negocio. Estas fases permiten identificar, analizar, rediseñar, implementar y monitorear los procesos, asegurando su alineación con los objetivos estratégicos de la organización. El ciclo de vida de BPM incluye:

- **Identificación de procesos:** Esta fase consiste en reconocer, delimitar y relacionar los procesos relevantes para la organización. Su resultado es una arquitectura de procesos que proporciona una visión general de los procesos existentes y sus interrelaciones.
- **Descubrimiento de procesos:** También conocida como modelado del proceso actual (*as-is process modeling*), esta fase documenta el estado actual de los procesos relevantes, generalmente mediante modelos gráficos.
- **Análisis de procesos:** Aquí se identifican y documentan los problemas asociados al proceso actual, así como su impacto en el desempeño organizacional. Estos problemas se priorizan según su efecto potencial y el esfuerzo requerido para resolverlos.
- **Rediseño de procesos:** En esta fase se proponen cambios al proceso con el fin de resolver los problemas detectados y alcanzar los objetivos de desempeño. El rediseño considera múltiples opciones de cambio, las cuales son evaluadas y combinadas para obtener un modelo de proceso mejorado (*to-be process model*).
- **Implementación de procesos:** Consiste en llevar a cabo los cambios necesarios para transitar del proceso actual al rediseñado. La implementación abarca la gestión del cambio organizacional y la automatización de procesos mediante sistemas de información.
- **Monitoreo de procesos:** Una vez implementado el proceso rediseñado, se recopilan y analizan datos relevantes para evaluar su desempeño. Los hallazgos permiten identificar cuellos de botella, errores recurrentes o desviaciones, generando nuevas iniciativas de mejora y reiniciando el ciclo.

A continuación, en la Figura 5 se visualiza el ciclo de vida de BPM.

Figura 5. Ciclo de vida de la Administración de Procesos de Negocio.



Nota. Tomado de Fundamentals of Business Process Management (2nd ed., p. 23), por Dumas et al., 2018, Springer.

2.1.7 Modelo y Notación de Procesos de Negocio (BPMN)

La norma ISO/IEC 19510:2013 (2013, p. 1) define el Modelo y Notación de Procesos de Negocio (BPMN, por sus siglas en inglés: *Business Process Model and Notation*) como un estándar desarrollado por el *Object Management Group* (OMG) que proporciona una notación gráfica diseñada para ser fácilmente comprensible por todos los usuarios de negocio, desde los analistas que elaboran los primeros borradores de los procesos, hasta los desarrolladores técnicos encargados de implementar dichos procesos, así como por los gestores responsables de supervisarlos.

El propósito fundamental de BPMN consiste en establecer un puente estandarizado que cierre la brecha existente entre el diseño de los procesos de negocio y su ejecución tecnológica. Además, asegura que los lenguajes de ejecución de procesos, como WS-BPEL (*Web Services*

Business Process Execution Language), se representen mediante una notación orientada al negocio.

Este estándar consolida las mejores prácticas de la comunidad de modelado de procesos de negocio, integrando múltiples enfoques previos en una notación única y formal. BPMN facilita la construcción de tres tipos principales de diagramas: procesos (*orchestration*), colaboraciones (*collaboration*) y coreografías (*choreography*). A través de estos diagramas, promueve la comunicación clara entre analistas, desarrolladores, clientes y proveedores, favoreciendo la transparencia y eficiencia en la gestión de procesos.

Asimismo, BPMN establece un mecanismo formal que garantiza la portabilidad e interoperabilidad de los modelos de procesos entre distintas herramientas y plataformas, asegurando su correcta interpretación y ejecución en diferentes entornos tecnológicos.

Por otra parte, Stryker & Belcic (2024) destacan que el Modelo y Notación de Procesos de Negocio (BPMN) constituye el estándar global para modelar procesos de negocio y forma parte esencial de la disciplina de Administración de Procesos de Negocio (BPM). Los autores enfatizan que BPMN permite a los interesados visualizar los procesos, facilitando la simplificación de los flujos de trabajo y resolviendo las ambigüedades propias de las especificaciones textuales, mediante una representación gráfica precisa de la secuencia de actividades y los flujos de información necesarios para completar un proceso. Además, subrayan que la especificación BPMN 2.0.1, actualmente mantenida por el *Object Management Group* (OMG), ha sido publicada como estándar internacional bajo la norma ISO/IEC 19510:2013. Este enfoque práctico complementa la perspectiva técnica de la norma, al resaltar la utilidad de BPMN para mejorar la eficiencia operativa, adaptarse a nuevas circunstancias y fortalecer la ventaja competitiva de las organizaciones.

2.1.7.1 Valor del Modelo y Notación de Procesos de Negocio (BPMN)

Stryker & Belcic (2024) destacan que el Modelo y Notación de Procesos de Negocio (BPMN) aporta un valor significativo al proporcionar un lenguaje común para modelar procesos de negocio, el cual resulta comprensible para todos los interesados en la organización. Este lenguaje estandarizado facilita la participación de los analistas de procesos responsables de crear y perfeccionar los procesos, de los desarrolladores técnicos encargados de implementarlos y de los usuarios de negocio que supervisan y gestionan su ejecución. Estos tres grupos constituyen actores clave en la optimización de las operaciones empresariales. La especificación BPMN fue diseñada para asistir a las organizaciones en distintos ámbitos.

- Alcanzar acuerdos ágiles sobre los procesos actuales y futuros mediante modelos claros, no ambiguos.
- Fomentar la participación de los interesados a través de notaciones gráficas expresivas.
- Facilitar el análisis y la mejora de las operaciones mediante la reingeniería de procesos de negocio
- Construir una biblioteca de flujos de procesos, definiciones de casos y reglas de negocio para la capacitación de nuevos empleados.

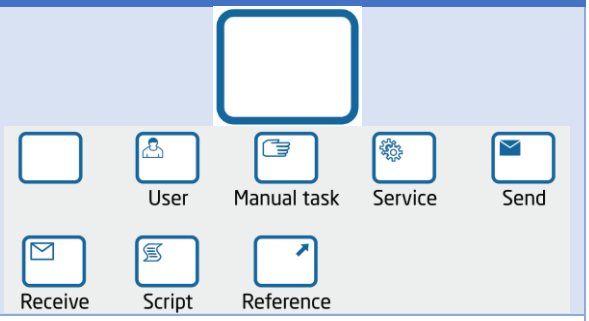
- Reducir las brechas de comunicación mediante un lenguaje común entre analistas de negocio, desarrolladores y otros interesados.
- Orientar los esfuerzos de automatización de procesos de negocio.
- Coordinar estrategias de externalización de procesos.

2.1.7.2 Elementos básicos y gráficos de la notación BPMN

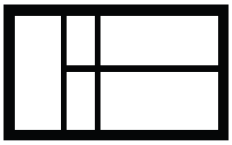
El lenguaje BPMN se basa en diagramas de flujo y notaciones gráficas. La notación estándar utilizada para representar elementos BPMN se divide en varias categorías para la diagramación. A continuación en la Tabla 2, se describen los elementos básicos y gráficos de la notación BPMN.

Tabla 2. Elementos fundamentales de la notación BPMN

Notación BPMN		
Objetos de flujo		
Eventos	Los eventos representan desencadenantes que inician, modifican o concluyen un proceso de negocio. Estos elementos permiten modelar situaciones o condiciones que afectan el flujo del proceso, ya sea para activarlo, interrumpirlo, alterar su ejecución o señalar su finalización.	**Eventos de inicio:** Señalan la instancia o el inicio de un proceso. No poseen ningún flujo de secuencia entrante 
		Eventos intermedios: Señalan situaciones que ocurren o llegan a presentarse durante el proceso, entre el evento de inicio y el de fin. Estos eventos tienen la función de capturar o lanzar un desencadenante y se ubican dentro del flujo de secuencia o en el borde de una actividad. 
		Eventos de fin: Señalan dónde concluye un proceso. Un proceso admite múltiples eventos de fin y estos no poseen flujos de secuencia salientes. 
Actividades	Acciones o tareas ejecutadas durante el proceso de negocio como respuesta a	**Tarea:** Actividad simple utilizada cuando el trabajo dentro del proceso no requiere un mayor nivel de detalle. BPMN establece distintos tipos de tareas.

Notación BPMN		
	eventos de inicio y requisito para eventos de fin. Estas acciones se clasifican como simples o complejas e incluyen subprocesos y repeticiones. Se representan como rectángulos de bordes redondeados.	 **Subproceso:** Actividad compuesta cuyo detalle se define mediante un flujo de otras actividades.  **Subproceso embebido:** Depende totalmente del proceso principal y no contiene piscinas ni carriles.  **Subproceso reutilizable:** Es un proceso independiente, similar a otro diagrama de procesos, que no depende del proceso principal. 
Compuertas	Puntos de decisión en el proceso que requieren elegir una opción para continuar. Se representan con un rombo y dirigen el flujo hacia al menos dos posibles resultados según la decisión tomada	**Compuerta exclusiva basada en datos:** Punto de decisión con dos o más flujos de salida, donde solo uno se toma tras evaluar una condición de negocio. También permite converger y unificar caminos alternativos. 
		Compuerta exclusiva basada en eventos: Elemento de divergencia que permite seleccionar un solo camino del proceso, pero basado en la ocurrencia de un evento, no en una condición de datos.

Notación BPMN		
		 **Compuerta paralela:** Elemento utilizado para crear flujos paralelos o sincronizar múltiples caminos paralelos en uno solo. El flujo continúa cuando todas las secuencias entrantes alcanzan la compuerta.  **Compuerta inclusiva:** Elemento que permite activar una o varias rutas entre varias disponibles, según datos del proceso. También posibilita sincronizar múltiples rutas divergentes en un solo flujo.  **Compuerta compleja:** Elemento utilizado para gestionar puntos de decisión complejos que no se controlan con otros tipos de compuertas. También permite unificar flujos de entrada mediante una expresión que determina cuál debe continuar. 
Piscinas y carriles		
Piscina	Contenedor de un único proceso. Su nombre se considera como el nombre del proceso y siempre debe existir al menos una piscina en el diagrama.	

Notación BPMN		
Carriles	Subdivisión de una piscina que representa un rol o un área organizacional dentro del proceso.	
Conectores		
Flujo de secuencia	Elemento que muestra el orden en que se ejecutan las actividades de un proceso. Representa la secuencia entre actividades, compuertas y eventos.	
Flujo de mensaje	Elemento que muestra el intercambio de mensajes entre dos entidades o procesos. Representa la comunicación, no el control del flujo.	
Asociación	Elemento utilizado para vincular información y artefactos con los objetos de flujo en un proceso.	
Artefactos		
Anotación	Elemento utilizado para brindar información adicional sobre el proceso con el fin de facilitar su comprensión.	
Grupo	Mecanismo visual que permite agrupar actividades con fines de documentación o análisis.	
Objeto de datos	Elemento que proporciona información sobre los datos que ingresan y salen de una actividad.	

Nota. La información presentada en la tabla se elaboró a partir de los contenidos de Stencil BPMN (2013) y Stryker - Belcic (2024).

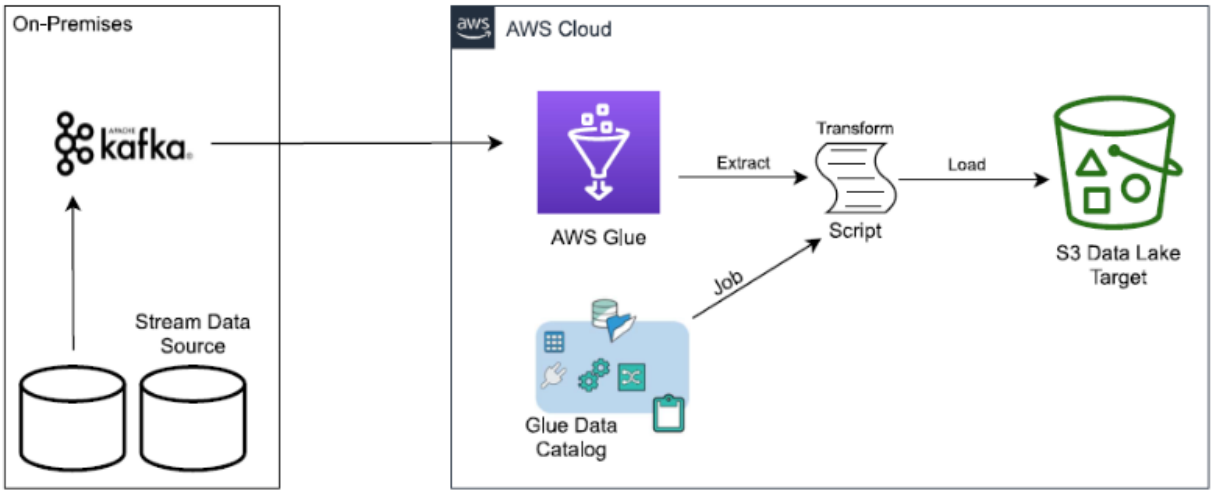
2.1.8 Proceso ETL (*Extract, Transform, Load*)

Según Amazon Web Services (s.f.), la extracción, transformación y carga (ETL) es un proceso mediante el cual se integran datos procedentes de diferentes orígenes en un repositorio central denominado almacenamiento de datos. Este proceso aplica un conjunto de reglas de negocio destinadas a limpiar, organizar y preparar los datos en bruto, con el objetivo de facilitar su almacenamiento, análisis y posterior aplicación en modelos de aprendizaje automático. Además, el proceso ETL permite satisfacer necesidades específicas de inteligencia empresarial, ya que posibilita la generación de informes, paneles de visualización y la identificación de oportunidades para reducir ineficiencias operativas.

La aplicación del proceso de extracción, transformación y carga (ETL) resulta esencial para las organizaciones, ya que estas gestionan datos estructurados y no estructurados procedentes de diversas fuentes, tales como sistemas de administración de la relación con el cliente (CRM), plataformas de pago en línea, sistemas de proveedores, sensores de dispositivos del Internet de las Cosas (IoT), redes sociales o sistemas internos de recursos humanos. Mediante ETL, es posible consolidar y transformar estos datos en bruto, generando un formato estructurado y accesible que facilite su análisis. Este proceso contribuye a convertir grandes volúmenes de información dispersa en insumos valiosos para la toma de decisiones estratégicas.

El proceso de extracción, transformación y carga (ETL) funciona mediante la transferencia periódica de datos desde un sistema de origen hacia un sistema de destino. Este procedimiento se estructura en tres etapas fundamentales. En primer lugar, se realiza la extracción de los datos relevantes contenidos en las bases de datos de origen. Posteriormente, dichos datos son sometidos a un proceso de transformación, mediante el cual se adaptan y preparan para su análisis. Finalmente, se procede con la carga de los datos transformados en el repositorio o base de datos de destino, donde quedan disponibles para su consulta y análisis estratégico. En la Figura 6, se visualiza un ejemplo de proceso ETL.

Figura 6. Ejemplo de proceso ETL, utilizando servicios de AWS



Nota. Tomado de Amazon Web Services (s.f.).

2.1.9 Prototipo de software

Según Suranto (2015, p. 150), un prototipo de software es una versión inicial del sistema diseñada para demostrar conceptos, explorar alternativas de diseño y comprender mejor el problema junto con sus posibles soluciones. Esta técnica resulta especialmente útil en las etapas tempranas del desarrollo, ya que permite a los ingenieros de software obtener y validar los requerimientos del sistema con los usuarios involucrados. El prototipado también contribuye a reducir costos y errores, al facilitar la detección temprana de inconsistencias o malentendidos antes de avanzar a fases más complejas del desarrollo.

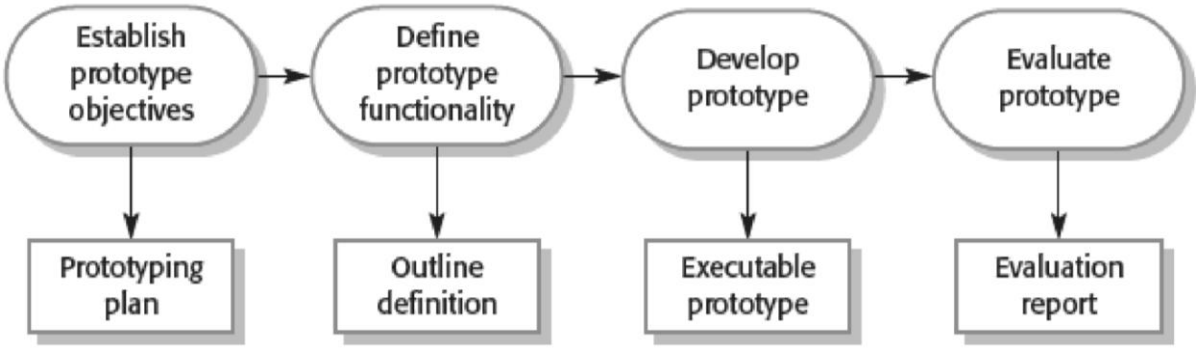
Los prototipos de software se emplean como herramienta clave durante el proceso de ingeniería de requerimientos. Su aplicación permite a los ingenieros de software obtener y validar los requerimientos del sistema directamente con los usuarios, lo cual resulta esencial para reducir ambigüedades y errores en etapas tempranas del desarrollo. Además, los prototipos facilitan la exploración de alternativas de diseño, así como la evaluación de la usabilidad de las interfaces propuestas.

En el diseño funcional, estos modelos preliminares permiten simular el comportamiento de características específicas del sistema, identificar posibles incompatibilidades entre componentes y observar el rendimiento de funciones combinadas. Así, se facilita la detección de errores, limitaciones o malentendidos antes de proceder a la implementación definitiva. También ofrecen un medio práctico para integrar retroalimentación de los usuarios, quienes interactúan con el prototipo y sugieren mejoras o requerimientos adicionales.

El prototipo cumple, por tanto, una doble función: por un lado, sirve como medio de validación temprana de los requerimientos; por otro, actúa como soporte en el diseño técnico, contribuyendo a mejorar la calidad del sistema y reducir los costos asociados a retrabajos en fases

posteriores del proyecto. A continuación en la Figura 7, se muestra un modelo general para el desarrollo de prototipos.

Figura 7. Proceso de desarrollo de un prototipo de software



Nota. La figura representa las cuatro etapas del desarrollo de un prototipo: establecimiento de objetivos, definición de funcionalidades, construcción del prototipo y evaluación final. Tomado de Suranto (2015).

Suranto (2015, pp. 151-152) distingue dos categorías principales en el desarrollo de prototipos: prototipos de baja fidelidad y prototipos de alta fidelidad. La elección entre uno u otro depende de los objetivos específicos del prototipo, el nivel de detalle requerido y la etapa del proceso de desarrollo en la que se utilice.

Los prototipos de baja fidelidad se caracterizan por su simplicidad y velocidad de elaboración. Generalmente consisten en representaciones visuales básicas, elaboradas con materiales accesibles como papel, pizarras o tarjetas. Este tipo de prototipo se emplea para representar ideas generales del diseño, facilitar la discusión inicial con los usuarios y fomentar procesos iterativos con bajo costo de producción. Sin embargo, no permite validar a fondo la interacción funcional del sistema ni garantizar la viabilidad de las funcionalidades simuladas. Su uso se limita a la etapa exploratoria del diseño, especialmente en requerimientos de interfaz de usuario.

En contraste, los prototipos de alta fidelidad presentan un aspecto y comportamiento más cercano al sistema final. Son desarrollados mediante herramientas de software especializadas y permiten una interacción realista por parte de los usuarios. Este tipo de prototipo resulta útil para validar funcionalidades específicas, verificar requerimientos técnicos y recoger retroalimentación detallada. Su implementación, aunque más costosa y demandante en términos técnicos, ofrece una experiencia cercana al producto final, lo que facilita decisiones informadas antes de la implementación definitiva.

Ambos tipos de prototipos cumplen funciones complementarias en el ciclo de vida del software. Mientras los de baja fidelidad permiten iterar rápidamente en fases tempranas, los de alta fidelidad aportan precisión y validación funcional en etapas más avanzadas del desarrollo.

2.1.10 Pruebas de prototipo

Según Wickramasinghe (2024), las pruebas de prototipo constituyen una fase esencial dentro del desarrollo de software, orientada a evaluar una versión preliminar e interactiva del producto con el fin de validar decisiones de diseño, experiencia de usuario y funcionalidad general antes de la implementación definitiva. Esta técnica permite presentar el prototipo a usuarios representativos y recopilar retroalimentación directa que facilita la identificación temprana de errores, deficiencias o mejoras necesarias.

Las pruebas de prototipo no solo permiten validar el diseño de una aplicación de software, sino que también benefician a todos los actores involucrados en el proyecto, incluyendo desarrolladores, analistas de negocio, clientes y usuarios finales. Esta estrategia desempeña un papel crucial a lo largo de todo el ciclo de desarrollo, desde la recopilación de requerimientos hasta la detección y corrección de defectos.

Entre las principales razones que justifican su importancia, se destaca su capacidad para ahorrar tiempo y evitar cambios costosos en fases avanzadas. Asimismo, promueve una mayor participación de los usuarios, lo que incrementa la confianza del cliente en el producto. Las pruebas permiten también recopilar requerimientos reales y obtener retroalimentación temprana, elementos fundamentales para orientar el desarrollo de forma efectiva.

Además, este tipo de evaluación contribuye a resolver conflictos funcionales en etapas iniciales, mejorando la alineación entre los objetivos del sistema y las necesidades del usuario. A través de esta metodología, se refuerza la experiencia del usuario y se favorece la construcción de un producto de alta calidad que cumpla con las expectativas del público final.

2.1.10.1 ¿Cómo se prueba un prototipo?

Según Wickramasinghe (2024), la prueba de un prototipo es un proceso secuencial que debe seguir una serie de pasos estructurados para garantizar su efectividad. Cada etapa tiene como objetivo validar aspectos específicos del diseño y funcionalidad del prototipo en relación con los requerimientos del usuario. El procedimiento incluye las siguientes fases:

- **Paso 1. Recopilar y analizar la información del usuario:** La primera etapa consiste en reunir datos relevantes proporcionados por los usuarios y analizarlos con detenimiento. Esta información incluye retroalimentación inicial y requerimientos funcionales. Una comprensión clara de las necesidades del usuario es esencial para las etapas posteriores del proceso.
- **Paso 2. Diseñar el prototipo:** Con base en los datos obtenidos, se procede a construir el prototipo que será objeto de prueba. Inicialmente, se elabora un diseño conceptual, al cual se le incorporan características más interactivas y representativas del producto final. Para incrementar la precisión de las pruebas, se sugiere utilizar datos reales, si es necesario.
- **Paso 3. Determinar qué se va a probar:** Antes de utilizar herramientas de prueba, el equipo debe definir con claridad los aspectos del prototipo que serán evaluados. Entre los elementos comunes se encuentran la funcionalidad según los requerimientos, la correcta

navegación entre páginas, la fluidez del flujo de trabajo y la ubicación adecuada de componentes como botones o etiquetas.

- **Paso 4. Crear un diseño preliminar:** Se desarrolla una versión simplificada del prototipo final, que sirve como referencia inicial para los usuarios. Esta versión consiste en bocetos a mano o esquemas en papel que representen distintas alternativas de diseño.
- **Paso 5. Elaborar escenarios de prueba:** En esta etapa, se diseñan escenarios que simulen situaciones reales de uso, considerando las funcionalidades iniciales del sistema, las expectativas del usuario y la retroalimentación previa. Los escenarios deben reflejar condiciones prácticas y ayudar a evaluar el comportamiento del sistema ante situaciones concretas.
- **Paso 6. Recoger retroalimentación del usuario:** Finalmente, se realiza una evaluación inicial con los usuarios finales. Se presenta el prototipo con el objetivo de recoger comentarios, observaciones y sugerencias. Esta información resulta fundamental para mejorar la versión evaluada, reducir ambigüedades y corregir defectos de requerimientos antes de avanzar con el desarrollo.

2.2 Herramientas y tecnologías

Esta sección describe las herramientas y tecnologías de automatización consideradas por el equipo de *Data Operations* de la empresa del sector financiero en Costa Rica. La selección de dichas herramientas se enmarca dentro del contexto de la automatización del proceso de carga de datos, con el propósito de identificar alternativas viables que respondan a los requerimientos técnicos y funcionales del proyecto.

2.2.1 Amazon Web Services

Según Amazon Web Services (2024, p. 1), AWS es una plataforma integral de servicios de computación en la nube que proporciona productos globales en áreas como cómputo, almacenamiento, bases de datos, redes, análisis, desarrollo de aplicaciones, herramientas de gestión, Internet de las Cosas (IoT), inteligencia artificial y aplicaciones empresariales. Estos servicios se ofrecen bajo demanda, con disponibilidad inmediata y precios ajustados al consumo.

Desde su lanzamiento en 2006, AWS ha permitido a las organizaciones reemplazar inversiones de capital en infraestructura por costos variables que escalan según la demanda del negocio. Esta capacidad de escalar recursos en cuestión de minutos, en lugar de semanas o meses, ha transformado la forma en que empresas de todos los tamaños acceden a capacidades tecnológicas, reduciendo significativamente el tiempo y costo para obtener resultados.

Actualmente, AWS proporciona una infraestructura confiable, escalable y de bajo costo que impulsa a cientos de miles de empresas en más de 190 países. Su modelo de servicio permite a organizaciones del sector privado, gubernamental y sin fines de lucro acceder a herramientas tecnológicas avanzadas sin necesidad de gestionar infraestructura física, favoreciendo así la agilidad, innovación y eficiencia operativa.

2.2.1.1 Amazon S3

Amazon Web Services (2025, pp. 1-6) describe *Amazon Simple Storage Service* (Amazon S3) como un servicio de almacenamiento de objetos que ofrece escalabilidad, alta disponibilidad, seguridad y rendimiento líderes en la industria. Está diseñado para almacenar grandes volúmenes de datos, siendo ampliamente utilizado en casos como *data lakes*, aplicaciones móviles, respaldo y recuperación, archivado, sistemas empresariales, dispositivos IoT y análisis de big data.

El modelo de almacenamiento de Amazon S3 organiza los datos en objetos dentro de buckets. Cada objeto incluye los datos en sí, sus metadatos y un identificador único (clave). Los buckets, como contenedores de almacenamiento, se crean en una región específica de AWS y ofrecen funcionalidades avanzadas de configuración, seguridad y gestión.

El sistema permite activar características como versionado de objetos (*S3 Versioning*), que conserva múltiples versiones de un archivo, facilitando la recuperación ante eliminaciones o modificaciones accidentales. También admite configuraciones de acceso privado mediante políticas, control de acceso a nivel de objeto y uso de puntos de acceso, lo que garantiza la seguridad y gobernanza de la información almacenada.

Este diseño, basado en principios de durabilidad y disponibilidad distribuidas, permite que las organizaciones gestionen sus datos con flexibilidad, minimizando los riesgos operativos y los costos de infraestructura

2.2.1.2 AWS Landing Zone

AWS Landing Zone es un entorno multi-cuenta bien arquitectado en Amazon Web Services (AWS), escalable y seguro, que establece una base estructurada desde la cual las organizaciones implementan sus cargas de trabajo y aplicaciones con confianza en la seguridad e infraestructura desplegada. Su configuración implica decisiones técnicas y estratégicas sobre la organización de cuentas, el diseño de red, la seguridad de los datos, así como la administración de accesos, todo ello alineado con los objetivos de expansión de la empresa.

Este entorno incorpora elementos clave como la gobernanza de recursos, la protección de datos, la gestión de identidades y el registro centralizado de eventos. Su implementación se realiza mediante AWS Control Tower, que automatiza la configuración inicial con controles predefinidos, o mediante una solución personalizada, diseñada para satisfacer necesidades específicas. La landing zone se fundamenta en un marco multi-cuenta, el cual asegura el aislamiento de recursos, mejora la seguridad y facilita el cumplimiento normativo a gran escala (Amazon Web Services, 2025, pp. 2-4).

2.2.1.3 AWS Step Functions

Según Amazon Web Services (2024, p. 21), AWS Step Functions es un servicio completamente gestionado que permite coordinar de forma visual los distintos componentes de una aplicación distribuida o basada en microservicios. A través de flujos de trabajo definidos paso a paso, este servicio facilita la construcción de soluciones escalables y modulares, sin requerir la gestión directa de infraestructura subyacente.

Step Functions proporciona una interfaz gráfica en la que se representan las distintas etapas del flujo como una secuencia lógica de eventos. Cada paso es monitoreado individualmente y, en caso de fallos, el sistema ejecuta reintentos automáticos y mantiene un seguimiento completo del estado, lo que facilita el control, la depuración y la trazabilidad.

El servicio permite modificar y extender flujos sin necesidad de escribir código adicional, lo que agiliza el mantenimiento y evolución de aplicaciones complejas. Esta capacidad lo convierte en una herramienta adecuada para automatizar procesos empresariales, orquestar tareas en la nube y gestionar integraciones entre servicios dentro del ecosistema AWS.

2.2.1.4 AWS Lambda

La documentación oficial de Amazon Web Services (2024, p. 40) describe AWS Lambda como un servicio de computación sin servidor (*serverless*) diseñado para ejecutar código en respuesta a eventos, sin necesidad de aprovisionar ni administrar servidores. Este modelo de ejecución se basa en el consumo real de recursos, lo que significa que el costo se calcula únicamente por el tiempo de ejecución efectivo del código.

AWS Lambda admite distintos lenguajes de programación y permite desencadenar funciones a partir de eventos provenientes de otros servicios de AWS, aplicaciones web, móviles o flujos automatizados. Una vez cargado el código, el entorno de Lambda se encarga de gestionar la ejecución, la disponibilidad y la escalabilidad, sin intervención adicional por parte del usuario.

Este servicio resulta especialmente útil en arquitecturas basadas en microservicios, procesamiento de datos en tiempo real, automatización de tareas y desarrollo de backends ligeros. Su integración nativa con múltiples servicios de AWS lo convierte en una herramienta estratégica para soluciones ágiles, modulares y orientadas a eventos.

2.2.1.5 Amazon Aurora

La documentación oficial de Amazon Web Services (2024, pp. 49-50) describe Amazon Aurora como un motor de base de datos relacional compatible con MySQL y PostgreSQL que combina la velocidad y disponibilidad de las bases de datos comerciales de alto nivel con la simplicidad y eficiencia en costos de las soluciones de código abierto.

Aurora ofrece un rendimiento significativamente superior: hasta cinco veces más rápido que MySQL estándar y tres veces más que PostgreSQL estándar. Este servicio es gestionado por *Amazon Relational Database Service* (RDS), lo que permite automatizar tareas administrativas como la provisión de hardware, configuración de la base de datos, actualizaciones y copias de seguridad.

Incorpora una arquitectura de almacenamiento distribuido, tolerante a fallos y autorreparable, que escala automáticamente hasta 128 TB por instancia de base de datos. Además, admite hasta 15 réplicas de lectura de baja latencia, recuperación en puntos específicos en el tiempo, respaldos continuos hacia Amazon S3 y replicación entre tres zonas de disponibilidad (AZs).

2.2.1.6 Amazon CloudWatch

Amazon CloudWatch es un servicio nativo de monitoreo y gestión ofrecido por Amazon Web Services (AWS), diseñado para brindar visibilidad operativa en tiempo real sobre los recursos de la nube, las aplicaciones y los flujos automatizados. Este servicio permite recolectar métricas clave, establecer alarmas, visualizar paneles personalizados y ejecutar acciones automatizadas en función del estado de los componentes monitoreados. Su utilidad radica en la capacidad para identificar comportamientos anómalos, analizar tendencias de rendimiento, optimizar el uso de recursos y garantizar la continuidad operacional de soluciones distribuidas (Amazon Web Services, p. 95). En el contexto de arquitecturas *serverless*, como las funciones Lambda y las Step Functions utilizadas en este proyecto, CloudWatch ofrece datos críticos para evaluar la eficiencia y precisión del flujo.

3 Marco Metodológico

El presente capítulo detalla el enfoque metodológico adoptado para el desarrollo del proyecto, definiendo criterios como el tipo de investigación, su alcance, las técnicas e instrumentos utilizados, así como el procedimiento seguido en cada una de las fases. Se establece una estructura metodológica alineada con los objetivos específicos del trabajo, permitiendo garantizar la validez de los resultados obtenidos. Además, se describe el proceso de operacionalización de variables y la forma en que se vinculan con los entregables y hallazgos alcanzados a lo largo del estudio.

3.1 Tipo de Investigación

De acuerdo con Hernández Sampieri, Fernández y Baptista (2014), existen dos tipos principales de investigación: la investigación básica, cuyo propósito es "producir conocimiento y teorías", y la investigación aplicada, que tiene como objetivo "resolver problemas" (p. xxiv). Huaire Inacio (2019) complementa esta idea al señalar que la investigación básica se orienta a generar conocimiento nuevo sobre un fenómeno o hecho, mientras que la investigación aplicada, frecuentemente derivada de la básica, busca ofrecer soluciones prácticas a problemas específicos, utilizando estrategias definidas (Huaire Inacio, 2019, p. 8).

El presente proyecto se clasifica como una investigación aplicada, ya que tiene como objetivo resolver un problema práctico relacionado con la mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros. Según Hernández Sampieri, Fernández y Baptista (2014), la investigación aplicada busca "resolver problemas" (p. xxiv), lo cual se alinea directamente con este proyecto, cuyo propósito es desarrollar una propuesta de solución tecnológica en forma de prototipo que elimine tareas manuales críticas, reduzca errores operativos y asegure la precisión de los datos maestros. Además, como menciona Huaire Inacio (2019), este tipo de investigación utiliza estrategias definidas para abordar necesidades concretas, como la integración de herramientas tecnológicas que permitan automatizar pasos de carga y transformación de datos.

En este contexto, el proyecto no solo responde a una necesidad organizacional específica del equipo de *Data Operations*, sino que también se proyecta hacia otras áreas de negocio que dependen de datos maestros confiables para operar con eficiencia. La mejora en la calidad, consistencia y disponibilidad de los datos impactará positivamente en procesos clave de toma de decisiones, auditoría, cumplimiento normativo y análisis estratégico. Así, el alcance de la solución trasciende el equipo, al fortalecer la infraestructura informacional de la organización. Este enfoque práctico reafirma que la investigación aplicada es el marco más adecuado para abordar la problemática identificada.

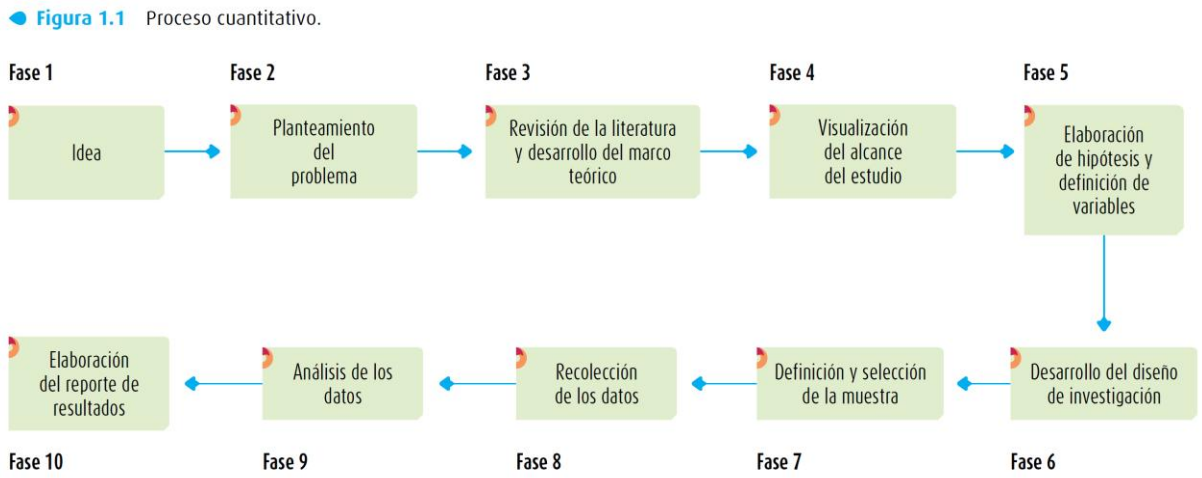
3.2 Enfoque de la Investigación

Los enfoques de investigación representan las estrategias generales que orientan el proceso de recolección y análisis de datos en un estudio. Según Hernández Sampieri, Fernández y Baptista (2014), "existen tres enfoques principales: cuantitativo, cualitativo y mixto", los cuales presentan características y objetivos específicos que responden a diferentes planteamientos y necesidades de investigación (p. 4). Estos enfoques ofrecen diversas perspectivas para abordar los problemas de

estudio, desde el análisis numérico y estadístico, hasta la interpretación de fenómenos a través de la observación y exploración profunda, e incluso combinando ambos métodos para lograr una comprensión más integral. A continuación, se describen en detalle cada uno de estos enfoques.

- **Enfoque cuantitativo:** Utiliza la recolección de datos para probar hipótesis con base en la medición numérica y el análisis estadístico, con el fin establecer pautas de comportamiento y probar teorías (p. 4). A continuación en la Figura 8, se visualiza el proceso cuantitativo en una investigación.

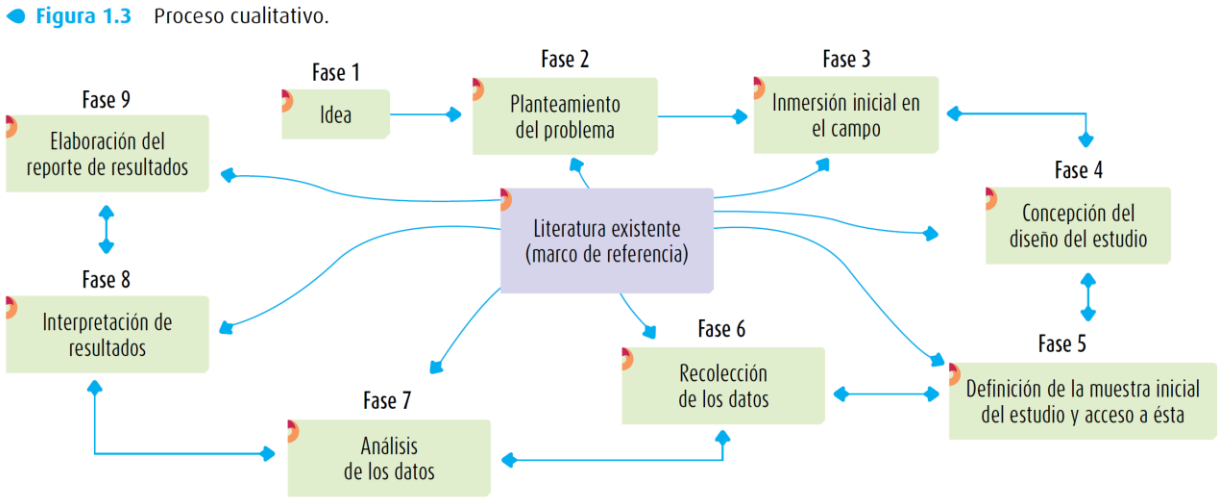
Figura 8. Proceso cuantitativo



Fuente: Hernández Sampieri, Fernández y Baptista (2014)

- **Enfoque cualitativo**: Utiliza la recolección y análisis de los datos para afinar las preguntas de investigación o revelar nuevas interrogantes en el proceso de interpretación (p. 7). A continuación, en la Figura 9 se visualiza el proceso cualitativo en una investigación.

Figura 9. Proceso cualitativo



Fuente: Hernández Sampieri, Fernández y Baptista (2014)

- El enfoque mixto combina procesos sistemáticos, empíricos y críticos de recolección y análisis de información tanto cuantitativa como cualitativa, con el fin de responder a planteamientos de investigación (p. 534). El enfoque mixto, entre otros aspectos, logra una perspectiva más amplia y profunda del fenómeno, ayuda a formular el planteamiento del problema con mayor claridad, produce datos más "ricos" y variados, potencia la creatividad teórica, apoya con mayor solidez las inferencias científicas y permite una mejor "exploración y explotación" de los datos. (p. 580)

El presente proyecto adopta un enfoque mixto, al integrar componentes cualitativos y cuantitativos para abordar de forma integral la mejora del proceso de carga de datos en una plataforma de Gestión de Datos Maestros (MDM). Desde la dimensión cualitativa, se analiza el proceso actual con el fin de identificar deficiencias, tareas manuales críticas y necesidades específicas del equipo de Data Operations. Este análisis permite comprender cómo las dinámicas internas afectan la eficiencia operativa y la calidad de los datos, además de ofrecer insumos clave para diseñar una solución tecnológica adaptada al contexto organizacional.

De forma complementaria, se incorporan métricas cuantitativas para evaluar el desempeño del prototipo. Entre ellas destacan la duración de los procesos, la cantidad de errores corregidos y la reducción de actividades manuales. Estas mediciones permiten validar el impacto real de la automatización propuesta. Según Hernández Sampieri, Fernández y Baptista (2014), el enfoque mixto combina procesos sistemáticos, empíricos, críticos, así como técnicas de recolección junto con análisis de datos cualitativos y cuantitativos. Esta integración permite una comprensión más profunda del fenómeno (p. 534). En este caso, dicha integración fortalece la capacidad del estudio para generar hallazgos útiles, transferibles y alineados con futuras iniciativas de mejora organizacional.

3.3 Diseño de la investigación

El diseño de investigación representa el plan estratégico que guía la recolección, análisis e interpretación de datos en un estudio. Según Hernández Sampieri, Fernández y Baptista (2014), entre los principales diseños cualitativos se encuentran la teoría fundamentada, el etnográfico, el narrativo, el fenomenológico y la investigación-acción, cada uno de los cuales responde a preguntas específicas y se adapta a las necesidades particulares de la investigación. A continuación en la Tabla 3, se describen estos diseños, destacando sus características.

Tabla 3. Diseños de investigación

Diseño	Información que proporciona	Pregunta de investigación
Teoría fundamentada	Categorías del proceso o fenómeno y sus vínculos. Teoría que explica el proceso o fenómeno (problema de investigación).	Preguntas sobre procesos y relaciones entre conceptos que conforman un fenómeno.
Etnográfico	Descripción y explicación de los elementos y categorías que integran al sistema social: historia y evolución, estructura (social, política, económica, etc.), interacciones, lenguaje, reglas y normas, patrones de conducta, mitos y ritos.	Preguntas sobre las características, estructura y funcionamiento de un sistema social (grupo, organización, comunidad, subcultura, cultura), desde una familia, hermandad o hinchada hasta una megaciudad.
Narrativo	Historias sobre procesos, hechos, eventos y experiencias, siguiendo una línea de tiempo, ensambladas en una narrativa general. Categorías relacionadas con tales historias y narrativa.	Preguntas orientadas a comprender una sucesión de eventos, a través de las historias o narrativas de quienes la vivieron (experiencias de vida bajo una secuencia cronológica). Eventos como una catástrofe, una elección, la biografía de un individuo, etcétera.
Fenomenológico	Experiencias comunes y distintas. Categorías que se presentan frecuentemente en las experiencias.	Preguntas sobre la esencia de las experiencias: lo que varias personas experimentan en común respecto a un fenómeno o proceso.
Investigación-acción	Diagnóstico de problemáticas sociales, políticas, laborales,	Preguntas sobre problemáticas o situaciones

Diseño	Información que proporciona	Pregunta de investigación
	económicas, etc., de naturaleza colectiva. Categorías sobre las causas y consecuencias de las problemáticas y sus soluciones.	de un grupo o comunidad (incluyendo cambios).

Nota: Adaptado de Hernández Sampieri, Fernández y Baptista (2014, p. 471)

El diseño de investigación-acción es el más adecuado para el presente proyecto, ya que este busca no solo comprender, sino también resolver una problemática específica relacionada con la mejora del proceso de carga de datos en una plataforma de Gestión de Datos Maestros (MDM). Este diseño se alinea perfectamente con la naturaleza del proyecto, dado que implica un análisis de la situación actual (estado *As-Is*), la identificación de áreas de mejora, y el desarrollo de una propuesta de solución concreta basada en herramientas tecnológicas para automatizar tareas críticas.

La investigación-acción promueve la intervención directa en el entorno organizacional, mediante un ciclo iterativo de diagnóstico, acción y evaluación. En este caso, se asegura la participación activa del equipo de *Data Operations* a través de espacios formales de retroalimentación, sesiones de validación funcional del prototipo y discusiones técnicas sobre su impacto. Estas instancias permiten incorporar el conocimiento operativo del equipo en la toma de decisiones, ajustar la solución según sus observaciones y validar su aplicabilidad dentro del flujo real de trabajo.

Este enfoque no solo facilita la transformación del proceso de carga de datos, sino que también genera un impacto directo en la eficiencia operativa y en la calidad de los datos, contribuyendo al logro de los objetivos del equipo y fomentando una cultura de mejora continua.

3.4 Fuentes de datos e información

A continuación, se presentan las fuentes de información clave para la elaboración del proyecto, clasificadas en dos categorías principales: fuentes primarias y fuentes secundarias.

3.4.1 Fuentes primarias

Las fuentes primarias se definen como aquellas que contienen información original, es decir, son de primera mano y representan el resultado directo de ideas, conceptos, teorías o investigaciones realizadas. Según Maranto Rivera y González Fernández (2015), “este tipo de fuentes contienen información original, es decir, son de primera mano, son el resultado de ideas, conceptos, teorías y resultados de investigaciones. Contienen información directa antes de ser interpretada o evaluada por otra persona” (p. 3). En la Tabla 4 se presentan las fuentes primarias a utilizar en la elaboración del proyecto.

Tabla 4. Fuentes de información primaria

Fuente de información	Importancia del documento para la investigación
Archivos con datos maestros proporcionados por el equipo de *Data Operations*	Estos archivos, entregados por el equipo de *Data Operations*, contienen registros maestros en formatos estructurados como CSV. Su análisis resultó clave para identificar inconsistencias, tareas manuales frecuentes y limitaciones técnicas presentes en el proceso actual de carga.
Plataforma de Gestión de Datos Maestros	La interacción directa con la plataforma utilizada por el equipo *Data Operations* permitió observar su funcionamiento, limitaciones técnicas, estructura de carga y lógica de validación. Esta observación fue clave para el análisis y la propuesta de solución.
Documentación interna del proceso actual.	Aunque no se trata de una interacción directa con el equipo, la revisión de documentos internos que describen las prácticas actuales del proceso constituye una fuente primaria. Esto incluye manuales técnicos, registros operativos y reportes previos.
Sesiones de trabajo con el equipo *Data Operations*	Estas sesiones fueron fundamentales para comprender el estado actual del proceso de carga de datos, identificar deficiencias, pasos manuales, y obtener detalles sobre las herramientas tecnológicas utilizadas. Además, permitieron recopilar información directa sobre los requerimientos específicos del equipo y la organización.

Nota. Elaboración propia (2025)

La participación activa de los actores clave se asegurará mediante espacios estructurados de interacción directa. Se prevé la realización de sesiones de trabajo con el equipo de *Data Operations*, en las que se discutirán los principales desafíos del proceso, se identificarán tareas susceptibles de automatización y se validarán escenarios operativos reales. Estas reuniones no solo permitirán recopilar información relevante, sino también involucrar a los participantes en la construcción colaborativa de soluciones. Adicionalmente, se revisarán documentos internos como reportes de errores frecuentes, manuales de operación y registros de cargas históricas, los cuales aportarán evidencia concreta sobre los puntos críticos del proceso actual. Esta combinación de interacción directa y revisión documental fortalecerá la calidad de la información obtenida y reflejará el carácter participativo propio del diseño de investigación-acción.

3.4.2 Fuentes secundarias

Las fuentes secundarias son aquellas que interpretan, analizan o sintetizan información previamente obtenida de fuentes primarias. Estas no contienen datos originales, sino que presentan un tratamiento derivado o complementario de información existente. Según Maranto Rivera y González Fernández (2015), "las fuentes secundarias son interpretaciones y análisis de información basada en fuentes primarias. Estas se presentan en forma de reseñas, resúmenes, críticas, interpretaciones o análisis de contenido previamente publicado" (p. 3). En la Tabla 5 se presentan las fuentes secundarias a utilizar en la elaboración del proyecto.

Tabla 5. Fuentes de información secundaria

Fuente de información	Importancia del documento para la investigación
Data Management Body of Knowledge (DMBOK)	Este marco de referencia proporciona las mejores prácticas y lineamientos para la gobernanza, calidad y gestión de datos, aspectos clave en la automatización del proceso de carga de datos en la plataforma de Gestión de Datos Maestros. Es esencial para estructurar el marco conceptual del proyecto y definir estándares de calidad aplicables.
Libros, artículos, foros y publicaciones sobre automatización de procesos y gestión de datos maestros	Ofrecen conocimientos teóricos, prácticos y contextuales que complementan el análisis específico del proceso de carga de datos.
Libros, artículos, foros y manuales sobre herramientas tecnológicas a utilizar.	Elementos cruciales para entender las capacidades y limitaciones de las herramientas tecnológicas utilizadas en el proyecto, sirviendo como guía técnica para implementar las soluciones automatizadas.

Nota: Elaboración propia (2025)

Como parte del proceso de análisis documental, se consultarán fuentes reconocidas que respalden tanto el marco teórico como los lineamientos técnicos del proyecto. Entre ellas se incluirán las versiones más recientes del *Data Management Body of Knowledge (DMBOK)*, artículos académicos indexados sobre automatización de procesos, y publicaciones especializadas en gestión de calidad de datos. Además, se revisarán manuales técnicos y documentación oficial de las herramientas tecnológicas que se consideren para la implementación del prototipo, tales como catálogos de funciones, especificaciones técnicas y casos de uso documentados. Esta selección permitirá sustentar conceptualmente las decisiones de diseño y asegurar que las soluciones propuestas se alineen con estándares reconocidos en el ámbito de los datos maestros.

3.5 Sujetos de investigación

Los sujetos de investigación son las personas, grupos o elementos sobre los cuales se recolecta información para responder a las preguntas planteadas en un estudio. Según Hernández Sampieri, Fernández y Baptista (2014), los sujetos representan la unidad de análisis de la

investigación y comprenden individuos, organizaciones, eventos o fenómenos específicos. La selección de los sujetos está directamente relacionada con el enfoque y los objetivos del estudio, ya que ellos son quienes proporcionan la información necesaria para generar hallazgos relevantes y fundamentar las conclusiones del proyecto. A continuación, en la Tabla 6 se especifican los sujetos de investigación del presente proyecto.

Tabla 6. Sujetos de investigación

Rol del sujeto	Años de experiencia en el rol	Caracterización del sujeto (diferentes responsabilidades y funciones del rol)	Justificación de la importancia de este sujeto para la investigación
Product Owner	7 años	Representado por *VP Manager – Data Governance*. Se encarga de hablar con el cliente, recopilar y transmitir los requerimientos del proyecto al equipo. Es el vínculo entre las necesidades del negocio y las soluciones técnicas propuestas.	Es esencial para identificar los requerimientos del proyecto, validar los objetivos de negocio, garantizar que las soluciones técnicas diseñadas cumplan con las expectativas y necesidades del cliente.
Scrum Master / Project Manager	4 años	Representado por *AVP – Delivery Manager*. Facilita la gestión del equipo, asegura que se cumplan los tiempos, presupuestos establecidos, y supervisa el avance del proyecto siguiendo las metodologías ágiles definidas.	Es clave para garantizar la correcta coordinación del equipo, gestionar riesgos, asegurar que el proyecto avance de manera ordenada, cumpla con los tiempos y recursos establecidos.
Senior Data Engineer	5 años	Representado por *AVP Manager - Software Engineering*. Se encarga del diseño, desarrollo y mantenimiento de *pipelines* de datos, asegurando la integración adecuada	Su rol es fundamental para implementar las conexiones con los recursos de datos, diseñar los *pipelines* y validar que los procesos sean eficientes y automatizados según

Rol del sujeto	Años de experiencia en el rol	Caracterización del sujeto (diferentes responsabilidades y funciones del rol)	Justificación de la importancia de este sujeto para la investigación
		de las fuentes y *datasets* para procesos que involucran datos.	los objetivos del proyecto.

Nota: Elaboración propia (2025)

Se estima que participarán tres sujetos de investigación, cada uno con un rol clave dentro del proceso objeto de estudio: *Product Owner*, *Scrum Master* / *Project Manager* y *Senior Data Engineer*. La selección se realizará mediante un muestreo intencional, orientado a incorporar únicamente a quienes tengan experiencia directa en la gestión, supervisión o ejecución del proceso de carga de datos en la plataforma de Gestión de Datos Maestros. Como criterio de inclusión, se considerará a las personas que hayan desempeñado activamente sus funciones durante al menos los últimos seis meses y que posean conocimiento técnico o funcional sobre el flujo actual. Se excluirá a individuos que no tengan participación directa en la operación o toma de decisiones del proceso en estudio. Esta selección busca asegurar la pertinencia de las perspectivas recogidas y fortalecer la validez de los hallazgos del proyecto.

3.6 Variables o categorías de la investigación

El concepto de variable se aplica a personas u otros seres vivos, objetos, hechos y fenómenos, los cuales adquieren diversos valores respecto de la variable referida. Las variables adquieren valor para la investigación científica cuando llegan a relacionarse con otras variables, es decir, si forman parte de una hipótesis o una teoría. En este caso, se les suele denominar constructos o construcciones hipotéticas (Hernández Sampieri, Fernández y Baptista 2014, p. 105). En la Tabla 7, 8, 9 y 10 se definen las variables relevantes a la investigación, junto con su tipo, los indicadores correspondientes y una descripción detallada de cada una.

Tabla 7. Variables de investigación para el objetivo específico #1

Objetivo especifico #1: Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros.			
Nombre de la variable	**Definición conceptual**	**Indicador**	**Definición instrumental**
VA-01: Estado actual del proceso de carga de datos (VI)	Conjunto de actividades manuales, procedimientos y herramientas utilizadas en la carga de datos en la plataforma de Gestión de Datos Maestros.	• Tiempo promedio de ejecución en cada etapa del proceso. • Cantidad de tareas manuales identificadas.	• Análisis detallado del flujo del proceso actual (*AS-IS*), documentando cada etapa mediante notación BPMN.

Objetivo especifico #1: Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros.			
			• Entrevistas con el equipo de *Data Operations* para identificar retos y tiempos involucrados.
VA-02: Deficiencias en la carga de datos (VD)	Factores específicos del proceso de carga de datos que generan demoras, errores o interrupciones, afectando la eficiencia operativa y la calidad de la información hacia la plataforma de Gestión de Datos Maestros.	• Cantidad de pasos identificados como ineficientes o repetitivos • Errores presentes durante la carga de datos. • Impacto de las deficiencias en los tiempos de procesamiento.	• Elaboración de un mapeo detallado del proceso actual (*AS-IS*), mediante notación BPMN para visualizar puntos de mejora. • Sesiones de revisión con el equipo de *Data Operations* para validar los problemas identificados.

Nota. Elaboración propia (2025)

Tabla 8. Variables de investigación para el objetivo específico #2

Objetivo especifico #2: Diseñar un nuevo proceso de carga de datos integrando herramientas de automatización y alineado con los requerimientos del equipo, con el fin del mejoramiento de la eficiencia del proceso.			
Nombre de la variable	**Definición conceptual**	**Indicador**	**Definición instrumental**
VA-03: Diseño del nuevo proceso de carga de datos. (VI)	Propuesta estructurada para la integración de herramientas de automatización en el proceso de carga de datos, estableciendo un flujo mejorado y alineado con las necesidades funcionales del equipo de *Data Operations*.	• Documentación del nuevo proceso de carga de datos (*To-Be*). • Nivel de integración con herramientas tecnológicas.	• Diagrama BPMN del nuevo flujo del proceso. • Matriz de integración
VA-04: Viabilidad del proceso diseñado (VD)	Grado en que el nuevo proceso diseñado es	• Alineamiento técnico del	• Matriz de alineamiento

Objetivo especifico #2: Diseñar un nuevo proceso de carga de datos integrando herramientas de automatización y alineado con los requerimientos del equipo, con el fin del mejoramiento de la eficiencia del proceso.			
	factible de implementar técnica y operativamente.	diseño con las necesidades identificadas en el proceso de carga de datos. • Nivel de aceptación del diseño por parte del equipo *Data Operations*.	entre diseño y necesidades funcionales. • Sesiones con el equipo *Data Operations*, dirigido a la revisión y validación formal del proceso diseñado.

Nota. Elaboración propia (2025)

Tabla 9. Variables de investigación para el objetivo específico #3

Objetivo especifico #3: Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta seleccionada y los requerimientos identificados.			
Nombre de la variable	**Definición conceptual**	**Indicador**	**Definición instrumental**
VA-06: Prototipo de solución automatizada (VI)	Desarrollo de un prototipo tecnológico que implemente la automatización de la carga de datos en la plataforma de Gestión de Datos Maestros, conforme a las funcionalidades diseñadas.	• Funcionalidades implementadas en el prototipo. • Correspondencia técnica con las necesidades funcionales establecidas.	• Implementación de las funcionalidades requeridas previstas en el diseño. • Validación técnica con el equipo de *Data Operations* sobre la operatividad del prototipo desarrollado.
VA-07: Alcance funcional del prototipo construido (VD)	Grado en que el prototipo incorpora las funcionalidades necesarias para automatizar el proceso de carga de datos, con base en el diseño técnico estructurado.	• Funcionalidades implementadas en la versión construida del prototipo. • Correspondencia entre los componentes del prototipo y las necesidades	• Revisión estructurada de la solución construida, con base en las funcionalidades esperadas. • Matriz de validación de

Objetivo especifico #3: Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta seleccionada y los requerimientos identificados.			
		funcionales establecidas.	funcionalidades implementadas.

Nota. Elaboración propia (2025)

Tabla 10. Variables de investigación para el objetivo específico #4

Objetivo especifico #4: Evaluar la efectividad del prototipo de la solución automatizada en términos de precisión, consistencia y reducción de tareas manuales en el proceso de carga de datos, utilizando métricas de desempeño para la determinación de su impacto en la eficiencia del proceso.			
Nombre de la variable	**Definición conceptual**	**Indicador**	**Definición instrumental**
VA-08 Prototipo de solución automatizada (VI)	Herramienta tecnológica desarrollada para la automatización del proceso de carga de datos en la plataforma de Gestión de Datos Maestros, con el fin de mejorar la consistencia de los datos y reducir las tareas manuales	• Implementación de funcionalidades automatizadas en el prototipo. • Reducción de tareas manuales en el proceso de carga de datos.	• Pruebas de funcionalidad en un entorno de simulación.
VA-09 Efectividad del prototipo en la carga de datos (VD)	Grado en que el prototipo automatizado mejora la precisión y consistencia de los datos, además de reducir las tareas manuales en el proceso de carga de datos.	• Porcentaje de registros cargados. • Tiempo total de procesamiento de datos comparado con el proceso manual.	• Comparación del tiempo de carga antes y después de la implementación del prototipo. • Métricas de desempeño para determinar la efectividad en la reducción de errores y tareas manuales.

Nota. Elaboración propia (2025)

Las variables definidas en las Tablas 7, 8, 9 y 10 serán revisadas por personas expertas del equipo técnico involucrado en el proceso de carga de datos, con el propósito de validar su coherencia con los objetivos de investigación y su aplicabilidad en el contexto organizacional. Esta revisión incluirá el análisis de los indicadores propuestos y de los instrumentos definidos para su medición, con el fin de asegurar su pertinencia conceptual, claridad técnica y utilidad práctica. Este procedimiento garantizará la calidad metodológica de las variables seleccionadas y la confiabilidad de los datos que se recolectarán durante las etapas correspondientes del estudio.

3.7 Técnicas e instrumentos de recolección de datos

Según Hernández Sampieri (2014, p. 198), una vez que se ha seleccionado el diseño de investigación adecuado y se ha definido la muestra conforme al problema de estudio y las hipótesis planteadas (en caso de que existan), la siguiente etapa corresponde a la recolección de datos relevantes sobre los atributos, conceptos o variables de las unidades de análisis o casos, tales como participantes, grupos, fenómenos, procesos u organizaciones.

Recolectar los datos implica elaborar un plan detallado de procedimientos orientados a reunir información con un propósito específico. Este plan debe considerar aspectos fundamentales como la identificación de las fuentes de donde se obtendrán los datos, ya sea a través de personas, observaciones, registros, documentos, archivos o bases de datos.

Asimismo, es necesario precisar la localización de dichas fuentes, que regularmente se encuentran dentro de la muestra seleccionada, así como definir el medio o método que se utilizará para recolectar los datos, asegurando que estos sean confiables, válidos y objetivos. Igualmente, debe establecerse la forma en que se prepararán los datos recolectados para ser analizados correctamente y ofrecer respuestas al planteamiento del problema.

El plan de recolección de datos se construye con base en varios elementos clave: las variables, conceptos o atributos definidos en el planteamiento del problema e hipótesis; las definiciones operacionales, que resultan esenciales para determinar el método de medición y fundamentar las inferencias; la muestra previamente seleccionada; y los recursos disponibles, como el tiempo, el apoyo institucional y los aspectos económicos.

En la Tabla 11, se exponen las técnicas de recolección de datos utilizadas durante el desarrollo del Trabajo Final de Graduación, por lo cual se presenta el nombre de la técnica, su definición conceptual y la importancia que tienen en el desarrollo del presente proyecto.

Tabla 11. Técnicas e instrumentos de recolección de datos

Técnica	Definición conceptual	Importancia en el proyecto
Revisión documental y registros (ver **Apéndice O**)	Esta técnica consiste en examinar los datos presentes en documentos ya existentes, como bases de datos, actas, informes, registros de asistencia, etc. Por lo tanto, lo más importante para este método es la habilidad para encontrar, seleccionar y analizar la información disponible. (Caro L, s.f.)	La revisión documental permitió analizar formatos, reportes de errores y documentación técnica del proceso de carga de datos. Fue clave para identificar deficiencias en la documentación del proceso y respaldar el diagnóstico inicial.
Entrevista semiestructurada (ver **Apéndice M**)	Según Hernández-Sampieri, Fernández-Collado y Baptista-Lucio (2014), la entrevista semiestructurada "se basa en una guía de asuntos o preguntas, y el entrevistador tiene la	Se utilizó para recopilar información detallada sobre el proceso de carga de datos desde la perspectiva del equipo de *Data Operations*. Las entrevistas facilitaron la

Técnica	Definición conceptual	Importancia en el proyecto
	libertad de introducir preguntas adicionales para precisar conceptos u obtener mayor información sobre los temas deseados" (p. 411).	identificación de tareas críticas, retos operativos, deficiencias actuales y validación de requerimientos funcionales para el nuevo proceso.
Análisis FODA (ver **Apéndice AA**)	El análisis FODA es una herramienta estratégica utilizada para evaluar factores internos (fortalezas y debilidades) y factores externos (oportunidades y amenazas) que afectan el desempeño de una organización. Su objetivo es generar estrategias que maximicen los elementos positivos y minimicen los negativos, proporcionando una base estructurada para la toma de decisiones empresariales. (Maguiña Rivero & Ugarriza Gross, 2016, p. 310)	Se empleó para analizar el proceso actual de carga de datos (*As-Is*) desde una perspectiva organizacional, permitiendo comprender de forma estructurada sus deficiencias, capacidades actuales y áreas potenciales de mejora.
Diagrama de Ishikawa (*Fishbone*) (ver **Apéndice AB**)	Los diagramas de Ishikawa (*Fishbone*) se utilizan para analizar relaciones causa-efecto, identificando diversas categorías de factores que contribuyen a un problema o resultado específico. Estos diagramas ayudan a estructurar sesiones de lluvia de ideas y fomentan un análisis exhaustivo de todas las posibles causas. (Skulimowski & Smętkowski, 2016, p. 85)	Esta técnica fue utilizada para identificar de manera estructurada las causas raíz de las deficiencias presentes en el proceso actual de carga de datos. Permitió agrupar factores críticos en categorías como herramientas, métodos, personal y datos, facilitando un análisis visual. Su aplicación fortaleció el diagnóstico inicial y proporcionó insumos clave para el diseño de la solución automatizada en fases posteriores.
Análisis comparativo	El análisis comparativo es un método de investigación, recolección y análisis de información que consiste en la comparación de dos o más procesos, documentos, conjuntos de datos u otros objetos. (QuestionPro, s.f.)	Fue esencial para comparar el proceso actual (*As-Is*) con el nuevo proceso propuesto (*To-Be*), y para evaluar el impacto del prototipo. Permitió medir mejoras en tiempo, precisión, reducción de errores y tareas manuales.
Matriz de trazabilidad de requerimientos (ver **Apéndice AC**)	Una matriz de trazabilidad de requerimientos es un instrumento que crea un mapeo claro entre cada requerimiento y	Permite verificar que cada requerimiento funcional identificado durante el análisis haya sido implementado en el

Técnica	Definición conceptual	Importancia en el proyecto
	sus elementos de diseño asociados, componentes de desarrollo, casos de prueba y otros entregables. Su objetivo principal es garantizar que todos los requerimientos definidos sean considerados, implementados y verificados a lo largo del ciclo de vida del proyecto. (Visure Solutions, 2021)	prototipo. Facilita el control del alcance, respalda la validación técnica del diseño y asegura coherencia entre lo planificado y lo construido. Su uso fortalece la documentación del desarrollo y contribuye a la toma de decisiones en fases posteriores del proyecto.
Pruebas de funcionalidad del prototipo	Las pruebas de funcionalidad en prototipos son evaluaciones exhaustivas realizadas para verificar que cada funcionalidad del prototipo opere según lo especificado y satisfaga las necesidades del usuario, permitiendo a los desarrolladores refinar y mejorar la funcionalidad antes de su implementación final. (FasterCapital, 2024)	Esta técnica permitió validar que el prototipo automatizado cumpliera con los requerimientos funcionales definidos previamente por el equipo de *Data Operations*.
Pruebas de desempeño del prototipo	Las pruebas de desempeño de prototipos son esenciales en el proceso de desarrollo de productos, ya que permiten evaluar aspectos como la usabilidad, funcionalidad y rendimiento antes de su lanzamiento al mercado. Estas pruebas facilitan la identificación de problemas y áreas de mejora, asegurando que el producto final cumpla con las expectativas de los usuarios y los estándares de calidad requeridos. (FasterCapital, 2024)	Se utilizaron para validar que el prototipo cumpliera con los requerimientos técnicos y funcionales, evaluando la precisión de la carga de datos, la consistencia de los registros, la reducción de tareas manuales y tiempo de ejecución.

Nota. Elaboración propia (2025)

3.8 Procedimiento metodológico de la Investigación

De acuerdo con Hernández Sampieri, Fernández y Baptista (2014), el desarrollo de una investigación requiere la planificación de un conjunto de fases sistemáticas que articulan la recolección, el análisis y la interpretación de los datos, en función de los objetivos planteados. Esta secuencia metodológica permite estructurar el estudio de manera lógica, asegurando que cada

etapa contribuya a generar conocimiento útil, confiable y coherente con el problema investigado. A continuación, se describen las fases y actividades correspondientes al presente proyecto.

3.8.1 Fase 1: Análisis de la situación actual del proceso de carga de datos

Esta fase tiene como propósito comprender a profundidad el estado actual del proceso de carga de datos, identificando deficiencias y tareas manuales críticas que impactan la gestión de datos maestros en la plataforma de gestión de datos maestros. A partir de este análisis, se establecerá una base de referencia que permitirá diseñar un nuevo proceso más eficiente. A continuación, se describen las actividades clave de la fase 1.

- **Revisión de documentación interna:** Se revisarán los manuales, reportes y registros técnicos existentes que describen el proceso actual y las prácticas de carga de datos, permitiendo identificar las áreas con mayor potencial de mejora.
- **Reuniones con el equipo de *Data Operations*:** Se realizarán entrevistas semiestructuradas con roles clave, como el *Product Owner* y el *Senior Data Engineer*, para recopilar información sobre los retos técnicos, pasos manuales y principales puntos críticos del proceso.
- **Mapeo del proceso actual (AS-IS):** Se elaborarán diagramas detallados utilizando herramientas de modelado, como BPMN, para representar visualmente el flujo de datos desde las fuentes de datos hasta la carga en la plataforma de gestión de datos maestros.
- **Identificación de métricas iniciales:** Se definirán métricas como tiempo promedio de procesamiento y cantidad de errores, las cuales servirán como línea base para evaluar el impacto de la solución propuesta en fases posteriores.

3.8.2 Fase 2: Diseño del nuevo proceso de carga de datos

Esta fase se orienta a proponer un nuevo flujo de trabajo (*To-Be*) que integre herramientas de automatización, que esté alineado con los requerimientos del equipo y las mejores prácticas de la industria. A continuación, se describen las actividades clave de la fase 2.

- **Identificación de oportunidades de mejora**: Se definirán las etapas del proceso que requieren modificaciones, ya sea por eliminación de tareas manuales, reestructuración del flujo de trabajo o integración de herramientas de automatización. Para esta actividad, se evaluará la realización de reuniones con miembros del equipo de *Data Operations*, con el fin de recopilar sus percepciones sobre los puntos críticos del proceso actual y validar las áreas con mayor potencial de mejora. Esta instancia participativa permitirá contrastar distintas perspectivas y enriquecer la propuesta del nuevo flujo de trabajo (*To-Be*).
- **Revisión de herramientas tecnológicas:** Se investigarán las herramientas y tecnologías disponibles, para evaluar su compatibilidad con las necesidades del proceso de carga de datos. Asimismo, se aplicará la técnica de entrevista semiestructurada dirigida al *Senior Data Engineer* del equipo, con el fin de obtener una valoración técnica sobre la viabilidad de las plataformas consideradas.

- **Creación de diagramas *To-Be*:** Utilizando herramientas de modelado como BPMN, se documentará el flujo de datos del nuevo proceso, incluyendo interacciones con fuentes de datos y herramientas de automatización.
- **Validación con el equipo *Data Operations*:** El diseño propuesto será presentado al equipo de Data Operations para asegurar que cumple con los requerimientos del equipo.

3.8.3 Fase 3: Desarrollo del prototipo de la solución automatizada

Esta fase tiene como propósito el desarrollo de un prototipo que automatice el proceso de carga de datos hacia la plataforma de Gestión de Datos Maestros, de acuerdo con la herramienta seleccionada y los requerimientos previamente identificados por el equipo de *Data Operations*. El enfoque principal estará en construir una solución que estructure y cargue los datos de forma automatizada, sustituyendo tareas manuales críticas detectadas en el análisis del proceso actual. A continuación se describen las actividades clave de la fase 3.

- **Desarrollo del prototipo:** Se desarrollará el prototipo utilizando la herramienta seleccionada, integrando funcionalidades que automaticen tareas tales como extracción y carga de los datos. Se garantizará que el desarrollo responda fielmente a los requerimientos funcionales y operativos definidos en la fase anterior.
- **Documentación del desarrollo:** Se generará la documentación del prototipo, incluyendo roles en el proceso, diagramas de flujo, scripts utilizados, configuraciones relevantes y guías básicas de uso, con el fin de facilitar su comprensión y futuras mejoras por parte del equipo.

3.8.4 Fase 4: Evaluación del prototipo de la solución automatizada

Esta fase tiene como propósito evaluar la efectividad del prototipo desarrollado en la fase anterior, mediante la aplicación de pruebas de desempeño que permitan analizar su funcionamiento en términos de precisión, consistencia y reducción de tareas manuales durante la carga de datos en la plataforma de Gestión de Datos Maestros. La evaluación se realizará utilizando métricas de desempeño que permitan determinar el impacto del prototipo en la eficiencia del proceso. A continuación, se describen las actividades clave de la fase 4.

- **Ejecución de pruebas funcionales:** Se realizarán pruebas controladas del prototipo para verificar que los datos se cargan correctamente en la plataforma, asegurando su fidelidad respecto a las fuentes originales y la correcta aplicación de reglas de validación.
- **Medición de métricas de desempeño:** Se recopilarán datos sobre precisión (porcentaje de registros correctos), consistencia (coherencia en los valores cargados) y reducción de intervención manual (cantidad de tareas automatizadas). Estas métricas permitirán evaluar la efectividad del prototipo.
- **Análisis comparativo con el proceso manual:** Se realizará una comparación estructurada entre el proceso actual (*As-Is*) y el proceso automatizado mediante el

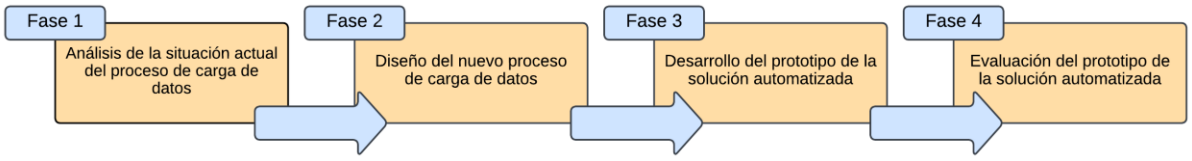
prototipo (*To-Be*), analizando el impacto logrado en términos de tiempos de ejecución y calidad de datos.

- **Validación con el equipo de *Data Operations*:** Los resultados obtenidos serán compartidos con los responsables del proceso para validar la efectividad percibida del prototipo y confirmar que este cumple con los requerimientos definidos.

3.8.5 Diagrama de propuesto para las fases del procedimiento metodológico

A continuación, en la Figura 10 se visualiza el diagrama propuesto para las fases del procedimiento metodológico.

Figura 10. Diagrama propuesto para las fases del procedimiento metodológico



Nota. Elaboración propia (2025)

3.9 Operacionalización de las variables o categorías.

A continuación, en la Tabla 12 se presenta la operacionalización de las variables o categorías.

Tabla 12. Operacionalización de las variables o categorías

Objetivo	Fase	Variable	Instrumentos	Sujetos de Investigación
Objetivo específico #1	I	Estado actual del proceso de carga de datos	• Entrevista semiestructurada • Revisión de documentos y registros • Notación BPMN para modelado del proceso *AS-IS* • Diagrama de Ishikawa (*Fishbone*)	• VP Manager – Data Governance (Product Owner) • AVP Manager – Software Engineering (Senior Data Engineer)
		Deficiencias en el proceso actual de carga de datos	• Entrevista semiestructurada	• VP Manager – Data Governance

Objetivo	Fase	Variable	Instrumentos	Sujetos de Investigación
			• Revisión de documentos y registros • Análisis FODA	(Product Owner) • AVP Manager – Software Engineering (Senior Data Engineer)
Objetivo específico #2	II	Diseño del nuevo proceso de carga de datos	• Notación BPMN para modelado del proceso *TO-BE*	• AVP Manager – Software Engineering (Senior Data Engineer) • VP Manager – Data Governance (Product Owner)
		Viabilidad del proceso diseñado	• Matriz de alineamiento entre diseño y necesidades funcionales. • Reuniones con equipo *Data Operations*.	• AVP Manager – Software Engineering (Senior Data Engineer) • VP Manager – Data Governance (Product Owner)
Objetivo específico #3	III	Prototipo de solución automatizada	• Notación BPMN para diseño técnico del flujo automatizado.	No aplica
		Alcance funcional del prototipo construido	• Matriz de validación de funcionalidades implementadas.	• AVP Manager – Software Engineering (Senior Data Engineer) • VP Manager – Data Governance (Product Owner)

Objetivo	Fase	Variable	Instrumentos	Sujetos de Investigación
Objetivo específico #4	IV	Prototipo de solución automatizada	• Pruebas de funcionalidad del prototipo	• AVP Manager – Software Engineering (Senior Data Engineer)
		Efectividad del prototipo en la carga de datos	• Pruebas de funcionalidad del prototipo • Pruebas de desempeño del prototipo • Análisis comparativo	• VP Manager – Data Governance (Product Owner) • AVP – Delivery Manager (Scrum Master / Project Manager) • AVP Manager – Software Engineering (Senior Data Engineer)

Nota. Elaboración propia (2025)

3.10 Tabla resumen del procedimiento metodológico de la investigación

A continuación, en la Tabla 13 se presenta la matriz de trazabilidad del procedimiento metodológico del Trabajo Final de Graduación.

Tabla 13. Matriz de trazabilidad del procedimiento metodológico del Trabajo Final de Graduación

Objetivo	Marco Conceptual	Metodología	Análisis de Resultados	Propuesta de solución	Conclusiones	Recomendaciones
Objetivo especifico #1	Sección 2.1.1 Sección 2.1.1.1 Sección 2.1.2 Sección 2.1.2.1 Sección 2.1.3 Sección 2.1.6 Sección 2.1.7 Sección 2.1.7.1 Sección 2.1.7.2	Sección 3.5 Sección 3.6 Sección 3.7 Sección 3.8.1 Sección 3.8.5	Sección 4.1	No aplica	Sección 6.1	Sección 7
Objetivo especifico #2	Sección 2.1.4 Sección 2.1.5 Sección 2.1.5.1 Sección 2.1.6 Sección 2.1.7 Sección 2.1.7.1 Sección 2.1.7.2	Sección 3.5 Sección 3.6 Sección 3.7 Sección 3.8.2 Sección 3.8.5	Sección 4.2	No aplica	Sección 6.2	Sección 7
Objetivo especifico #3	Sección 2.1.1 Sección 2.1.1.1 Sección 2.1.2	Sección 3.5 Sección 3.6 Sección 3.7 Sección 3.8.3 Sección 3.8.5	No aplica	Sección 5.1	Sección 6.3	Sección 7

Objetivo	Marco Conceptual	Metodología	Análisis de Resultados	Propuesta de solución	Conclusiones	Recomendaciones
	Sección 2.1.2.1 Sección 2.1.3 Sección 2.1.4 Sección 2.1.8 Sección 2.1.9 Sección 2.2 Sección 2.2.1					
Objetivo especifico #4	Sección 2.1.1 Sección 2.1.1.1 Sección 2.1.2 Sección 2.1.2.1 Sección 2.1.3 Sección 2.1.4 Sección 2.1.10 Sección 2.1.10.1 Sección 2.2 Sección 2.2.1	Sección 3.5 Sección 3.6 Sección 3.7 Sección 3.8.4 Sección 3.8.5	No aplica	Sección 5.2	Sección 6.4	Sección 7

Nota. Elaboración propia (2025)

4 Análisis de Resultados

En esta sección se detallan los resultados obtenidos a partir de la ejecución de las actividades descritas en el marco metodológico. El análisis se enfoca en evaluar la viabilidad de mejorar y automatizar el proceso de carga de datos en la plataforma de Gestión de Datos Maestros, mediante la identificación de deficiencias en la carga de datos, el diseño de un flujo optimizado y la validación de un prototipo funcional.

El estudio se organiza en fases, cada una orientada a examinar aspectos específicos del proceso. Esta estructura asegura una evaluación coherente con los objetivos definidos, manteniendo una estrecha relación entre la metodología aplicada y los hallazgos obtenidos.

Cada fase incluye un análisis riguroso de la información recopilada, seguido de una interpretación crítica que fundamenta las propuestas de mejora. Este capítulo representa el aporte analítico del proyecto, al demostrar que los resultados obtenidos justifican la intervención del proceso actual mediante soluciones basadas en automatización.

4.1 Fase 1: Análisis de la situación actual del proceso de carga de datos

La primera fase del análisis tiene como propósito examinar de forma detallada el estado actual del proceso de carga de datos en la plataforma de Gestión de Datos Maestros. Esta etapa responde al primer objetivo específico del proyecto: “Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros.”

Para cumplir con este propósito, se abordan dos variables fundamentales: el estado actual del proceso y las deficiencias en la carga de datos que impiden su correcto funcionamiento. La evaluación se enfoca en caracterizar el flujo existente, identificar los puntos críticos manuales y determinar los factores que generan errores e ineficiencia.

El desarrollo de esta fase se sustenta en la aplicación de técnicas metodológicas previamente definidas, tales como la revisión documental de registros y manuales internos, la realización de entrevistas semiestructuradas con miembros clave del equipo de *Data Operations*, y la elaboración de un Análisis FODA que permite visualizar, de manera estructurada, las fortalezas, debilidades, oportunidades y amenazas asociadas al proceso actual.

4.1.1 Descripción del proceso actual de carga de datos.

El proceso actual de carga de datos en la plataforma de Gestión de Datos Maestros presenta una serie de características operativas que evidencian una fuerte dependencia de la ejecución manual, fragmentación y dependencia de herramientas no especializadas, lo cual afecta significativamente la eficiencia, escalabilidad y calidad de los datos procesados. En la Tabla 14, se describen las etapas presentes en el flujo actual del proceso.

Tabla 14. Etapas del proceso actual de carga de datos

Etapa	Descripción
Obtención de archivos	Los datos provienen de cinco fuentes distintas y se obtienen mediante FTP, correos electrónicos o extracciones desde APIs.
Almacenamiento inicial	Los archivos se guardan manualmente en carpetas definidas dentro de un repositorio en SharePoint.
Procesamiento de datos	Un analista estructura manualmente los datos en Excel, corrige errores, valida duplicaciones y relaciona entidades.
Consolidación	Los datos procesados se unifican en un archivo final, que posteriormente es cargado en la base de datos mediante scripts de PostgreSQL.
Carga en la plataforma de Gestión de Datos Maestros	El archivo consolidado se carga en PostgreSQL, con la expectativa de que cumpla con los requisitos estructurales del sistema.

*Nota. Esta descripción se fundamenta en la información proporcionada por el Senior Data Engineer del equipo, durante una entrevista técnica detallada (ver **Apéndice N**). Elaboración propia (2025)*

El proceso actual de carga de datos se sostiene sobre una infraestructura técnica compuesta por herramientas de uso general, no especializadas, que limitan la eficiencia y escalabilidad del flujo de trabajo. La principal herramienta utilizada es Microsoft Excel, que funciona como el medio fundamental para el procesamiento y estructuración de los datos recibidos. A través de Excel, los analistas realizan tareas críticas como la identificación de valores faltantes, la validación de duplicaciones y la estandarización manual de los datos, a pesar de que esta herramienta no está diseñada para manejar volúmenes de información que alcanzan millones de registros, lo que la convierte en un factor limitante para la calidad y consistencia de los datos procesados.

Complementariamente, el almacenamiento inicial de los archivos se realiza en SharePoint, una plataforma que se utiliza como repositorio temporal para organizar manualmente los archivos descargados de las diversas fuentes. Este sistema depende totalmente de la correcta gestión por parte del analista, quien debe asegurar que cada archivo se almacene en la carpeta correspondiente, sin contar con mecanismos automáticos de validación o control de versiones.

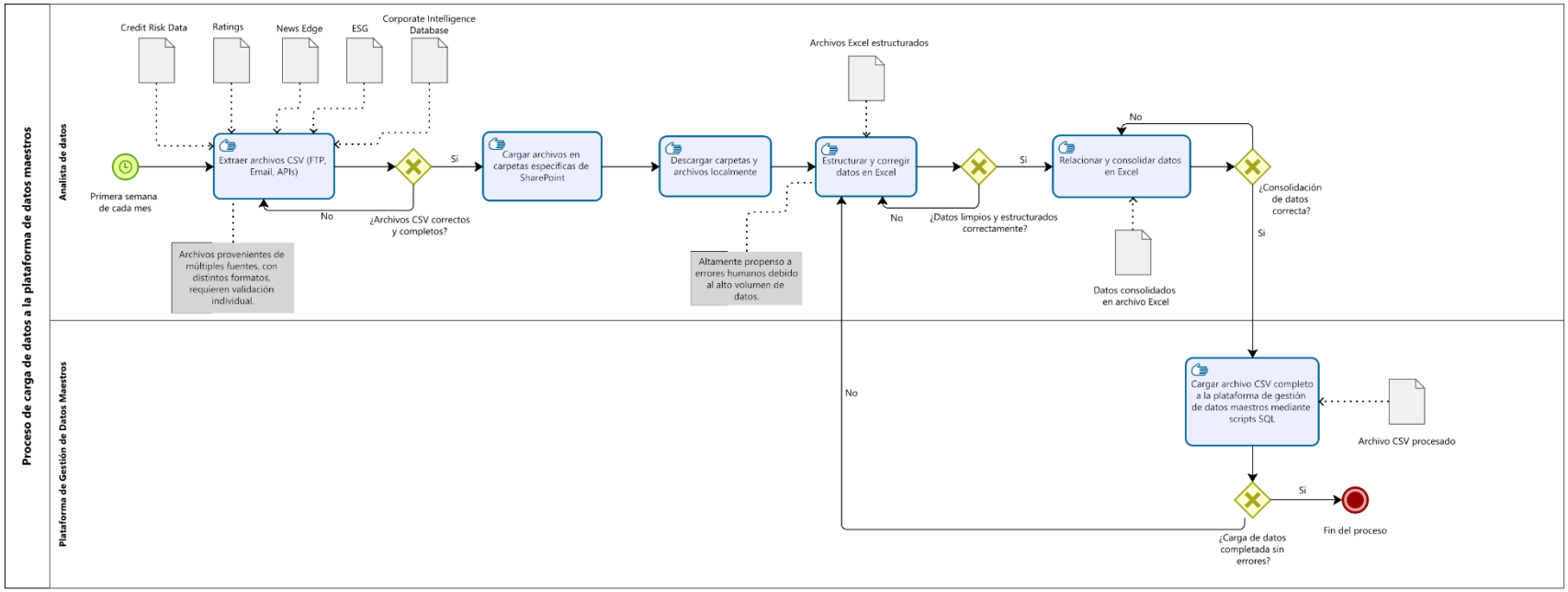
La carga final de los datos procesados se ejecuta mediante PostgreSQL, que opera como el motor de base de datos donde los archivos consolidados son integrados en la plataforma de Gestión de Datos Maestros. Esta carga se efectúa a través de scripts SQL desarrollados por el analista, los cuales, si bien automatizan la carga de los datos, dependen de que el archivo fuente haya sido

correctamente preparado y validado, debido a la ejecución manual segmentada de las etapas previas.

Adicionalmente, se identificó la existencia de una guía informal que orienta al analista sobre ciertos elementos a verificar dentro de los archivos antes de proceder con la carga. Esta guía carece de una estructura exhaustiva y no establece procedimientos normalizados, lo cual refuerza la dependencia del juicio individual de quien ejecuta las tareas. (ver **Apéndice P** y **Anexo I:** ***Data Review Checklist for Master Data Loading Process***). La falta de lineamientos formales compromete la estandarización del proceso y aumenta el riesgo de variabilidad en los resultados (ver **Apéndice N**).

A continuación, en la Figura 11 se visualiza mediante notación BPMN el diagrama *As-Is* que conforma el proceso actual de carga de datos en la plataforma de Gestión de Datos Maestros.

Figura 11. Diagrama As-Is del proceso de carga de datos en la plataforma de Gestión de Datos Maestros



Nota. Elaboración propia (2025)

El diagrama *As-Is* ilustra de manera detallada el flujo actual del proceso de carga de datos en la plataforma de Gestión de Datos Maestros. Este proceso inicia con la extracción de archivos CSV provenientes de cinco fuentes externas, cada una con formatos diversos que requieren validación individual. Una vez verificada la integridad de los archivos, estos se cargan manualmente en carpetas específicas dentro de SharePoint, desde donde se descargan nuevamente para su procesamiento local.

Posteriormente, los datos se estructuran y corrigen en Excel, tarea propensa a errores debido al elevado volumen de información. La validación de la limpieza y estructura de los datos es crítica, ya que de ello depende la correcta consolidación en un archivo final también manejado en Excel. Este archivo, una vez validado, es cargado a la plataforma de datos mediante scripts SQL.

El diagrama resalta varios puntos críticos:

- El 100% de las actividades del proceso son manuales (6 de 6 tareas identificadas).
- La falta de automatización en validaciones.
- El riesgo operativo derivado de la dependencia total del analista.
- La repetición de pasos en caso de errores, lo cual afecta la eficiencia del proceso.

El diagrama *As-Is* confirma las deficiencias, reforzando la necesidad de reestructurar el flujo mediante herramientas que permitan la automatización y reducción de tareas manuales.

4.1.2 Métricas operativas del proceso de carga de datos

Con el propósito de caracterizar de forma objetiva el estado actual del proceso de carga de datos, se recopilaron diversas métricas operativas a través de la entrevista técnica con el *Senior Data Engineer* (ver **Apéndice N**). Estas métricas permiten evidenciar el nivel de intervención humana, la carga operativa, las deficiencias que afectan la eficiencia y calidad del proceso. Los datos obtenidos ofrecen una visión cuantitativa que respalda el diagnóstico realizado, permitiendo establecer relaciones directas entre el volumen de datos, los tiempos de ejecución, y la frecuencia de errores, aspectos clave para fundamentar la necesidad de intervención. A continuación, la Tabla 15 detalla las principales métricas relevantes al proceso analizado.

Tabla 15. Métricas identificadas del proceso de carga de datos

Métrica	Valor aproximado/Observación
Tiempo total del proceso	Aproximadamente 3.6 días hábiles por ciclo (29 horas por analista), de los cuales 20 horas se destinan a tareas de validación y corrección ejecutadas manualmente.
Tiempos por etapa	Descarga: 3 horas; Validación y corrección manual: 20 horas; Procesamiento y consolidación: 4 horas; Carga en la plataforma MDM: 2 horas.

Métrica	Valor aproximado/Observación
Volumen de datos por ciclo	Entre 50,000 y varios millones de filas por archivo; los archivos más grandes se dividen en lotes.
Errores detectados por ciclo	Errores en cada ciclo: duplicaciones, campos faltantes, inconsistencias en identificadores.
Escalabilidad	Un aumento del 10-20% en volumen incrementa proporcionalmente el tiempo del proceso.

*Nota. Las métricas identificadas se fundamentan en la información proporcionada por el Senior Data Engineer del equipo, durante una entrevista técnica detallada (ver **Apéndice N**). Elaboración propia (2025).*

4.1.3 Problemática técnica, deficiencias y riesgos del proceso actual

El proceso de carga de datos revela limitaciones técnicas y operativas que comprometen la eficiencia, calidad y sostenibilidad del flujo actual. Uno de los principales retos consiste en manejar grandes volúmenes de datos mediante herramientas inadecuadas. Microsoft Excel, utilizada como herramienta principal para la estructuración y validación, no está diseñada para procesar millones de filas. Esta condición genera cuellos de botella, incrementa la carga operativa sobre el equipo, además de elevar significativamente el riesgo de errores, tanto por las limitaciones de la herramienta como por la intervención manual necesaria (ver **Apéndice N**).

A esta limitación se suma la fuerte dependencia del conocimiento empírico del analista. En ausencia de controles estandarizados, la calidad final de los datos depende de la experiencia individual del encargado. Esta variabilidad, inaceptable en un entorno que requiere consistencia, limita la capacidad del proceso para escalar y adaptarse a mayores exigencias.

Estas condiciones técnicas reflejan deficiencias estructurales que afectan cada etapa del proceso. La intervención manual constante incrementa la probabilidad de errores, especialmente bajo volúmenes crecientes. Entre los errores más comunes destacan las duplicaciones de registros, originadas por la falta de estandarización en identificadores provenientes de distintas fuentes, y los campos faltantes (*NAs*) en columnas esenciales, lo que compromete la integridad de la información. Asimismo, la presencia de formatos inconsistentes en nombres, fechas o tipos de datos dificulta la consolidación adecuada.

La ausencia de estandarización agrava el problema. No existen validaciones automatizadas tras la carga, lo que obliga a revisiones visuales, propensas a omisiones. Esta falta de verificación sistemática reduce la fiabilidad del proceso. Además, el equipo de *Data Operations* dispone únicamente de una guía informal, que ofrece orientaciones básicas sin cubrir todas las situaciones posibles (ver **Apéndice P** y **Anexo I**). La carencia de lineamientos formales amplifica la dependencia de habilidades individuales, disminuyendo la capacidad de auditar o replicar el procedimiento de manera consistente (ver **Apéndice N**).

Estas deficiencias en la operación actual evidencian la necesidad de transformar el proceso hacia un modelo automatizado, estandarizado y con capacidad de crecimiento. Las ineficiencias detectadas están alineadas con las variables metodológicas analizadas, en particular el estado actual del proceso (VA-01) y las deficiencias en la carga de datos (VA-02). Indicadores clave, como la cantidad de tareas manuales, errores recurrentes y el impacto en los tiempos, se reflejan con claridad en cada ciclo operativo. Este análisis proporciona la base para justificar, en fases posteriores, el diseño de soluciones enfocadas en la automatización y mejora continua del proceso.

4.1.4 Impacto en la calidad y consistencia de los datos

El proceso actual de carga de datos presenta limitaciones en cuanto a la calidad y consistencia de la información que se gestiona. La dependencia casi exclusiva de la intervención manual por parte del colaborador (analista) en todas las etapas críticas genera una alta probabilidad de errores que afectan directamente la integridad de los datos maestros.

De acuerdo con la entrevista brindada al *Senior Data Engineer*, en la Tabla 16 se identificaron los siguientes factores clave que deterioran la calidad y consistencia.

Tabla 16. Problemas críticos identificados en el proceso de carga de datos

Problema identificado	Descripción
Duplicaciones de registros	Por falta de estandarización en identificadores únicos, dificultan las consolidación adecuada.
Campos faltantes en columnas	Valores NAs recurrentes; subjetividad en su manejo incrementa el riesgo de inconsistencia.
Inconsistencias en estandarización	Estándares manuales para formatos diversos (fechas, nombres), resultados variables entre ciclos.
Falta de validaciones post-carga	No existen reglas automáticas que verifiquen calidad; errores persisten si no son detectados previamente.

*Nota. Los problemas identificados se fundamentan en la información proporcionada por el Senior Data Engineer del equipo, durante una entrevista técnica detallada (ver **Apéndice N**). Elaboración propia (2025).*

Los problemas expuestos en la Tabla 16 generan consecuencias significativas para la operación del proceso. La precisión de los datos maestros se encuentra comprometida, lo que limita su valor para otras áreas que dependen de información confiable en sus actividades estratégicas. La falta de exactitud afecta la calidad de los análisis derivados, lo que origina decisiones basadas en datos erróneos o incompletos.

Asimismo, la coherencia entre las fuentes de datos se deteriora debido a la ausencia de mapeos fiables en los identificadores. Esta inconsistencia impide una consolidación efectiva, produce redundancias y genera conflictos que deben resolverse manualmente, lo que incrementa la carga operativa. La confiabilidad del proceso se reduce al depender de un único analista, quien

asume toda la responsabilidad sobre la calidad de los datos sin apoyo de mecanismos automatizados. Esta dependencia representa un riesgo operativo considerable, ya que cualquier error u omisión pasa desapercibido y afecta el resultado final.

Además, las limitaciones actuales fomentan un aumento en los reprocesos, pues los errores detectados obligan a repetir tareas y corregir datos en varias etapas. Este retrabajo extiende los tiempos de entrega y limita la capacidad del equipo para responder con agilidad a nuevas demandas. En conjunto, estas consecuencias refuerzan la necesidad de implementar controles automatizados junto con procedimientos estandarizados que reduzcan riesgos y mejoren la calidad de los datos maestros.

Como parte de la evidencia recopilada durante la revisión documental, se presenta una plantilla utilizada para registrar los errores detectados durante el proceso manual de carga de datos. Esta herramienta, que no cuenta con una estandarización formal a nivel de equipo, permite observar la frecuencia, naturaleza y gravedad de las fallas que afectan la calidad de los datos maestros. La plantilla se incluye en el **Apéndice Q** y **Anexo II** como respaldo de los hallazgos obtenidos.

4.1.5 Diagrama Ishikawa (*Fishbone*) de la situación actual del proceso de carga de datos

El presente diagrama Ishikawa constituye una herramienta clave para profundizar en el análisis del proceso actual de carga de datos. Su valor dentro de la Fase 1 radica en la identificación y clasificación sistemática de las causas raíz que originan la ineficiencia y los errores detectados. Esta representación gráfica sintetiza la información obtenida mediante entrevistas y revisión documental, además de evidenciar la interacción entre factores técnicos, operativos y organizacionales que inciden directamente en los problemas analizados.

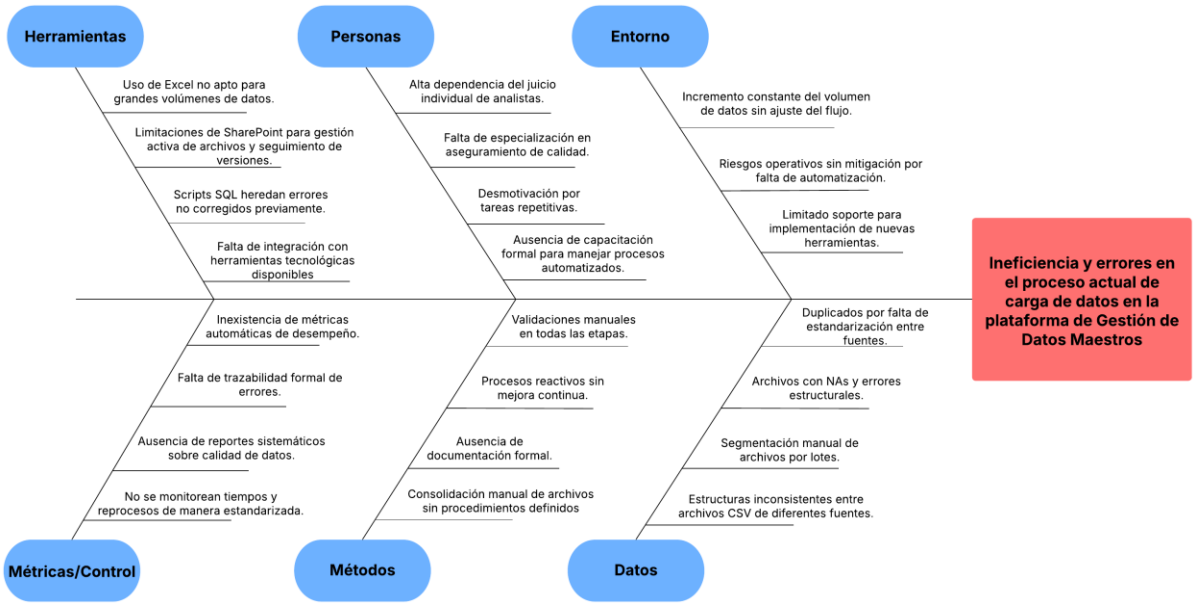
A diferencia del análisis FODA, que ofrece una visión estratégica, el Diagrama Ishikawa descompone los problemas operativos en dimensiones específicas: Herramientas, Personas, Entorno, Métodos, Datos y Métricas/Control. Estas dimensiones, alineadas al marco metodológico, facilitan la evaluación detallada del proceso y destacan áreas críticas que requieren intervención inmediata.

Entre los aportes más relevantes se encuentran:

- La visualización estructurada de causas técnicas y organizativas que afectan la calidad y consistencia de los datos.
- La identificación de relaciones causales no abordadas con anterioridad, como el impacto de la falta de métricas en la propagación de errores.
- El respaldo gráfico que valida las deficiencias expuestas, facilitando la transición hacia la fase de diseño del proceso mejorado (situación ideal).
- La profundización en elementos omitidos o subestimados, tales como la ausencia de mejora continua y la falta de monitoreo sistemático, factores clave para sustentar la automatización.

En conjunto, este diagrama permite una comprensión más precisa de la problemática y fortalece la argumentación técnica que justifica la transformación del proceso. Su integración en la Fase 1 enriquece el análisis y establece una base sólida para un rediseño efectivo, orientado a las necesidades reales del equipo *Data Operations*. Lo explicado anteriormente, se visualiza a continuación en la Figura 12.

Figura 12. Diagrama de Ishikawa del proceso actual de carga de datos



Nota. Elaboración propia (2025)

4.1.6 Análisis FODA de la situación actual del proceso de carga de datos

El presente análisis FODA en la Tabla 17, sintetiza los hallazgos de la Fase 1, evaluando internamente las capacidades y limitaciones del proceso de carga de datos, así como los factores externos que influyen en su desempeño. Esta visión estratégica fundamenta la necesidad de transformación operativa hacia un modelo más eficiente y automatizado.

Tabla 17. Análisis FODA del proceso actual de carga de datos

Análisis FODA	
Fortalezas	**Oportunidades**
Experiencia del equipo operativo: El equipo de *Data Operations* posee un conocimiento profundo del flujo actual y cuenta con habilidades técnicas suficientes para resolver problemas emergentes durante el proceso. **Conocimiento acumulado de errores recurrentes**: La experiencia adquirida facilita	**Disponibilidad de herramientas de automatización**: Soluciones tecnológicas compatibles con la infraestructura actual (ej. Amazon Web Services) ofrecen alternativas viables para mejorar el flujo de trabajo. **Apoyo organizacional para iniciativas de mejora**: La organización se encuentra abierta

Análisis FODA	
la identificación temprana de patrones de error y soluciones prácticas. **Flexibilidad en ajustes inmediatos**: La estructura manual del proceso permite realizar correcciones directas sin dependencia de desarrollos tecnológicos externos. **Acceso a herramientas conocidas:** La disponibilidad de Excel, SharePoint y SQL facilita la manipulación inmediata de datos sin requerir nuevas licencias ni capacitación.	a propuestas que contribuyan a reducir tiempos y minimizar errores operativos. **Espacio para estandarización formal**: La creación de procedimientos documentados y validaciones automatizadas permitiría disminuir la dependencia de criterios individuales. **Demanda creciente de datos confiables**: La necesidad organizacional de información precisa y oportuna justifica inversiones orientadas a mejorar la eficiencia operativa.
Debilidades	**Amenazas**
Intervención manual en todas las etapas: Cada fase del proceso depende de acciones por parte de analistas, lo cual incrementa la posibilidad de errores y ralentiza el flujo. **Ausencia de estandarización formal**: No existen procedimientos documentados ni validaciones sistemáticas que aseguren consistencia y trazabilidad en los datos. **Uso de herramientas limitadas para grandes volúmenes de datos**: Excel y SharePoint presentan restricciones técnicas al manejar archivos extensos, afectando la capacidad operativa. **Dependencia de conocimientos empíricos**: La calidad del proceso recae en el juicio del analista, lo que introduce variabilidad en los resultados entre ciclos. **Falta de monitoreo automatizado**: No se cuenta con mecanismos de alerta y métricas en tiempo real que permitan detectar fallas de manera proactiva.	**Incremento constante en el volumen de datos**: El crecimiento de registros y fuentes sobrepasa la capacidad del proceso actual, generando riesgos de saturación operativa. **Riesgos altos de integridad de datos**: Un error humano sin detección oportuna compromete la confiabilidad de los datos maestros, afectando decisiones críticas. **Obsolescencia funcional del proceso actual**: La permanencia de un flujo manual limita la capacidad de adaptación frente a nuevas exigencias regulatorias y de negocio. **Riesgo de desmotivación del equipo operativo**: La carga laboral repetitiva, junto con la falta de mejora continua en el proceso, podría generar una disminución en la motivación de los analistas. **Riesgo por pérdida de conocimiento clave**: La falta de documentación formal impide garantizar la continuidad del proceso en caso de rotación del personal.

Nota. Elaboración propia (2025)

4.2 Fase 2. Diseño del nuevo proceso de carga de datos

La segunda fase de esta investigación tiene como propósito el diseño de un nuevo proceso de carga de datos que integre herramientas de automatización, alineado con los requerimientos funcionales del equipo *Data Operations*. Esta etapa responde al objetivo específico #2, mediante la elaboración de una propuesta estructurada orientada a reducir tareas manuales, mejorar los tiempos de ejecución y fortalecer la gestión de datos maestros mediante un flujo optimizado.

El desarrollo de esta fase inicia con la identificación de oportunidades de mejora en el proceso actual, enfocándose en puntos críticos que presentan ineficiencias o alta dependencia de

la intervención manual. Posteriormente, se revisan las herramientas tecnológicas disponibles, considerando su compatibilidad con los sistemas existentes. Seguidamente, se diseña conceptualmente el nuevo flujo de trabajo.

Cada actividad se vincula directamente con las variables VA-03 (Diseño del nuevo proceso de carga de datos) y VA-04 (Viabilidad del proceso diseñado), evaluadas mediante indicadores como la documentación del proceso, el nivel de integración con herramientas tecnológicas y la validación de la factibilidad técnica y operativa del diseño propuesto. Los instrumentos utilizados incluyen revisión documental, entrevistas semiestructuradas, matrices comparativas, además de la construcción de diagramas BPMN.

Esta fase establece los lineamientos de la solución planteada, posibilitando visualizar un proceso mejorado, validado y viable, enfocado en resolver las problemáticas identificadas en la etapa anterior.

4.2.1 Identificación de oportunidades de mejora

La identificación de oportunidades de mejora constituye una etapa clave en el desarrollo del nuevo proceso de carga de datos, ya que permite analizar las deficiencias actuales y establecer las bases para su reestructuración. Durante la Fase #1 se evidenciaron múltiples problemáticas asociadas a la alta dependencia de tareas manuales, fragmentación en el flujo de trabajo y ausencia de mecanismos automáticos para validar, consolidar y cargar los datos maestros. Estas limitaciones afectan directamente la eficiencia operativa del equipo *Data Operations*, alargando los tiempos de ejecución y aumentando el riesgo de errores.

Con el fin de profundizar en el diagnóstico, se realizó una entrevista con el *Senior Data Engineer* del equipo, cuyos aportes permitieron precisar las áreas críticas que requieren intervención prioritaria (véase **Apéndice R**). A partir del análisis conjunto de los hallazgos de la Fase #1 y la entrevista, se identificaron las siguientes oportunidades de mejora:

- **Automatización de la obtención de datos**: Actualmente, la información es descargada manualmente desde métodos como *File Transfer Protocol*, correos electrónicos y APIs. Esta actividad representa un punto crítico por su alta repetitividad y riesgo de errores. Se propone automatizar la extracción de datos hacia un repositorio centralizado, lo que permitirá iniciar el procesamiento sin intervención manual.
- **Unificación y estandarización de datos**: Las tareas de consolidación se realizan de forma manual, dificultando la aplicación de estándares consistentes. Es necesario automatizar este proceso, permitiendo que la unificación de la información siga reglas predefinidas que aseguren su integridad y coherencia.
- **Procesamiento manual de datos**: La revisión y validación de datos son ejecutadas por los analistas, lo cual consume una cantidad significativa de tiempo y recursos. Se identificó la oportunidad de automatizar estas funciones mediante el desarrollo de procesos lógicos que reproduzcan de forma automática las tareas actualmente realizadas.
- **Detección y corrección de errores frecuentes**: Se presentan discrepancias comunes, tales como campos vacíos, duplicados y formatos inconsistentes. La incorporación de

validaciones automáticas permitirá detectar estos errores de forma inmediata y reducir significativamente su ocurrencia.

- **Generación de reportes y monitoreo**: El proceso actual carece de mecanismos automáticos de monitoreo. Se requiere implementar registros automáticos (logs) que faciliten la supervisión de cada ejecución y permitan generar reportes periódicos para evaluar el desempeño del flujo.

Adicionalmente, se estableció la necesidad de mantener ciertas integraciones con sistemas actuales, tales como la base de datos PostgreSQL, utilizada en la carga final de datos, y las fuentes originales, las cuales seguirán siendo utilizadas, aunque con un enfoque distinto en la forma de extracción.

Estas oportunidades de mejora representan áreas clave para la transformación del proceso, enfocándose en la reducción de la intervención manual, el fortalecimiento de la calidad de los datos y una mayor eficacia en el uso de los recursos del equipo. La identificación detallada de estos aspectos permitirá orientar el diseño conceptual del nuevo flujo de trabajo hacia una solución viable y técnicamente sustentada.

A partir de estas observaciones, en la entrevista se mencionaron métricas clave que facilitarán la evaluación objetiva del impacto del nuevo proceso en comparación con el actual. Estas métricas permitirán establecer parámetros objetivos para medir el éxito del nuevo proceso una vez diseñado y validar su viabilidad tanto técnica como operativa.

- Reducción en la cantidad de discrepancias en los datos cargados, tales como campos vacíos, duplicados y errores de formato.
- Disminución del tiempo total de ejecución del proceso, cuyo objetivo es pasar de un plazo actual de hasta tres días a una duración inferior a un día.
- Reducción significativa de tareas manuales, con una meta inicial de automatización estimada en un 60% del flujo operativo actual.
- Mejora en la calidad de los datos, mediante validaciones automáticas que aseguren la integridad y consistencia de la información procesada.
- Implementación de mecanismos de monitoreo, a través de la generación automática de registros y reportes que faciliten el seguimiento continuo del desempeño del proceso.

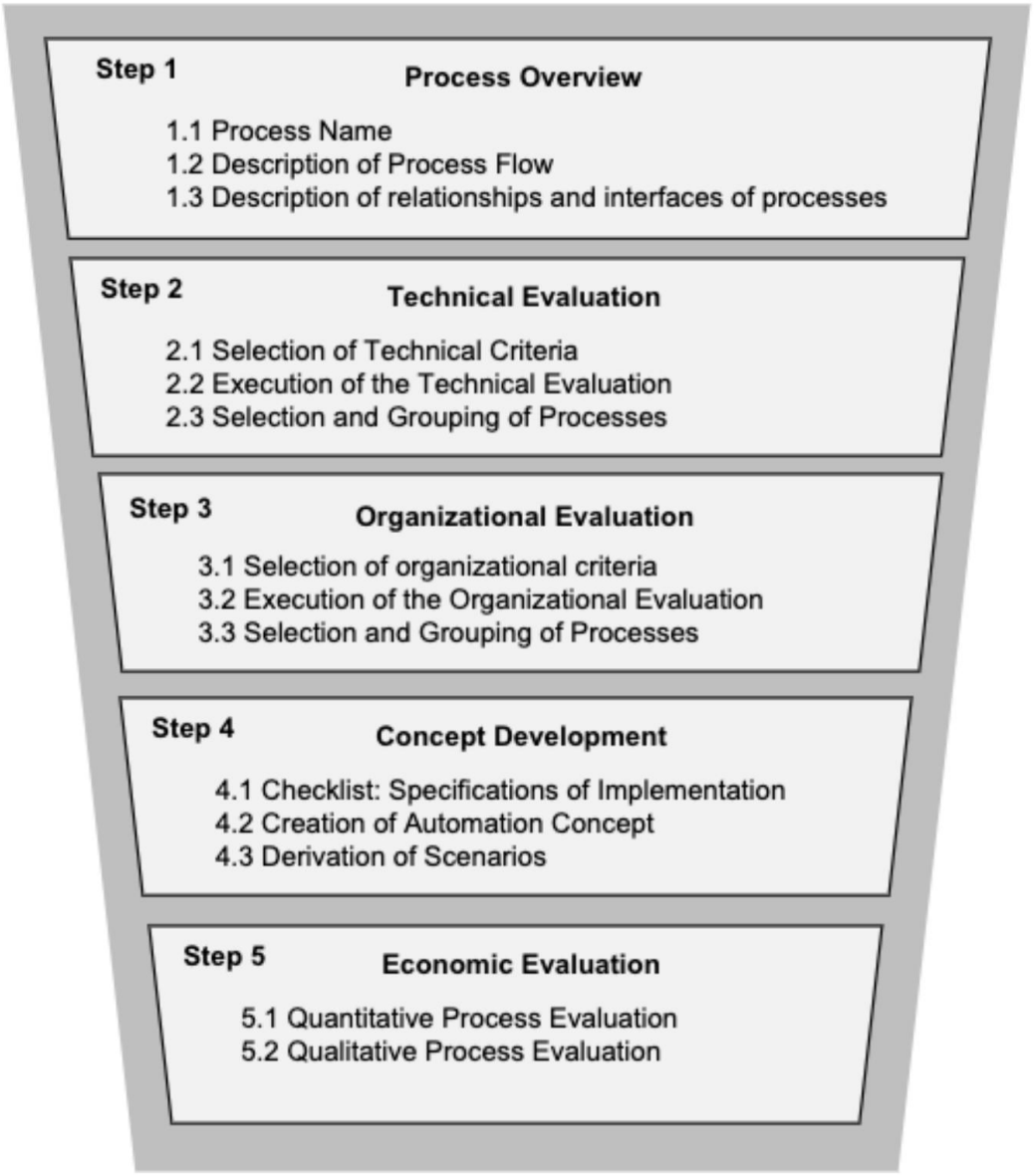
4.2.2 Visión general del nuevo proceso de carga de datos.

La visión general del nuevo proceso de carga de datos parte de la necesidad de estructurar un flujo de trabajo más eficiente, confiable y técnicamente sustentado, en respuesta a las limitaciones detectadas en el proceso actual. La propuesta busca establecer una estructura lógica que reduzca la intervención manual, fortalezca la calidad de los datos y garantice un flujo de ejecución controlado.

Para guiar el análisis de las actividades actuales y estructurar el diseño del nuevo proceso, se utilizará el método PROTEOCE, el cual propone una evaluación sistemática basada en cinco etapas: visión general del proceso, evaluación técnica, evaluación organizacional, desarrollo de

conceptos y evaluación económica (Nguyen et al., 2021, p. 437). Este enfoque metodológico permite considerar de manera integral los factores técnicos, organizativos y económicos que influyen en la viabilidad de automatizar un proceso, asegurando que la propuesta resultante responda de manera efectiva a las necesidades identificadas en el contexto actual de operación. A continuación en la Figura 13, se visualizan los pasos que comprenden el método PROTEOCE.

Figura 13. Estructura del método PROTEOCE



Nota. Tomado de Automation? Yes … But Where to Begin?, por Nguyen, T et al (2021)

La revisión del flujo operativo vigente, realizada en la Fase #1 de esta investigación, junto con los aportes obtenidos en la entrevista al *Senior Data Engineer* (véase **Apéndice R**), permitió

identificar las actividades críticas que impactan de manera directa la eficiencia, la calidad y la trazabilidad del proceso. A partir de este análisis, se seleccionaron aquellas actividades que, por su naturaleza repetitiva, su propensión a errores manuales o su influencia en los tiempos de ejecución, resultan prioritarias para su automatización.

En la Tabla 18, se incluyen las actividades relevantes del proceso de carga de datos identificadas para automatización.

Tabla 18. Actividades relevantes para la automatización del proceso de carga de datos

Actividades relevantes para la automatización del proceso de carga de datos		
Actividad	**Actores**	**Tipo de actividad**
Descargar archivos desde SharePoint, correos electrónicos, APIs y *File Transfer Protocol*.	Analista de *Data Operations*.	Manual, mediante diferentes canales.
Consolidar archivos de distintas fuentes en hojas de cálculo.	Analista de *Data Operations*.	Manual, mediante herramienta tecnológica (Microsoft Excel)
Estandarizar columnas y estructuras de los archivos recibidos.	Analista de *Data Operations*.	Manual, mediante herramienta tecnológica (Microsoft Excel)
Verificar la integridad de los datos (campos vacíos, duplicados, formatos incorrectos).	Analista de *Data Operations*.	Manual, mediante herramienta tecnológica (Microsoft Excel)
Realizar la carga de datos en la plataforma de Gestión de Datos Maestros mediante scripts SQL.	Analista de *Data Operations*.	Manual, mediante herramienta tecnológica (PostgreSQL)
Realizar controles posteriores a la carga revisando registros cargados.	Analista de *Data Operations*.	Manual, se utiliza plantilla (ver **Anexo I**)
Notificar errores detectados durante el control de carga a los responsables de origen.	Analista de *Data Operations*.	Manual, se utiliza guía /plantilla (ver **Anexo II**)

Nota. Elaboración propia (2025)

La selección de estas actividades se fundamentó en criterios claramente definidos. Para definir los criterios clave que guían la automatización del proceso de carga de datos, se tomó como referencia la categorización propuesta por Taulli (2020, pp. 45-46) en *The Robotic Process Automation Handbook*, la cual identifica las características que hacen que un proceso sea adecuado para su automatización. Se estableció que, de los trece criterios planteados, nueve aplican de forma directa al proceso actual. Estos criterios son: trabajo tedioso, pérdida de tiempo, repetitividad,

frecuencia, existencia de reglas claras, definición precisa de los procesos, alto volumen de datos, propensión al error y manejo de información sensible.

La Tabla 19 detalla los criterios aplicables, evidenciando cómo cada uno de ellos se manifiesta en las actividades actuales del proceso de carga de datos. Este análisis permite justificar de manera técnica la selección de actividades para su rediseño y posterior automatización, fortaleciendo la alineación de la propuesta con las necesidades operativas identificadas.

Tabla 19. Criterios clave para la automatización del proceso de carga de datos

Criterios clave para la automatización del proceso de carga de datos	
Trabajo tedioso	El proceso implica tareas altamente repetitivas: descargas manuales, consolidación de archivos en Excel, revisiones de datos fila por fila.
Pérdida de tiempo	El proceso actual tarda hasta 3 días, consumiendo recursos que podrían ser asignados a tareas de mayor valor.
Repetitivo	Las acciones son rutinarias y siguen siempre el mismo patrón: descarga → consolidación → validación → carga.
Frecuencia	La carga de datos ocurre de forma mensual (o incluso más frecuente en algunos casos), lo cual justifica la automatización.
Basado en reglas	Validaciones como "no debe haber campos vacíos", "formato de fecha estándar" o "estructura de columnas fija" siguen reglas claras.
Procesos claramente definidos	Aunque el proceso actual es manual, los pasos están definidos de forma concreta, lo cual facilita su modelado para automatización.
Alto volumen	Maneja grandes cantidades de registros y archivos provenientes de diversas fuentes, lo cual hace que el procesamiento manual sea insostenible.
Propenso al error	Se detectan errores manuales frecuentes: datos inconsistentes, errores de formato, cargas incompletas.
Datos sensibles	La información que se gestiona es crítica para la operación interna y la calidad de los datos maestros, lo que requiere procesos controlados y seguros.

Nota. Los criterios fueron tomados de The Robotic Process Automation Handbook (Taulli, 2020, pp. 45–46).

4.2.3 Evaluación técnica del nuevo proceso de carga de datos

La segunda etapa del método PROTEOCE corresponde a la evaluación técnica del proceso actual, con el objetivo de determinar si las actividades identificadas en la etapa previa cumplen con los requisitos mínimos que permitan su automatización de forma viable. Este paso resulta esencial para garantizar que el rediseño posterior no solo sea deseable desde una perspectiva operativa, sino también factible desde el punto de vista tecnológico, evitando así inversiones ineficientes o propuestas insostenibles.

Según lo establecido por Nguyen et al. (2021), la evaluación técnica consiste en definir un conjunto de criterios estandarizados que permitan clasificar las actividades como técnicamente aptas para automatización. En el contexto de este proyecto, se consideraron cinco criterios específicos, adaptados tanto del estado del arte en automatización como de la revisión realizada en la Fase #1 y la entrevista con el *Senior Data Engineer* (véase **Apéndice R**).

A continuación, en la Tabla 20 se presentan los criterios técnicos aplicados al proceso actual de carga de datos maestros.

Tabla 20. Criterios técnicos del proceso actual de carga de datos

Criterios técnicos para la automatización	
Criterio Técnico	**Justificación**
Repetitividad	Las tareas se repiten mensualmente siguiendo un mismo patrón operativo.
Basado en reglas	La verificación de estructuras, formatos y campos responde a reglas predefinidas y estandarizadas.
Estabilidad del flujo	No se presentan cambios frecuentes en la lógica del proceso ni en las fuentes de origen de los datos.
Volumen de información significativo	Se gestionan archivos extensos, con alta densidad de registros que dificultan el procesamiento manual.
Accesibilidad tecnológica	Las fuentes de datos están disponibles mediante *SharePoint*, correo electrónico, *APIs* y *FTPs*, con herramientas que ofrecen conectividad directa.

Nota. Elaboración propia (2025)

En la Tabla 20, se determina que todas las actividades priorizadas en la etapa anterior cumplen con estos cinco criterios, lo cual permite concluir que el proceso actual es técnicamente viable para su automatización. Esta validación constituye la base para avanzar hacia la evaluación organizacional, donde se analizará el impacto operativo de la propuesta y el grado de alineación con las funciones del equipo *Data Operations*.

4.2.4 Evaluación organizacional del nuevo proceso de carga de datos

Una vez validada la viabilidad técnica de las actividades actuales mediante la evaluación anterior, se procede a analizar su valor estratégico y operativo dentro del entorno organizacional. Esta etapa, correspondiente a la Evaluación Organizacional del método PROTEOCE, tiene como finalidad determinar en qué medida la automatización del proceso contribuye de forma efectiva al cumplimiento de los objetivos organizacionales y al fortalecimiento de la operación (Nguyen et al., 2021).

La Evaluación Organizacional resulta esencial para asegurar que los recursos destinados a la automatización se enfoquen en procesos que no solo son técnicamente viables, sino también relevantes para la eficiencia global, el aprovechamiento de los recursos humanos y la mejora de la calidad de la información. Para definir los criterios organizacionales utilizados en esta evaluación, se consideraron los hallazgos obtenidos durante la Fase #1 de diagnóstico del proceso de carga de datos, así como los aportes recopilados en la entrevista con el *Senior Data Engineer* (véase **Apéndice R**), lo que permitió adaptar los criterios al contexto real de operación del equipo *Data Operations*.

Para llevar a cabo esta evaluación, se definieron criterios específicos adaptados al contexto de la carga de datos maestros, a los cuales se asignaron pesos relativos según su importancia estratégica. Posteriormente, se aplicó la fórmula matemática propuesta en el método PROTEOCE para calcular el nivel de prioridad del proceso evaluado.

4.2.4.1 Definición de criterios organizacionales

Para orientar la priorización de la automatización del proceso de carga de datos maestros, se establecieron los criterios organizacionales que reflejan las características específicas del entorno operativo y las necesidades estratégicas del equipo *Data Operations*. La selección de estos criterios responde a la necesidad de evaluar no solo la viabilidad técnica, sino también el valor estratégico y operativo del proceso (Nguyen et al., 2021). A continuación en la Tabla 21, se definen los criterios organizacionales.

Tabla 21. Criterios organizacionales para su evaluación

Criterios organizacionales a evaluar	
Criterio organizacional	**Descripción**
Nivel de uso del proceso	Frecuencia con la que el proceso es ejecutado dentro de las operaciones habituales del equipo *Data Operations*.
Impacto en la operación	Grado en que la calidad o eficiencia del proceso influye en resultados críticos de la organización, como la gestión de datos maestros.
Dependencia de recursos humanos	Nivel de esfuerzo manual requerido para completar el proceso, afectando la disponibilidad del personal para tareas de mayor valor agregado.

Valor estratégico para la organización	Contribución del proceso a la mejora de la calidad de los datos, a la eficiencia interna y al cumplimiento de objetivos estratégicos del área de datos.
Facilidad de adopción organizacional	Grado de familiaridad del equipo con las herramientas involucradas y la facilidad de transición hacia flujos de trabajo automatizados.

Nota. Elaboración propia (2025)

4.2.4.2 Asignación de pesos

Con el propósito de reflejar la importancia relativa de cada criterio organizacional en la evaluación del proceso de carga de datos maestros, se procedió a asignar un peso específico a cada uno de ellos. Esta ponderación responde a la necesidad de priorizar aquellos aspectos que tienen un mayor impacto en la operación y en la estrategia del equipo *Data Operations*. El método PROTEOCE establece que los pesos deben ser definidos de forma proporcional, asegurando que su suma total alcance el 100%, lo cual permite aplicar la fórmula de puntuación organizacional de manera estandarizada (Nguyen et al., 2021).

La asignación de pesos se fundamentó en el análisis crítico del contexto operativo actual, considerando factores como la frecuencia de uso del proceso, el impacto en resultados críticos, la carga de trabajo manual implicada, el valor estratégico para la organización y la facilidad de adopción de nuevas soluciones tecnológicas. La Tabla 22 presenta la distribución de pesos asignada a cada criterio.

Tabla 22. Distribución de pesos asignada a cada criterio organizacional

Distribución de pesos asignada a cada criterio	
Criterio organizacional	**Peso asignado (%)**
Nivel de uso del proceso	20%
Impacto en la operación	30%
Dependencia de recursos humanos	20%
Valor estratégico para la organización	20%
Facilidad de adopción organizacional	10%

Nota. Elaboración propia (2025)

Estos pesos reflejan una priorización estratégica, otorgando mayor relevancia al impacto operativo del proceso y a su nivel de uso dentro de las operaciones diarias. La dependencia de recursos humanos y el valor estratégico también poseen una ponderación significativa, dada su influencia directa en la eficiencia del equipo. Finalmente, la facilidad de adopción tecnológica, aunque importante, recibe un peso menor en comparación con los demás criterios, dado que el equipo cuenta con familiaridad previa con las herramientas base.

4.2.4.3 Evaluación del proceso según los criterios

Una vez definidos los criterios organizacionales y sus respectivos pesos, se procede a evaluar el proceso de carga de datos maestros en función de su desempeño respecto a cada criterio. La evaluación se realizó utilizando una escala de valoración de tres niveles:

- 1: Bajo cumplimiento del criterio
- 2: Cumplimiento intermedio o parcial
- 3: Alto cumplimiento del criterio

Esta metodología, propuesta en el método PROTEOCE (Nguyen et al., 2021), permite asignar una puntuación objetiva a cada criterio y, posteriormente, calcular un índice de prioridad organizacional para el proceso analizado. La Tabla 23 presenta la puntuación asignada al proceso de carga de datos para cada criterio organizacional.

Tabla 23. Puntuación asignada al proceso de carga de datos para cada criterio organizacional

Puntuación asignada al proceso de carga de datos para cada criterio organizacional		
Criterio organizacional	**Peso asignado (%)**	**Puntuación asignada**
Nivel de uso del proceso	20%	3
Impacto en la operación	30%	3
Dependencia de recursos humanos	20%	3
Valor estratégico para la organización	20%	3
Facilidad de adopción organizacional	10%	2

Nota. Elaboración propia (2025)

Justificación de las puntuaciones asignadas:

- El nivel de uso del proceso recibió la máxima puntuación (3) al tratarse de un proceso recurrente y crítico en el flujo de trabajo mensual del equipo.
- El impacto en la operación también obtuvo puntuación alta (3), debido a que la calidad del proceso de carga afecta directamente la integridad de los datos maestros utilizados por otras áreas estratégicas.
- La dependencia de recursos humanos fue evaluada con la máxima puntuación (3), dado el nivel actual de intervención manual requerida.
- El valor estratégico para la organización fue igualmente evaluado con (3), ya que la automatización de este proceso liberaría capacidad operativa para actividades de mayor valor, como la gestión de calidad de datos.
- La facilidad de adopción organizacional recibió una puntuación de (2), ya que, aunque existe familiaridad con las herramientas base, la adopción de flujos automáticos en herramientas tecnológicas disponibles implica una curva de aprendizaje moderada.

4.2.4.4 Cálculo del Score organizacional

Con las puntuaciones asignadas y los pesos definidos para cada criterio organizacional, se procede al cálculo del *Score* organizacional del proceso de carga de datos maestros. Este cálculo se realizó aplicando la fórmula matemática establecida en el método PROTEOCE (Nguyen et al., 2021).

$$Score_i = \sum_{j=1}^{n} p_{ij} \times w_j$$

Dónde:

- $Score_i$ = Puntuación organizacional total del proceso i
- p_{ij} = Puntuación asignada al proceso i en el criterio j.
- w_j = Peso relativo asignado al criterio j.
- n = Número total de criterios evaluados.

A continuación, en la Tabla 24 se muestra la aplicación práctica del cálculo.

Tabla 24. Score organizacional

Score organizacional			
Criterio organizacional	**Peso asignado (%)**	**Puntuación asignada**	**Producto (Peso * Puntuación)**
Nivel de uso del proceso	20%	3	$0.20 \times 3 = 0.60$
Impacto en la operación	30%	3	$0.30 \times 3 = 0.90$
Dependencia de recursos humanos	20%	3	$0.20 \times 3 = 0.60$
Valor estratégico para la organización	20%	3	$0.20 \times 3 = 0.60$
Facilidad de adopción organizacional	10%	2	$0.10 \times 2 = 0.20$

Nota. Elaboración propia (2025)

Interpretación del *Score* obtenido:

El proceso de carga de datos maestros alcanzó un *Score* organizacional total de 2.90 sobre un máximo de 3.00, lo cual indica que presenta una prioridad organizacional muy alta para ser automatizado. Esta puntuación confirma que la automatización de este proceso no solo es técnicamente viable, sino también estratégica para mejorar la eficiencia, liberar recursos y fortalecer la calidad de los datos en la organización.

4.2.5 Desarrollo de conceptos

El desarrollo de conceptos constituye la cuarta etapa del método PROTEOCE, cuyo objetivo principal es formular un concepto claro y estructurado que guíe el diseño del nuevo

proceso de carga de datos maestros. Esta etapa toma como base los resultados obtenidos en las evaluaciones técnica y organizacional, buscando definir las especificaciones funcionales y los escenarios de solución que permitan materializar el rediseño del proceso.

En el contexto de esta investigación, el concepto de automatización se fundamentará en el uso de servicios de Amazon Web Services (AWS), en virtud de la validación realizada durante la entrevista al *Senior Data Engineer* (véase **Apéndice R**). Durante dicha entrevista, se destacó la existencia de una alianza estratégica entre la organización y AWS, así como la disponibilidad de infraestructura, licenciamiento y soporte técnico necesarios para implementar soluciones basadas en esta plataforma. Además, se reconoció que AWS proporciona un conjunto de herramientas maduras y compatibles con las necesidades del proceso de carga de datos, facilitando la integración con fuentes de datos, y ofreciendo capacidades robustas para la transformación, validación, orquestación y monitoreo de datos.

Si bien no se realizó una evaluación exhaustiva de plataformas alternativas, se consideró de forma preliminar la posibilidad de implementar la automatización del flujo en entornos como Microsoft Azure o Google Cloud Platform, ambos reconocidos ampliamente en la industria. Estas plataformas ofrecen servicios equivalentes para almacenamiento, ejecución de funciones sin servidor y orquestación de procesos; no obstante, presentan restricciones más marcadas en sus versiones gratuitas. El *Free Tier* de Azure impone limitaciones significativas en duración, almacenamiento y cantidad de ejecuciones mensuales, mientras que Google Cloud requiere configuraciones adicionales para alcanzar una orquestación comparable a la que proporciona AWS Step Functions.

Esta valoración fue discutida directamente con el *Senior Data Engineer* (véase **Apéndice R**) de la organización, quien confirmó que, si bien técnica y conceptualmente las tres plataformas permiten alcanzar objetivos similares, AWS representa la alternativa más robusta, accesible y alineada con el nivel de madurez requerido para el desarrollo de un prototipo académico funcional. Por ello, se concluyó que Amazon Web Services era la opción más viable para validar la propuesta técnica del proyecto sin incurrir en costos adicionales.

La definición del concepto de automatización incluirá la construcción de un *checklist* de especificaciones técnicas y organizativas, así cómo la formulación de la propuesta de flujo de trabajo automatizado del proceso de carga de datos, siguiendo el enfoque metodológico propuesto por Nguyen et al. (2021).

4.2.5.1 Checklist de requerimientos para la automatización

Con base en la evaluación técnica, la evaluación organizacional y los aportes recopilados durante la entrevista, se definió un conjunto de especificaciones técnicas y organizativas que deben ser consideradas en el diseño del nuevo proceso de carga de datos maestros. Estas especificaciones constituyen el punto de partida para estructurar el concepto de automatización, asegurando que las condiciones reales de operación y los requerimientos funcionales del equipo *Data Operations* se vean reflejados en la propuesta de solución.

La Tabla 25 presenta el *checklist* de requerimientos elaborado.

Tabla 25. Checklist de requerimientos para la automatización

Requerimientos para la automatización		
Área	**Especificación**	**Herramienta a utilizar**
Orquestación de extracción	Coordinar la descarga automática de archivos desde múltiples fuentes externas hacia la zona Bronze del bucket de almacenamiento.	AWS Step Functions
Extracción de datos	Ejecutar funciones independientes que extraen archivos desde las fuentes definidas.	AWS Lambda
Almacenamiento inicial	Centralizar los archivos sin procesar en una zona de almacenamiento seguro y segmentado por capas.	Amazon S3
Transformación de datos	Validar automáticamente la estructura de cada archivo, eliminar duplicados y estandarizar los formatos de forma modular.	AWS Lambda
Validación de consistencia	Interrumpir el flujo en caso de errores estructurales detectados durante la transformación.	AWS Step Functions
Almacenamiento estructurado	Guardar los archivos validados en una zona diferenciada para preparación de consolidación.	Amazon S3
Consolidación de datos	Integrar los archivos transformados bajo un esquema unificado mediante un proceso de consolidación centralizado.	AWS Glue
Almacenamiento final	Conservar los datos consolidados listos para la carga estructurada.	Amazon S3
Carga en base de datos	Insertar automáticamente los datos consolidados en la base de datos relacional utilizada por la plataforma de gestión de datos maestros.	AWS Lambda / Amazon Aurora

Requerimientos para la automatización		
Publicación en plataforma de Gestión de datos Maestros	Acceder a los datos desde la base estructurada para su posterior validación y publicación dentro de la plataforma MDM.	Plataforma de Gestión de Datos Maestros / Amazon Aurora
Monitoreo del proceso	Registrar logs de ejecución, errores y métricas de rendimiento en cada etapa del flujo automatizado.	Amazon CloudWatch

Nota. Elaboración propia (2025)

4.2.5.2 Formulación del concepto de automatización

El concepto de automatización del proceso de carga de datos maestros se estructura a partir de las especificaciones técnicas y organizativas definidas en el *checklist* previo. El objetivo principal es transformar el flujo de trabajo actual, altamente manual y susceptible a errores, en una secuencia automatizada, controlada y trazable, apoyada en los servicios de Amazon Web Services (AWS).

El flujo conceptual del nuevo proceso comprende las siguientes etapas.

1. Extracción de datos

La extracción de archivos provenientes de fuentes como *Credit Risk Data, Ratings, News Edge, ESG* y *Corporate Intelligence Database* se ejecuta mediante una AWS Step Function denominada "Source to Landing", la cual orquesta múltiples AWS Lambda Functions que descargan los datos y los almacenan en bruto en la zona Bronze de un bucket Amazon S3, que actúa como *Landing Zone*. En caso de fallos durante la extracción, la ejecución se detiene y los errores se registran automáticamente en Amazon CloudWatch

2. Transformación de datos

Cada archivo depositado en Bronze es procesado en Silver por una serie de funciones Lambda que realizan limpieza, validación de campos, eliminación de duplicados y estandarización de estructuras. Esta fase ocurre en la zona Silver de Amazon S3, asegurando que los datos cumplan con los requisitos estructurales mínimos.

Si un archivo no cumple con las condiciones de validación, se genera un log en CloudWatch y el proceso no avanza hasta que se subsane la inconsistencia.

3. Consolidación de datos

Una vez validados individualmente, los datos se consolidan mediante un AWS Glue Job, el cual integra bajo un estándar unificado los distintos archivos transformados y estructurados por las múltiples funciones Lambda ejecutadas previamente. Esta consolidación representa la preparación final de los datos, que son almacenados en la zona Gold del mismo bucket S3.

Esta actividad es gestionada por la Step Function "Landing to Staging", la cual evalúa que todos los datasets requeridos estén disponibles antes de ejecutar el Glue Job. Si el proceso de consolidación falla, se genera una entrada de error en Amazon CloudWatch y se interrumpe la ejecución hasta que el incidente sea diagnosticado.

4. Carga en base de datos estructurada

Los datos consolidados en Gold son transferidos mediante una función Lambda a Amazon Aurora (PostgreSQL), base de datos relacional que actúa como repositorio final estructurado. En esta etapa se verifica automáticamente si la carga fue exitosa; de lo contrario, se genera una entrada de error en CloudWatch.

5. Publicación y acceso a los datos maestros en la plataforma

Una vez almacenados en Aurora, los datos son consultados por la plataforma de Gestión de Datos Maestros, donde se realiza la validación interna final para su publicación. Si se detectan inconsistencias o fallos en la lectura, se revisan los registros de ejecución correspondientes en CloudWatch.

6. Monitoreo automatizado del flujo

Todas las actividades clave, incluyendo extracción, transformación, consolidación y carga son monitoreadas en tiempo real mediante Amazon CloudWatch, el cual genera logs detallados, métricas de ejecución, alertas de error y reportes de control.

7. Orquestación del flujo

La ejecución secuencial y lógica del proceso completo está gestionada mediante dos AWS Step Functions:

- **"Source to Landing"**: gestiona la extracción y almacenamiento inicial en Bronze.
- **"Landing to Staging"**: coordina la transformación en Silver, consolidación en Gold y preparación para carga.

Estas funciones permiten rutas condicionales y control total sobre la ejecución de Lambda y Glue, en sincronía con los criterios de éxito definidos para cada etapa.

De acuerdo con lo expuesto anteriormente, se espera que este concepto de automatización permita reducir significativamente los tiempos de ejecución, minimizar la intervención manual, mejorar la calidad de los datos maestros y fortalecer el control operativo, en cumplimiento de los objetivos planteados en esta investigación.

Finalmente, para garantizar la protección de la información manipulada en cada una de las etapas del flujo automatizado, se establecen medidas de seguridad basadas en control de accesos. La protección de la información procesada se sustenta en un modelo de control de accesos que combina las políticas internas de la organización con los mecanismos nativos de Amazon Web Services. La autenticación de usuarios y equipos técnicos se gestiona mediante Okta Verify,

solución corporativa que permite aplicar autenticación multifactor (MFA) y gestión centralizada de identidades bajo políticas de acceso definidas por rol.

En el entorno de AWS, las acciones permitidas sobre los recursos son controladas mediante AWS Identity and Access Management (IAM). Estas configuraciones limitan el acceso a buckets de Amazon S3, funciones Lambda, procesos de consolidación en Glue, estructuras almacenadas en Aurora y registros operativos en CloudWatch. Cada recurso dispone de permisos diferenciados según su nivel de sensibilidad o criticidad operativa.

Las políticas aplicadas responden al principio de privilegios mínimos. Solo los componentes autorizados del sistema, junto con los usuarios autenticados mediante Okta, tienen permiso para interactuar con los datos. Esta combinación asegura confidencialidad en el acceso, integridad en el tratamiento de la información y trazabilidad en cada etapa del proceso.

4.2.5.3 Selección del enfoque final

Como resultado del desarrollo conceptual realizado, se estableció un enfoque único y validado para la automatización del proceso de carga de datos maestros, basado en el cumplimiento de las especificaciones técnicas y organizativas definidas previamente. Este enfoque se apoya en el ecosistema de herramientas de Amazon Web Services (AWS), seleccionado por su compatibilidad con las necesidades operativas del equipo Data Operations, su disponibilidad organizacional y su capacidad para integrar de forma robusta las diferentes etapas del proceso.

La solución conceptual propuesta contempla la automatización de la extracción, transformación, consolidación, carga y monitoreo de los datos mediante servicios como Amazon S3, AWS Glue, Amazon Aurora, AWS Lambda, Amazon CloudWatch y AWS Step Functions, orquestando un flujo continuo, seguro y trazable de principio a fin. Esta estructura responde a las limitaciones identificadas en el análisis inicial y se alinea plenamente con los objetivos establecidos para esta investigación.

Al no haberse planteado múltiples escenarios alternativos, se concluye que el enfoque seleccionado representa la opción más viable y estratégica, tanto desde el punto de vista técnico como organizacional. Este concepto será la base para el diseño detallado del nuevo proceso en las etapas siguientes.

4.2.6 Evaluación económica

La etapa final del método PROTEOCE contempla la realización de una evaluación económica destinada a determinar la factibilidad financiera de la propuesta de automatización, a través de análisis cuantitativos como el cálculo del periodo de recuperación de la inversión (*Payback Period*) y evaluaciones cualitativas de beneficios operativos (Nguyen et al., 2021).

No obstante, en el marco metodológico establecido para este Trabajo Final de Graduación, el análisis de la viabilidad económica de la propuesta ha sido programado para ser desarrollado en el Capítulo 5: Propuesta de solución, conforme a las directrices oficiales que indican la inclusión de un análisis costo-beneficio integral en etapas posteriores del proyecto. Según estas directrices, en el capítulo correspondiente se incluirá una evaluación detallada de los costos de

implementación, los beneficios esperados y las razones financieras que sustenten la viabilidad de la solución planteada.

Por lo tanto, en esta Fase #2 se omite la aplicación de la evaluación económica prevista en el método PROTEOCE, reservando el análisis financiero para las etapas finales del Trabajo Final de Graduación, donde se abordará de forma estructurada y completa.

4.2.7 Diagrama *To-Be* del proceso de carga de datos maestros

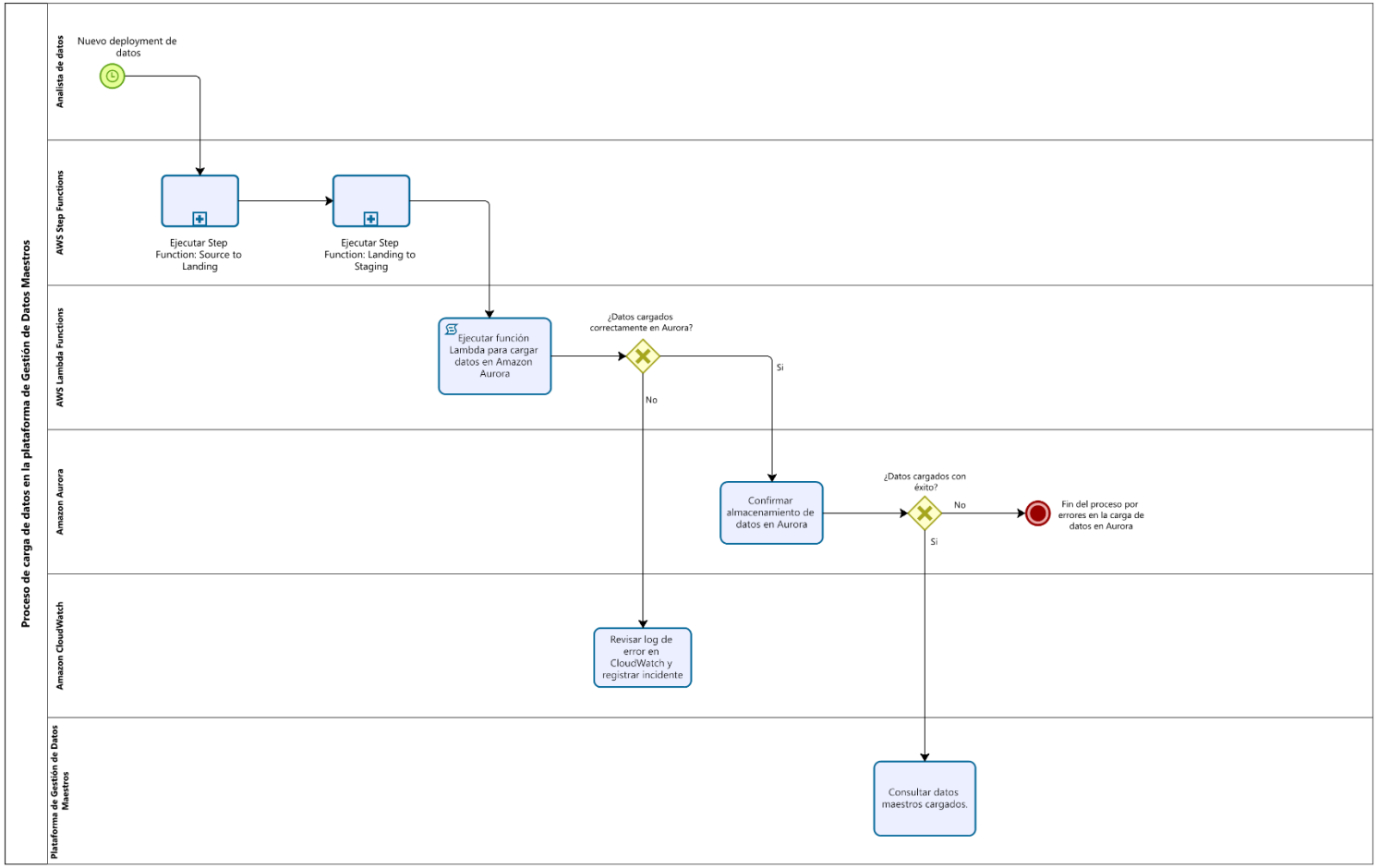
La representación gráfica del proceso *To-Be* corresponde al modelo detallado del flujo automatizado previamente formulado en la sección **4.2.5.2**. Este diagrama fue construido utilizando la notación BPMN 2.0, estándar internacional para la documentación de procesos de negocio, lo cual permite plasmar de manera clara y estructurada las actividades, decisiones, actores técnicos y herramientas involucradas en cada etapa del nuevo flujo.

El modelo refleja fielmente la arquitectura definida sobre servicios de Amazon Web Services, incluyendo la ejecución secuencial de tareas orquestadas mediante AWS Step Functions, el procesamiento distribuido mediante AWS Lambda, la consolidación centralizada a través de AWS Glue y el almacenamiento segmentado en zonas Bronze, Silver y Gold dentro de Amazon S3. Asimismo, se incorporan compuertas de decisión que representan puntos críticos de validación lógica en el proceso, y anotaciones que evidencian el monitoreo técnico automatizado gestionado mediante Amazon CloudWatch.

Cada *lane* del diagrama corresponde a un componente específico del entorno, ya sea una función técnica, un servicio de orquestación, un sistema de almacenamiento o una plataforma de publicación como MDM. La estructura general responde a los principios de modularidad, trazabilidad y escalabilidad definidos en el diseño conceptual, permitiendo visualizar con precisión la interacción entre cada elemento del sistema.

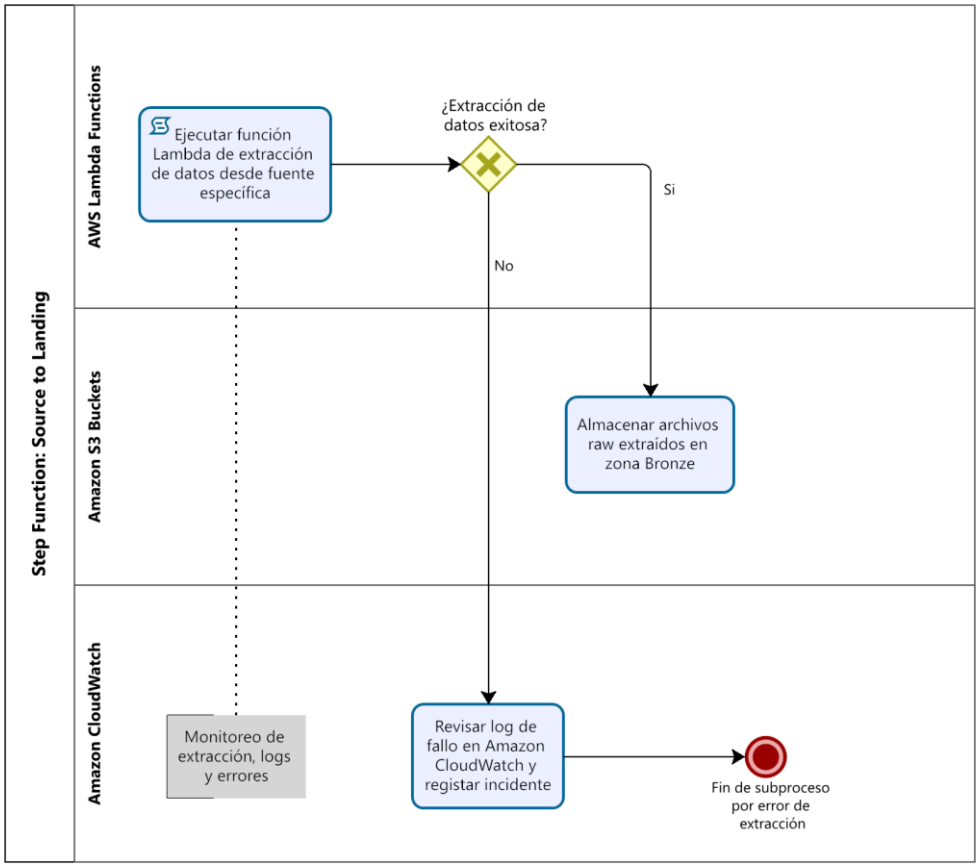
El diagrama *To-Be*, por tanto, no solo representa la secuencia operativa del nuevo proceso automatizado, sino que constituye un instrumento clave para validar su viabilidad técnica, delimitar responsabilidades operativas y facilitar su posterior implementación. A continuación en la Figura 14 se visualiza el diagrama *To-Be* del proceso de carga de datos maestros. En la Figura 15 y 16 se observan los subprocesos correspondientes.

Figura 14. Diagrama To-Be del proceso de carga de datos maestros



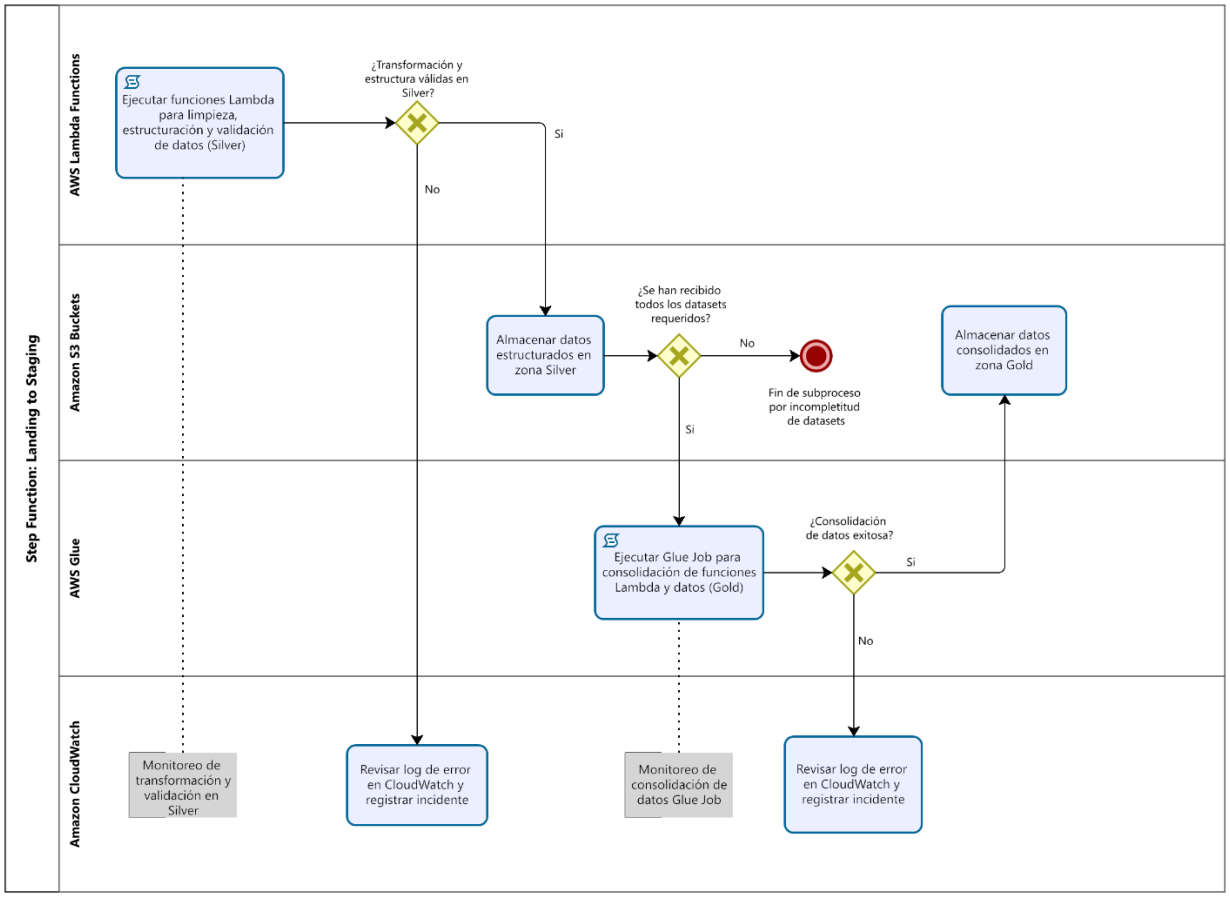
Nota. Elaboración propia (2025)

Figura 15. Subproceso: Ejecutar Step Function: Source to Landing



Nota. Elaboración propia (2025)

Figura 16. Subproceso: Ejecutar Step Function: Landing to Staging



Nota. Elaboración propia (2025)

4.2.8 Matriz requerimientos vs Diseño

Con el propósito de validar que cada requerimiento funcional identificado durante el análisis fue debidamente resuelto dentro del diseño técnico propuesto, se construye la matriz Requerimientos vs. Diseño. Este instrumento permite verificar la cobertura de necesidades mediante la comparación directa entre las especificaciones operativas levantadas y los componentes implementados en el flujo automatizado representado en el diagrama *To-Be*.

Esta metodología encuentra respaldo en el enfoque de validación temprana de requerimientos mediante modelos de diseño, planteado por Stachtiari et al. (2018, pp. 50–67), quienes afirman que al comparar los requerimientos con un modelo técnico derivado es posible detectar de forma anticipada inconsistencias o vacíos en la solución. Si el diseño no satisface las propiedades esperadas, este debe ajustarse o bien reformularse parte del requerimiento. Tal enfoque refuerza el uso de este tipo de matrices como herramienta de alineación entre lo conceptual y lo implementado.

En el contexto del presente proyecto, esta matriz cumple una función clave dentro de la Fase 2 del análisis de resultados, al permitir demostrar de manera estructurada cómo cada requerimiento ha sido operacionalizado en decisiones tecnológicas concretas mediante servicios de Amazon Web Services. Asimismo, su construcción permite validar el cumplimiento del objetivo específico #2, el cual busca rediseñar el proceso de carga de datos incorporando automatización, control y trazabilidad. A continuación en la Tabla 26, se muestra la matriz requerimientos vs diseño.

Tabla 26. Matriz requerimientos vs diseño

Matriz requerimientos vs diseño		
Área	**Especificación del requerimiento**	**Solución implementada en el diseño *To-Be***
Orquestación de extracción	Coordinar la descarga automática de archivos desde múltiples fuentes externas hacia la zona Bronze del bucket de almacenamiento.	Step Function *Source to Landing* orquesta la ejecución de múltiples funciones Lambda, programando el orden y verificando el éxito de cada extracción antes de avanzar.
Extracción de datos	Ejecutar funciones independientes que extraen archivos desde las fuentes definidas.	Cada fuente tiene asignada una *Lambda Function* independiente que descarga el archivo y lo deposita en la zona Bronze del bucket S3.
Almacenamiento inicial	Centralizar los archivos sin procesar en una zona de almacenamiento seguro y segmentado por capas.	Archivos extraídos se almacenan en Amazon S3 bajo la carpeta correspondiente a la zona Bronze, siguiendo la arquitectura *medallion*.
Transformación de datos	Validar automáticamente la estructura de cada archivo, eliminar	Funciones Lambda realizan la limpieza, verificación de estructura, eliminación de duplicados y

Matriz requerimientos vs diseño		
	duplicados y estandarizar los formatos de forma modular.	estandarización de formatos en la zona Silver.
Validación de consistencia	Interrumpir el flujo en caso de errores estructurales detectados durante la transformación.	Validaciones lógicas distribuidas en el subproceso *Landing to Staging* aseguran la consistencia de los datos antes de avanzar entre etapas críticas.
Almacenamiento estructurado	Guardar los archivos validados en una zona diferenciada para preparación de consolidación.	Datos que pasan la validación son almacenados en Amazon S3 bajo la zona Silver, separando claramente los archivos procesados de los sin transformar.
Consolidación de datos	Integrar los archivos transformados bajo un esquema unificado mediante un proceso de consolidación centralizado.	AWS Glue Job toma los archivos de la zona Silver y realiza una consolidación unificada, integrando los datos bajo un esquema común y preparándolos para carga.
Almacenamiento final	Conservar los datos consolidados listos para la carga estructurada.	Datos consolidados son almacenados en la zona Gold del bucket S3, en espera de su transferencia a la base de datos relacional.
Carga en base de datos	Insertar automáticamente los datos consolidados en la base de datos relacional utilizada por la plataforma de gestión de datos maestros.	Función Lambda toma los datos desde la zona Gold y los inserta en Amazon Aurora, donde se almacenan con estructura relacional.
Publicación en plataforma de Gestión de Datos Maestros	Acceder a los datos desde la base estructurada para su posterior validación y publicación dentro de la plataforma MDM.	La plataforma MDM accede a Amazon Aurora y consume directamente los datos estructurados para fines de validación y publicación.
Monitoreo del proceso	Registrar logs de ejecución, errores y métricas de rendimiento en cada etapa del flujo automatizado.	Amazon CloudWatch supervisa de forma continua cada Lambda, Glue Job y Step Function, generando logs detallados, métricas de rendimiento y alertas ante errores en tiempo real.

Nota. Elaboración propia (2025)

4.2.9 Matriz de integración

Con el objetivo de validar que el diseño técnico del proceso automatizado responde efectivamente a los requerimientos funcionales establecidos, se presenta a continuación la matriz de integración. Este instrumento fue elaborado de forma propia como respuesta a la necesidad de estructurar la relación entre las herramientas tecnológicas utilizadas y las etapas específicas del flujo *To-Be*, permitiendo evaluar de forma explícita el nivel de integración alcanzado en cada componente del diseño.

La construcción de esta matriz se fundamenta metodológicamente en el enfoque de validación temprana de requisitos mediante modelos de diseño, propuesto por Stachtiari et al. (2018, pp. 50–67). Este enfoque, conocido como *correctness-by-construction*, plantea que los requerimientos del sistema deben verificarse en etapas tempranas del desarrollo, estableciendo correspondencias claras con los elementos técnicos que los implementan. Aunque el artículo no propone una matriz como instrumento formal, su lógica de verificación estructurada sirve como base conceptual para el modelo desarrollado en esta investigación.

En este contexto, la matriz permite operacionalizar el indicador "Nivel de integración con herramientas tecnológicas", correspondiente a la variable VA-03 del presente proyecto. Cada fila de la tabla documenta una etapa del proceso *To-Be*, la herramienta de AWS utilizada, el tipo de integración aplicada y el rol funcional que dicha herramienta desempeña en el cumplimiento del diseño automatizado. A continuación en la Tabla 27, se presenta la matriz de integración.

Tabla 27. Matriz de integración

Etapa del proceso automatizado	Herramienta tecnológica utilizada	Tipo de integración	Responsabilidad funcional	Observación técnica
Orquestación de extracción	AWS Step Functions	Secuencial / condicional	Coordina la ejecución de funciones Lambda para extraer datos de múltiples fuentes.	Evalúa resultados y controla rutas condicionales lógicas según éxito o error.
Extracción de datos	AWS Lambda	Modular / individual	Extracción de archivos desde cada fuente externa.	Cada fuente tiene asignada una Lambda dedicada.
Almacenamiento inicial	Amazon S3 - zona Bronze	Directa	Recibe archivos en su formato original.	Archivos se segmentan por fuente y fecha; mantiene trazabilidad.
Transformación de datos	AWS Lambda	Automatizada y distribuida	Limpieza, validación estructural y estandarización.	Aplicada individualmente a cada archivo en zona Silver.
Validación de consistencia	AWS Step Functions / Amazon CloudWatch	Condicional / monitoreo automatizado	Interrumpe el flujo si se detectan errores en los datos.	Compuertas de decisión controlan el avance en función de reglas de calidad.
Almacenamiento estructurado	Amazon S3 - zona Silver	Directa	Guarda archivos validados listos para consolidación.	Mantiene separación lógica respecto a datos sin procesar.
Consolidación de datos	AWS Glue Jobs	Integración centralizada	Integra múltiples archivos estructurados bajo un esquema común.	Se activa solo si todos los *datasets* requeridos están disponibles.
Almacenamiento final	Amazon S3 - zona Gold	Directa	Conserva los datos listos para carga.	Resultado directo del proceso de

				consolidación en Glue.
Carga en base de datos	AWS Lambda + Amazon Aurora	Automatizada	Inserta los datos “limpios” en la base estructurada utilizada por la plataforma MDM.	Verifica éxito de carga mediante compuerta condicional.
Publicación en plataforma de MDM	Amazon Aurora + Plataforma MDM	Acceso estructurado	La plataforma accede y publica los datos validados.	Aurora actúa como fuente de datos maestros para la plataforma MDM.
Monitoreo del proceso	Amazon CloudWatch	Transversal / continuo	Supervisa métricas, errores y ejecución en cada etapa.	Genera alertas en tiempo real y logs técnicos detallados.

Nota. Elaboración propia (2025)

5 Propuesta de Solución

En este capítulo se detalla la propuesta de solución, la cual tiene como objetivo abordar la problemática identificada en el proceso de carga de datos hacia la plataforma de Gestión de Datos Maestros de la organización. La solución se enfoca en la automatización progresiva de dicho proceso mediante el desarrollo de un prototipo funcional, con el propósito de reducir la dependencia de tareas manuales, minimizar errores operativos y mejorar la trazabilidad de los datos.

El capítulo corresponde a la tercera fase del proyecto y presenta la forma en que se concreta la solución diseñada en etapas anteriores, integrando las herramientas tecnológicas seleccionadas y aplicando los principios metodológicos definidos. A partir de estos lineamientos, se plantea un flujo automatizado que da respuesta directa a las brechas previamente diagnosticadas, junto con un análisis general de viabilidad técnica y financiera que permite evaluar los beneficios esperados frente a los recursos requeridos para su implementación.

5.1 Fase 3. Desarrollo del prototipo de la solución automatizada

La tercera fase del proyecto corresponde al desarrollo del prototipo de una solución automatizada orientada a transformar el proceso actual de carga de datos en la plataforma de Gestión de Datos Maestros (MDM) de la organización. Esta fase representa un punto de inflexión metodológico, al pasar de las etapas analíticas y de diseño conceptual a una instancia constructiva, en la que se materializa técnicamente la solución propuesta con base en los requerimientos previamente identificados.

El desarrollo del prototipo responde al cumplimiento del objetivo específico #3, el cual plantea: *"Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta seleccionada y los requerimientos identificados."* En este sentido, la fase se orienta a demostrar, mediante componentes funcionales y técnicamente viables, cómo se propone automatizar un proceso que hasta ahora se ha caracterizado por una fuerte dependencia de tareas manuales, con altos riesgos de error, duplicidad e ineficiencia operativa.

Esta fase también permite operacionalizar dos variables clave definidas en la matriz metodológica del proyecto. En primer lugar, la variable independiente *"Prototipo de solución automatizada"*, entendida como el conjunto de componentes tecnológicos diseñados para ejecutar el flujo de carga de datos de forma automática. Esta variable se analiza mediante la implementación concreta de funcionalidades automatizadas, cuyo grado de avance y alineación con el diseño propuesto es representado mediante diagramas técnicos y descripciones funcionales. En segundo lugar, se incorpora la variable dependiente *"Alcance funcional del prototipo construido"*, la cual evalúa en qué medida el prototipo refleja y cumple los requerimientos definidos en fases anteriores, a través de una matriz de trazabilidad que vincula cada funcionalidad implementada con su requerimiento original.

La fase no se limita a presentar una solución idealizada, sino que busca representar de forma verosímil y coherente cómo se vería operativamente el nuevo proceso automatizado, a través de una simulación técnica construida con código *dummy* y datos genéricos. Dicha simulación fue desarrollada utilizando servicios disponibles en Amazon Web Services (AWS), dentro del entorno gratuito, con el fin de preservar la confidencialidad de los recursos organizacionales sin comprometer la estructura lógica y funcional del proceso propuesto. Este enfoque permite validar la factibilidad técnica del modelo diseñado, analizar su alineación con los requerimientos establecidos y sentar las bases para una futura implementación real bajo condiciones controladas. Adicionalmente, se detallan los avances obtenidos, las limitaciones encontradas, las decisiones técnicas adoptadas y las áreas en las que se prevé ampliar el desarrollo en fases posteriores.

De este modo, la Fase 3 constituye una etapa clave dentro de la estructura metodológica del proyecto, al integrar los hallazgos del diagnóstico, el diseño conceptual validado y la estrategia tecnológica definida, dando lugar a una propuesta funcional, replicable y evaluable, que aporta una solución concreta a la problemática planteada desde el inicio del Trabajo Final de Graduación.

5.1.1 Descripción general del prototipo

El prototipo desarrollado en esta fase representa una solución técnica simulada que responde directamente a la necesidad de automatizar el proceso de carga de datos en la plataforma de Gestión de Datos Maestros (MDM). Su diseño fue estructurado bajo una arquitectura lógica por capas (bronze, silver y gold), la cual organiza el flujo de datos desde su origen hasta su almacenamiento estructurado, con el propósito de eliminar tareas manuales, minimizar errores y garantizar la trazabilidad en cada etapa del proceso.

El propósito del prototipo dentro del flujo de carga de datos es servir como modelo funcional del nuevo proceso propuesto. A través de esta simulación, se busca evidenciar cómo una solución automatizada podría ejecutar de forma ordenada y controlada tareas críticas como la extracción de archivos desde múltiples fuentes, la transformación y validación de los datos, su consolidación bajo un estándar común, la carga en una base de datos relacional y el monitoreo integral de la ejecución. Todo ello bajo un esquema técnico que respeta la segmentación lógica y las dependencias funcionales entre etapas.

Para construir este prototipo se utilizaron servicios de Amazon Web Services (AWS), todos operando bajo los límites del *Free Tier*. Entre ellos, se destaca el uso de AWS Lambda para ejecutar funciones de extracción, transformación y carga de datos; Amazon S3 como repositorio principal segmentado por capas; AWS Step Functions para la orquestación de los flujos; Amazon Aurora como base de datos relacional donde se consolidan los datos estructurados; y Amazon CloudWatch como herramienta de monitoreo para registrar logs de ejecución, errores y métricas de control. Cabe señalar que, en el entorno organizacional real, la consolidación de archivos en la zona Gold se ejecutaría mediante AWS Glue, según los requerimientos definidos. Sin embargo, debido a que este servicio no se encuentra incluido dentro del Free Tier, su funcionalidad fue simulada en el prototipo mediante funciones Lambda combinadas con la biblioteca Pandas, sin alterar la lógica técnica ni el propósito funcional de la etapa de consolidación.

El prototipo responde a la problemática diagnosticada en fases anteriores al ofrecer una alternativa técnica concreta frente a las limitaciones observadas en el proceso actual. Mientras que el procedimiento vigente depende de tareas manuales dispersas, formatos no estandarizados y validaciones informales, la solución simulada propone un flujo integral, controlado y basado en automatización, que permite reducir la carga operativa, mejorar la calidad de los datos maestros, así cómo habilitar capacidades de auditoría y trazabilidad.

Es fundamental aclarar que el prototipo fue desarrollado bajo condiciones de prueba controladas, utilizando fragmentos de código representativos, configuraciones simuladas y datos *dummy* que emulan el comportamiento funcional del flujo automatizado. Esta decisión respondió a las restricciones de acceso a los entornos productivos y a la naturaleza confidencial de los datos reales, garantizando la protección de la información organizacional. A pesar de su carácter simulado, el prototipo permitió validar rigurosamente la lógica de automatización diseñada, su secuencia funcional, las interacciones entre componentes y su alineación con los requerimientos definidos, lo cual proporciona evidencia técnica sólida para sustentar su viabilidad.

No obstante, en un entorno organizacional real, se espera la ejecución concurrente de todas las funciones Lambda asociadas a cada fuente de datos, el procesamiento de archivos de gran volumen, la ejecución simultánea de múltiples cargas, así como la integración continua con herramientas de monitoreo, auditoría y control bajo políticas corporativas. Estas condiciones implican un nivel de exigencia superior que no fue replicado íntegramente en el prototipo, por lo cual los resultados deben interpretarse como una validación preliminar que demuestra la factibilidad técnica del prototipo propuesto y sienta las bases para su futura implementación en ambientes reales controlados.

5.1.2 Arquitectura lógica del flujo automatizado

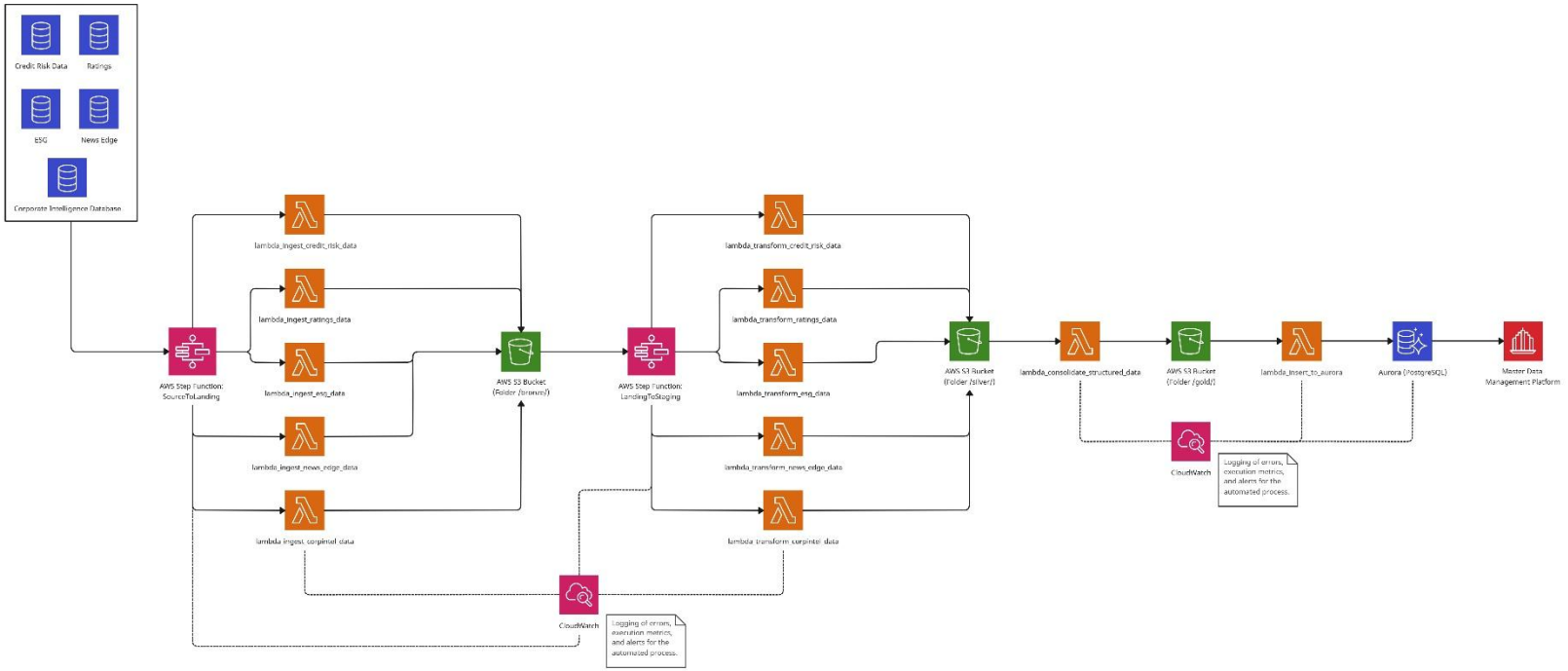
Esta sección presenta la arquitectura técnica que sustenta el prototipo desarrollado para automatizar el proceso de carga de datos en la plataforma de Gestión de Datos Maestros. A través de un diagrama estructurado con servicios nativos de Amazon Web Services (AWS), se representa de forma visual y secuencial el recorrido que siguen los datos desde su extracción inicial hasta su almacenamiento final en una base de datos relacional estructurada.

La arquitectura se organiza bajo el enfoque por capas definido en el diseño conceptual del proceso, segmentando el flujo en tres zonas funcionales: Bronze, Silver y Gold. Cada una de estas zonas cumple un propósito específico en la manipulación de los archivos CSV extraídos desde fuentes externas, y se apoya en funciones Lambda, almacenamiento en Amazon S3, mecanismos de orquestación con AWS Step Functions y monitoreo centralizado con Amazon CloudWatch. La carga final se ejecuta en una instancia simulada de Amazon Aurora (PostgreSQL), desde donde los datos son consumidos por la plataforma de Gestión de Datos Maestros.

Todas las herramientas utilizadas pertenecen al ecosistema de servicios disponibles en el plan gratuito de AWS (*Free Tier*), lo cual permitió construir una simulación funcional sin incurrir en costos adicionales ni comprometer información sensible. No obstante, es importante señalar que en el entorno organizacional real, la consolidación de los datos en la zona Gold se encuentra

diseñada para ejecutarse mediante AWS Glue, según los requerimientos definidos. Dado que este servicio no está incluido en el Free Tier, su funcionalidad fue replicada en el prototipo mediante funciones Lambda combinadas con la biblioteca Pandas, conservando la lógica de consolidación definida y respetando la secuencia técnica establecida en el flujo. A continuación, en la Figura 17, se visualiza la arquitectura lógica del flujo de carga de datos automatizado, integrada con los servicios de Amazon Web Services utilizados en el diseño técnico de la solución. Para visualizar el diagrama en una mejor resolución, acceder al siguiente **enlace**.

Figura 17. Arquitectura lógica del proceso de carga de datos automatizado mediante AWS



Nota. Elaboración propia (2025)

5.1.3 Descripción detallada por etapas del proceso automatizado

Esta sección detalla, de forma estructurada, cada una de las etapas que conforman el flujo automatizado propuesto para la carga de datos en la plataforma de Gestión de Datos Maestros. El objetivo es exponer la lógica técnica implementada en cada fase, las herramientas utilizadas y el comportamiento funcional que define el tránsito de los archivos desde su origen hasta su integración final en la base de datos estructurada.

Cada subsección presenta el propósito específico de la etapa correspondiente, el componente responsable de ejecutarla, los mecanismos de control integrados y la forma en que se gestionan errores o eventos críticos. Además, se vincula cada decisión técnica con los requerimientos funcionales definidos en fases previas del proyecto, reforzando la trazabilidad entre el diseño conceptual y la solución implementada.

5.1.3.1 Extracción de datos

La primera etapa del flujo automatizado corresponde al proceso de extracción de datos desde diversas fuentes externas utilizadas por la organización, entre ellas: Credit Risk Data, Ratings, ESG, News Edge y Corporate Intelligence Database. Estas fuentes proporcionan archivos en formato CSV que contienen información crítica para la gestión de datos maestros, y cuya recolección se realizaba tradicionalmente de forma manual mediante descargas independientes, consolidación local y transporte informal hacia los sistemas internos. Esta situación generaba altos niveles de dependencia operativa, vulnerabilidad ante errores y ausencia de trazabilidad en los registros procesados.

Con el propósito de atender los requerimientos definidos en la Fase 2 del proyecto, la solución propuesta automatiza esta etapa a través de una función de orquestación implementada con AWS Step Functions, denominada ***SourceToLanding*** la cual coordina la ejecución de múltiples funciones AWS Lambda independientes, cada una encargada de extraer archivos desde una fuente específica. Las funciones Lambda se programan para ejecutarse en paralelo, conectarse a las ubicaciones predefinidas de cada fuente y trasladar los archivos obtenidos hacia una carpeta temporal de almacenamiento inicial ubicada en Amazon S3, específicamente en la zona Bronze del bucket correspondiente.

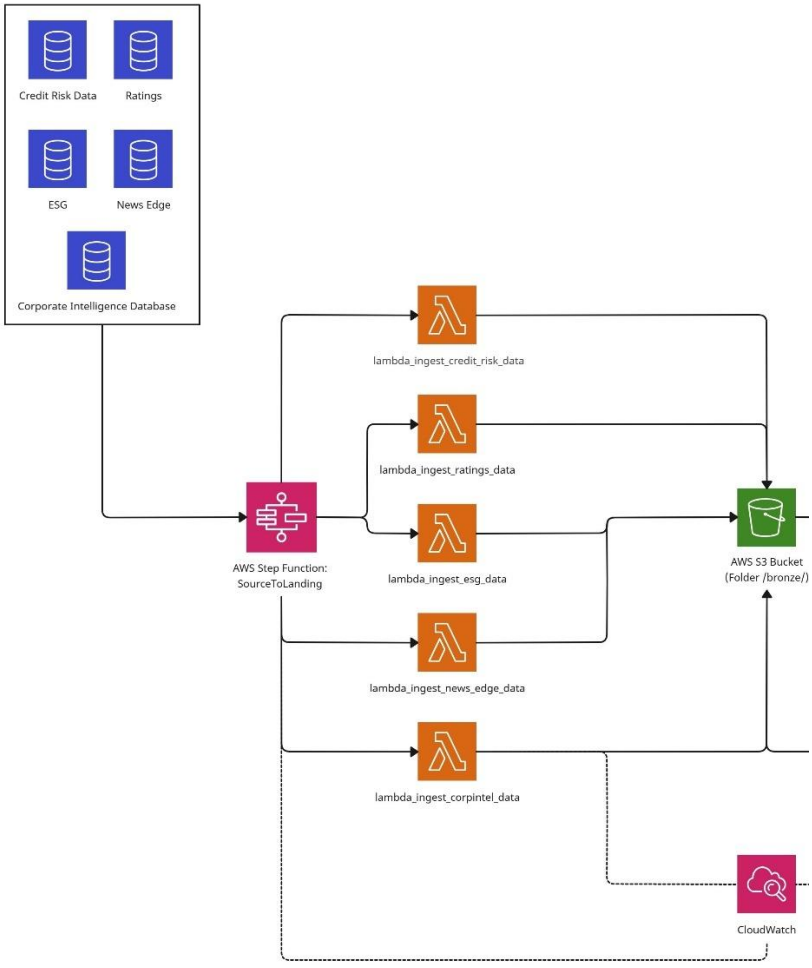
Cada función Lambda de extracción fue definida de manera modular y especializada, siguiendo una nomenclatura estandarizada que permite identificar la fuente que atiende: *lambda_ingest_credit_risk_data, lambda_ingest_ratings_data, lambda_ingest_esg_data,* entre otras. Esta decisión de diseño responde a la necesidad de mantener un control granular sobre el comportamiento individual de cada conexión, facilitando la trazabilidad, depuración y escalabilidad futura del flujo. Además, al implementar la lógica en funciones separadas, se permite una mayor flexibilidad ante cambios en la estructura, periodicidad o disponibilidad de las fuentes de origen.

Desde el punto de vista técnico, cada función Lambda utiliza conectores simulados para representar el acceso a las fuentes externas, evitando exponer credenciales o rutas reales por motivos de confidencialidad. Los archivos extraídos son validados superficialmente a nivel de

formato y luego enviados al bucket S3 en la ruta designada como /bronze/, donde quedan disponibles para las siguientes etapas del proceso. Todo este mecanismo opera bajo los límites del AWS Free Tier, sin requerir servidores ni procesos residentes, lo cual demuestra la viabilidad del modelo incluso en entornos con recursos limitados.

En términos de monitoreo y control, cada ejecución de función Lambda se encuentra conectada a Amazon CloudWatch, donde se registran los logs de entrada, errores de ejecución, métricas de duración y cantidad de archivos transferidos. En caso de fallas durante la extracción (por ejemplo, archivo corrupto, error de conexión o tiempo de espera excedido), la Step Function detiene la ejecución del flujo, notifica el error mediante CloudWatch y deja el evento disponible para su análisis posterior. Esto permite establecer un sistema de trazabilidad completa desde el origen de los datos, cumpliendo con uno de los requerimientos más importantes detectados durante el diagnóstico inicial del proceso. En la Figura 18, se observa la etapa de extracción de datos en la arquitectura lógica diagramada.

Figura 18. Etapa de extracción de datos del flujo automatizado



Nota. Elaboración propia (2025)

5.1.3.2 Transformación y validación estructural

Una vez finalizado el proceso de extracción, los archivos almacenados en la zona Bronze ingresan a la etapa de transformación, cuya finalidad es garantizar la calidad estructural mínima requerida antes de avanzar en el flujo automatizado. Esta fase contempla tareas de limpieza, validación y estandarización de los archivos CSV provenientes de las distintas fuentes externas.

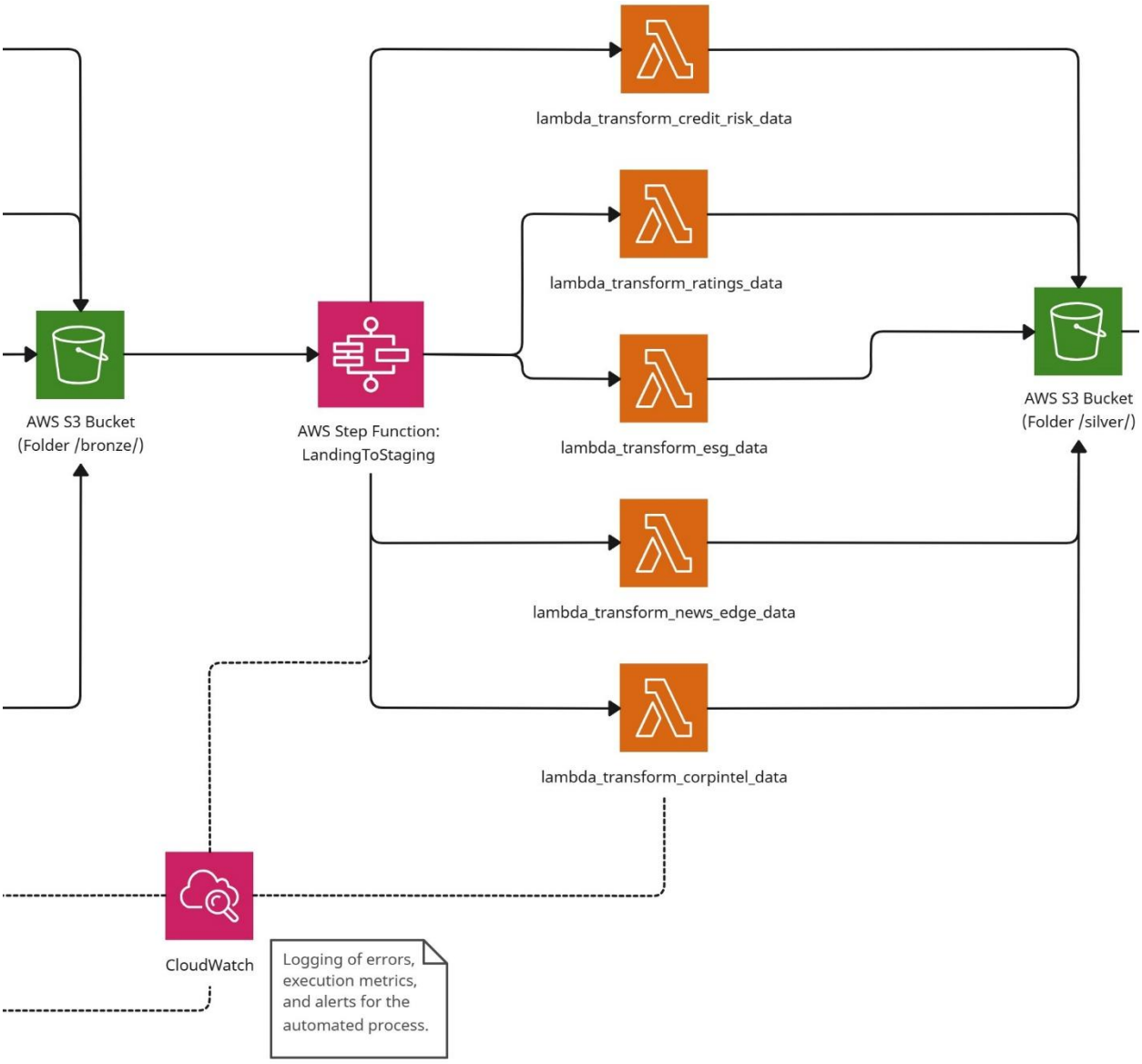
Cada archivo ubicado en la carpeta /bronze/ del bucket Amazon S3 es procesado mediante funciones AWS Lambda, programadas para detectar inconsistencias estructurales, eliminar registros duplicados y unificar formatos de acuerdo con las reglas definidas durante la fase de diseño. El procesamiento se realiza de forma aislada por archivo, manteniendo la trazabilidad de cada origen y evitando la mezcla prematura de datos. La salida generada por cada función Lambda se almacena en la carpeta /silver/, representando así la zona en la que se resguardan únicamente los archivos que han superado exitosamente la etapa de validación.

Las funciones Lambda utilizadas en esta fase fueron desarrolladas con el uso de la biblioteca *Pandas*, integrada en entornos de ejecución adaptados para el Free Tier de AWS. Cada función cuenta con lógica modular que permite aplicar reglas específicas por fuente, manteniendo la escalabilidad y la facilidad de mantenimiento del flujo. Los nombres asignados a estas funciones, como *lambda_transform_credit_risk_data* o *lambda_transform_esg_data*, reflejan su especialización por fuente de datos.

Durante la ejecución, se verifica la estructura del archivo (encabezados, cantidad de columnas, tipos de dato, entre otros) y se aplican controles para identificar registros vacíos o inconsistentes. Si un archivo incumple las condiciones establecidas, el flujo no continúa hacia la siguiente etapa. En su lugar, se genera un log de error en Amazon CloudWatch, que incluye información sobre el tipo de falla, el archivo afectado y la hora de ocurrencia. Esta capacidad de detección temprana evita que datos erróneos contaminen etapas posteriores del flujo.

La segmentación clara entre la zona Bronze (archivos crudos) y la zona Silver (archivos estructurados) permite mantener un control riguroso sobre el estado de cada conjunto de datos. Esta decisión responde directamente a los requerimientos funcionales establecidos en fases anteriores, en los que se destacó la necesidad de contar con un mecanismo formal de validación estructural antes de consolidar los datos. La implementación realizada garantiza dicho control, sin depender de intervención manual y de herramientas fuera del entorno simulado. En la Figura 19, se visualiza la etapa de transformación y validación estructural en el diagrama de arquitectura lógica.

Figura 19. Etapa de transformación y validación estructural



Nota. Elaboración propia (2025)

5.1.3.3 Consolidación de datos

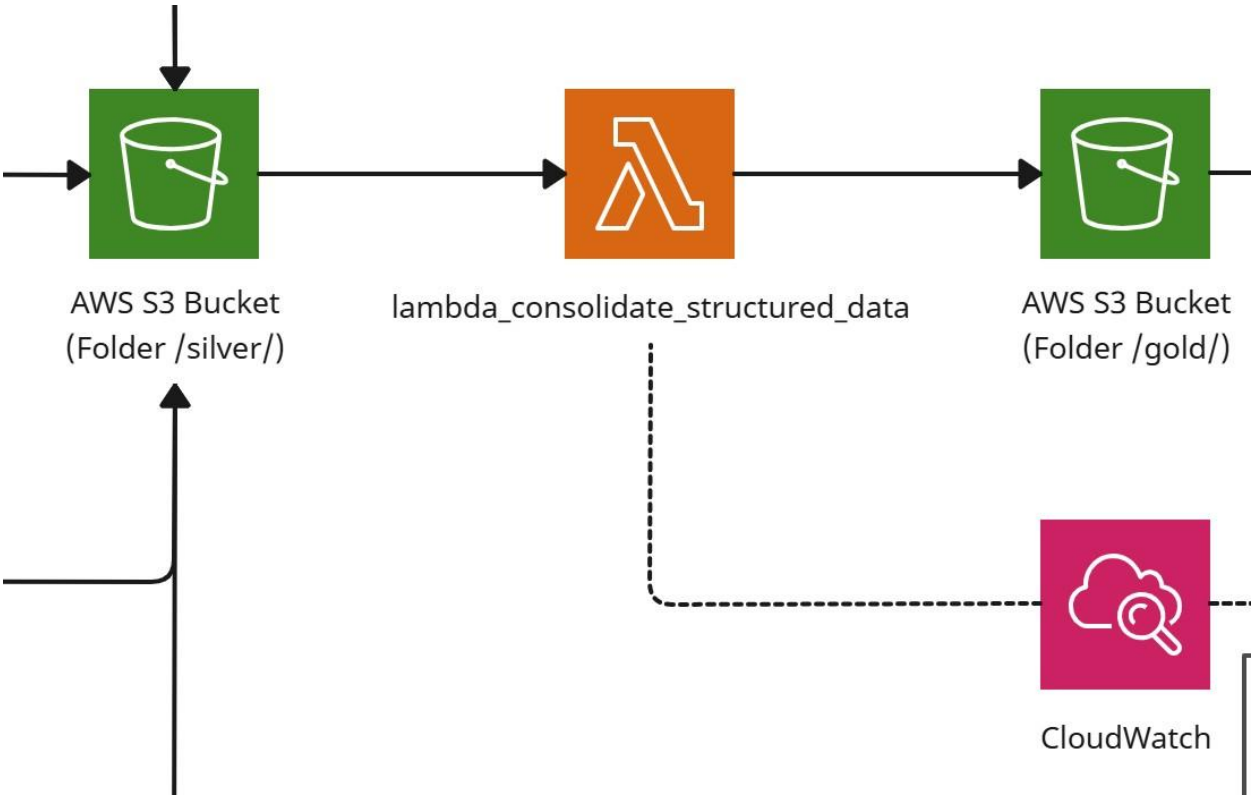
La consolidación representa una de las etapas más críticas del flujo automatizado, ya que integra los archivos previamente validados en una estructura común que será utilizada como base para la carga final. Esta fase permite unificar el contenido proveniente de distintas fuentes bajo un esquema estandarizado, asegurando la consistencia semántica y estructural de los datos que serán insertados en la base relacional. En el entorno organizacional real, esta tarea se ejecutará mediante un proceso gestionado por AWS Glue, herramienta especializada para la preparación y transformación de grandes volúmenes de datos.

Dado que AWS Glue no está disponible dentro del plan Free Tier de AWS, la funcionalidad de consolidación fue simulada en el prototipo utilizando una función AWS Lambda, desarrollada con la biblioteca Pandas. Esta función, denominada *lambda_consolidate_structured_data*, se encarga de leer los archivos ubicados en la zona Silver del bucket Amazon S3, concatenarlos bajo un esquema único y escribir el resultado consolidado en la zona Gold. El proceso se ejecuta de forma automática, sin intervención humana, garantizando que únicamente se integren archivos que hayan superado las etapas previas de validación.

La lógica aplicada en esta etapa evalúa la coherencia de las columnas entre archivos, aplica filtros para remover duplicados globales y reorganiza los datos conforme al modelo requerido por la base relacional. Esta consolidación es desencadenada por la función de orquestación ***LandingToStaging***, una AWS Step Function que verifica previamente la disponibilidad de todos los archivos esperados en la zona Silver. Si alguno de los archivos requeridos no ha sido generado correctamente o no cumple con las condiciones mínimas, la ejecución se detiene y se genera una alerta en Amazon CloudWatch, indicando el origen del problema.

Al finalizar esta etapa, los datos consolidados se almacenan en la carpeta /gold/, listos para ser transferidos hacia la base de datos estructurada. Esta separación lógica entre zonas Silver y Gold asegura una transición controlada, donde únicamente los datos listos para su persistencia definitiva acceden al entorno transaccional. A pesar de que la consolidación fue simulada con Lambda, el flujo técnico y la lógica operativa mantienen total equivalencia con la solución definida para el entorno productivo, en el cual AWS Glue asumiría esta función con características adicionales de rendimiento y escalabilidad. En la Figura 20, se visualiza la etapa específica de consolidación de datos en el diagrama de arquitectura lógica.

Figura 20. Etapa de consolidación de datos



Nota. Elaboración propia (2025)

5.1.3.4 Carga en la base de datos relacional

Una vez completada la consolidación de los archivos validados, el siguiente paso en el flujo automatizado consiste en transferir los datos hacia una base de datos relacional estructurada, donde quedarán disponibles para su posterior consulta y validación desde la plataforma de Gestión de Datos Maestros. Esta etapa representa el cierre técnico del flujo de transformación y preparación, marcando el punto de ingreso de los datos al entorno transaccional.

Para simular esta funcionalidad en el prototipo, se utilizó Amazon Aurora con compatibilidad PostgreSQL, servicio de base de datos relacional incluido en el plan Free Tier de AWS. La carga de datos fue gestionada mediante una función AWS Lambda, denominada *lambda_insert_to_aurora*, encargada de leer los archivos ubicados en la zona Gold del bucket Amazon S3, convertirlos a estructuras tabulares y ejecutar las sentencias SQL necesarias para insertarlos en una tabla relacional previamente definida. Esta tabla simula el modelo estructural que utiliza la plataforma de Gestión de Datos Maestros para organizar y validar los registros.

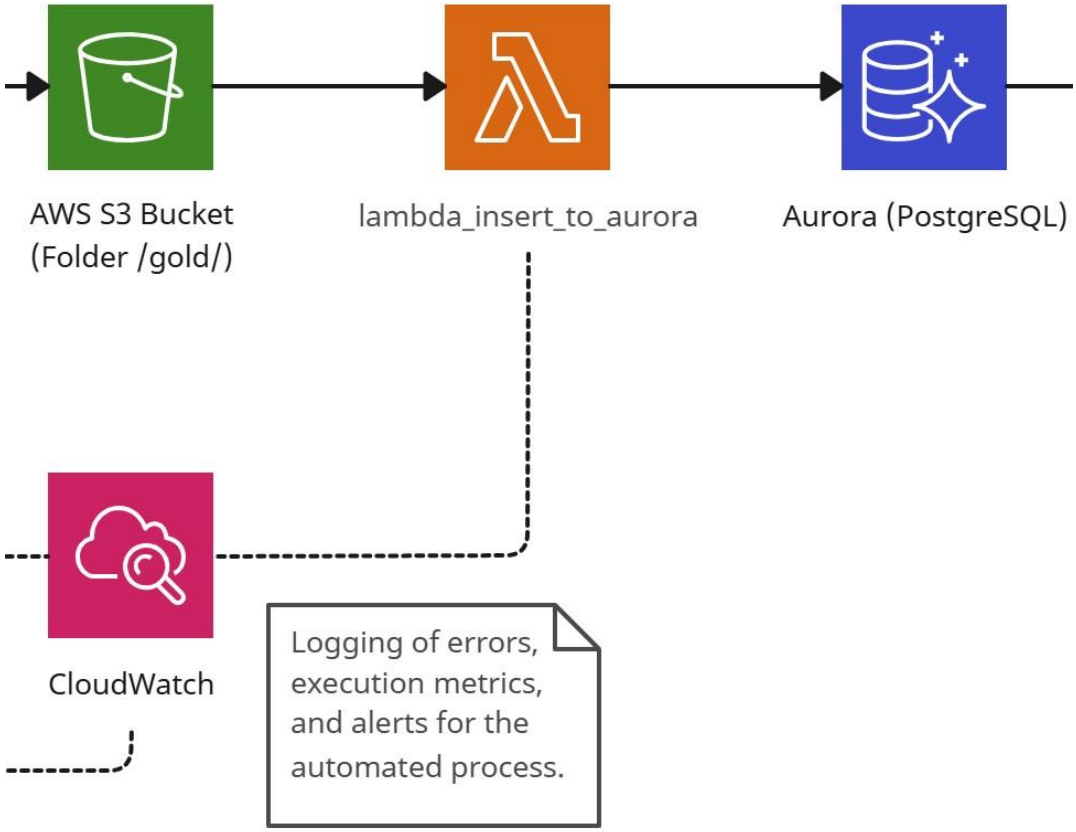
La función Lambda de carga incluye controles básicos para verificar que el archivo de entrada cumpla con la estructura esperada y que los datos no se encuentren duplicados en la tabla de destino. Además, en caso de fallo durante el proceso de inserción (por ejemplo, por error en la conexión, tipo de dato incompatible o inconsistencia en el archivo fuente), se genera una entrada

detallada en Amazon CloudWatch, lo que permite trazar el error sin interrumpir de forma silenciosa la ejecución.

Esta etapa asegura la persistencia de los datos en un entorno relacional, lo cual habilita posteriormente su visualización y validación desde la plataforma MDM, aunque en este prototipo dicha plataforma no se conecta directamente a Aurora por razones de confidencialidad. A nivel lógico, se garantiza que únicamente los datos previamente estructurados y consolidados sean insertados en la base de datos, cumpliendo así con el principio de integridad que rige la carga técnica hacia el repositorio oficial de datos maestros.

El diseño de esta función fue concebido para reflejar la lógica real del proceso de carga, adaptado a un entorno simulado, pero compatible con la arquitectura esperada en un contexto corporativo. La separación clara entre el procesamiento previo (zonas Bronze, Silver y Gold) y la inserción en base de datos permite preservar la trazabilidad del flujo completo, garantizando que la etapa de carga no se vea afectada por errores heredados de fases anteriores. La Figura 21 ilustra de manera específica la etapa correspondiente a la carga de datos en la base de datos relacional, tal como se representa dentro del diagrama de arquitectura lógica del flujo automatizado.

Figura 21. Etapa de carga en la base de datos relacional



Nota. Elaboración propia (2025)

5.1.3.5 Publicación simulada en la plataforma de Gestión de Datos Maestros

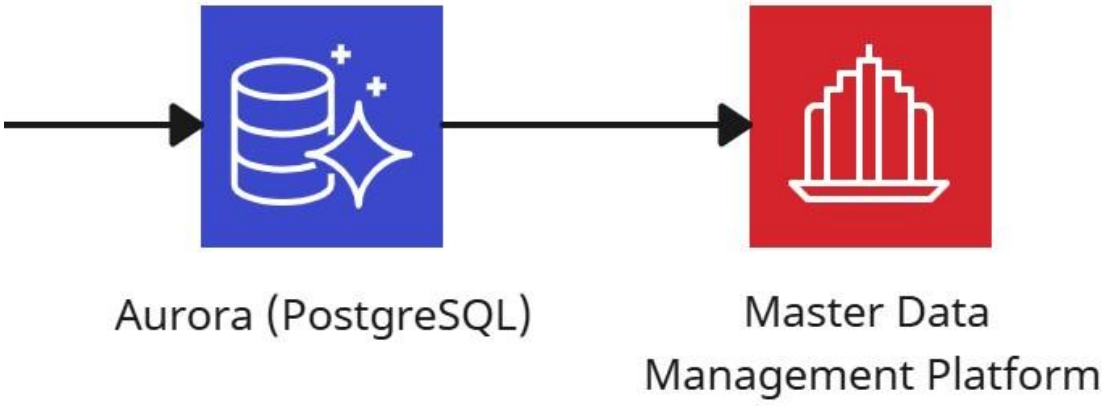
La etapa de publicación constituye el punto de acceso desde el cual los datos estructurados, ya almacenados en la base de datos relacional, se encuentran disponibles para su validación y eventual integración dentro de la plataforma de Gestión de Datos Maestros (MDM). En el entorno operativo de la organización, esta interacción se realiza mediante conexiones directas entre la base de datos y la plataforma MDM, las cuales permiten la visualización, verificación y edición de los registros antes de su publicación definitiva.

Debido a restricciones de confidencialidad y al acceso limitado a los entornos productivos de la organización, esta etapa fue representada en el prototipo como una simulación técnica. Se asumió que los datos cargados en Amazon Aurora son consultados por la plataforma MDM, sin establecer una conexión real entre ambas. Esta simulación se sustenta en la premisa de que la estructura de la tabla relacional en Aurora replica fielmente el esquema requerido por la plataforma MDM para la ingesta y validación de datos maestros.

La validación interna que realiza el equipo responsable de la plataforma consiste en revisar los datos cargados, verificar la completitud de los campos requeridos, confirmar la ausencia de registros duplicados y aplicar reglas de negocio específicas. Si bien estas acciones no fueron implementadas dentro del prototipo por razones técnicas y organizacionales, su representación lógica fue considerada en el diseño del flujo, garantizando que el punto de integración hacia MDM se encuentra estructuralmente definido y técnicamente viable.

En términos de trazabilidad, cualquier intento de lectura por parte de la plataforma se considera exitoso si los datos están disponibles en la base de datos Aurora sin inconsistencias estructurales. Esta validación conceptual refuerza la hipótesis de que el flujo completo es funcional y extensible hacia entornos productivos, siempre que se respeten los mismos principios de estructuración y consolidación de datos aplicados en las etapas previas. La Figura 22 presenta esta etapa dentro del diagrama de arquitectura lógica, mostrando la transición entre la base de datos estructurada y la plataforma MDM como un flujo de consulta controlado.

Figura 22. Etapa de publicación de simulada en la plataforma de Gestión de Datos Maestros



Nota. Elaboración propia (2025)

5.1.3.6 Monitoreo automatizado con Amazon CloudWatch

El monitoreo del flujo automatizado representa un componente fundamental para garantizar la trazabilidad, el control operativo y la detección oportuna de errores durante la ejecución del proceso de carga de datos. En el entorno simulado del prototipo, esta función fue implementada mediante Amazon CloudWatch, servicio de observabilidad nativo de AWS que permite recopilar métricas, registrar logs y generar alertas en tiempo real, sin necesidad de configuración adicional fuera del ecosistema utilizado.

Cada una de las funciones AWS Lambda integradas al flujo automatizado fue configurada para registrar sus eventos relevantes en CloudWatch. Esto incluye información sobre el inicio y fin de ejecución, duración de los procesos, errores durante la extracción o transformación, validaciones fallidas, e incluso resultados de inserción en la base de datos. De igual manera, las AWS Step Functions orquestadoras (***SourceToLanding*** y ***LandingToStaging***) reportan el estado de cada transición, el éxito o fallo de cada paso y la lógica condicional aplicada en cada rama del flujo.

Esta configuración permite identificar con precisión en qué etapa se produce una falla, qué archivo la provocó y cuál fue la causa técnica del error. Este nivel de detalle resulta indispensable para garantizar la confiabilidad del flujo, facilitar el diagnóstico de incidentes y permitir mejoras continuas en futuras versiones de la solución. A nivel organizacional, este tipo de monitoreo proporciona una base para implementar alertas automáticas, *dashboards* operativos o auditorías periódicas sobre el desempeño del proceso de carga.

En el contexto del prototipo, el uso de CloudWatch refuerza la validez técnica del diseño, al demostrar que el flujo no solo automatiza tareas operativas, sino que también incorpora mecanismos robustos de supervisión. Esta capacidad es esencial en procesos críticos como la gestión de datos maestros, donde la integridad junto con la trazabilidad de los registros resultan indispensables para la toma de decisiones y el cumplimiento de normativas internas.

5.1.3.7 Orquestación del flujo con AWS Step Functions

La orquestación del flujo automatizado constituye un componente esencial dentro del diseño técnico de la solución propuesta. Su propósito es coordinar de forma lógica, estructurada y secuencial la ejecución de las distintas funciones Lambda que intervienen en el proceso de carga de datos. Esta coordinación asegura que cada etapa del flujo (extracción, transformación, consolidación y carga) se realice únicamente cuando se hayan cumplido las condiciones necesarias, minimizando riesgos operativos y errores derivados de ejecuciones fuera de secuencia.

En el prototipo desarrollado, esta orquestación fue implementada mediante AWS Step Functions, servicio nativo de Amazon Web Services que permite construir flujos de trabajo definidos por estados. Se diseñaron dos funciones principales de orquestación: ***SourceToLanding***, encargada de gestionar la extracción y almacenamiento inicial de los archivos en la zona Bronze; y ***LandingToStaging***, responsable de coordinar las etapas de transformación, consolidación y carga hasta la base de datos relacional.

Cada Step Function define una secuencia de tareas ejecutadas en orden condicional. Por ejemplo, ***SourceToLanding*** activa funciones Lambda independientes por cada fuente de datos, permitiendo su ejecución en paralelo, pero garantizando que el flujo avance únicamente cuando todas hayan finalizado correctamente. Por su parte, ***LandingToStaging*** incorpora verificaciones sobre la existencia de archivos en la zona Silver antes de activar el proceso de consolidación, y no permite la carga en la base relacional si la consolidación no se ejecuta con éxito.

Las transiciones entre tareas, los manejos de errores y los puntos de espera fueron definidos utilizando el lenguaje de definición de estados (Amazon States Language), lo que aporta claridad, control y facilidad de mantenimiento al flujo diseñado. Además, cada Step Function está integrada con Amazon CloudWatch, permitiendo la trazabilidad completa de cada ejecución, el monitoreo de duración y el registro de errores por etapa.

La implementación de Step Functions refuerza la lógica del modelo automatizado al garantizar que las dependencias funcionales entre procesos estén explícitamente controladas y auditadas. Esta arquitectura por estados no solo mejora la resiliencia de la solución, sino que también la hace extensible ante futuras integraciones o nuevas fuentes de datos.

5.1.4 Evidencia funcional del prototipo en el entorno de Amazon Web Services

El desarrollo del prototipo técnico se materializó mediante la implementación estructurada de un flujo automatizado dentro del entorno de Amazon Web Services (AWS). Esta implementación refleja con precisión las etapas críticas de un proceso moderno de integración de datos, orientado a la gobernanza y trazabilidad requeridas por plataformas de Gestión de Datos Maestros. Las funciones técnicas, agrupadas en secuencias lógicas a través de Step Functions, ejecutan tareas que abarcan desde la recolección inicial de archivos hasta su inserción controlada en una base de datos relacional.

Cada componente del flujo fue diseñado, desplegado y probado dentro de los límites del nivel gratuito (Free Tier) ofrecido por AWS. El objetivo fue validar la viabilidad técnica de un mecanismo automatizado de carga de datos que no dependa de intervención manual para garantizar calidad, consistencia y completitud. La evidencia funcional no solo respalda el cumplimiento de los requerimientos planteados en el diseño conceptual, también demuestra que la arquitectura implementada resulta replicable y extensible a entornos productivos bajo condiciones similares.

Con el fin de concentrar los esfuerzos en una demostración funcional controlada, se eligió trabajar únicamente con la fuente de datos ESG. Esta decisión permitió simplificar las pruebas iniciales sin sacrificar el nivel de complejidad técnica requerido. Se utilizó un archivo JSON denominado *esg_data_sample*,. El archivo contiene un conjunto de registros ficticios diseñados para simular condiciones reales, tanto en variedad de indicadores como en estructura relacional. Cada registro representa un reporte mensual asociado a una empresa específica, incluyendo métricas ambientales, sociales y de gobernanza.

Los datos contenidos en *esg_data_sample* cubren seis períodos consecutivos para cinco compañías distintas, generando un total de 90 registros. Estos datos fueron construidos con valores de prueba suficientemente heterogéneos para verificar el correcto funcionamiento de las funciones Lambda de transformación, consolidación y carga. La existencia de valores atípicos, registros con diferentes fechas y diversas métricas numéricas permitió evaluar la robustez del flujo implementado ante escenarios comunes en la gestión de datos empresariales.

Esta sección documenta, en orden secuencial, el comportamiento funcional del prototipo desde la extracción inicial hasta la simulación de publicación, evidenciando su comportamiento real en un entorno completamente basado en servicios de AWS.

5.1.4.1 Etapa: Extracción de datos

La etapa inicial del flujo automatizado consistió en la extracción de los registros provenientes de la fuente ESG. Para efectos de demostración del prototipo funcional, se implementó una función AWS Lambda denominada *lambda_ingest_esg_data*, encargada de simular la recepción de datos externos y almacenarlos en formato estructurado dentro del bucket Amazon S3 correspondiente a la zona *Bronze*. Esta función representa el punto de ingreso de datos al entorno de procesamiento, siguiendo los lineamientos definidos para la segmentación por zonas del *data lake*.

La implementación de esta función se realizó en el lenguaje Python, y su propósito principal fue almacenar en S3 los eventos recibidos como objetos JSON, generando un nombre de archivo único basado en la fecha y hora de ejecución. La Figura 23 ilustra el fragmento de código de la lógica aplicada:

Figura 23. Lógica de almacenamiento en el Amazon S3 Bucket

```
# Generar nombre de archivo con timestamp
timestamp = datetime.datetime.now().strftime("%Y%m%d_%H%M%S")
filename = f"{BRONZE_FOLDER}registro_esg_{timestamp}.json"

# Subir el archivo al bucket
s3.put_object(
    Bucket=BUCKET_NAME,
    Key=filename,
    Body=json.dumps(event),
    ContentType='application/json'
)
```

Nota. Elaboración propia (2025)

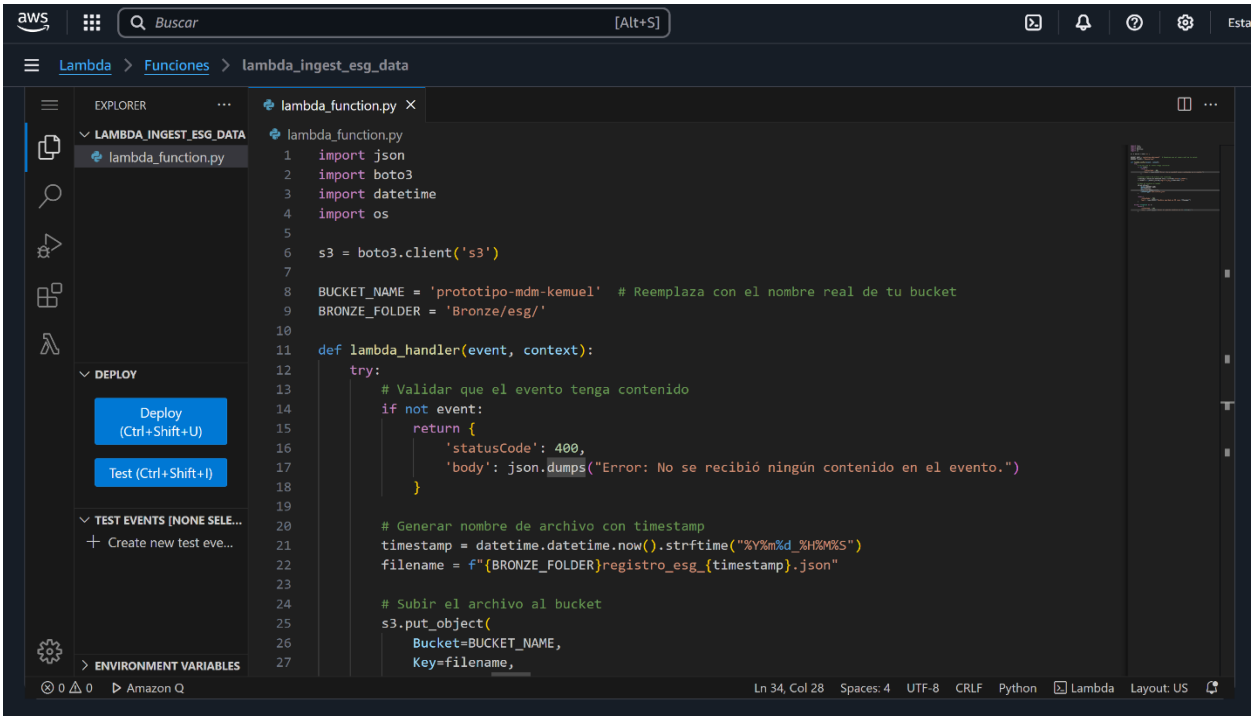
Durante la ejecución, la función valida que exista un evento entrante, genera dinámicamente un nombre de archivo con marca temporal y realiza la carga del contenido en la carpeta Bronze/esg/ del bucket prototipo-mdm-kemuel. El archivo resultante es almacenado bajo el formato registro_esg_YYYYMMDD_HHMMSS.json.

Esta configuración permite simular de manera realista un escenario de integración con una fuente externa, manteniendo los principios de trazabilidad y estructuración de datos que caracterizan a un proceso ETL formal. Además, el uso de un archivo JSON como formato de

entrada responde a la necesidad de contar con una estructura flexible, legible y fácilmente manipulable en etapas posteriores del flujo.

La función fue desplegada correctamente en el entorno de AWS Lambda, con permisos específicos para interactuar con Amazon S3 y realizar operaciones de escritura. La configuración completa de esta etapa se muestra en la Figura 24, en la cual se aprecia la interfaz de AWS Lambda y el código fuente responsable de la operación de extracción. En la Figura 25 se muestra el S3 Bucket creado.

Figura 24. Interfaz de AWS Lambda + Código fuente de extracción



Nota. Elaboración propia (2025)

Figura 25. Amazon S3 Bucket creado

Nota. Elaboración propia (2025)

5.1.4.2 Etapa: Transformación de datos y validación de consistencia

Una vez completada la etapa de extracción, el flujo automatizado continúa con la transformación de los registros almacenados en la zona *Bronze*. Esta operación es ejecutada mediante la función Lambda *lambda_transform_esg_data*, diseñada para validar la estructura de los datos ingresados y generar un nuevo conjunto de registros limpios y listos para consolidación. El resultado de esta operación se almacena en la carpeta Silver/esg/, que corresponde a la zona de datos transformados.

La función fue desarrollada en Python, y su lógica se centra en tres acciones principales: la lectura de archivos en formato JSON desde el bucket de origen, la verificación estructural de cada registro, y la posterior escritura de los registros válidos en una nueva ubicación. La validación se ejecuta campo por campo, conforme a una lista de atributos requeridos que incluye indicadores ambientales, sociales y de gobernanza tales como emisiones de carbono, índice de diversidad, inversión comunitaria y cumplimiento de auditoría. A continuación en la Figura 26, se ilustra un fragmento del código que realiza esta validación.

Figura 26. Lógica de validación de atributos

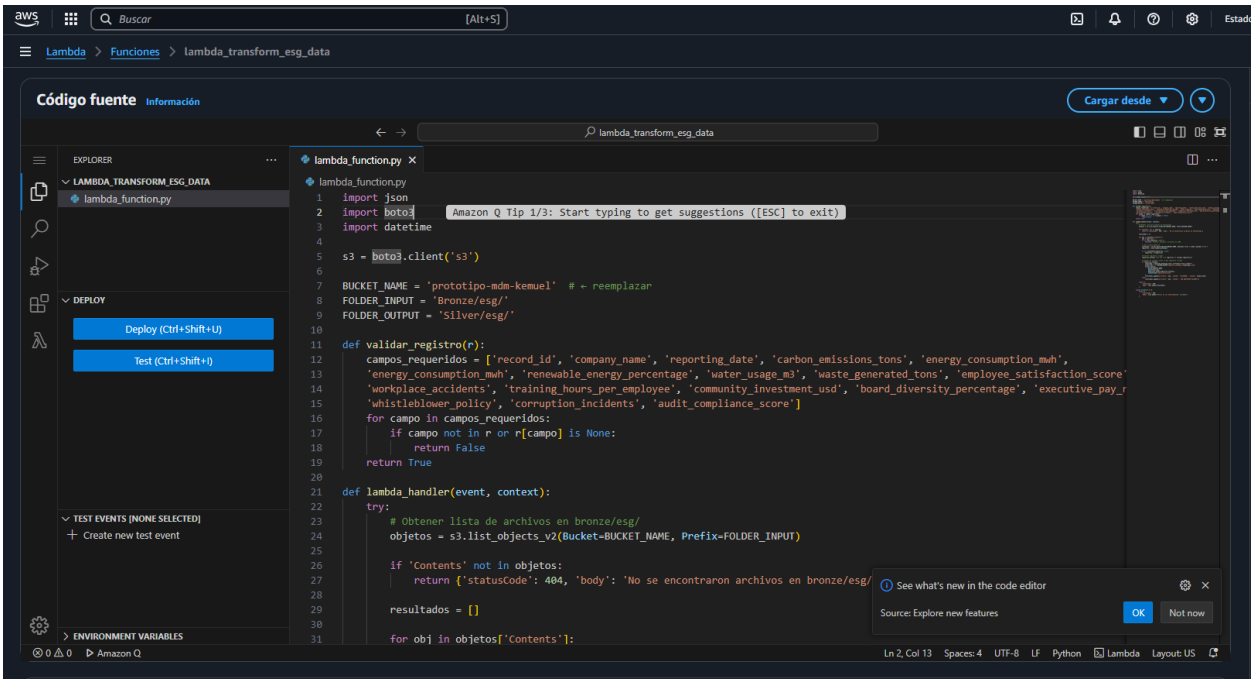
```
def validar_registro(r):
    campos_requeridos = ['record_id', 'company_name', 'reporting_date', 'carbon_emissions_tons', 'energy_consumption_mwh',
    'energy_consumption_mwh', 'renewable_energy_percentage', 'water_usage_m3', 'waste_generated_tons', 'employee_satisfaction_score'
    'workplace_accidents', 'training_hours_per_employee', 'community_investment_usd', 'board_diversity_percentage', 'executive_pay_r
    'whistleblower_policy', 'corruption_incidents', 'audit_compliance_score']
    for campo in campos_requeridos:
        if campo not in r or r[campo] is None:
            return False
    return True
```

Nota. Elaboración propia (2025)

Una vez validado el contenido, la función consolida únicamente los registros completos en un nuevo archivo JSON, cuyo nombre se genera dinámicamente con marca temporal (registro_validado_YYYYMMDD_HHMMSS.json). Esta segmentación garantiza que los datos en la zona *Silver* cumplan con los estándares requeridos para su uso posterior en procesos de consolidación y análisis.

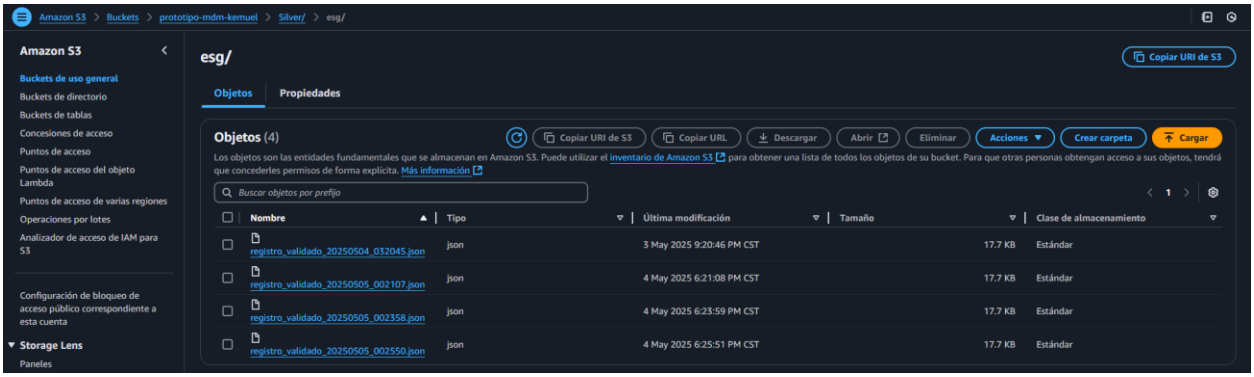
Desde una perspectiva técnica, la transformación permite filtrar ruidos o registros incompletos sin interrumpir el flujo general, promoviendo una separación clara entre datos brutos y datos listos para procesamiento. La ejecución de esta función fue orquestada a través de la Step Function ***LandingToStaging***, que asegura su correcta secuenciación dentro del flujo completo. La Figura 27 muestra el entorno de implementación de esta función dentro de AWS Lambda, así como el código fuente que permite realizar la validación estructural de manera automatizada y escalable. En la Figura 28 se observa el archivo almacenado en la carpeta *Silver* del S3 Bucket.

Figura 27. Interfaz de AWS Lambda + Código fuente de transformación y validación



Nota. Elaboración propia (2025)

Figura 28. Archivo almacenado en la carpeta Silver



Nota. Elaboración propia (2025)

5.1.4.3 Etapa: Consolidación de datos

Una vez validados estructuralmente, los registros almacenados en la zona *Silver* avanzan hacia la etapa de consolidación en la carpeta *Gold*. Esta zona representa el repositorio final del flujo ETL, en el cual convergen los datos depurados y listos para su análisis o integración posterior. Su función principal consiste en centralizar todos los archivos procesados desde múltiples fuentes, facilitando la trazabilidad, la integridad del *dataset* consolidado y la preparación para su posterior consumo.

En un entorno real, la zona Gold representa el punto de consolidación final en el que convergen todas las fuentes de datos procesadas, transformadas y validadas durante las etapas anteriores del flujo. Esta consolidación tiene como propósito unificar los registros provenientes de diversas fuentes heterogéneas, en este caso: *ESG, Credit Risk Data, Ratings, News Edge* y *Corporate Intelligence Database* en un único archivo estructurado, listo para su posterior análisis, carga o publicación hacia la plataforma de Gestión de Datos Maestros (MDM).

En el desarrollo del prototipo, esta consolidación fue simulada mediante una función Lambda enfocada únicamente en la fuente ESG, cuyo flujo consiste en copiar directamente un archivo validado desde la carpeta *Silver/esg* hacia *Gold/esg*. Esta operación resulta funcional y adecuada para efectos demostrativos.

La consolidación integral requeriría recorrer secuencialmente o en paralelo cada una de las carpetas correspondientes a las cinco fuentes mencionadas. Para cada fuente, se debería identificar los archivos más recientes o válidos, extraer los registros y unificarlos en una estructura homogénea. Además, sería necesario incluir metadatos por cada registro que indiquen su procedencia, facilitando así trazabilidad y control de calidad. La lógica debe contemplar también la armonización de esquemas entre las distintas fuentes, el manejo de duplicados y la validación de integridad de los datos.

Esta aproximación garantizaría una consolidación robusta y alineada con las exigencias de un entorno productivo, donde la interoperabilidad de datos, la calidad y la trazabilidad resultan elementos críticos. En este sentido, el diseño actual del prototipo, aunque limitado a una sola fuente, establece las bases necesarias para escalar hacia una implementación real de mayor alcance y complejidad.

La consolidación se llevó a cabo mediante la función Lambda *lambda_consolidate_esg_data*, que implementa una lógica de copia directa del archivo validado desde la carpeta Silver/esg/ hacia Gold/esg/, generando un nuevo nombre con marca temporal. Este comportamiento se ilustra en la Figura 29, con el siguiente fragmento de código.

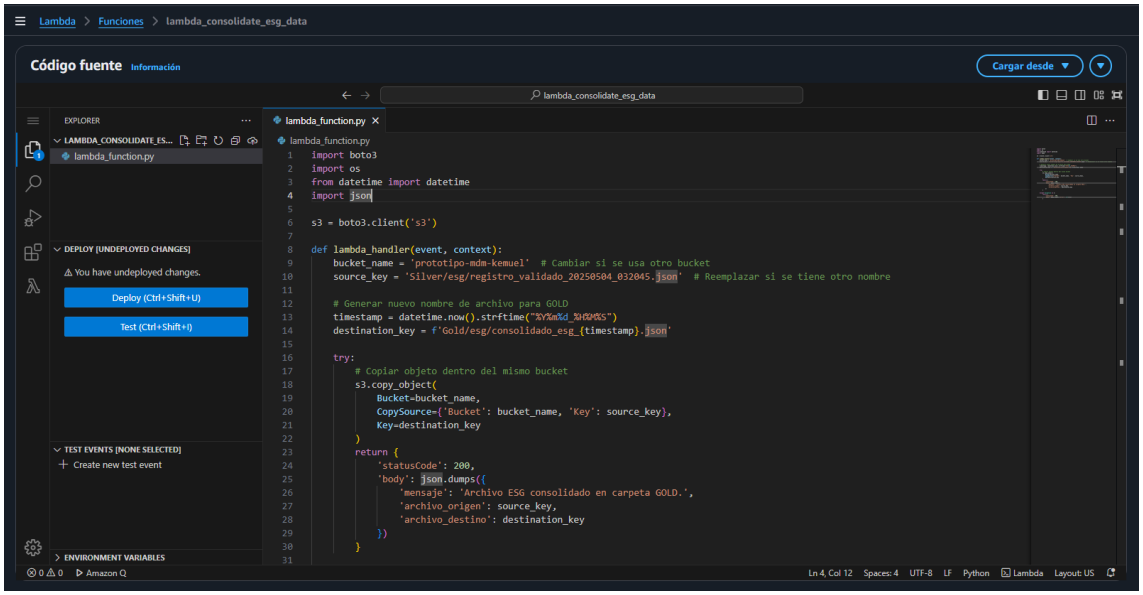
Figura 29. Lógica de consolidación de datos

```python
def lambda_handler(event, context):
    bucket_name = 'prototipo-mdm-kemuel'  # Cambiar si se usa otro bucket
    source_key = 'Silver/esg/registro_validado_20250504_032045.json'  # Reemplazar si se tiene otro nombre

    # Generar nuevo nombre de archivo para GOLD
    timestamp = datetime.now().strftime("%Y%m%d_%H%M%S")
    destination_key = f'Gold/esg/consolidado_esg_{timestamp}.json'

    try:
        # Copiar objeto dentro del mismo bucket
        s3.copy_object(
            Bucket=bucket_name,
            CopySource={'Bucket': bucket_name, 'Key': source_key},
            Key=destination_key
        )
        return {
            'statusCode': 200,
            'body': json.dumps({
                'mensaje': 'Archivo ESG consolidado en carpeta GOLD.',
                'archivo_origen': source_key,
                'archivo_destino': destination_key
            })
        }
```

Nota. Elaboración propia (2025)

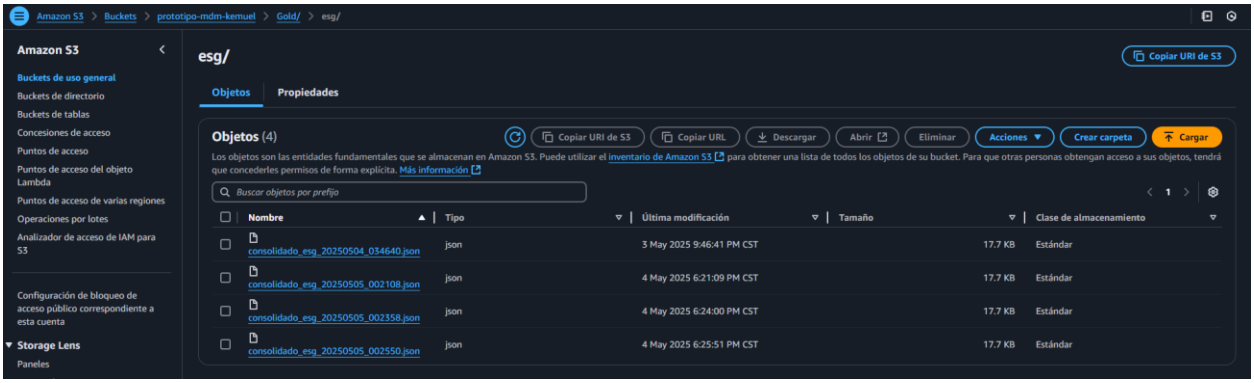
Este paso, aunque no transforma el contenido del archivo, es esencial para garantizar un repositorio único, organizado y alineado con las buenas prácticas de diseño de flujos ETL. Su ejecución fue coordinada desde la máquina de estados ***LandingToStaging***, lo que asegura que esta consolidación ocurra de forma controlada, solo tras la validación exitosa de los datos. La Figura 30 presenta el entorno de configuración de esta función Lambda dentro de AWS, evidenciando la integración directa con Amazon S3 como mecanismo de almacenamiento en la arquitectura del prototipo. La Figura 31 visualiza el archivo almacenado en la carpeta Gold del S3 Bucket.

Figura 30. Interfaz de AWS Lambda + Código fuente de consolidación de datos



Nota. Elaboración propia (2025)

Figura 31. Archivo almacenado en la carpeta Gold



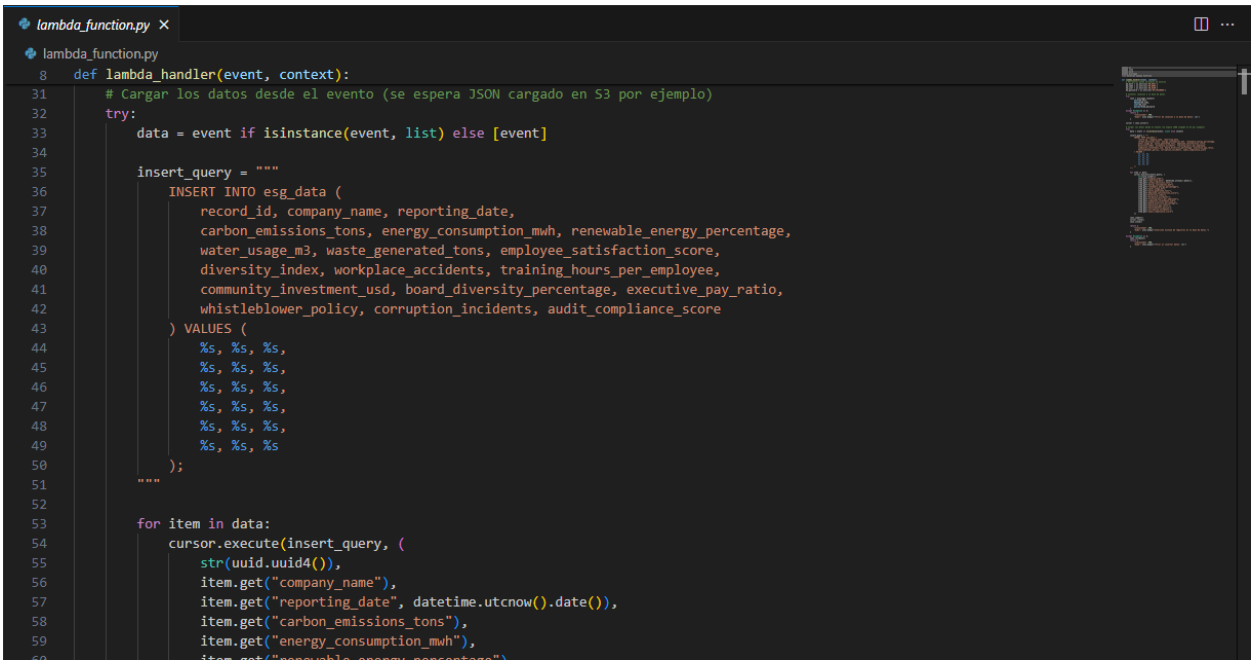
Nota. Elaboración propia (2025)

5.1.4.4 Etapa: Carga en base de datos relacional

La presente etapa del flujo automatizado corresponde a la carga estructurada de los datos en la base de datos relacional Aurora, lo cual permite su almacenamiento permanente y acceso controlado para procesos de validación y análisis posteriores. Esta acción representa el cierre lógico del flujo automatizado de carga de datos maestros, asegurando que la información procesada desde las capas iniciales del flujo sea finalmente persistida en una tabla compatible con esquemas relacionales tipo PostgreSQL.

Para ello, se implementó la función Lambda *insert_to_aurora_v2*, desarrollada en Python y equipada con la biblioteca psycopg2, responsable de establecer la conexión con el clúster Aurora PostgreSQL. Esta función recibe un conjunto de registros en formato JSON (provenientes de la zona Gold del bucket S3), y realiza la inserción de cada uno en la tabla *esg_data* a través de una sentencia SQL parametrizada. La estructura de dicha tabla fue diseñada para albergar los 18 campos validados a lo largo del flujo, incluyendo indicadores ambientales, sociales y de gobernanza. Adicionalmente, cada registro es enriquecido con un *record_id* único generado mediante UUID, lo que fortalece la trazabilidad y la integridad referencial. El siguiente fragmento visualizado en la Figura 32, resume la lógica central utilizada en esta función.

Figura 32. Lógica de inserción de datos en Aurora



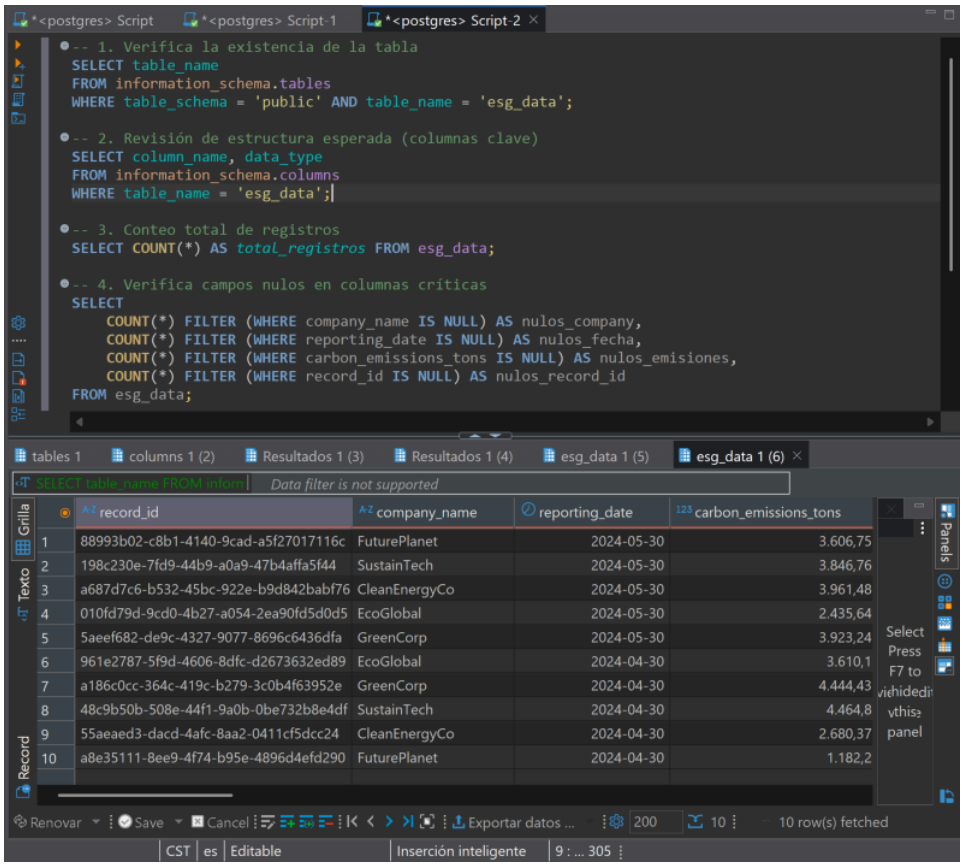
```python
lambda_function.py
def lambda_handler(event, context):
    # Cargar los datos desde el evento (se espera JSON cargado en S3 por ejemplo)
    try:
        data = event if isinstance(event, list) else [event]

        insert_query = """
            INSERT INTO esg_data (
                record_id, company_name, reporting_date,
                carbon_emissions_tons, energy_consumption_mwh, renewable_energy_percentage,
                water_usage_m3, waste_generated_tons, employee_satisfaction_score,
                diversity_index, workplace_accidents, training_hours_per_employee,
                community_investment_usd, board_diversity_percentage, executive_pay_ratio,
                whistleblower_policy, corruption_incidents, audit_compliance_score
            ) VALUES (
                %s, %s, %s,
                %s, %s, %s,
                %s, %s, %s,
                %s, %s, %s,
                %s, %s, %s,
                %s, %s, %s
            );
        """

        for item in data:
            cursor.execute(insert_query, (
                str(uuid.uuid4()),
                item.get("company_name"),
                item.get("reporting_date", datetime.utcnow().date()),
                item.get("carbon_emissions_tons"),
                item.get("energy_consumption_mwh"),
```

Nota. Elaboración propia (2025)

Esta instrucción fue ejecutada por cada objeto del evento recibido, utilizando identificadores únicos (uuid4) para la clave primaria *record_id*. El procesamiento contempló conversiones de tipos, validación de campos obligatorios y gestión de errores mediante try/except con retroceso de la transacción en caso de excepción (conn.rollback()).

La validación de la carga de datos se efectuó a través del cliente *DBeaver*, herramienta que permitió conectarse directamente al clúster Aurora y consultar la tabla *esg_data*. Allí se confirmó la existencia de los registros previamente transformados y consolidados, con estructura completa y sin inconsistencias. Esta verificación sirvió como evidencia tangible de la correcta persistencia de los datos en el sistema relacional. A continuación en la Figura 33, se muestra la verificación estructural de la tabla *esg_data* mediante cliente DBeaver.

Figura 33. Datos cargados en Aurora



Nota. Elaboración propia (2025)

Por motivos asociados al modelo de facturación de AWS, es importante señalar que la ejecución de consultas SQL sobre instancias Aurora implica un consumo de recursos que excede los límites establecidos por el plan Free Tier. Considerando que este trabajo se enmarca dentro del desarrollo de un prototipo funcional, se descartó la realización de consultas intensivas desde el entorno de AWS, dado que la comprobación por medio de herramientas locales resultó suficiente para validar la integridad del flujo automatizado sin incurrir en costos innecesarios.

5.1.4.5 Etapa: Simulación de publicación en plataforma de Gestión de Datos Maestros

La etapa final del flujo automatizado consiste en la simulación de la publicación de los datos en la plataforma de Gestión de Datos Maestros (MDM). En un entorno corporativo real, esta fase implicaría la conexión directa entre la base de datos relacional Aurora y la plataforma MDM utilizada por la organización, permitiendo la validación, edición y publicación formal de los registros cargados. Sin embargo, debido a restricciones propias del entorno de pruebas y por motivos de confidencialidad, esta conexión no fue implementada en el prototipo funcional.

Como alternativa, se optó por simular esta interacción técnica bajo un enfoque conceptual estructurado. Se asumió que los datos insertados correctamente en la tabla *esg_data* de Aurora

quedarían disponibles para lectura por parte de la plataforma MDM, replicando fielmente las condiciones necesarias para su validación. Esta suposición se basa en la correcta estructuración, consolidación e integridad de los datos obtenidos durante todas las etapas anteriores del flujo automatizado.

El esquema relacional definido, junto con la validación de consistencia efectuada mediante consultas SQL y revisión en *DBeaver*, permiten afirmar que los datos cumplen con los requisitos fundamentales para ser consumidos por una plataforma MDM: unicidad de identificadores (*record_id*), completitud en campos críticos, y representación estandarizada de indicadores ESG. Asimismo, se consideró la existencia lógica de un punto de acceso desde MDM hacia la tabla *esg_data*, aunque no se configuró una interfaz de integración real debido al alcance acotado del prototipo.

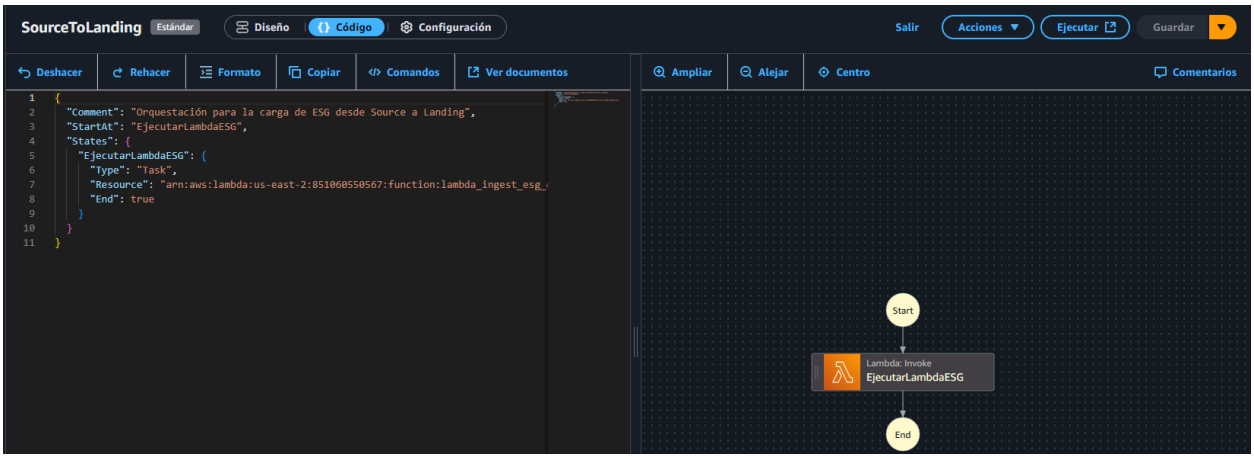
Desde una perspectiva funcional, esta simulación resulta válida para demostrar la viabilidad del flujo completo. La disposición final de los registros en Aurora representa el estado más avanzado del flujo de datos automatizado y responde a la necesidad de asegurar trazabilidad, calidad estructural y disponibilidad para futuras integraciones. Este diseño también deja abierta la posibilidad de extender el prototipo hacia escenarios reales de publicación y gobierno de datos, sin necesidad de reestructurar el modelo actual.

5.1.4.6 Etapa: Orquestación del flujo automatizado de carga de datos

La automatización de procesos mediante flujos orquestados representa una de las principales fortalezas del entorno AWS. En el contexto del presente prototipo, la primera fase del flujo fue implementada utilizando una máquina de estado desarrollada con AWS Step Functions, la cual coordina el paso inicial del proceso: la extracción de datos desde una fuente externa simulada y su posterior almacenamiento en la zona Bronze del bucket S3.

La máquina de estado denominada ***SourceToLanding*** consta de una única tarea representada por el estado *EjecutarLambdaESG*, cuya función es invocar de forma controlada la función Lambda *lambda_ingest_esg_data*. Esta función actúa como punto de ingreso de los datos ESG al entorno *cloud*, validando que el evento recibido contenga información estructurada y generando dinámicamente un archivo con formato JSON, que posteriormente es almacenado en la ruta Bronze/esg/ dentro del bucket S3 prototipo-mdm-kemuel. El código de definición de la Step Function en formato JSON se presenta a continuación en la Figura 34.

Figura 34. Código de definición de la AWS Step Function - SourceToLanding



Nota. Elaboración propia (2025)

Este flujo comienza en el estado EjecutarLambdaESG y finaliza una vez que la función Lambda ha ejecutado su tarea con éxito. En caso de fallos, la propia Step Function permite capturar errores para su posterior depuración, aunque en el prototipo se trabajó bajo una estructura lineal simple, sin bloques de manejo de excepciones integrados en el JSON de definición.

Cabe destacar que, por tratarse de una versión funcional simplificada, la máquina de estado ejecuta únicamente una función Lambda correspondiente a la fuente ESG. No obstante, en un entorno organizacional real, esta misma Step Function sería extendida para orquestar de forma paralela las cinco funciones Lambda asociadas a cada fuente de datos: *Credit Risk Data, Ratings, ESG, News Edge* y *Corporate Intelligence Database*. Cada una de estas tareas sería ejecutada como un estado independiente, permitiendo así el ingreso simultáneo de múltiples flujos de datos al entorno de almacenamiento inicial, y garantizando la integridad del proceso en escenarios de mayor volumen y complejidad operativa.

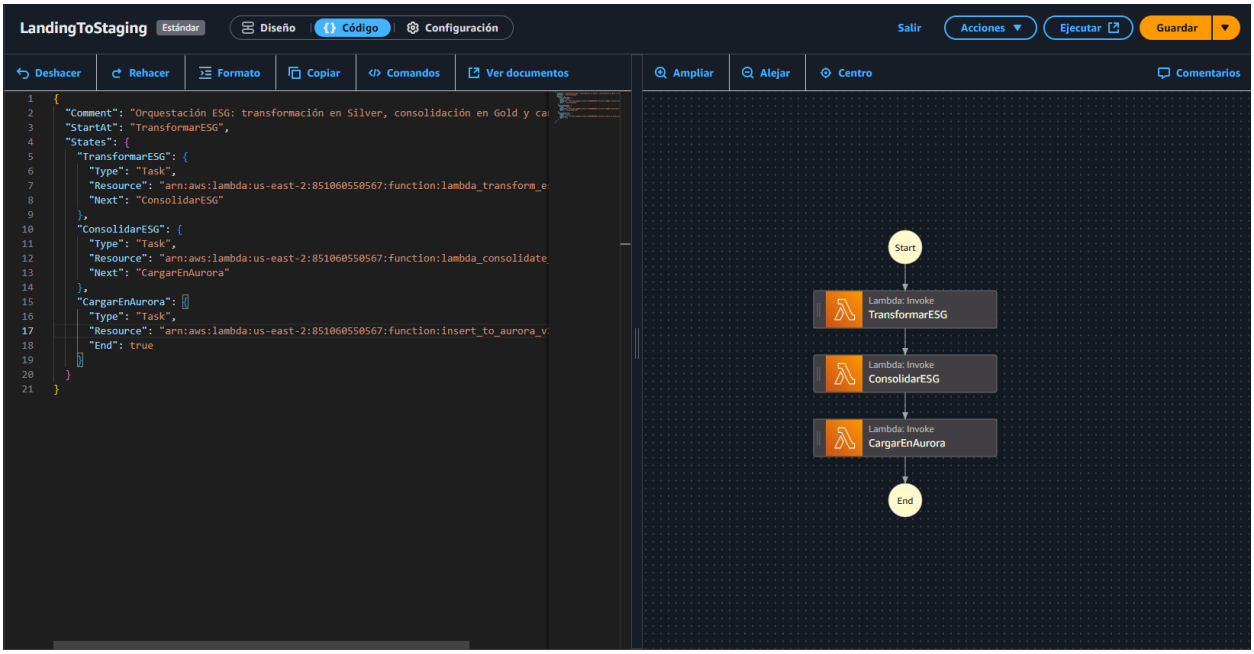
La segunda máquina de estado diseñada dentro del entorno de Amazon Web Services (***LandingToStaging***) tiene como objetivo coordinar las operaciones críticas de transformación, consolidación y carga de los datos ESG, asegurando la progresión estructurada desde la zona Silver hacia la base de datos relacional Aurora. Esta etapa representa el núcleo funcional del flujo de procesamiento, ya que aplica reglas de validación, controla la integridad de los datos y los prepara para su disponibilidad futura en la plataforma de Gestión de Datos Maestros. La máquina de estado ***LandingToStaging*** se compone de tres tareas secuenciales.

- **TransformarESG**: esta primera tarea invoca la función Lambda *lambda_transform_esg_data*, la cual lee todos los archivos contenidos en la carpeta Bronze/esg/, valida que los registros cumplan con la estructura esperada y los escribe en formato JSON dentro de la carpeta Silver/esg/. Durante esta validación estructural se revisan campos clave como *record_id*, *company_name*, *reporting_date* y distintos indicadores ESG, descartando cualquier entrada incompleta o malformada.

- **ConsolidarESG**: una vez validados, los datos se consolidan mediante la función *lambda_consolidate_esg_data*. Esta Lambda toma como entrada un archivo específico de la carpeta Silver y lo copia directamente en la carpeta Gold/esg/, donde se ubican los archivos listos para ser cargados en la plataforma MDM. Aunque en este prototipo la consolidación se limita a una única fuente (ESG), en el entorno organizacional real esta tarea integraría múltiples archivos provenientes de todas las fuentes de datos definidas: *Credit Risk Data, Ratings, ESG, News Edge* y *Corporate Intelligence Database*.
- **CargarEnAurora**: esta etapa finaliza el flujo de integración con la ejecución de la función *insert_to_aurora_v2*, responsable de insertar los datos validados en la base de datos relacional Amazon Aurora. La función se conecta utilizando el motor PostgreSQL y emplea la biblioteca psycopg2 para ejecutar sentencias SQL parametrizadas, garantizando una inserción segura y estructurada. Durante la ejecución, se recorren los registros contenidos en el archivo consolidado y se insertan en la tabla *esg_data*, generando un *record_id* único mediante la librería *uuid*.

La definición técnica de la máquina de estado se define en la Figura 35.

Figura 35. Código de definición de la AWS Step Function – LandingToStaging



Nota. Elaboración propia (2025)

Al igual que en la máquina de estado *SourceToLanding*, por fines de simplicidad y control en esta fase inicial, el prototipo se limitó a ejecutar únicamente las funciones asociadas a la fuente ESG. En un entorno organizacional completo, este flujo incorporaría múltiples ramas de ejecución paralela que orquestarían la transformación y consolidación de todas las fuentes relevantes, permitiendo así una carga hacia la plataforma MDM.

5.1.4.7 Etapa: Monitoreo integral

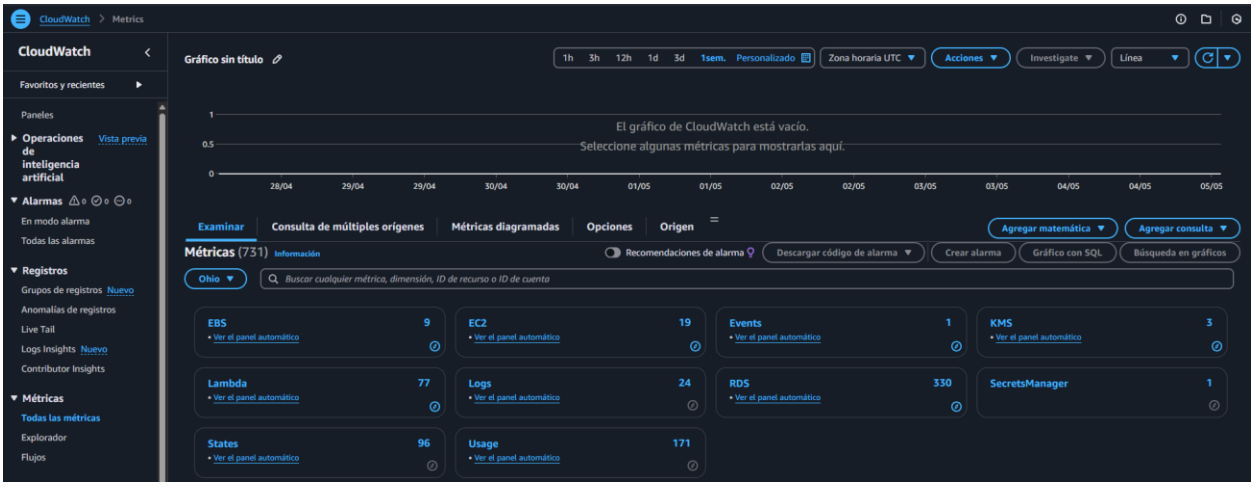
La última fase del prototipo automatizado se centró en establecer un entorno de monitoreo integral que permitiera supervisar en tiempo real la ejecución de cada una de las funciones Lambda, el comportamiento de las máquinas de estado y los eventos asociados a la base de datos Aurora. Para ello, se utilizó Amazon CloudWatch, la solución nativa de AWS para recolección y visualización de métricas operativas y de rendimiento.

Desde el panel central de CloudWatch se habilitó la recopilación de métricas en las siguientes categorías relevantes:

- **Lambda**: Incluye tiempos de ejecución, número de invocaciones, errores y tasas de éxito por función.
- **States**: Permite rastrear el estado de cada ejecución de Step Functions, determinando si fue completada correctamente, fallida o en espera.
- **Logs**: Reúne los registros detallados generados por cada función Lambda, ofreciendo trazabilidad completa de los mensajes de entrada, salidas, errores y validaciones.
- **RDS**: Aunque no se realizaron consultas activas por motivos presupuestarios del Free Tier, el monitoreo de instancias Aurora PostgreSQL quedó disponible para futuras pruebas.

Asimismo, es importante destacar que tanto AWS Lambda como Step Functions exponen sus propias métricas directamente en sus interfaces, lo que facilita la visualización rápida del estado de cada invocación, incluyendo duración, éxito, error, tiempo de espera y frecuencia de ejecución. Estas métricas no solo permiten diagnósticos ágiles, sino que también alimentan automáticamente los tableros y alarmas de Amazon CloudWatch, fortaleciendo la trazabilidad y robustez del prototipo desde una perspectiva operativa. En la Figura 36, se observa la interfaz de Amazon CloudWatch.

Figura 36. Interfaz de Amazon CloudWatch



Nota. Tomado de Amazon Web Services (2025)

5.1.5 Configuración de permisos y roles IAM en la arquitectura del prototipo

Aunque la gestión de identidades y permisos no constituye una etapa funcional directa del flujo de carga de datos maestros, resulta indispensable para garantizar la ejecución segura y controlada del prototipo dentro del entorno de Amazon Web Services (AWS). En este contexto, la correcta configuración de roles IAM (*Identity and Access Management*) representa un componente crítico para el funcionamiento exitoso de las funciones Lambda, los flujos de Step Functions y la conexión a servicios como Amazon S3 o Amazon Aurora.

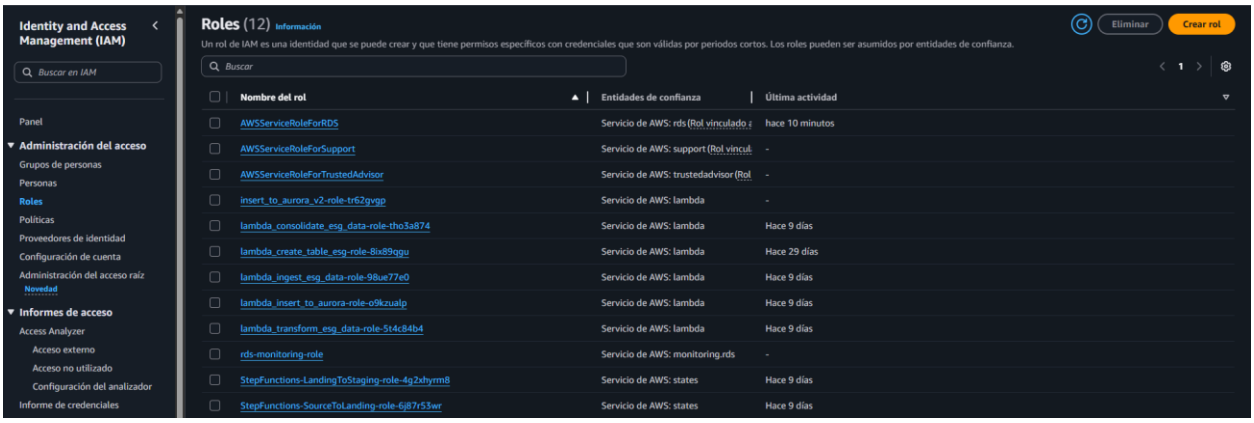
Durante la construcción del prototipo, se definió una estrategia de permisos segmentada por funcionalidad, siguiendo el principio de privilegios mínimos, recomendado por las buenas prácticas de seguridad en la nube. Cada componente que requiere interacción con otros servicios; por ejemplo, una función Lambda que debe escribir en un *bucket* S3 o cargar registros en la base de datos relacional (Aurora PostgreSQL), cuenta con un rol IAM independiente que delimita claramente las acciones que está autorizado a ejecutar.

En total, se crearon 12 roles de IAM personalizados, incluyendo:

- Roles exclusivos para cada función Lambda (por ejemplo, ***lambda_ingest_esg_data***, ***lambda_transform_esg_data***, ***lambda_consolidate_esg_data***, ***insert_to_aurora_v2***, entre otros).
- Roles específicos para las Step Functions ***SourceToLanding*** y ***LandingToStaging***, con permisos delegados para invocar únicamente las funciones Lambda autorizadas dentro de su flujo de orquestación.
- Un rol de monitoreo (***rds-monitoring-role***) que permite a los servicios de Amazon RDS reportar métricas en CloudWatch.
- Roles integrados de servicio como ***AWSServiceRoleForRDS*** o ***AWSServiceRoleForSupport***, que facilitan la operación nativa de servicios gestionados.

Esta segmentación no solo facilita la trazabilidad de errores y auditorías de seguridad, sino que también permite revocar o ajustar permisos de forma granular en caso de actualizaciones futuras. En síntesis, aunque esta gestión no es una fase visible para el usuario final del proceso automatizado, constituye una base estructural obligatoria para el correcto despliegue, orquestación y ejecución segura de cada componente del prototipo de solución. Su documentación resulta, por tanto, esencial dentro del diseño técnico del prototipo. En la Figura 37 se visualiza la configuración de permisos y roles IAM creados.

Figura 37. Configuración de permisos y roles IAM en la arquitectura del prototipo



Nota. Elaboración propia (2025)

5.1.6 Matriz de validación de requerimientos vs prototipo implementado

Con el objetivo de asegurar la trazabilidad entre los requerimientos funcionales definidos en la fase 2 del proyecto y las funcionalidades desarrolladas dentro del prototipo, se construyó una matriz de validación que detalla, para cada área del flujo automatizado, la correspondencia con los componentes implementados en el entorno de Amazon Web Services. La matriz correspondiente, incluida en la Tabla 28, permite verificar de manera sistemática que cada especificación fue contemplada técnica y funcionalmente, ya sea mediante su ejecución directa o por medio de simulación lógica dentro del alcance establecido para el entorno de pruebas.

Tabla 28. Matriz requerimientos vs prototipo

Matriz de requerimientos vs prototipo funcional				
Área funcional	**Especificación del requerimiento**	**Herramienta definida**	**Implementación en el prototipo**	**Estado**
Orquestación de extracción	Coordinar la descarga automática de archivos desde múltiples fuentes externas hacia la zona Bronze del bucket de almacenamiento.	AWS Step Functions	Implementada mediante la Step Function ***SourceToLanding***, la cual orquesta la ejecución de la función Lambda *lambda_ingest_esg_data*. En este prototipo se usó únicamente una fuente de datos (ESG) para efectos de demostración.	Cumplido
Extracción de datos	Ejecutar funciones independientes que extraen archivos desde las fuentes definidas.	AWS Lambda	Función *lambda_ingest_esg_data* implementada exitosamente. Extrae datos en formato JSON y los almacena en la carpeta Bronze.	Cumplido
Almacenamiento inicial	Centralizar los archivos sin procesar en una zona de almacenamiento	Amazon S3	Los archivos sin procesar se almacenan en la zona **Bronze/esg/** del bucket prototipo-mdm-kemuel. La	Cumplido

Matriz de requerimientos vs prototipo funcional				
	seguro y segmentado por capas.		estructura por capas (Bronze, Silver, Gold) fue aplicada correctamente.	
Transformación de datos	Validar automáticamente la estructura de cada archivo, eliminar duplicados y estandarizar los formatos de forma modular.	AWS Lambda	Función *lambda_transform_esg_data* transforma y valida la estructura de los registros. Verifica campos críticos definidos y transfiere los datos válidos a la carpeta Silver.	Cumplido
Validación de consistencia	Interrumpir el flujo en caso de errores estructurales detectados durante la transformación.	AWS Step Functions	La lógica de validación se encuentra embebida dentro de la Lambda. Si no hay registros válidos, no se genera salida. El flujo es controlado por la Step Function ***LandingToStaging***, que gestiona estas condiciones sin propagar fallos innecesarios.	Cumplido
Almacenamiento estructurado	Guardar los archivos validados en una zona diferenciada para preparación de consolidación.	Amazon S3	Al finalizar la validación estructural, los archivos válidos se almacenan correctamente en la carpeta **Silver/esg/**.	Cumplido
Consolidación de datos	Integrar los archivos transformados bajo un esquema unificado mediante un proceso de consolidación centralizado.	AWS Lambda	Para este prototipo, se usó una función Lambda (*lambda_consolidate_esg_data*) como alternativa a AWS Glue, dada la simplicidad y unicidad de la fuente ESG. Simula la consolidación de múltiples	Cumplido

Matriz de requerimientos vs prototipo funcional				
			archivos en un único archivo en Gold.	
Almacenamiento final	Conservar los datos consolidados listos para la carga estructurada.	Amazon S3	Los datos consolidados se almacenan en la carpeta **Gold/esg/** bajo una estructura estandarizada, con nombres únicos por *timestamp*.	Cumplido
Carga en base de datos	Insertar automáticamente los datos consolidados en la base de datos relacional utilizada por la plataforma de gestión de datos maestros.	AWS Lambda / Amazon Aurora	Se implementó la función *insert_to_aurora_v2*, que realiza la conexión con la base Aurora PostgreSQL e inserta los registros. La función es orquestada por la Step Function ***LandingToStaging***.	Cumplido
Publicación en plataforma MDM	Acceder a los datos desde la base estructurada para su posterior validación y publicación dentro de la plataforma MDM.	Amazon Aurora / Plataforma MDM	Esta etapa fue simulada conceptualmente. Se asumió que la tabla *esg_data* en Aurora refleja el modelo requerido por la plataforma MDM, y se validó la estructura mediante consultas SQL en DBeaver.	Simulado
Monitoreo del proceso	Registrar logs de ejecución, errores y métricas de rendimiento en cada etapa del flujo automatizado.	Amazon CloudWatch	CloudWatch registró la actividad de todas las funciones Lambda y de las Step Functions utilizadas. Se confirmó la visibilidad de métricas, errores y logs.	Cumplido

Nota. Elaboración propia (2025)

5.1.7 Análisis de riesgos de la solución

El análisis de riesgos es una práctica fundamental en la gestión de proyectos, cuyo objetivo es identificar y preparar respuestas ante eventos inciertos que podrían afectar negativamente el cumplimiento de los objetivos del proyecto. De acuerdo con la Guía del PMBOK (Project Management Institute, 2017), el riesgo se define como "un evento o condición incierta que, de ocurrir, tiene un efecto positivo o negativo en uno o más objetivos del proyecto" (PMI, 2017, p. 395).

Dentro de este enfoque, el análisis de riesgos permite reconocer de forma proactiva las amenazas más relevantes, estimar la probabilidad de su ocurrencia, cuantificar su impacto y establecer planes de acción apropiados para su mitigación o contingencia. Este proceso se integra en el grupo de procesos de planificación, como parte de la gestión de los riesgos del proyecto, y se apoya en herramientas como análisis cualitativos, cuantitativos, matrices de impacto y mapas de calor.

Para este Trabajo Final de Graduación, se ha adoptado una estructura simplificada y funcional para el análisis de riesgos, considerando la naturaleza de un prototipo en fase de pruebas y el contexto del entorno AWS utilizado. Esta metodología, sigue lineamientos sólidos del PMBOK, proporcionando un balance adecuado entre rigurosidad y aplicabilidad para una solución académica. Se busca así garantizar que los riesgos técnicos, operativos y estratégicos sean gestionados de manera eficaz, sin sobrecargar la estructura del prototipo con herramientas innecesarias.

5.1.7.1 Categoría de riesgos

La categorización de riesgos constituye una práctica esencial en la gestión de proyectos, ya que permite identificar áreas específicas que presentan mayor exposición a la incertidumbre, facilitando una respuesta más eficaz y focalizada. De acuerdo con el PMBOK Guide (Project Management Institute, 2017), esta técnica se utiliza para agrupar los riesgos por fuentes comunes, facilitando su análisis estructurado y la asignación de responsabilidades de gestión (p. 409).

En el caso del prototipo funcional diseñado para automatizar el proceso de carga de datos en la plataforma de Gestión de Datos Maestros, la categorización no se limita a los desafíos del entorno de pruebas. Más bien, se proyecta hacia el escenario operativo de la organización, anticipando los riesgos que emergerían durante la eventual transición a un entorno de producción. No obstante, las pruebas ejecutadas en el entorno AWS Free Tier permitieron identificar riesgos potenciales que orientan esta clasificación. Se definieron cinco categorías principales de riesgo:

- **Riesgos tecnológicos**: esta categoría contempla incompatibilidades entre servicios en la nube, diferencias entre versiones de entornos de ejecución (por ejemplo, la incompatibilidad temporal de Python 3.12 con bibliotecas utilizadas en AWS Lambda), así como fallos derivados de dependencias no soportadas. También incluye posibles errores en la integración entre funciones Lambda y bases de datos relacionales como Amazon Aurora.

- **Riesgos operativos**: se refieren a fallos en la ejecución del flujo automatizado, errores en la validación estructural de datos, interrupciones por archivos incompletos o inconsistencias en el esquema de los registros. Estas fallas afectan la trazabilidad y la confiabilidad de los datos transformados y consolidados.
- **Riesgos de seguridad**: asociados a accesos no autorizados a datos almacenados en Amazon S3 o Aurora, gestión deficiente de credenciales a través de variables de entorno, y configuración inadecuada de políticas de acceso (IAM). Aunque estos aspectos se controlaron en el prototipo, su criticidad aumenta considerablemente en un entorno productivo.
- **Riesgos de disponibilidad de servicios en la nube**: incluyen posibles interrupciones, latencias o degradación en los servicios de AWS que componen la solución. Aunque algunos límites fueron evidentes durante las pruebas en Free Tier, como la imposibilidad de realizar consultas SQL desde el entorno gratuito, esta categoría considera el comportamiento esperado bajo condiciones operativas reales.
- **Riesgos de escalabilidad futura**: aunque no se manifestaron durante el prototipo, se anticipa que al integrar múltiples fuentes de datos y aumentar la frecuencia de ejecución, surjan desafíos relacionados con concurrencia, control de versiones, balance de carga y supervisión de errores. Esta categoría adquiere relevancia en la planificación de un despliegue organizacional a mayor escala.

5.1.7.2 Identificación de riesgos

La identificación de riesgos es un proceso sistemático que permite reconocer y documentar los eventos potenciales que podrían afectar negativamente los objetivos de un proyecto. Según el PMBOK Guide (Project Management Institute, 2017), esta etapa constituye la base del análisis posterior y requiere considerar tanto factores internos como externos al proyecto (pp. 405–406). Para el desarrollo del prototipo de automatización en Amazon Web Services, la identificación de riesgos se realizó tomando en cuenta las condiciones técnicas reales observadas durante las pruebas, así como las características del entorno organizacional futuro en la empresa. A continuación en la Tabla 29, se enumeran los principales riesgos identificados.

Tabla 29. Riesgos identificados

Riesgos identificados		
ID-Riesgo	**Nombre**	**Descripción**
RT – 01	Incompatibilidad del entorno Python en AWS Lambda	Durante el desarrollo se evidenció que la versión 3.12 de Python aún no cuenta con soporte completo en AWS Lambda para ciertas bibliotecas como *psycopg2*, lo cual limitó temporalmente la ejecución

Riesgos identificados		
		de funciones de carga en bases de datos relacionales.
RT – 02	Dependencia de librerías externas no compiladas nativamente	La necesidad de utilizar librerías con componentes binarios representa un riesgo si estas no se encuentran precompiladas para el entorno de ejecución de Lambda, generando posibles fallos en la ejecución del flujo automatizado.
RO – 01	Ausencia de validación estructural en la capa de entrada	Si los archivos cargados en la zona Bronze no respetan el esquema definido, el proceso de transformación en Silver podría fallar, interrumpiendo el flujo general de consolidación y carga de datos.
RO – 02	Pérdida de datos durante la consolidación	Una configuración incorrecta en la función que copia archivos desde Silver hacia Gold podría derivar en la sobrescritura accidental de archivos o en la consolidación incompleta de registros.
RS – 01	Configuración errónea de credenciales o permisos IAM	El mal manejo de variables de entorno, roles IAM o políticas de acceso podría impedir que las funciones Lambda interactúen con los recursos asignados, generando fallas críticas en la ejecución.
RA – 01	Latencias o interrupciones en servicios de AWS	La ejecución del flujo depende de múltiples servicios (S3, Lambda, Step Functions, Aurora). Una interrupción puntual en alguno de ellos podría dejar el proceso en estado

Riesgos identificados		
		inconsistente o inactivo temporalmente.
RE – 01	Aumento en volumen de datos no gestionado	Aunque el prototipo funciona correctamente con una sola fuente de datos (ESG), la transición a un entorno productivo con múltiples fuentes simultáneas requiere estrategias de paralelización, control de versiones y escalabilidad que, si no se planifican correctamente, podrían comprometer la estabilidad del sistema.

Nota. Elaboración propia (2025)

5.1.7.3 Definición de probabilidad e impacto de riesgos

Con el propósito de establecer una base objetiva para la evaluación de riesgos en el prototipo de automatización del proceso de carga de datos, se definieron escalas cualitativas para medir tanto la probabilidad de ocurrencia como el impacto de cada riesgo identificado. Esta clasificación permite priorizar los riesgos de forma estructurada y facilita la toma de decisiones para su mitigación, de acuerdo con las recomendaciones del PMBOK Guide (Project Management Institute, 2017, pp. 413–414).

Escala de probabilidad:

- **Baja (1):** El evento es poco probable y no se ha presentado en las pruebas realizadas.
- **Media (2):** Existe una posibilidad moderada de ocurrencia, basada en limitaciones del entorno actual (por ejemplo, el Free Tier o compatibilidades de versión).
- **Alta (3):** El evento es probable o ya ha ocurrido durante la ejecución del prototipo.

Escala de impacto:

- **Bajo (1):** El evento no afecta significativamente la funcionalidad del flujo; es posible corregirlo sin reestructuración técnica.
- **Medio (2):** El evento afecta parcialmente el flujo automatizado, requiriendo ajustes en la configuración, lógica o entorno.
- **Alto (3):** El evento impide la ejecución del prototipo o compromete su integridad lógica, generando una falla estructural.

Esta doble escala facilita la construcción de una matriz de riesgos, que será utilizada en la siguiente sección para representar visualmente el nivel de exposición de cada riesgo mediante su

ubicación dentro de una matriz de impacto vs. probabilidad. El uso de esta metodología garantiza trazabilidad, transparencia y claridad en el análisis de amenazas técnicas y operativas relacionadas con el prototipo.

5.1.7.4 Matriz de probabilidad e impacto de riesgos

Una vez definidos los niveles de probabilidad e impacto, se procede a la construcción de una matriz de riesgos que permite visualizar, de forma estructurada, el nivel de exposición asociado a cada riesgo identificado en el prototipo de automatización del flujo de carga de datos. Esta matriz, incluida en la Tabla 30, combina las escalas cualitativas establecidas previamente para determinar el nivel de riesgo resultante, categorizado como bajo, moderado o alto, según la siguiente lógica:

Tabla 30. Matriz de probabilidad e impacto de riesgos

Matriz de probabilidad e impacto de riesgos			
Impacto / Probabilidad	**Baja (1)**	**Media (2)**	**Alta (3)**
Bajo (1)	Bajo	Bajo	Moderado
Medio (2)	Bajo	Moderado	Alto
Alto (3)	Moderado	Alto	Alto

Nota. Elaboración propia (2025)

5.1.7.5 Análisis cuantitativo de riesgos

El análisis cuantitativo de riesgos tiene como propósito asignar una medida numérica al nivel de exposición ante los eventos identificados, con el fin de facilitar su priorización y la toma de decisiones informadas. Para este propósito, se empleó un modelo de ponderación simple, en el cual la probabilidad y el impacto de cada riesgo fueron evaluados en una escala de 1 a 3. Posteriormente, se multiplicaron ambos valores para obtener un índice de riesgo (IR), cuya fórmula es la siguiente:

$$Índice\ de\ riesgo\ (IR) = Probabilidad\ \times Impacto$$

El IR resultante permite clasificar cada riesgo de acuerdo con el siguiente criterio:

- **1 a 3:** Riesgo Bajo
- **4 a 6:** Riesgo Moderado
- **7 a 9:** Riesgo Alto

La Tabla 31, resume los resultados cuantitativos obtenidos para los riesgos analizados.

Tabla 31. Análisis cuantitativo de riesgos

Análisis cuantitativo de riesgos				
ID - Riesgo	**Probabilidad (1-3)**	**Impacto (1-3)**	$\boldsymbol{IR = P \times I}$	**Clasificación**
RT-01	3	3	9	Alto
RT-02	2	3	6	Moderado
RO-01	2	2	4	Moderado
RO-02	3	2	6	Moderado

Análisis cuantitativo de riesgos				
RS-01	2	3	6	Moderado
RA-01	2	1	2	Bajo
RE-01	3	2	6	Moderado

Nota. Elaboración propia (2025)

El riesgo con mayor puntuación (IR = 9) fue la incompatibilidad entre la versión de Python utilizada en AWS Lambda con librerías necesitadas (RT-01), lo cual impidió temporalmente realizar la carga directa a Aurora desde el flujo automatizado. Aunque este evento no representó una falla estructural del prototipo, sí implicó ajustes técnicos y pruebas adicionales. Este análisis permitió establecer prioridades claras en la formulación del plan de respuesta, especialmente para aquellos riesgos que comprometen la funcionalidad crítica del prototipo.

5.1.7.6 Mapa de calor / nivel de riesgos

Para facilitar la interpretación visual del impacto de los riesgos identificados en el prototipo, se construyó un mapa de calor con base en la escala del análisis cuantitativo. Esta matriz clasifica los riesgos en tres niveles.

- **Bajo (Verde):** Puntaje de 1 a 3
- **Medio (Amarillo):** Puntaje de 4 a 6
- **Alto (Rojo):** Puntaje de 7 a 9

A continuación, en la Tabla 32, se presenta la matriz.

Tabla 32. Mapa de calor de riesgos

Impacto / Probabilidad	1 (Baja)	2 (Media)	3 (Alta)
3 (Alta)		RT-02, RS-01	RT-01
2 (Media)		RO-01	RO-02, RE-01
1 (Baja)		RA-01	

Nota. Elaboración propia (2025)

La matriz de impacto y probabilidad permite visualizar la concentración de los riesgos del prototipo según su severidad estimada. Como se observa, el riesgo RT-01, relacionado con la incompatibilidad del entorno Python 3.12 en AWS Lambda, se posiciona en la zona roja, lo cual lo clasifica como un riesgo alto que requiere atención inmediata antes de escalar a producción.

En la categoría de riesgo medio (zona amarilla), se ubican RT-02 y RS-01, asociados a librerías externas no compiladas y configuración incorrecta de permisos IAM, respectivamente, así como RO-01, vinculado a la validación estructural de archivos en la capa de entrada.

Los riesgos RO-02 (pérdida de datos en consolidación) y RE-01 (gestión de grandes volúmenes en producción) también se sitúan en una zona media-alta, reflejando que, si bien no son críticos en el prototipo actual, escalarán en relevancia al integrarse más fuentes de datos.

Finalmente, el riesgo RA-01, relacionado con interrupciones de servicios AWS, se ubica en el cuadrante verde de bajo impacto y baja probabilidad, dado que no se registraron fallos en la infraestructura durante la fase de pruebas.

5.1.7.7 Análisis cualitativo de riesgos

El análisis cualitativo de riesgos permite evaluar con mayor profundidad la naturaleza de cada riesgo identificado, su origen, factores desencadenantes y posibles consecuencias, con el fin de priorizar su tratamiento de acuerdo con su criticidad dentro del prototipo. Esta evaluación complementa el análisis cuantitativo mediante una interpretación basada en la comprensión del entorno técnico de Amazon Web Services (AWS) y las características específicas del flujo automatizado diseñado. En la Tabla 33, se especifica el resultado del análisis cualitativo de riesgos.

Tabla 33. Análisis cualitativo de riesgos

Análisis cualitativo de riesgos			
ID – Riesgo	**Justificación**	**Prioridad**	**Recomendación**
RT-01	Este riesgo se considera crítico, ya que afecta directamente la ejecución de funciones clave, como la carga en Aurora. Aunque se identificó durante la fase de pruebas, su impacto estructural exige monitoreo constante y validación previa del entorno antes del despliegue en producción.	Alta	Restringir el uso de versiones no certificadas por AWS para producción.
RT-02	Este riesgo posee un impacto técnico moderado, ya que compromete la portabilidad y estabilidad de las funciones Lambda cuando dependen de bibliotecas complejas.	Media	Sustituir dependencias críticas por paquetes compatibles con *Lambda Layers* o contenedores.
RO-01	Este riesgo podría derivar en errores	Media	Mantener validadores estructurales activos

Análisis cualitativo de riesgos			
	silenciosos que afectan la calidad de los datos procesados. Es especialmente sensible si no se controlan adecuadamente los formatos en la zona Bronze.		como etapa inicial del proceso.
RO-02	Aunque no se materializó en el prototipo, este riesgo posee relevancia operativa en entornos productivos por la posible sobrescritura de archivos en Gold.	Alta	Implementar control de versiones y validaciones de integridad.
RS-01	Se trata de un riesgo técnico con alta recurrencia en proyectos basados en AWS, especialmente en prototipos. Su aparición podría bloquear ejecuciones críticas del flujo automatizado.	Media	Usar políticas IAM con privilegios mínimos y revisiones periódicas de roles.
RA-01	Aunque de baja probabilidad durante la prueba, sigue siendo una amenaza en contextos reales donde múltiples servicios concurren.	Baja	Diseñar mecanismos de reintento y alertas ante fallos en servicios *core*.
RE-01	Su relevancia aumenta al escalar a múltiples fuentes simultáneas. Si no se aborda, comprometería la integridad y el rendimiento del proceso.	Alta	Planificar escalabilidad desde el diseño, aplicando procesamiento paralelo y control de cargas.

Nota. Elaboración propia (2025)

El análisis cualitativo permitió identificar con claridad los riesgos críticos que afectan la estabilidad, seguridad y escalabilidad del prototipo desarrollado. Si bien algunos riesgos se presentaron exclusivamente en el entorno de pruebas (como la incompatibilidad con ciertos entornos de ejecución), otros representan desafíos estructurales en la transición hacia una solución productiva. Esta evaluación proporciona una base sólida para priorizar acciones correctivas y establecer mecanismos preventivos, asegurando que la solución mantenga su funcionalidad en condiciones reales de operación.

5.1.7.8 Plan de respuesta a riesgos

En concordancia con las buenas prácticas establecidas por el *PMBOK Guide* (Project Management Institute, 2017, p. 413), el plan de respuesta a riesgos define acciones concretas para reducir la probabilidad de ocurrencia o mitigar el impacto de cada riesgo identificado. Este plan es fundamental para fortalecer la resiliencia técnica del flujo automatizado en entornos productivos. A continuación en la Tabla 34, se expone el plan de respuesta a riesgos.

Tabla 34. Plan de respuesta a riesgos

Plan de respuesta a riesgos	
ID – Riesgo	**Estrategia de respuesta**
RT-01	Migrar las funciones críticas que dependen de bibliotecas sensibles a versiones de Python oficialmente soportadas por AWS Lambda (por ejemplo, 3.11 o inferiores). Alternativamente, utilizar contenedores personalizados en AWS Lambda que permitan un entorno controlado con todas las dependencias requeridas.
RT-02	Empaquetar todas las bibliotecas críticas junto con la función Lambda utilizando entornos de desarrollo replicables (Docker). Validar que cada paquete esté precompilado para la arquitectura x86_64 compatible con Lambda antes del despliegue.
RO-01	Mantener y mejorar la función Lambda de transformación que valida el esquema estructural. Incorporar reglas de negocio adicionales, pruebas unitarias sobre estructuras de ejemplo y rechazo automático de archivos corruptos o incompletos.
RO-02	Implementar una capa de control de versiones en la zona Gold, evitando sobrescritura accidental de archivos. Además, generar registros de auditoría (*audit logs*) que permitan rastrear el origen de cada consolidación y verificar su completitud.
RS-01	Aplicar políticas de *principle of least privilege* para cada función Lambda. Auditar y revisar periódicamente los permisos IAM asignados. Establecer controles de acceso centralizados mediante roles predefinidos y utilizar AWS Secrets Manager para credenciales sensibles.
RA-01	Diseñar el flujo automatizado con tolerancia a fallos y reintentos (*Retry* y *Catch* en *Step Functions*). Establecer mecanismos de alerta temprana mediante Amazon CloudWatch y realizar pruebas de resiliencia en distintos puntos del pipeline.

RE-01	Diseñar estrategias de escalabilidad horizontal para las funciones de extracción y transformación (por ejemplo, dividir grandes archivos en bloques procesables). Integrar mecanismos de control de concurrencia y segmentación por fuente para evitar cuellos de botella en ambientes productivos.

Nota. Elaboración propia (2025)

5.1.8 Análisis costo-beneficio

El análisis costo-beneficio representa una herramienta fundamental para evaluar la viabilidad económica de una propuesta de solución tecnológica en el entorno organizacional. Su objetivo principal es contrastar los costos asociados a la implementación del proyecto con los beneficios tangibles e intangibles que este generaría, permitiendo así justificar su ejecución con base en criterios financieros y estratégicos.

En este Trabajo Final de Graduación, se propone la automatización del proceso de carga de datos en la plataforma de Gestión de Datos Maestros (MDM) de la organización, actualmente caracterizado por una alta dependencia de actividades manuales, dispersión de fuentes de datos, validaciones no sistematizadas y ausencia de mecanismos integrados de monitoreo. Esta situación ha sido ampliamente documentada en la **Fase 1** del análisis de resultados, donde se identificaron ineficiencias operativas que incrementan el riesgo de errores, prolongan los tiempos de entrega y dificultan la escalabilidad del proceso.

Aunque el análisis se construye exclusivamente sobre el entorno organizacional real, sin considerar las limitaciones propias del prototipo, debe destacarse que el desarrollo funcional realizado en Amazon Web Services constituye una base técnica valiosa desde la cual iniciar la futura implementación. Este prototipo evidencia la viabilidad técnica del flujo automatizado propuesto y ofrece un marco inicial estructurado sobre el cual escalar la solución hacia ambientes productivos. De esta forma, se busca estimar de manera realista el impacto económico de adoptar la solución automatizada, considerando tanto los costos actuales del proceso manual como los ahorros proyectados derivados de su transformación digital. Asimismo, se incorporan indicadores financieros relevantes que permiten evidenciar la conveniencia y sostenibilidad de la implementación propuesta.

5.1.8.1 Costos laborales directos del proceso actual

Dentro del análisis de costos del proceso actual, uno de los componentes más representativos corresponde a los costos laborales directos estimados, es decir, aquellos asociados al tiempo que los colaboradores del equipo *Data Operations* deben dedicar a la ejecución de las distintas etapas del proceso de carga de datos. Esta estimación no implica que los analistas estén dedicados exclusivamente al proceso, sino que se calcula con base en un volumen de trabajo equivalente a dos jornadas completas. Se utiliza el salario mínimo mensual como unidad referencial para establecer un parámetro base en el análisis costo-beneficio.

De acuerdo con la **Lista de Salarios Mínimos del Ministerio de Trabajo y Seguridad Social para el año 2025 (ver Anexo III)**, el salario mínimo mensual correspondiente a un profesional con título universitario de licenciatura es de ¢784.139,53 colones. Este perfil se ajusta

al nivel académico del personal actualmente asignado al proceso, compuesto por dos analistas de datos a tiempo completo (jornada laboral de 40 horas semanales, equivalente a 8 horas por día hábil). La Tabla 35 resume el cálculo mensual de los costos laborales directos asociados al proceso actual.

Tabla 35. Cálculo mensual de los costos laborales directos asociados al proceso manual

Cálculo mensual de los costos laborales directos asociados al proceso manual.			
Concepto	**Monto por persona**	**Cantidad de personas**	**Total mensual**
Salario base mensual	¢784.139,53	2	¢1.568.279,06

Nota. Elaboración propia (2025)

Este valor representa el piso mínimo de inversión mensual que la organización debe asumir para ejecutar el proceso manualmente, sin incluir factores como horas extraordinarias, cargas patronales, ni los efectos económicos de posibles errores, retrabajos o ineficiencias asociadas a la ausencia de automatización.

En Costa Rica, los empleadores deben asumir un conjunto de contribuciones sociales obligatorias conocidas como cargas patronales, las cuales representan un 26,67% adicional sobre el salario base mensual de cada trabajador (BGA, 2024). Estas cargas incluyen aportes a la CCSS, INA, FODESAF, IMAS, Banco Popular, INS, y fondos de pensión según lo dispuesto por la Ley de Protección al Trabajador. En la Tabla 36, se resume lo explicado anteriormente.

Tabla 36. Costo laboral mensual incluyendo cargas patronales

Cálculo del costo por analista	
Concepto	**Monto (₡)**
Salario base mensual	784.139,53
Cargas patronales (26,67%)	209.174,31
Costo total mensual	993.313,84
Cálculo total para dos analistas	
Concepto	**Monto (₡)**
Costo total por analista	993.313,84
Costo para dos analistas	1.986.627,68

Nota. Elaboración propia (2025)

De acuerdo a lo expuesto en la Tabla 36, el proceso manual de carga de datos actualmente requiere de dos analistas trabajando tiempo completo, lo que representa un costo laboral mensual de ₡1.986.627,68 para la organización, considerando únicamente salarios base y cargas patronales. Este monto no incluye otros costos asociados como:

- Retrabajos por errores humanos
- Tiempos muertos por validaciones manuales
- Supervisión adicional del proceso

- Costos indirectos por demoras en la publicación de datos

Estos aspectos, aunque no cuantificados directamente en esta sección, elevan aún más el impacto financiero del modelo manual y refuerzan la necesidad de transitar hacia una solución automatizada.

5.1.8.2 Costo total por ciclo del proceso actual

Según el diagnóstico realizado en la Fase 1, se desglosa lo siguiente:

- **Duración total por ciclo**: 3,6 días laborables (29 horas laborales en total por persona).
- **Tiempo dedicado a validación y corrección manual**: 20 horas por persona.
- **Descarga de archivos**: 3 horas por ciclo.
- **Procesamiento y consolidación**: 4 horas por analista, con intervención manual en su totalidad.
- **Carga en la plataforma de Gestión de Datos Maestros**: Estimada en 2 horas por analista.

Dado que los 2 analistas participan activamente durante todo el proceso, podemos suponer una distribución estimada de horas por etapa del proceso, tal como se presenta en la Tabla 37:

Tabla 37. Tiempo estimado por analista en cada etapa del proceso actual

Tiempo estimado por analista en cada etapa del proceso actual		
Etapa	**Tiempo estimado por analista**	**Total horas (2 analistas)**
Descarga de archivos	3 horas	6 horas
Validación y corrección manual	20 horas	40 horas
Procesamiento y consolidación	4 horas	8 horas
Carga en la plataforma MDM	2 horas	4 horas
Total estimado por ciclo	29 horas	58 horas

Nota. Elaboración propia (2025)

Cálculo del costo por hora:

Como ya fue determinado, el costo mensual por analista con cargas sociales incluidas es de ₡993.313,84. Bajo una jornada laboral de 22 días al mes, 8 horas por día, se tiene:

- **Total horas trabajadas por mes**: 176 h
- **Costo por hora** ≈ ₡993.313,84 ÷ 176 ≈ ₡5.642,67 por hora

De acuerdo con lo expuesto anteriormente, en la Tabla 38 se desglosa el costo por ciclo del proceso actual de carga de datos.

Tabla 38. Costo por ciclo del proceso actual de carga de datos

Costo por ciclo del proceso actual de carga de datos	
Detalle	**Valor**
Total horas por ciclo (2 personas)	58 horas
Costo por hora	₡5.642,67
Costo total por ciclo	₡327.275,06
Ciclos estimados por mes	1
Costo mensual total	₡327.275,06

Nota. Elaboración propia (2025)

A pesar de tratarse de un proceso que se ejecuta únicamente una vez al mes, el costo operativo directo por cada ciclo manual de carga asciende a más de ₡327.000 mensuales, solo en tiempo invertido por el personal analista. Este monto refuerza la importancia de considerar alternativas automatizadas, como la propuesta planteada en el prototipo, que ofrece una base técnica sólida sobre la cual escalar una solución más eficiente y sostenible a nivel organizacional.

5.1.8.3 Estimación de costos del nuevo proceso automatizado

La propuesta de automatización contempla un plan de implementación con una duración máxima de dos meses, durante los cuales se llevarán a cabo las actividades de configuración, desarrollo, pruebas y despliegue de todos los componentes que conforman el flujo automatizado de carga de datos. Dado que los analistas actualmente responsables del proceso manual poseen conocimiento funcional y técnico sobre las estructuras de datos, se estima que ellos mismos tienen la capacidad de liderar el desarrollo de la automatización. Por lo tanto, para efectos de estimación de costos, se considera que ambos analistas dedicarán el 100% de su jornada laboral durante dos meses a este plan de implementación.

Con base en los cálculos anteriores, el costo mensual por analista, incluyendo cargas sociales (26,67%), asciende a ₡993.313,84. El costo total del plan, considerando ambos recursos durante dos meses, se resume a continuación en la Tabla 39:

Tabla 39. Costo total del plan de implementación

Costo total del plan de implementación		
Concepto	**Valor por unidad**	**Total estimado**
Salario mensual por analista	₡993.313,84	-
Costo por 2 meses por 1 analista	₡993.313,84 × 2	₡1.986.627,68
Costo total para 2 analistas	₡1.986.627,68 × 2	₡3.973.255,36

Nota. Elaboración propia (2025)

Una vez implementado el flujo automatizado en un entorno organizacional real, el proyecto incurre en costos asociados al uso de servicios en la nube provistos por Amazon Web Services (AWS). Durante la reunión sostenida con el *Senior Data Engineer (*ver **Apendice T***)*, se estimó que una ejecución completa del flujo automatizado incluyendo extracción, transformación,

consolidación y carga en base de datos generaría un costo aproximado de entre USD 100 y 150 por ciclo completo. Este rango consideraría el consumo integrado de los servicios Amazon S3, AWS Lambda, AWS Glue, AWS Step Functions, Amazon Aurora y Amazon CloudWatch.

Con base en esta estimación, se define un valor promedio mensual conservador de USD 125, considerando que el flujo se ejecuta una vez por mes (el ciclo completo de carga de datos). Se resumen los valores proyectados de la siguiente forma en la Tabla 40:

Tabla 40. Estimación de costo promedio mensual y anual del ciclo de carga de datos

Costo promedio mensual y anual del ciclo de carga de datos		
Concepto	**Frecuencia**	**Costo estimado**
Ejecución completa del flujo automatizado	1 vez al mes	$125
Costo mensual estimado	-	$125
Costo anual proyectado	12 ciclos por año	$1.500

Nota. Elaboración propia (2025)

Este monto representa el costo base de operación mensual del sistema automatizado en AWS, y sustituye por completo el tiempo operativo invertido por los analistas en el proceso manual, con una relación costo-beneficio considerablemente favorable. Además, esta infraestructura es escalable, por lo que la adición de nuevas fuentes de datos (como *Ratings, News Edge* o *Credit Risk*) se realiza sin necesidad de rediseñar el modelo técnico actual.

Una vez desplegado el flujo automatizado en el entorno organizacional, será necesario establecer un esquema mínimo de mantenimiento operativo, supervisión de ejecución y soporte técnico, con el objetivo de asegurar la continuidad del proceso, mitigar posibles fallos y permitir ajustes menores ante cambios en las fuentes de datos o reglas de validación.

Aunque el diseño propuesto está orientado a ser altamente autónomo y requiere intervención humana mínima, el entorno organizacional demanda prácticas estándar de gobernanza técnica, las cuales incluyen:

- Revisión periódica de ejecuciones
- Documentación de cambios o eventos técnicos.
- Soporte ante fallos inesperados
- Revisión y ajuste de permisos

Con base en estos requerimientos, se estima que un analista deberá dedicar aproximadamente 6 a 8 horas por mes a tareas de mantenimiento y monitoreo del flujo. Esta estimación contempla únicamente tareas rutinarias de soporte.

Dado que el recurso encargado de este soporte corresponde al mismo perfil profesional utilizado en el resto del análisis (licenciatura universitaria, jornada de 40 horas), se mantiene el costo por hora ya calculado: ₡5.642,67 colones por hora. A continuación, en la Tabla 41 se resumen los costos de soporte estimados.

Tabla 41. Estimación de costos de soporte

Estimación de costos de soporte	
Concepto	**Valor estimado**
Horas promedio mensuales de soporte	7 horas
Costo por hora del analista	₡5.642,67
Costo mensual de mantenimiento	7 × ₡5.642,67 = ₡39.498,69
Costo anual estimado	₡39.498,69 × 12 = ₡473.984,28

Nota. Elaboración propia (2025)

Este monto representa una fracción menor respecto al costo actual del proceso manual, y garantiza la sostenibilidad técnica del sistema automatizado en un entorno productivo. A diferencia del modelo anterior, en el que se requería una dedicación mensual de al menos 58 horas operativas, el flujo automatizado reduce dicha carga en más de un 85%, trasladando el esfuerzo hacia una supervisión más estratégica y puntual.

Para asegurar una adopción efectiva del nuevo flujo automatizado, se contempla una capacitación inicial dirigida a los analistas responsables del proceso. Aunque la propuesta de solución fue diseñada para operar de manera autónoma, se considera indispensable que el equipo posea el conocimiento del entorno AWS utilizado.

La capacitación podría ejecutarse mediante dos sesiones técnicas internas de 2 horas cada una, impartidas por el *Senior Data Engineer*. Alternativamente, se valora la posibilidad de brindar acceso a una certificación breve en fundamentos de AWS (Cloud Practitioner), la cual otorgaría mayor autonomía técnica a los usuarios.

Para efectos de este análisis, se considera un escenario conservador de capacitación interna con los siguientes supuestos:

- **Duración total:** 4 horas de capacitación (2 sesiones × 2 h)
- **Costo hora del capacitador interno:** ₡10.000 (referencia del mercado interno)
- **Número de analistas capacitados:** 2

A continuación, en la Tabla 42 se resumen los costos de capacitación.

Tabla 42. Estimación de costos de capacitación

Estimación de costos de capacitación		
Concepto	**Cálculo**	**Monto estimado**
Duración total de las sesiones	4 horas	-
Costo por hora de capacitación interna	₡10.000	-
Costo total de la capacitación	4 h × ₡10.000 = ₡40.000	₡40.000
Capacitación por 2 analistas	₡40.000 × 2	₡80.000 (único pago)

Nota. Elaboración propia (2025)

5.1.8.3.1 Cálculo del Retorno de la Inversión (ROI)

El retorno sobre la inversión (ROI) es una herramienta financiera esencial que permite a las organizaciones evaluar la rentabilidad de sus iniciativas y tomar decisiones estratégicas fundamentadas. Esta métrica cuantifica la relación entre el beneficio neto obtenido y el costo inicial de una inversión, generalmente expresada como un porcentaje, lo cual facilita la comparación entre distintas alternativas de inversión y la identificación de aquellas que ofrecen mayores beneficios relativos. Su aplicación en la toma de decisiones abarca diversas áreas, como el desarrollo de productos, mejora de procesos, la expansión de mercados, las fusiones, adquisiciones, y la mejora operativa. Además, el monitoreo continuo del ROI permite ajustar las estrategias frente a cambios del entorno, promoviendo una gestión proactiva de riesgos y asegurando la sostenibilidad del crecimiento empresarial. Comprender y aplicar esta métrica no solo contribuye a evaluar la viabilidad financiera de proyectos específicos, sino que también fortalece la planificación estratégica organizacional (Southern Illinois University Carbondale, 2024).

El principal beneficio cuantificable se deriva de la eliminación de tareas manuales que, en el modelo actual, consumen un total de 58 horas (ver sección **5.1.7.2**) por mes entre dos analistas. Al automatizar completamente el flujo, este esfuerzo operativo se reduce a tareas mínimas de supervisión, lo que representa un ahorro económico directo en tiempo hombre.

- **Horas ahorradas por mes**: 58 horas
- **Horas ahorradas por año**: 58 × 12 = 696 horas
- **Costo por hora de los analistas**: ₡5.642,67
- **Beneficio económico anual estimado**:

$$696\ horas \times ₡5.642{,}67 = ₡3.927.298{,}32$$

En la Tabla 43, se muestra el costo total de implementación y operación del sistema automatizado en su primer año.

Tabla 43. Estimación de costos del primer año

Costos estimados del primer año	
Concepto	**Monto (₡)**
Implementación (2 meses)	₡3.973.255,36
Servicios AWS (USD 1.500 × ₡540)	₡810.000,00
Mantenimiento y soporte anual	₡473.984,28
Capacitación inicial (único pago)	₡80.000,00
Costo total año 1	**₡5.337.239,64**

Nota. Elaboración propia (2025)

El Retorno de Inversión (ROI) se calcula con la fórmula:

$$ROI = (\frac{Beneficio\ anual - Costo\ total}{Costo\ total}) \times 100$$

Sustituyendo los valores:

$$ROI = \left(\frac{₡3.927.298{,}32 - ₡5.337.239{,}64}{₡5.337.239{,}64}\right) \times 100 = -26{,}42\%$$

Durante el primer año, el proyecto presenta un ROI negativo de -26,42%, lo cual es esperado debido a que en esta etapa se incurre en los costos de implementación inicial (dos meses completos de dedicación exclusiva del personal), así como en los pagos únicos por capacitación. Esto implica que aún no se alcanza el punto de equilibrio financiero en el primer año.

No obstante, a partir del segundo año, el modelo cambia drásticamente. Al eliminar los costos iniciales y mantener únicamente los costos recurrentes (servicios en AWS y soporte técnico), los beneficios anuales superan ampliamente los egresos, lo que transforma el ROI en positivo y acelera la recuperación de la inversión. Este comportamiento es característico de proyectos tecnológicos de automatización; presentan una inversión inicial moderada, pero su retorno se incrementa progresivamente conforme se estabiliza la operación y se escala el proceso.

A partir del segundo año, el proyecto ya no incurre en los costos únicos de implementación y capacitación. Por tanto, los costos se reducen únicamente a los elementos operativos recurrentes: servicios en la nube y mantenimiento técnico mensual. A continuación en la Tabla 44, se resumen los costos proyectados para el segundo año.

Tabla 44. Proyección de costos del segundo año

Estimación de costos del segundo año	
Concepto	**Monto (₡)**
Servicios AWS anuales	₡810.000,00
Mantenimiento y soporte anual	₡473.984,28
Costo total año 2	**₡1.283.984,28**

Nota. Elaboración propia (2025)

Cálculo del ROI del segundo año:

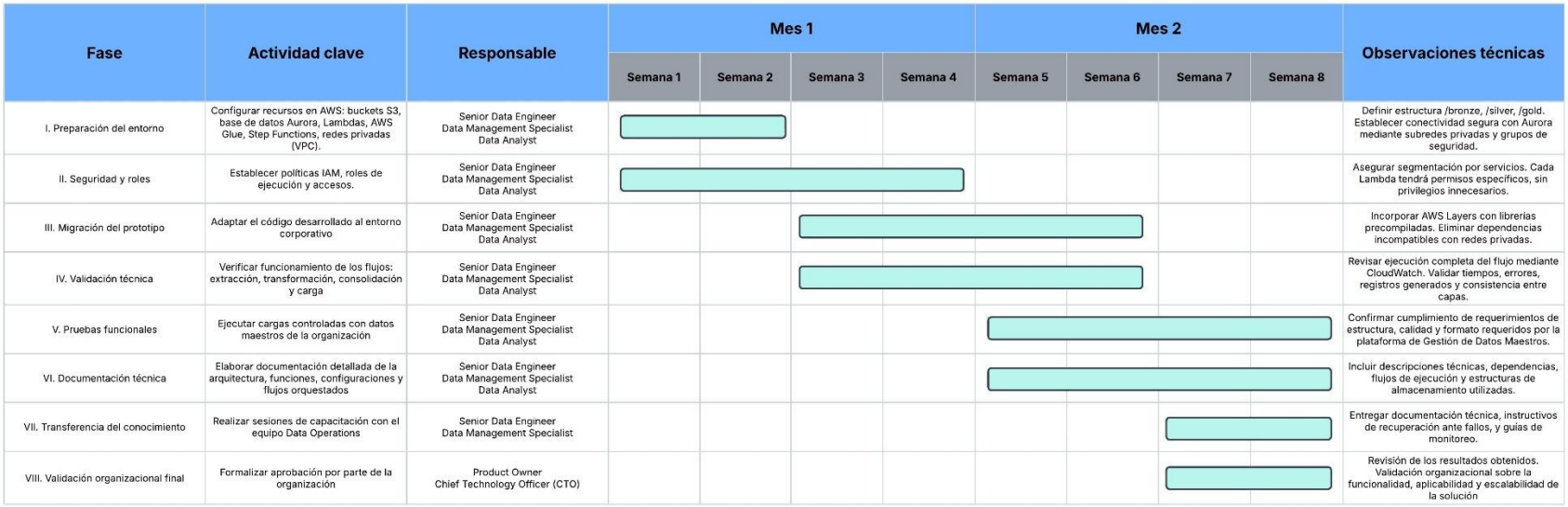
$$ROI = \left(\frac{₡3.927.298{,}32 - ₡1.283.984{,}28}{₡1.283.984{,}28}\right) \times 100 = 205{,}87\%$$

A partir del segundo año, la solución automatizada muestra un Retorno de Inversión positivo del 205,87%, lo cual indica que por cada colón invertido se recuperan más del triple en beneficios operativos. Este resultado confirma que, tras superada la etapa inicial de implementación, la solución automatizada no solo se vuelve autosostenible, sino altamente rentable

5.1.9 Hoja de ruta de implementación de la propuesta de solución

La siguiente hoja de ruta visualizada en la Figura 38, proyecta la implementación técnica del prototipo automatizado del proceso de carga de datos hacia la plataforma de Gestión de Datos Maestros (MDM) de la organización. El plan contempla un periodo de dos meses, conforme a la estimación del *Senior Data Engineer* (ver **Apéndice W**) del equipo *Data Operations*. Las actividades se alinean con los flujos validados previamente y responden a las condiciones reales del entorno corporativo.

Figura 38. Diagrama de Gantt - Hoja de ruta de implementación de la solución



Fase	Actividad clave	Responsable	Mes 1				Mes 2				Observaciones técnicas
			Semana 1	Semana 2	Semana 3	Semana 4	Semana 5	Semana 6	Semana 7	Semana 8	
I. Preparación del entorno	Configurar recursos en AWS: buckets S3, base de datos Aurora, Lambdas, AWS Glue, Step Functions, redes privadas (VPC).	Senior Data Engineer Data Management Specialist Data Analyst									Definir estructura /bronze, /silver, /gold. Establecer conectividad segura con Aurora mediante subredes privadas y grupos de seguridad.
II. Seguridad y roles	Establecer políticas IAM, roles de ejecución y accesos.	Senior Data Engineer Data Management Specialist Data Analyst									Asegurar segmentación por servicios. Cada Lambda tendrá permisos específicos, sin privilegios innecesarios.
III. Migración del prototipo	Adaptar el código desarrollado al entorno corporativo	Senior Data Engineer Data Management Specialist Data Analyst									Incorporar AWS Layers con librerías precompiladas. Eliminar dependencias incompatibles con redes privadas.
IV. Validación técnica	Verificar funcionamiento de los flujos: extracción, transformación, consolidación y carga	Senior Data Engineer Data Management Specialist Data Analyst									Revisar ejecución completa del flujo mediante CloudWatch. Validar tiempos, errores, registros generados y consistencia entre capas.
V. Pruebas funcionales	Ejecutar cargas controladas con datos maestros de la organización	Senior Data Engineer Data Management Specialist Data Analyst									Confirmar cumplimiento de requerimientos de estructura, calidad y formato requeridos por la plataforma de Gestión de Datos Maestros.
VI. Documentación técnica	Elaborar documentación detallada de la arquitectura, funciones, configuraciones y flujos orquestados	Senior Data Engineer Data Management Specialist Data Analyst									Incluir descripciones técnicas, dependencias, flujos de ejecución y estructuras de almacenamiento utilizadas.
VII. Transferencia del conocimiento	Realizar sesiones de capacitación con el equipo Data Operations	Senior Data Engineer Data Management Specialist									Entregar documentación técnica, instructivos de recuperación ante fallos, y guías de monitoreo.
VIII. Validación organizacional final	Formalizar aprobación por parte de la organización	Product Owner Chief Technology Officer (CTO)									Revisión de los resultados obtenidos. Validación organizacional sobre la funcionalidad, aplicabilidad y escalabilidad de la solución

Nota. Elaboración propia (2025)

5.2 Fase 4. Evaluación del prototipo de la solución automatizada

La presente fase tiene como finalidad determinar la efectividad del prototipo desarrollado en la etapa anterior, en relación con su desempeño funcional en la carga de datos hacia la plataforma de Gestión de Datos Maestros. Esta evaluación se fundamenta en el cumplimiento del objetivo específico #4, que establece como propósito evaluar la efectividad del prototipo de la solución automatizada en términos de precisión, consistencia y reducción de tareas manuales, mediante el uso de métricas de desempeño que permitan determinar su impacto en la eficiencia del proceso.

El análisis se estructuró en torno a las variables definidas metodológicamente. Por un lado, se consideró la variable independiente **VA-08: Prototipo de solución automatizada**, entendida como una herramienta tecnológica desarrollada para automatizar el proceso de carga de datos, con el fin de reducir la intervención manual y mejorar la consistencia de los registros. Por otro lado, se abordó la variable dependiente **VA-09: Efectividad del prototipo en la carga de datos**, la cual se define como el grado en que la solución propuesta logra mejorar la precisión, consistencia y eficiencia del proceso.

Para evaluar estas variables, se ejecutaron pruebas funcionales controladas que permitieron verificar el comportamiento técnico del prototipo frente a escenarios simulados. Estas pruebas incluyeron la recopilación y análisis de métricas clave, tales como el porcentaje de registros correctamente cargados, el nivel de coherencia en los datos y la cantidad de tareas manuales eliminadas. El desempeño del prototipo fue contrastado con el proceso manual documentado en fases anteriores, con el fin de identificar el impacto real de la automatización en términos operativos. Finalmente, los resultados fueron validados por el equipo de *Data Operations*, con el objetivo de confirmar que la solución propuesta responde adecuadamente a las necesidades funcionales y técnicas del proceso.

5.2.1 Criterios de evaluación e indicadores

La evaluación del prototipo automatizado se estructuró con base en tres criterios fundamentales: precisión en la carga de datos, consistencia en los registros almacenados y reducción de tareas manuales durante el proceso. Estos criterios permitieron traducir operativamente los elementos definidos en el marco metodológico, facilitando la recolección y análisis de evidencia empírica que sustenta el cumplimiento del objetivo específico #4.

El primer criterio, precisión, se refiere a la correcta inserción de los datos en la plataforma Aurora, sin alteraciones o pérdidas respecto a los registros originales provenientes de la fuente ESG. Este criterio se midió a través del porcentaje de registros que fueron cargados exitosamente, considerando como válidos aquellos que cumplieron íntegramente con las condiciones estructurales y semánticas requeridas por la plataforma.

El segundo criterio, consistencia, se vincula con la coherencia de los valores cargados en relación con las reglas de validación predefinidas. Se evaluó a través de la revisión de los campos transformados y consolidados durante la etapa intermedia (Silver) y final (Gold), verificando la

correcta aplicación de las funciones Lambda responsables de las operaciones de transformación y consolidación.

El tercer criterio, reducción de tareas manuales, se orientó a identificar el grado en que el prototipo reemplazó procesos operativos previamente realizados por el equipo de *Data Operations*. Este aspecto fue abordado mediante una comparación estructurada entre el flujo manual documentado en la fase uno y el flujo automatizado implementado en la fase tres. Se documentaron las tareas eliminadas o transformadas en pasos automáticos a través de Lambda Functions y Step Functions.

A continuación en la Tabla 45, se presenta un cuadro resumen que sintetiza los indicadores definidos para cada criterio, así como su respectiva unidad de medida y técnica de recolección.

Tabla 45. Criterios de evaluación e indicadores

Criterios de evaluación e indicadores			
Criterio de evaluación	**Indicador**	**Unidad de medida**	**Técnica de recolección**
Precisión	Porcentaje de registros correctamente cargados	Porcentaje (%)	Consultas de validación en DBeaver sobre base Aurora
Consistencia	Coherencia de los valores cargados	Coincidencias respecto a las reglas	Comparación entre dataset original y archivo consolidado
Reducción de tareas manuales	Tareas eliminadas o transformadas	Número de tareas identificadas	Análisis comparativo entre flujos *As-Is* y *To-Be*
Tiempo de ejecución del proceso	Duración total del flujo automatizado	Tiempo (segundos, minutos u horas)	Análisis comparativo entre tiempo de ejecución actual vs propuesta de solución automatizada.

Nota. Elaboración propia (2025)

Esta estructura de indicadores facilita una evaluación objetiva, fundamentada sobre la funcionalidad y desempeño del prototipo desarrollado.

5.2.2 Método de medición y entorno de prueba

La evaluación del prototipo se llevó a cabo en un entorno de simulación configurado en el ecosistema Amazon Web Services (AWS), utilizando exclusivamente los recursos disponibles en el plan Free Tier. El flujo de trabajo automatizado fue probado empleando la fuente de datos ESG, seleccionada por su estructura representativa y su valor como insumo de validación funcional. La arquitectura técnica desplegada incluyó un bucket S3 estructurado en zonas Bronze, Silver y Gold,

una base de datos Amazon Aurora PostgreSQL Serverless v2, funciones Lambda para cada etapa del proceso, y dos Step Function que coordinó la secuencia de ejecución.

Durante la ejecución del prototipo, se recopilaron datos que permitieron evaluar la precisión y consistencia de la carga. La validación se realizó mediante consultas directas a la base de datos Aurora utilizando la herramienta DBeaver, dado que la plataforma MDM no se integró directamente en esta fase, por limitaciones técnicas asociadas al entorno de pruebas. Esta herramienta facilitó la revisión de los registros insertados, permitiendo confirmar la fidelidad de los datos respecto a la fuente original, así como la correcta aplicación de las reglas de transformación.

Para la evaluación de la consistencia, se analizaron los archivos generados en las zonas Silver y Gold del bucket S3. La inspección se enfocó en verificar que las transformaciones aplicadas por la función *lambda_transform_esg_data* y la consolidación realizada por *lambda_consolidate_esg_data* cumplieran con los requerimientos establecidos en la lógica de negocio. Esta revisión se apoyó en la comparación directa entre los archivos de entrada y salida, garantizando la integridad del proceso.

Finalmente, la medición de la reducción de tareas manuales se fundamentó en un análisis comparativo entre el flujo operativo actual (*As-Is*) y el proceso automatizado (*To-Be*). Se identificaron las tareas originalmente ejecutadas por el equipo de *Data Operations* y se documentaron aquellas que fueron reemplazadas por funciones automatizadas, con evidencia generada a partir de logs de ejecución y archivos procesados. Esta evaluación permitió establecer con claridad los impactos logrados en términos de eficiencia operativa, sin necesidad de recurrir a herramientas específicas de medición temporal.

5.2.3 Resultados de pruebas funcionales

Las pruebas funcionales del prototipo se realizaron en condiciones controladas, utilizando como insumo un archivo JSON estructurado con datos de la fuente ESG. Estas pruebas permitieron evaluar el comportamiento del flujo automatizado desde la carga inicial en la zona Bronze hasta la inserción final de registros en la base de datos Aurora PostgreSQL.

5.2.3.1 Descripción del archivo fuente

El archivo *esg_data_sample.json* (ver **Anexo IV**) contiene una colección estructurada de registros vinculados al desempeño ambiental, social y de gobernanza (ESG) de diversas empresas. Cada elemento dentro del archivo representa un reporte periódico generado por una organización determinada, identificado mediante los campos *company_name* y *reporting_date*. Este conjunto de datos fue utilizado como insumo principal para probar el prototipo desarrollado en AWS, simulando el escenario de extracción, transformación, consolidación y carga final en la base de datos Aurora PostgreSQL. La estructura de cada registro incluye los siguientes atributos clave.

- **Indicadores ambientales**:
    - **carbon_emissions_tons**: Emisiones de carbono en toneladas métricas.
    - **energy_consumption_mwh**: Consumo energético en megavatios hora.

  - **renewable_energy_percentage**: Porcentaje del consumo energético proveniente de fuentes renovables.
  - **water_usage_m3**: Uso de agua en metros cúbicos.
  - **waste_generated_tons**: Generación de residuos en toneladas.
- **Indicadores sociales:**
  - **employee_satisfaction_score**: Nivel de satisfacción de los empleados, medido en una escala numérica.
  - **diversity_index**: Índice de diversidad en la fuerza laboral.
  - **workplace_accidents**: Número de accidentes laborales registrados.
  - **training_hours_per_employee**: Horas de capacitación por empleado.
  - **community_investment_usd**: Inversión en iniciativas comunitarias, expresada en dólares estadounidenses.
- **Indicadores de gobernanza**:
  - **board_diversity_percentage**: Porcentaje de diversidad en la junta directiva.
  - **executive_pay_ratio**: Relación entre la remuneración ejecutiva y la media organizacional.
  - **whistleblower_policy**: Existencia o no de una política de denuncias internas.
  - **corruption_incidents**: Cantidad de incidentes de corrupción reportados.
  - **audit_compliance_score**: Puntaje asignado a la empresa en auditorías de cumplimiento.
- **Control de integridad**:
  - **record_id**: Identificador único de cada registro.

El conjunto incluye datos de múltiples compañías como *GreenCorp*, *EcoGlobal*, *SustainTech*, *CleanEnergyCo* y *FuturePlanet*, lo que permitió simular un entorno de datos heterogéneo y verificar la capacidad del prototipo para manejar entradas variadas en términos de volumen, consistencia y formato. Esta diversidad de registros resultó fundamental para validar el comportamiento del flujo automatizado en escenarios representativos del entorno real de la plataforma de Gestión de Datos Maestros.

Cabe destacar que se definió como insumo principal un archivo estructurado en formato JSON (ver **Anexo VII**). Esta elección respondió a criterios técnicos, funcionales y metodológicos vinculados con la naturaleza del flujo automatizado, las capacidades del entorno de ejecución en AWS y las exigencias del prototipo en términos de integridad estructural de los datos.

En primer lugar, el formato JSON (*JavaScript Object Notation*) ofrece una representación estructurada y jerárquica de los datos, lo que resulta especialmente útil en procesos de transformación y consolidación que requieren mantener la integridad semántica entre campos relacionados. Cada objeto JSON encapsula múltiples atributos con precisión, sin necesidad de establecer delimitadores externos como ocurre en el caso de los archivos CSV. Esta característica facilita una manipulación más directa de los datos mediante lenguajes de programación compatibles con AWS Lambda, como Python, reduciendo la probabilidad de errores derivados de estructuras planas o ambigüedades en los encabezados.

En segundo lugar, el ecosistema AWS incorpora soporte nativo para el tratamiento de archivos JSON en diversos servicios, incluyendo S3, Lambda y Glue. En el contexto del Free Tier utilizado para la implementación del prototipo, esta compatibilidad representa una ventaja significativa, ya que permite ejecutar flujos de procesamiento sin requerir transformaciones intermedias del formato. La carga de archivos JSON directamente en las zonas Bronze y su posterior manipulación en Silver y Gold se realizó de manera fluida, con total compatibilidad con los motores de *parsing* utilizados por las funciones Lambda.

Además, la naturaleza del archivo seleccionado, compuesto por registros complejos que incluyen datos numéricos, booleanos y cadenas de texto, favorece una estructura como la proporcionada por JSON, que permite definir cada campo con claridad y acceder a su contenido de forma directa, sin necesidad de interpretar tipos o realizar conversiones manuales. Este aspecto contribuyó a preservar la fidelidad de los datos originales en todo el flujo, asegurando que los valores cargados en la base de datos Aurora mantuvieran su estructura original.

Por último, desde una perspectiva metodológica, el uso de JSON reflejó mejor el tipo de insumos que el equipo de *Data Operations* gestionaría en escenarios reales, en los cuales el intercambio de información estructurada entre plataformas suele realizarse mediante APIs que utilizan este mismo formato. Incorporar JSON en el entorno de pruebas permitió replicar con mayor realismo las condiciones operativas del proceso de carga que se busca automatizar.

5.2.3.2 Prueba funcional - Etapa de extracción de datos

La primera prueba funcional realizada durante la evaluación del prototipo correspondió a la etapa de extracción de datos en la arquitectura definida. Esta función tiene como propósito recibir un objeto JSON con registros ESG y almacenarlo en la carpeta Bronze del bucket S3, iniciando así el flujo automatizado diseñado para el proceso de carga de datos en la plataforma de Gestión de Datos Maestros.

La función evaluada se denomina *lambda_ingest_esg_data* y fue desarrollada en Python, utilizando el SDK de AWS Boto3 para interactuar con el servicio S3. Su lógica consiste en validar la existencia de un evento de entrada, generar un nombre de archivo dinámico con marca de tiempo (registro_esg_YYYYMMDD_HHMMSS.json) y almacenar el contenido recibido en formato JSON dentro del bucket prototipo-mdm-kemuel, específicamente en la ruta Bronze/esg/.

El resumen técnico de la ejecución mostró una duración total de 308.04 ms, con un uso máximo de 88 MB de memoria, dentro del límite de 128 MB configurados. La función fue desplegada en su versión más reciente ($LATEST), sin presentar errores ni advertencias. Esta información validó el comportamiento esperado de la etapa de extracción, asegurando que el prototipo inició correctamente el proceso automatizado desde el almacenamiento en la zona Bronze. En la Figura 39, se visualiza el log de ejecución.

Figura 39. Log de ejecución - Etapa de extracción de datos

Nota. Tomado de Amazon Web Services (2025)

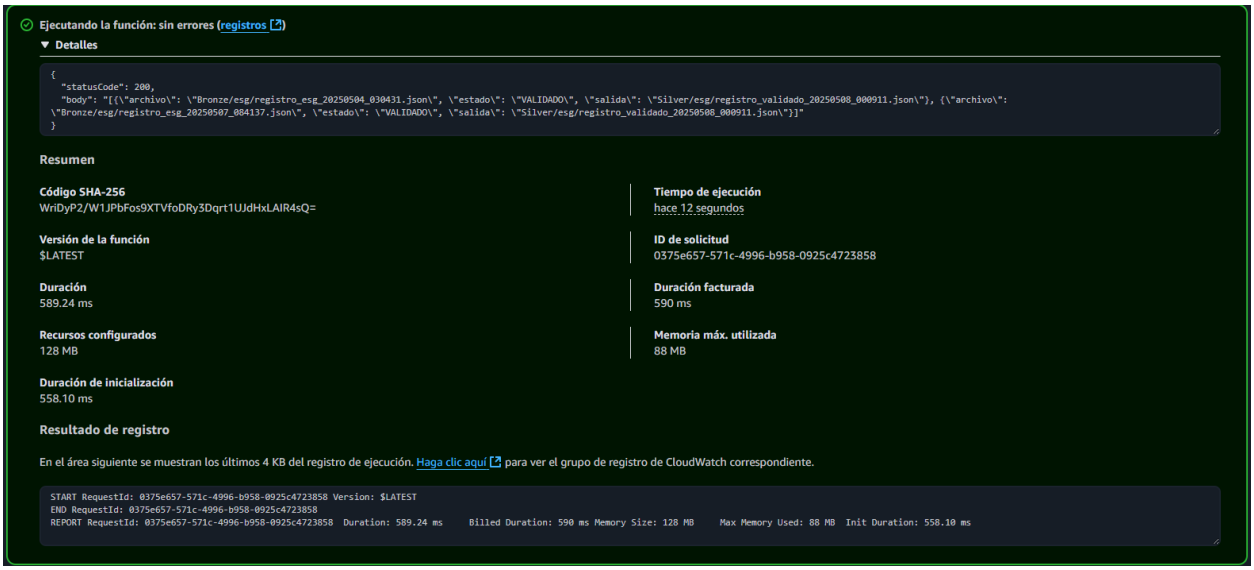
5.2.3.3 Prueba funcional - Etapa de transformación y validación de consistencia de datos

La segunda prueba funcional correspondió a la etapa de transformación y validación de datos, cuya finalidad consiste en procesar los archivos almacenados en la zona Bronze, filtrar únicamente los registros válidos y trasladarlos a la carpeta Silver del bucket S3. Esta operación representa un paso clave en el proceso de aseguramiento de la calidad antes de proceder con la consolidación y carga final.

La función Lambda utilizada en esta etapa se denominó *lambda_transform_esg_data*. Fue desarrollada en Python y diseñada para iterar sobre los archivos contenidos en la carpeta Bronze/esg/, validar cada registro según la existencia de campos obligatorios, y escribir los resultados en formato JSON en la carpeta Silver/esg/. Entre los campos validados se incluyen identificadores únicos (*record_id*), fechas (*reporting_date*), métricas ambientales, sociales y de gobernanza, asegurando así que solo los registros completos pasen al siguiente nivel del flujo.

El proceso se ejecutó sin errores, con una duración total de 589.24 ms y un uso máximo de 88 MB de memoria, valores que se mantienen dentro del umbral definido para el entorno de pruebas. La inicialización de la función tomó 558.10 ms, evidenciando una respuesta eficiente y adecuada para entornos como AWS Lambda. Esta prueba confirmó que la etapa de validación y transformación se ejecuta con precisión, filtrando adecuadamente los registros completos, trasladándolos a una capa intermedia que garantiza integridad estructural y semántica antes de la consolidación. En la Figura 40, se observa del log de ejecución.

Figura 40. Log de ejecución – Etapa de transformación y validación de datos



Nota. Tomado de Amazon Web Services (2025)

5.2.3.4 Prueba funcional – Etapa de consolidación de datos

La tercera prueba funcional correspondió a la etapa de consolidación de datos validados, cuya finalidad es trasladar los registros previamente transformados desde la carpeta Silver hacia la carpeta Gold del bucket S3. Esta operación representa el cierre de la fase de preparación de datos, dejándolos listos para su posterior carga en la base de datos Aurora PostgreSQL.

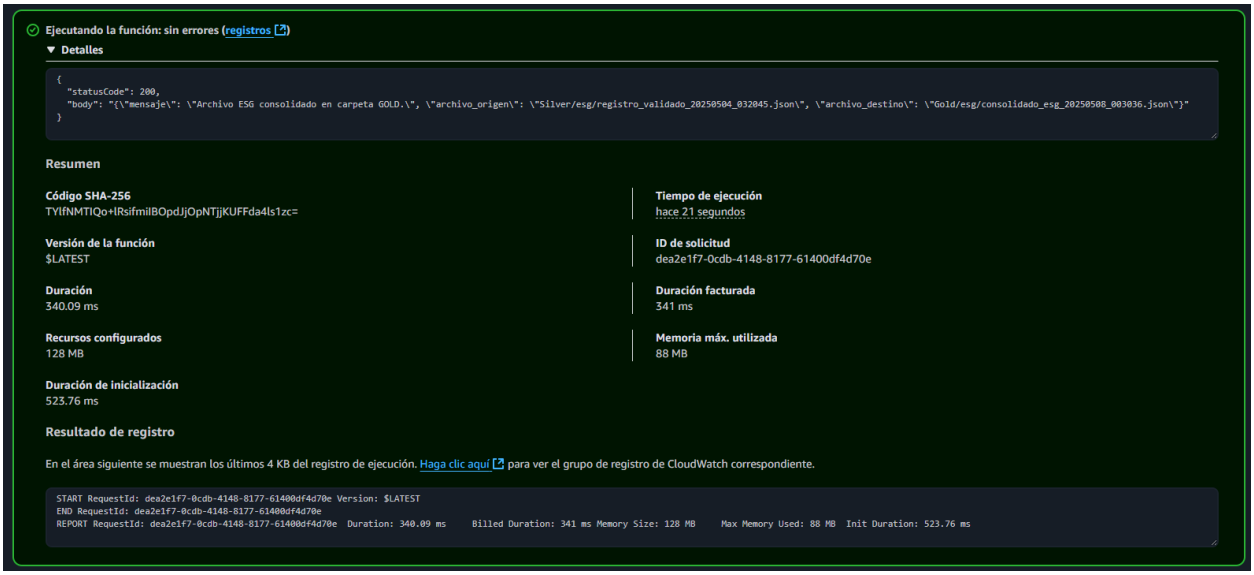
La función utilizada para esta etapa se denominó *lambda_consolidate_esg_data*. Fue diseñada para ejecutar una copia directa de un archivo específico desde la ubicación Silver/esg/ hacia la ruta de destino Gold/esg/, asignándole un nombre único con base en la marca de tiempo *(consolidado_esg_YYYYMMDD_HHMMSS.json*). Esta lógica fue implementada en Python utilizando el método *copy_object* del cliente Boto3, lo cual permitió realizar la operación sin necesidad de descargar o modificar el contenido original del archivo.

Esta decisión técnica respondió al alcance del prototipo, centrado únicamente en una fuente de datos única (ESG). Bajo estas condiciones, la consolidación no requirió fusiones y agregaciones entre múltiples archivos o estructuras heterogéneas. Sin embargo, se reconoce que en un entorno de implementación real, donde convergen las diversas fuentes *ESG, Credit Risk Data, Ratings, News Edge y Corporate Intelligence Database*, la lógica actual deberá adaptarse. La función *lambda_consolidate_esg_data* tendrá que incorporar un mecanismo que permita combinar múltiples archivos validados, armonizar estructuras de datos, aplicar reglas de priorización y garantizar una salida consolidada coherente con las necesidades de análisis e integración del negocio.

El proceso se ejecutó sin errores, con una duración total de 340.09 ms y un uso máximo de 88 MB de memoria, dentro de los parámetros establecidos para el entorno del prototipo. La

inicialización de la función tomó 523.76 ms, reflejando un comportamiento eficiente. Esta prueba confirmó que el flujo automatizado ejecuta correctamente la consolidación bajo condiciones de prueba controladas, dejando claro que, ante una ampliación del alcance funcional, esta etapa deberá evolucionar hacia una lógica más compleja y escalable. En la Figura 41, se observa del log de ejecución.

Figura 41. Log de ejecución – Etapa de consolidación de datos



Nota. Tomado de Amazon Web Services (2025)

5.2.3.5 Prueba funcional - Etapa de carga en base de datos Aurora PostgreSQL

La etapa final del flujo automatizado corresponde a la carga de los registros consolidados en la base de datos Aurora PostgreSQL, donde los datos quedan disponibles para ser consumidos por la plataforma de Gestión de Datos Maestros. Esta operación representa el cierre del flujo automatizado definido, validando que la información almacenada en la capa Gold haya sido correctamente cargada en el modelo relacional de la base de datos.

La función Lambda diseñada para esta tarea fue denominada *insert_to_aurora_v2*. Fue implementada en Python, utilizando la librería *psycopg2* para establecer conexión con la base de datos y ejecutar sentencias SQL de inserción. Esta función extrae cada campo clave del archivo consolidado, genera identificadores únicos (*record_id*) mediante la librería *uuid*, y ejecuta una inserción por cada registro individual. Los datos procesados corresponden a las métricas ESG previamente validadas y consolidadas.

El proceso mostró un desempeño eficiente, con una duración total de 15.86 ms, un uso máximo de 30 MB de memoria, y una inicialización rápida de 75.09 ms. Estos valores reflejan la liviandad del procedimiento de carga cuando los datos ya han sido estructurados y depurados adecuadamente en fases anteriores del flujo. Esta prueba concluyó satisfactoriamente el recorrido funcional del prototipo, demostrando que el sistema automatizado logra insertar registros ESG en

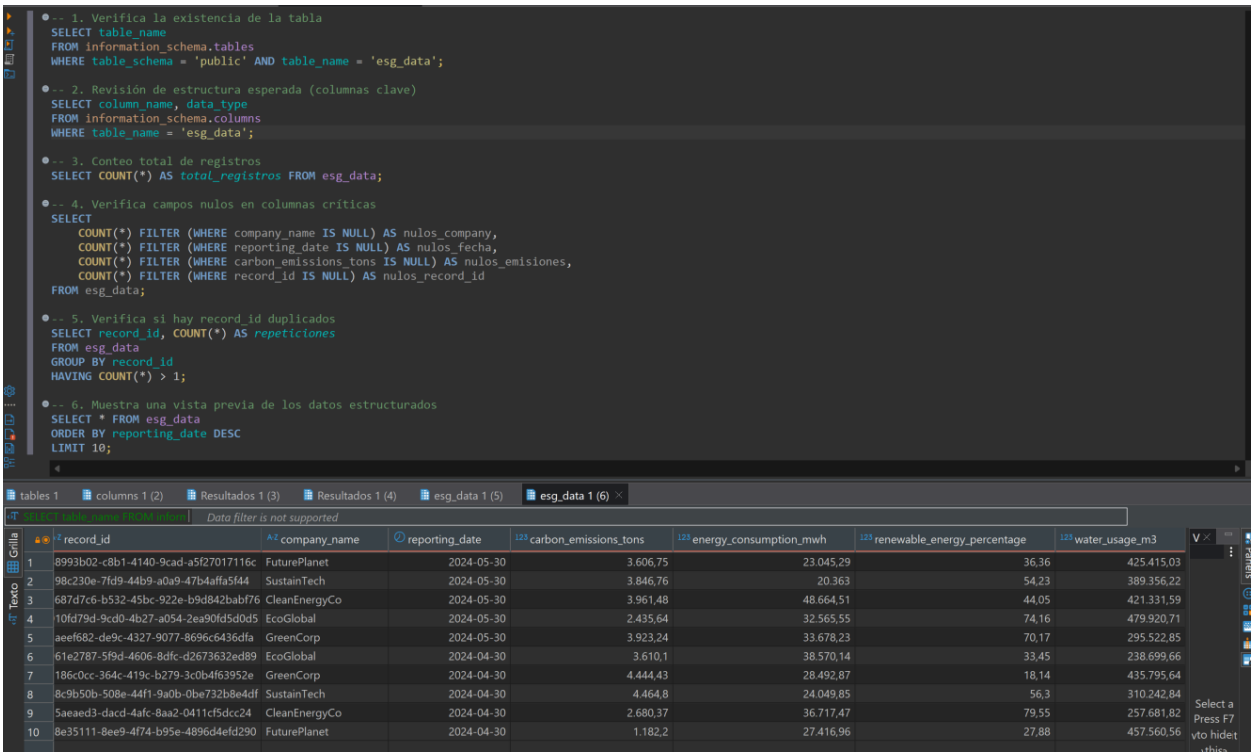
una base de datos relacional sin intervención manual, cerrando el circuito completo desde la extracción hasta la carga/inserción de datos.

Para verificar la carga exitosa, se utilizó DBeaver como herramienta cliente, ejecutando una serie de sentencias SQL que permitieron validar aspectos clave como:

- Existencia de la tabla esg_data y su estructura.
- Conteo total de registros insertados.
- Ausencia de valores nulos en campos críticos.
- No duplicidad en los identificadores record_id.
- Visualización de una muestra de registros cargados ordenados por fecha.

En la Figura 42, se visualizan los registros cargados en Aurora mediante la interfaz de DBeaver, lo que respalda la validación técnica realizada durante la prueba.

Figura 42. Datos/Registros cargados en Aurora



Nota. Elaboración propia (2025)

5.2.3.6 Prueba funcional – Orquestación del flujo automatizado

Como parte de la arquitectura automatizada diseñada en este prototipo, se incluyó el uso de AWS Step Functions para orquestar de manera secuencial y estructurada las diferentes funciones Lambda del flujo. La primera Step Function evaluada fue denominada ***SourceToLanding***, cuya responsabilidad principal consiste en coordinar la ejecución de la función

lambda_ingest_esg_data, encargada de recibir y almacenar los datos ESG en la carpeta Bronze del bucket S3.

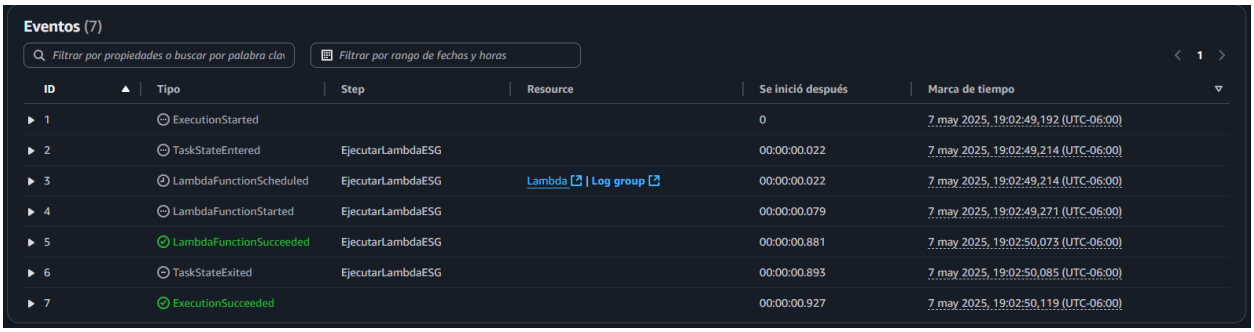
Esta Step Function fue definida mediante Amazon States Language (ASL), iniciando con el estado EjecutarLambdaESG, de tipo Task, el cual invoca directamente el ARN de la función Lambda responsable de la extracción. Una vez completada la tarea, el flujo finaliza con el estado End: true, sin ramificaciones adicionales, dado que esta prueba específica contempló únicamente la automatización de una fuente.

Durante la ejecución, la Step Function fue invocada desde la consola de AWS y se completó sin errores. La traza de eventos registró una duración total de aproximadamente 927 milisegundos, distribuidos entre siete eventos consecutivos.

- Inicio de ejecución (*ExecutionStarted*)
- Entrada al estado EjecutarLambdaESG
- Agendamiento de la función Lambda
- Inicio de la función Lambda
- Ejecución exitosa de la función Lambda (*LambdaFunctionSucceeded*)
- Salida del estado
- Finalización exitosa de la Step Function (*ExecutionSucceeded*)

La Figura 43 evidencia el registro de eventos de esta ejecución.

Figura 43. Log de ejecución – AWS Step Function SourceToLanding



Nota. Tomado de Amazon Web Services (2025)

Es importante destacar que, aunque en este prototipo la Step Function gestiona exclusivamente una única función Lambda correspondiente a la fuente ESG, en un entorno organizacional real su estructura se extendería para ejecutar de forma simultánea todas las funciones Lambda de extracción de cada fuente. Esta ejecución paralela garantizaría la recolección sincronizada de los datos requeridos por la organización, reduciendo los tiempos de extracción y facilitando el tratamiento posterior en fases subsiguientes del flujo. La evaluación de esta etapa confirmó la solidez del mecanismo de orquestación y su potencial escalabilidad dentro de escenarios productivos más amplios.

La segunda Step Function evaluada fue denominada ***LandingToStaging***, diseñada para coordinar de forma secuencial las tres etapas intermedias y finales del flujo ESG: transformación de registros en la zona Silver, consolidación en la carpeta Gold y carga definitiva en la base de datos Aurora PostgreSQL.

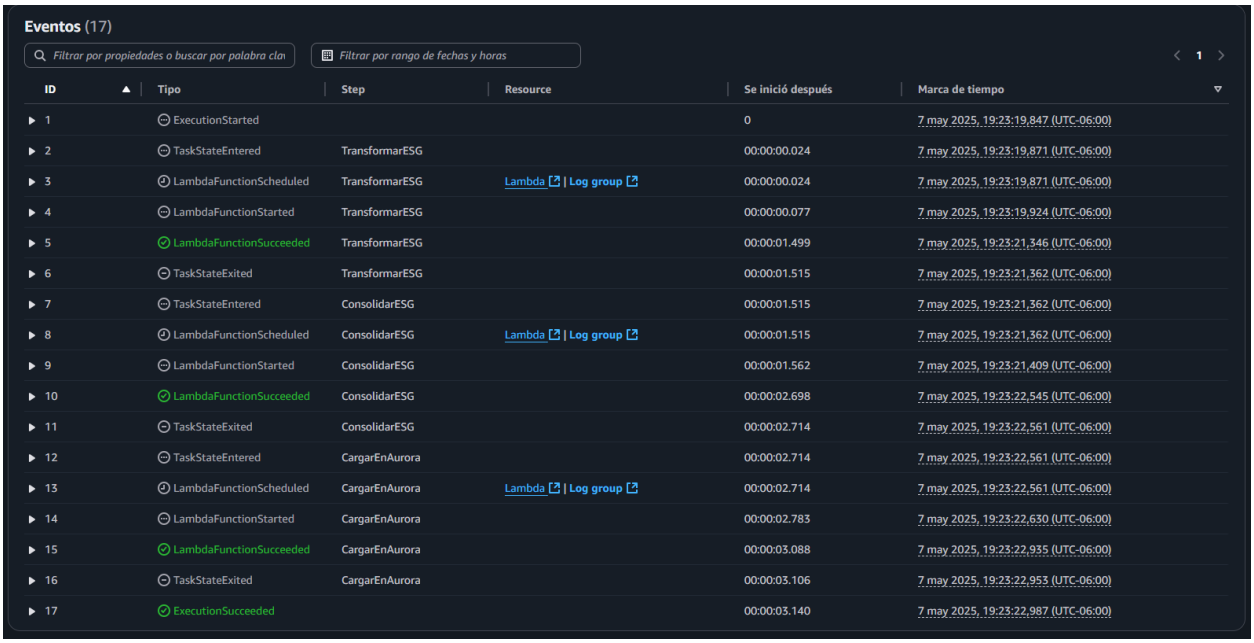
Esta Step Function inicia con el estado TransformarESG, encargado de invocar la función *lambda_transform_esg_data*. Una vez transformados y validados los registros, el flujo continúa con el estado ConsolidarESG, que ejecuta *lambda_consolidate_esg_data*, trasladando los datos depurados hacia la capa Gold. Finalmente, se ejecuta el estado CargarEnAurora, asociado a la función *insert_to_aurora_v2*, responsable de insertar los datos en la tabla *esg_data* de la base Aurora PostgreSQL, concluyendo así el recorrido completo de los datos desde su validación hasta su carga.

Durante la ejecución de la prueba, la orquestación funcionó de manera estable y sin errores. Se registraron 17 eventos consecutivos que reflejan el paso ordenado por cada uno de los estados definidos. El flujo se completó en 3.140 segundos, según los registros de tiempo proporcionados por la consola de AWS Step Functions. Se incluyen eventos de inicio y salida para cada tarea, así como confirmaciones de éxito para las funciones Lambda involucradas.

- **TransformarESG** → función *lambda_transform_esg_data* ejecutada con éxito.
- **ConsolidarESG** → función *lambda_consolidate_esg_data* ejecutada con éxito.
- **CargarEnAurora** → función *insert_to_aurora_v2* ejecutada con éxito.

La Figura 44 muestra el historial completo de eventos.

Figura 44. Log de ejecución - AWS Step Function LandingToStaging



Eventos (17)

Filtrar por propiedades o buscar por palabra cla | Filtrar por rango de fechas y horas | 1

ID	Tipo	Step	Resource	Se inició después	Marca de tiempo
1	ExecutionStarted			0	7 may 2025, 19:23:19,847 (UTC-06:00)
2	TaskStateEntered	TransformarESG		00:00:00.024	7 may 2025, 19:23:19,871 (UTC-06:00)
3	LambdaFunctionScheduled	TransformarESG	Lambda | Log group	00:00:00.024	7 may 2025, 19:23:19,871 (UTC-06:00)
4	LambdaFunctionStarted	TransformarESG		00:00:00.077	7 may 2025, 19:23:19,924 (UTC-06:00)
5	LambdaFunctionSucceeded	TransformarESG		00:00:01.499	7 may 2025, 19:23:21,346 (UTC-06:00)
6	TaskStateExited	TransformarESG		00:00:01.515	7 may 2025, 19:23:21,362 (UTC-06:00)
7	TaskStateEntered	ConsolidarESG		00:00:01.515	7 may 2025, 19:23:21,362 (UTC-06:00)
8	LambdaFunctionScheduled	ConsolidarESG	Lambda | Log group	00:00:01.515	7 may 2025, 19:23:21,362 (UTC-06:00)
9	LambdaFunctionStarted	ConsolidarESG		00:00:01.562	7 may 2025, 19:23:21,409 (UTC-06:00)
10	LambdaFunctionSucceeded	ConsolidarESG		00:00:02.698	7 may 2025, 19:23:22,545 (UTC-06:00)
11	TaskStateExited	ConsolidarESG		00:00:02.714	7 may 2025, 19:23:22,561 (UTC-06:00)
12	TaskStateEntered	CargarEnAurora		00:00:02.714	7 may 2025, 19:23:22,561 (UTC-06:00)
13	LambdaFunctionScheduled	CargarEnAurora	Lambda | Log group	00:00:02.714	7 may 2025, 19:23:22,561 (UTC-06:00)
14	LambdaFunctionStarted	CargarEnAurora		00:00:02.783	7 may 2025, 19:23:22,630 (UTC-06:00)
15	LambdaFunctionSucceeded	CargarEnAurora		00:00:03.088	7 may 2025, 19:23:22,935 (UTC-06:00)
16	TaskStateExited	CargarEnAurora		00:00:03.106	7 may 2025, 19:23:22,953 (UTC-06:00)
17	ExecutionSucceeded			00:00:03.140	7 may 2025, 19:23:22,987 (UTC-06:00)

Nota. Tomado de Amazon Web Services (2025)

Esta Step Function permitió verificar que las tres etapas críticas del proceso fueron gestionadas en una sola cadena lógica y sincronizada. Además, su diseño secuencial asegura la ejecución condicionada: la carga en Aurora solo se activa si las fases de transformación y consolidación concluyen correctamente. Esta condición refuerza el control del flujo y mitiga riesgos asociados a la propagación de errores hacia capas finales del proceso.

Desde una perspectiva organizacional, esta estructura se expandiría para incluir múltiples rutas paralelas, ejecutando de manera simultánea todas las funciones Lambda correspondientes a las distintas fuentes de datos gestionadas por la organización. Cada fuente seguiría su propio flujo de transformación, consolidación y carga, permitiendo así una carga de datos distribuida pero coordinada dentro del mismo entorno de ejecución. La evaluación de esta Step Function confirma que el prototipo está preparado para escalar hacia escenarios más complejos, conservando control, trazabilidad y estabilidad en cada transición del flujo automatizado.

5.2.4 Interpretación de resultados de pruebas funcionales

Una vez ejecutadas y documentadas las pruebas funcionales del prototipo automatizado, esta sección tiene como propósito interpretar los resultados obtenidos con base en los criterios de evaluación definidos en el marco metodológico, así cómo en la **sección 5.2.1**. La interpretación se estructura en torno a tres dimensiones fundamentales: precisión, consistencia y reducción de tareas manuales, cada una respaldada por indicadores específicos, técnicas de recolección de datos y evidencias extraídas del entorno de simulación implementado en AWS.

El análisis presentado no se limita a una descripción cuantitativa de los resultados, sino que incorpora una valoración crítica sobre la efectividad del prototipo, atendiendo a su capacidad para cumplir con los requerimientos funcionales, mejorar la calidad del proceso de carga de datos y reducir la intervención manual del equipo *Data Operations*. Esta interpretación es clave para validar el cumplimiento del objetivo específico #4 y para proyectar el potencial del prototipo en un contexto organizacional más amplio.

5.2.4.1 Precisión

La precisión se define en este proyecto como el porcentaje de registros correctamente cargados en la base de datos Aurora PostgreSQL, reflejando la capacidad del flujo para mantener la fidelidad de los datos desde su origen hasta su destino. Este criterio fue evaluado mediante consultas directas en DBeaver, verificando la integridad estructural y semántica de los registros cargados.

Durante la situación actual documentada en la **Fase #1**, la carga de datos presenta múltiples deficiencias que comprometen la precisión del proceso. El flujo depende casi por completo de tareas manuales realizadas en Excel, donde se detectaron errores frecuentes como:

- Duplicación de registros por falta de estandarización de identificadores.
- Presencia de campos vacíos o incompletos (valores NAs).
- Falta de validaciones automáticas tras la carga, lo que permite la carga de datos erróneos en la plataforma.
- Repetición de pasos ante errores detectados en etapas posteriores, generando retrabajos y aumentando la probabilidad de inconsistencias.

Estos problemas se reflejan directamente en los tiempos estimados por analista en cada etapa del proceso, detallados en la sección **4.1.2** y **5.1.7.2**. La tabla muestra que el equipo requiere 40 horas por ciclo (dos analistas) únicamente en validación y corrección manual, además de otras 18 horas distribuidas entre descarga, procesamiento y carga. En total, el ciclo completo demanda 58 horas de esfuerzo humano, gran parte de las cuales están destinadas a tareas orientadas exclusivamente a controlar y corregir errores.

En contraste, las pruebas funcionales realizadas en la Fase #4 demuestran una transformación radical en la precisión del proceso. El prototipo automatizado, estructurado en funciones Lambda y orquestado mediante AWS Step Functions, logró una tasa de 100% de precisión en la inserción de registros ESG. Este resultado fue verificado a través de consultas SQL ejecutadas en DBeaver, que confirmaron la ausencia de campos nulos, duplicidades y errores estructurales en la tabla *esg_data*. Cada etapa del flujo (extracción, validación, consolidación y carga) se ejecutó con éxito, respetando la integridad de los datos maestros en todas las etapas.

Además de la verificación estructural y semántica realizada mediante consultas SQL en DBeaver, se integró el monitoreo automatizado de métricas técnicas mediante Amazon CloudWatch, con el fin de validar la precisión del flujo desde una perspectiva operativa. Entre las métricas registradas, la más directamente vinculada al criterio de precisión fue ***Errors: Sum***, la

cual representa la cantidad total de fallos generados durante la ejecución de cada función Lambda del flujo automatizado.

Durante el período de pruebas, todas las funciones involucradas (*lambda_ingest_esg_data*, *lambda_transform_esg_data*, *lambda_consolidate_esg_data* e *insert_to_aurora_v2*) reportaron un valor de error igual a cero. Este resultado implica que no se produjo ningún fallo durante el procesamiento, validación o carga de datos en ninguna de las etapas del flujo. La ausencia total de errores no solo respalda la estabilidad técnica de la solución, sino que constituye un indicador empírico de que la precisión alcanzada en la carga de datos fue del 100%, confirmando que los datos fueron procesados íntegramente desde su origen hasta la base de datos Aurora PostgreSQL sin pérdida. En la Figura 45, se visualiza la métrica anteriormente explicada.

Figura 45. Métrica: Errors: Sum

Nota. Tomado de Amazon CloudWatch (2025)

Esta evidencia técnica de Amazon CloudWatch, complementa y fortalece la validación realizada mediante consultas en DBeaver (PostgreSQL). En conjunto, ambas estrategias de evaluación respaldan la afirmación de que el prototipo desarrollado garantiza una precisión operativa sustancialmente superior al proceso actual, al eliminar en un 100% los errores de carga.

Adicional a la validación realizada a nivel de funciones Lambda, se monitoreó el comportamiento de las máquinas de estado que orquestan el flujo automatizado, utilizando las métricas de Amazon CloudWatch. En particular, la métrica ***ExecutionsSucceeded: Sum*** permitió verificar cuántas veces se completaron con éxito las ejecuciones de las Step Functions ***SourceToLanding*** y ***LandingToStaging***, las cuales conforman el núcleo del flujo automatizado.

La Figura 46 muestra que ambas Step Functions se ejecutaron correctamente durante el periodo de pruebas, registrando dos ejecuciones exitosas en ***SourceToLanding*** y una ejecución exitosa en ***LandingToStaging***, sin reportar errores en ninguna instancia. Esta métrica confirma que la lógica de orquestación fue ejecutada con precisión, garantizando que los datos transitaran de forma íntegra desde su punto de origen hasta su carga en la base de datos.

Figura 46. Métrica: ExecutionsSucceeded: Sum

Nota. Tomado de Amazon CloudWatch (2025)

La ausencia de ejecuciones fallidas en estas máquinas de estado refuerza la validación funcional del prototipo, ya que el éxito de cada Step Function depende directamente de que cada función Lambda interna procese los datos correctamente. Por tanto, el monitoreo de ***ExecutionsSucceeded*** representa una evidencia técnica indirecta pero robusta de que la precisión fue mantenida a lo largo de todo el flujo automatizado.

Adicionalmente, se monitoreó la métrica ***ExecutionsFailed: Sum***, la cual reportó una única ejecución fallida en la Step Function ***LandingToStaging*** durante las pruebas iniciales del prototipo. La Figura 47 valida que esta ejecución no generó efectos negativos en la calidad de los datos, ya que el fallo detuvo el flujo antes de alcanzar la etapa de carga, impidiendo así la carga de datos incompleta o incorrecta. El diseño modular del prototipo permitió identificar el origen del fallo, corregir la lógica y volver a ejecutar el flujo con éxito en los intentos siguientes, como se evidenció en la métrica ***ExecutionsSucceeded: Sum***. Por tanto, esta métrica no representa una debilidad del sistema, sino una manifestación del control preventivo que ofrece la arquitectura automatizada, al evitar errores silenciosos y preservar la precisión en cada ejecución validada.

Figura 47. ExecutionsFailed: Sum

Nota. Tomado de Amazon CloudWatch (2025)

Esta comparación evidencia que el prototipo no solo mejora la precisión, sino que reduce de forma significativa la necesidad de validación manual, eliminando más de 40 horas de trabajo por ciclo. La estructura técnica implementada permite preservar la calidad de los datos mediante validaciones automatizadas, flujos definidos y trazabilidad en cada etapa. Sin embargo, la validación manual no desaparece por completo. Los analistas deben ejecutar los componentes del flujo automatizado, revisar su comportamiento durante la operación y atender cualquier resultado anómalo. También corresponde a los analistas supervisar el entorno, ajustar la lógica cuando cambien las fuentes de datos o las reglas de validación del negocio, y resolver eventuales fallos. En este contexto, la intervención humana se orienta hacia tareas de control puntual y mantenimiento, dejando atrás actividades repetitivas propensas a errores.

5.2.4.2 Consistencia

La consistencia de los datos se refiere a la coherencia de los valores cargados en cada etapa del flujo, garantizando que los registros conserven su estructura lógica y semántica conforme avanzan desde el origen hasta su consolidación en la base de datos. Este criterio se evaluó mediante la comparación entre el *dataset* original proporcionado por la fuente ESG y los archivos generados en las zonas Silver y Gold del bucket S3.

Durante la situación actual, descrita en la **Fase #1**, la consistencia presenta un alto nivel de vulnerabilidad. El proceso depende de archivos Excel que se editan manualmente, sin mecanismos que verifiquen la estructura, los tipos de datos ni las reglas de validación. Esta ausencia de controles propicia alteraciones accidentales y errores no detectados que comprometen la calidad de los datos antes de su carga en la plataforma MDM.

En las pruebas funcionales realizadas en el prototipo, se comprobó que la transformación, consolidación y carga de los registros se ejecutan sin alteraciones en la semántica ni en la estructura de los datos. La función *lambda_transform_esg_data* validó los campos requeridos y descartó registros incompletos, mientras que *lambda_consolidate_esg_data* unificó los datos válidos sin

modificar su contenido. Posteriormente, *insert_to_aurora_v2* cargó los registros en la base de datos Aurora PostgreSQL manteniendo los valores en su forma original.

Esta interpretación se respalda mediante el análisis directo entre los archivos. El *dataset* original (ver **Anexo IV**) contenía múltiples registros de distintas compañías. El archivo Silver (ver **Anexo V**) reflejó una versión filtrada con todos los campos requeridos validados correctamente. Finalmente, el archivo Gold (ver **Anexo VI**) preservó los mismos registros con sus valores intactos, lo que confirma que no hubo pérdida de información o alteración estructural durante el proceso. Esta comparación ratifica que el prototipo respeta la consistencia de los datos en todo el recorrido, desde la fuente hasta la consolidación. En síntesis, el flujo automatizado no solo mitiga los riesgos asociados a la edición manual, sino que establece un sistema de validación temprana que asegura la homogeneidad, integridad y coherencia de los datos cargados en la plataforma MDM.

5.2.4.3 Reducción de tareas manuales

Este criterio se refiere a la disminución de actividades operativas realizadas manualmente por los analistas de *Data Operations* en el proceso de carga de datos. La evaluación se fundamentó en la comparación entre el flujo actual (*As-Is*), descrito en la **Fase 1**, y el flujo automatizado (To-Be), documentado en la **Fase 2** y **3**.

En la situación actual, el proceso requiere intervención humana en cada una de sus cinco etapas: obtención de archivos desde distintas fuentes, almacenamiento manual en SharePoint, procesamiento en Excel, consolidación de múltiples archivos y carga final en la plataforma MDM mediante scripts SQL. Este enfoque, aunque funcional, expone el proceso a errores humanos, duplicación de esfuerzos y tiempos prolongados de ejecución.

El prototipo desarrollado automatiza integralmente las cinco etapas mencionadas. La función *lambda_ingest_esg_data* asume la extracción inicial, *lambda_transform_esg_data* ejecuta la validación estructural, *lambda_consolidate_esg_data* unifica los registros, y *insert_to_aurora_v2* realiza la carga final en la base de datos Aurora PostgreSQL. Estas funciones se orquestan mediante Step Functions, lo que elimina la necesidad de intervención manual y permite ejecutar el proceso de forma estructurada y trazable.

No obstante, afirmar que se ha eliminado el 100% de las tareas manuales resultaría inexacto. Si bien el prototipo sustituye tareas manuales repetitivas y propensas al error, también introduce nuevas formas de intervención humana. El analista debe ejecutar el flujo, además de revisar los resultados. En caso de fallos, el analista debe interpretar logs, depurar errores y actualizar funciones. Estas tareas, aunque menos operativas, siguen siendo manuales y exigen un nivel técnico mayor que el proceso anterior.

Por tanto, la intervención humana no desaparece, sino que se transforma. En lugar de ejecutar pasos mecánicos como la validación en Excel o la carga manual mediante scripts, el equipo se enfoca ahora en tareas de control, supervisión y mantenimiento del proceso. Este cambio representa una mejora sustancial en términos de eficiencia, pero también redefine el perfil

operativo requerido para sostener el flujo automatizado. Se estima que el prototipo logró reducir aproximadamente entre un 80% y un 90% de las tareas manuales principales, concentrando la participación humana en actividades más estratégicas y menos propensas al error.

5.2.4.4 Tiempo de ejecución del proceso

El tiempo de ejecución constituye un criterio fundamental para evaluar el impacto del prototipo en términos de eficiencia operativa. Este indicador permite determinar si el flujo automatizado es capaz de reducir la duración total del proceso de carga de datos en comparación con el método manual actualmente utilizado por el equipo de *Data Operations*.

Durante la **Fase 1** se identificó que el proceso actual requiere hasta 58 horas por ciclo, considerando la intervención simultánea de dos analistas. Este tiempo incluye la extracción, procesamiento, consolidación y carga de los datos desde diversas fuentes, se ve afectado por la fragmentación del flujo, la validación manual y la ausencia de herramientas de automatización.

En contraste, las pruebas funcionales realizadas en el prototipo automatizado evidenciaron una reducción sustancial en el tiempo total del flujo. La ejecución completa, coordinada mediante AWS Step Functions, se dividió en dos segmentos principales:

- La Step Function **SourceToLanding**, encargada de la ingestión inicial del archivo, registró un tiempo total de 927 milisegundos.
- La Step Function **LandingToStaging**, responsable de la transformación, consolidación y carga en la base de datos, tuvo una duración de 3.14 segundos.

Cada función Lambda dentro de este segundo segmento presentó los siguientes tiempos individuales:

- *lambda_transform_esg_data*: 589.24 milisegundos
- *lambda_consolidate_esg_data*: 340.09 milisegundos
- *insert_to_aurora_v2*: 15.86 milisegundos

Este comportamiento fue corroborado mediante la métrica ***ExecutionTime: Average*** de Amazon CloudWatch, que monitorea la duración media de ejecución de cada máquina de estado. La Figura 48 muestra que la Step Function ***SourceToLanding*** mantuvo tiempos de ejecución estables por debajo de 1 segundo, mientras que ***LandingToStaging***, encargada de la transformación, consolidación y carga, presentó una duración media entre 1.5 y 3 segundos durante el período de prueba. Esta diferencia es coherente con la complejidad relativa de las etapas involucradas en cada flujo, y respalda los resultados previamente registrados al nivel de funciones Lambda. Ambas ejecuciones se completaron con éxito en menos de 4.1 segundos, lo cual confirma la eficiencia general del flujo automatizado bajo condiciones de simulación controlada.

Figura 48. Métrica: ExecutionTime: Average

Nota. Tomado de Amazon CloudWatch (2025)

No obstante, si se proyecta el comportamiento observado durante las pruebas funcionales a la totalidad de las 12 funciones contempladas en el diseño completo del prototipo, y se asume un tiempo promedio de ejecución de entre 300 y 600 milisegundos por función, el tiempo total estimado del flujo completo se ubicaría en un rango de 3.6 a 7.2 segundos. A este valor debe añadirse el tiempo medio de ejecución de las dos Step Functions encargadas de orquestar el flujo, el cual se ubicó entre 2.8 y 3.9 segundos, según lo registrado en Amazon CloudWatch. En consecuencia, el tiempo total proyectado del flujo automatizado, en condiciones normales y con las 12 funciones operativas, oscilaría entre 6.4 - 11.1 segundos

Si bien las pruebas funcionales del prototipo evidenciaron tiempos de ejecución significativamente bajos, entre 6.4 - 11.1 segundos en todo el flujo desde la extracción hasta la carga final, resulta fundamental contrastar este comportamiento con las capacidades máximas que ofrece la plataforma de implementación. Según la documentación oficial de AWS (2025), cada función Lambda cuenta con un tiempo de ejecución máximo configurable de 900 segundos (15 minutos). Considerando que el flujo automatizado está compuesto por doce funciones Lambda distribuidas en cuatro etapas secuenciales; extracción (5 funciones), transformación (5 funciones), consolidación (1 función) y carga en la base de datos (1 función), el tiempo máximo teórico acumulado bajo condiciones extremas sería de hasta 10,800 segundos, equivalentes a 180 minutos o 3 horas. Esta estimación se obtiene multiplicando el tiempo límite de ejecución por función por la cantidad total de funciones.

Esta proyección establece un límite técnico realista para evaluar la escalabilidad del prototipo en un entorno organizacional con mayores volúmenes de datos. Aún en ese escenario extremo, el flujo automatizado representaría una reducción del tiempo de ejecución de aproximadamente 94.83% con respecto al proceso actual de 58 horas, de acuerdo con la fórmula:

$$Reducción\ (\%) = (\frac{TiempoActual - TiempoAutomatizado}{TiempoActual}) \times 100$$

$$Reducción\ (\%) = \left(\frac{58 - 3}{58}\right) \times 100 = \left(\frac{55}{58}\right) \times 100 \approx 94.83\%$$

Además de los límites individuales de las funciones Lambda, es fundamental considerar el tiempo máximo permitido por el servicio que orquesta todo el flujo: AWS Step Functions. De acuerdo con la documentación oficial de AWS (2025), una Step Function de tipo estándar permite una ejecución continua de hasta 1 año por instancia, lo cual proporciona un margen más que suficiente para coordinar flujos compuestos por múltiples fases y funciones Lambda, incluso en el escenario organizacional de alto volumen de datos y complejidad. En el contexto de este prototipo, las funciones ***SourceToLanding*** y ***LandingToStaging*** fueron diseñadas para estructurar el proceso en cuatro etapas secuenciales. Aún considerando una ejecución total de hasta 3 horas en el caso extremo de alcanzar el límite máximo de las 12 funciones Lambda involucradas, la duración acumulada permanece ampliamente dentro de los márgenes operativos de las Step Functions. Este contraste confirma que, desde el punto de vista de la arquitectura de orquestación, el prototipo se encuentra técnicamente preparado para escalar sin limitaciones de tiempo impuestas por la infraestructura de la empresa.

Por tanto, si bien el prototipo demostró una ejecución ágil y técnicamente eficiente en condiciones de prueba, el análisis del tiempo de ejecución no debe interpretarse de forma aislada. Será necesario validar su comportamiento con datos reales y bajo escenarios de carga intensiva para determinar si mantiene niveles adecuados de rendimiento en entornos productivos. En conclusión, el prototipo automatizado supera ampliamente al proceso actual en términos de tiempo, tanto en escenarios simulados como proyectados. Considerando que el proceso manual requiere hasta 58 horas por ciclo y que el flujo automatizado, en su límite teórico máximo, podría alcanzar hasta 3 horas de ejecución acumulada, se estima una reducción del tiempo de ejecución de aproximadamente 94.83%. A medida que se avance hacia una eventual implementación organizacional, el monitoreo de desempeño deberá consolidarse como un componente esencial en la evaluación técnica continua, permitiendo identificar oportunidades de ajuste y garantizar la sostenibilidad operativa de la solución automatizada.

6 Conclusiones

El presente capítulo expone los hallazgos más relevantes obtenidos durante el desarrollo del Trabajo Final de Graduación, con base en el análisis de resultados de las cuatro fases del proyecto. Las conclusiones presentadas tienen como propósito sintetizar el cumplimiento del objetivo general y los objetivos específicos, así como evidenciar que los entregables definidos en el proyecto fueron alcanzados de manera satisfactoria.

Las conclusiones se organizan por cada objetivo específico, permitiendo identificar con claridad los resultados derivados del análisis de la situación actual, el diseño de la solución propuesta, el desarrollo del prototipo automatizado y la evaluación de su desempeño en un entorno técnico de simulación. Esta estructura garantiza una trazabilidad directa entre las metas planteadas y los logros alcanzados, reafirmando la coherencia metodológica del proyecto.

6.1 Conclusiones del objetivo específico 1

Respecto al objetivo específico #1, el cual consistió en *"Analizar la situación actual del proceso de carga de datos para la identificación de deficiencias que afectan la carga de datos en la plataforma de gestión de datos maestros"*, se concluye lo siguiente:

- El proceso actual de carga de datos se encuentra compuesto en su totalidad por tareas manuales, representando el 100% del esfuerzo operativo por ciclo, distribuido en aproximadamente 18 horas para la obtención, transformación y carga de los archivos, así como 40 horas adicionales para labores de validación. Esta estimación se documenta en la sección **4.1.2**, a partir del levantamiento de información con el equipo *Data Operations*.
- Las herramientas empleadas; como Microsoft Excel, SharePoint y conexiones FTP, carecen de mecanismos automáticos de control, trazabilidad y monitoreo, lo cual incrementa significativamente la probabilidad de errores humanos y necesidad de retrabajo. Esta situación es descrita en la sección **4.1.1**, y es reafirmada por las observaciones recogidas durante las entrevistas semiestructuradas.
- El flujo actual no contempla ninguna estrategia formal de monitoreo o estandarización, lo que genera dependencia directa de criterios individuales y dificulta la auditoría del proceso. Esta limitación estructural queda evidenciada en la sección **4.1.3** y se consolida en el análisis FODA de la sección **4.1.6**, donde se señalan debilidades clave como la ausencia de control automatizado y la inexistencia de un repositorio único para la trazabilidad.
- La modelación del proceso *As-Is* mediante notación BPMN 2.0, presentada en la sección **4.1.1**, permitió concluir que el proceso actual presenta una alta fragmentación, múltiples tareas redundantes y una secuencia de validaciones manuales que no se encuentran integradas. Estos hallazgos evidencian deficiencias estructurales que justifican la necesidad de automatizar el proceso para mejorar su eficiencia y trazabilidad.
- La dependencia casi total del procesamiento manual en el flujo actual genera consecuencias directas sobre la calidad, consistencia y confiabilidad de los datos maestros. Según lo evidenciado en la sección **4.1.4**, los principales problemas identificados; duplicaciones, campos faltantes, formatos inconsistentes y ausencia de validaciones post-carga, comprometen la integridad de la información y dificultan su consolidación. Estas

deficiencias aumentan la carga operativa por reprocesos, elevan el riesgo de decisiones basadas en datos incorrectos y reducen la trazabilidad, al no contar con mecanismos automatizados que aseguren la integridad de los registros en cada etapa del proceso

6.2 Conclusiones del objetivo específico 2

En relación con el objetivo específico #2, el cual planteó *"Diseñar un nuevo proceso de carga de datos integrando herramientas de automatización y alineado con los requerimientos del equipo, con el fin del mejoramiento de la eficiencia del proceso"*, se concluye lo siguiente:

- El diseño del nuevo flujo de trabajo (modelo *To-Be*) respondió a una estructura modular, basada en una arquitectura por capas (Bronze, Silver y Gold) sobre la plataforma Amazon Web Services (AWS). Esta propuesta quedó formalmente documentada en el diagrama BPMN incluido en la sección **4.2.7**, donde se reflejan las interacciones entre componentes técnicos como AWS Lambda, Step Functions, S3 y Aurora, así como puntos de control lógico y monitoreo mediante CloudWatch.
- La reestructuración del proceso se fundamentó en una evaluación detallada de las deficiencias técnicas y organizativas identificadas en la Fase 1. A partir de estas, se formularon cinco oportunidades de mejora clave, incluyendo la reducción de tareas manuales, la implementación de validaciones automáticas, la mejora de la calidad de los datos, la trazabilidad operativa y la disminución de tiempos de ejecución. Estas oportunidades fueron validadas mediante entrevista al *Senior Data Engineer*, tal como se evidencia en las secciones **4.2.1** y **4.2.2**.
- La aplicación del método PROTEOCE como marco de análisis permitió estructurar la propuesta desde dimensiones técnicas, organizacionales y funcionales, garantizando que el diseño final no se limitara a un flujo idealizado, sino que respondiera a las condiciones operativas reales del equipo *Data Operations*. Esta metodología se detalla en la secciones **4.2.2**, **4.2.3**, **4.2.4**, **4.2.5** y guió la documentación sistemática de requerimientos, herramientas, interacciones y condiciones técnicas del entorno de automatización propuesto.
- El diseño del nuevo proceso permitió automatizar un total de once tareas clave, distribuidas desde la extracción de datos hasta la publicación en la plataforma de Gestión de Datos Maestros. Estas automatizaciones comprenden la orquestación del flujo, validación estructural, consolidación, carga y registro de métricas operativas. Su identificación se encuentra documentada en la sección **4.2.5.1**, mientras que su implementación conceptual se detalla en la sección **4.2.5.2**. Este enfoque redujo de forma significativa la intervención manual, fortaleció la trazabilidad técnica del flujo y mejoró la eficiencia operativa del proceso.
- El diseño del nuevo proceso fue presentado al equipo de *Data Operations* para su validación, quien corroboró que el modelo respondía de forma efectiva a los requerimientos funcionales y limitaciones identificadas en el entorno actual. Esta validación se documenta en la sección **4.2.5**, y constituyó un insumo esencial para garantizar la viabilidad técnica del proceso automatizado.

- La construcción de la matriz requerimientos vs diseño, documentada en la sección **4.2.8**, permitió validar sistemáticamente que cada requerimiento funcional identificado en el análisis fue resuelto mediante componentes técnicos específicos del modelo To-Be. Esta matriz evidenció la cobertura completa de necesidades operativas como orquestación, transformación, validación, consolidación, carga y monitoreo, a través de servicios de AWS.
- La matriz de integración, presentada en la sección **4.2.9**, permitió demostrar con precisión cómo cada etapa del flujo automatizado diseñado se encuentra alineada funcionalmente con servicios específicos de AWS. Esta herramienta documenta no solo la secuencia operativa del modelo *To-Be* sino también el tipo de integración técnica empleada, la responsabilidad funcional asignada a cada componente y las observaciones técnicas relevantes sobre su ejecución.

6.3 Conclusiones del objetivo específico 3

Con respecto al objetivo específico #3, el cuál formula *"Desarrollar un prototipo de solución automatizada para la carga de datos en la plataforma de gestión de datos maestros, basado en la herramienta seleccionada y los requerimientos identificados"*, se concluye lo siguiente:

- El prototipo fue desarrollado utilizando exclusivamente herramientas incluidas en el plan Free Tier de Amazon Web Services (AWS), replicando la arquitectura definida en la Fase 2. Esta implementación incluyó la estructuración del almacenamiento en zonas Bronze, Silver y Gold en Amazon S3, la automatización de tareas mediante funciones Lambda y la orquestación lógica con AWS Step Functions, según lo descrito en las secciones **5.1.2** y **5.1.3**.
- El desarrollo técnico contempló la creación de funciones Lambda independientes para cada etapa del flujo automatizado: extracción, transformación, consolidación y carga. Estas funciones fueron diseñadas siguiendo principios de modularidad y responsabilidad única, lo que facilita su mantenimiento y futuras extensiones, tal como se documenta en la secciones **5.1.2**, **5.1.3** y **5.1.4**.
- La lógica implementada en cada componente del prototipo se alinea con los requerimientos funcionales definidos en las fases anteriores. Las funciones Lambda desarrolladas replican el comportamiento esperado del flujo automatizado propuesto, utilizando datos simulados que representaron estructuras y comportamientos comunes del entorno organizacional. Esta correspondencia técnica entre diseño y ejecución se establece en la sección **5.1.5**.
- El prototipo resultante es técnicamente funcional, modular y adaptable. Su estructura permite incorporar nuevas fuentes de datos o integrar servicios adicionales sin alterar la lógica general del flujo. Además, su diseño facilita la incorporación progresiva a entornos organizacionales, según lo demostrado a lo largo de la **Fase 3**.

6.4 Conclusiones del objetivo específico 4

En relación con el objetivo específico #4, el cuál expone *"Evaluar la efectividad del prototipo de la solución automatizada en términos de precisión, consistencia y reducción de tareas manuales en el proceso de carga de datos"*, se concluye lo siguiente:

- La evaluación funcional del prototipo fue estructurada con base en cuatro criterios definidos: precisión, consistencia, tiempo de ejecución del proceso y reducción de tareas manuales. Cada criterio se operacionalizó mediante indicadores específicos, técnicas de recolección de datos y fuentes de verificación documentadas en la sección **5.2.1**.
- En relación con la precisión, se comprobó que los datos fueron insertados correctamente en la base de datos Aurora sin alteraciones estructurales ni semánticas. El análisis realizado mediante consultas SQL en DBeaver validó que el 100% de los registros esperados fueron cargados, cumpliendo con los requerimientos del proceso. Esta verificación técnica se documenta en las secciones **5.2.3.5** y **5.2.4.1**.
- En cuanto a la consistencia, se confirmó que los registros transformados en la zona Silver y consolidados en Gold conservaron la lógica de negocio predefinida. Las funciones Lambda ejecutadas aplicaron reglas de limpieza, validación de tipos de dato y normalización, garantizando coherencia en los valores cargados. Esta evidencia se detalla en las secciones **5.2.3.3**, **5.2.3.4** y **5.2.4.2**, que incluye ejemplos específicos y trazabilidad de los datos procesados.
- Respecto a la reducción de tareas manuales, el análisis comparativo entre el proceso *As-Is* y el proceso automatizado *To-Be* evidenció que el prototipo reemplazó entre un 80% y un 90% de las tareas manuales originalmente ejecutadas por el equipo de *Data Operations*. Actividades como la extracción, transformación, validación estructural, consolidación y carga fueron delegadas a componentes de AWS, reduciendo significativamente la intervención humana directa, tal como se documenta en la sección **5.2.4.3**.
- La ejecución del prototipo completo se realizó en un tiempo estimado de 4.1 segundos, con registro exitoso de 17 eventos en las Step Functions implementadas. Esta duración refuerza la eficiencia operativa del proceso automatizado y respalda su aplicabilidad en entornos reales, siempre que el volumen de datos y concurrencia se mantengan dentro de condiciones similares. Este resultado se detalla en la sección **5.2.4.4**.
- En un escenario extremo de alta carga y complejidad operativa, el flujo automatizado podría alcanzar una duración máxima teórica de hasta 3 horas, considerando que cada una de las doce funciones Lambda ejecuta durante el límite superior permitido por AWS (900 segundos por función). Esta proyección técnica, documentada en la sección **5.2.4.4**, establece un margen realista para la evaluación de escalabilidad de la solución en entornos organizacionales con grandes volúmenes de datos. Incluso bajo estas condiciones límite, el prototipo automatizado representaría una reducción del 94.83% en el tiempo de ejecución respecto al proceso actual, que requiere hasta 58 horas por ciclo.
- Finalmente, la validación del prototipo por parte del equipo de *Data Operations* confirmó que la solución automatizada responde adecuadamente a los requerimientos funcionales establecidos en las fases previas. Esta aceptación organizacional refuerza la viabilidad

tanto técnica cómo operativa del proceso propuesto y fue documentada en las minutas de validación incluidas en los apéndices (ver secciones **9.21**, **9.24** y **9.25**).

7 Recomendaciones

Este capítulo presenta un conjunto de recomendaciones formuladas a partir del análisis integral del proceso actual, el diseño técnico de la solución automatizada, el desarrollo del prototipo funcional y los resultados obtenidos durante su evaluación. Las sugerencias expuestas se fundamentan en la comprensión detallada del proceso de carga de datos y en la evidencia generada a lo largo de las distintas fases del proyecto.

En primer lugar, se plantean recomendaciones orientadas a la adopción de la solución técnica desarrollada, considerando su viabilidad estructural, su correspondencia con los requerimientos especificados y su potencial para reemplazar el modelo operativo actual. Posteriormente, se abordan aspectos relacionados con el proceso de implementación, incluyendo la planificación por fases, la preparación del entorno organizacional y la incorporación de mecanismos de seguimiento técnico. Finalmente, se identifican elementos que no fueron abordados directamente dentro del alcance de este estudio, pero que podrían representar oportunidades relevantes para fortalecer la automatización y extender su impacto hacia otras dimensiones del ciclo de vida de los datos maestros.

A continuación, se detallan las recomendaciones propuestas con base en los hallazgos obtenidos y la solución desarrollada.

- Se recomienda adoptar el flujo automatizado propuesto como modelo base para reemplazar el proceso actual de carga de datos en la plataforma de gestión de datos maestros. La arquitectura técnica desarrollada, validada mediante la ejecución de pruebas controladas en Amazon Web Services (AWS), replicó con fidelidad el comportamiento esperado del flujo, integrando componentes como Lambda, Step Functions, Amazon S3, Aurora y CloudWatch. Esta solución demostró ser funcional, modular, trazable y alineada con los requerimientos levantados por el equipo de *Data Operations*.
- Incorporar activamente a los analistas del equipo de *Data Operations* en la implementación de la solución automatizada, considerando su perspectiva operativa como insumo clave para ajustar decisiones técnicas y funcionales. Su participación facilita la apropiación del nuevo proceso, promueve una transición fluida y refuerza la alineación con la realidad del entorno operativo.
- La automatización del proceso de carga reducirá la carga operativa manual, lo que permitirá a los analistas enfocar sus esfuerzos en tareas estratégicas propias de su rol, como el análisis de calidad de datos o la generación de reportes especializados. Esta redistribución del trabajo favorece el aprovechamiento del conocimiento técnico del equipo, además de fortalecer su contribución dentro del ciclo de vida de los datos.
- Con el fin de mitigar riesgos operativos, se recomienda iniciar la adopción del flujo automatizado con una única fuente de datos, utilizando datos reales en un entorno controlado. Esta fase piloto permitiría identificar ajustes necesarios antes de una implementación a mayor escala, además de validar el comportamiento de los componentes y la estructura de los datos frente a casos reales del entorno organizacional.

- La implementación del proceso automatizado debe ejecutarse bajo un enfoque por fases. Se recomienda diseñar un plan que contemple actividades específicas como la habilitación de entornos productivos en AWS, la configuración de permisos de acceso, la integración con las fuentes de datos existentes y la capacitación del equipo *Data Operations* en el uso y mantenimiento del proceso automatizado.
- Es fundamental incorporar una estrategia de gestión del cambio, orientada a reducir la resistencia organizacional, alinear a los actores clave y promover la apropiación organizacional de la solución. La capacitación técnica y funcional del equipo de *Data Operations*, junto con el establecimiento de una matriz de roles y responsabilidades, facilitará la sostenibilidad del sistema a largo plazo.
- Se sugiere, además, implementar un sistema de monitoreo continuo del proceso, mediante alertas técnicas, *dashboards* operativos y métricas clave, tales como tiempo medio de ejecución, porcentaje de registros válidos por carga y número de errores detectados por fase. Este conjunto de indicadores permitirá evaluar de forma objetiva el rendimiento y la estabilidad de la solución una vez puesta en marcha.
- Aunque el alcance de esta propuesta se centró en la automatización del proceso de carga de datos, se identifica como una oportunidad futura la integración con herramientas empresariales de gestión de calidad de datos (DQM). Explorar estas integraciones permitiría fortalecer los controles de calidad, asegurar la conformidad normativa y mejorar la trazabilidad de los datos desde su origen hasta su uso final.
- La arquitectura propuesta está diseñada para integrar múltiples fuentes (*Credit Risk Data*, *Ratings*, *ESG*, *News Edge*, *Corporate Intelligence Database*). Se recomienda definir desde ahora un conjunto formal de criterios de escalabilidad que incluyan: estructuras requeridas en cada archivo, condiciones de calidad para su carga, reglas de transformación comunes y excepciones permitidas. Establecer estos criterios garantizará que futuras cargas de datos mantengan la integridad del flujo sin requerir rediseños.
- Se sugiere evaluar la incorporación de mecanismos de recuperación automática ante fallos, como políticas de reversión (*rollback*) y control de versiones de archivos procesados, así como la aplicación de pruebas de estrés que permitan simular escenarios de alta concurrencia o volúmenes masivos de datos. Estas medidas contribuirían a fortalecer la resiliencia del flujo automatizado en un entorno organizacional real.
- Si bien el enfoque del proyecto se centró en la automatización técnica del proceso de carga de datos, se recomienda avanzar hacia la incorporación progresiva de prácticas formales de gobernanza de datos. Esto incluye definir políticas de acceso a los datos procesados, establecer lineamientos de calidad, asignar roles de responsabilidad sobre cada etapa del flujo automatizado y asegurar el cumplimiento de estándares corporativos en la gestión de los datos maestros.

8 Referencias

Amazon Web Services. (2024). *Overview of Amazon Web Services*.

Amazon Web Services. (2025). *Amazon Simple Storage Service: User guide*. https://docs.aws.amazon.com/AmazonS3/latest/userguide/Welcome.html

Amazon Web Services. (2024). *AWS Lambda Developer Guide*. https://docs.aws.amazon.com/lambda/latest/dg/welcome.html

Amazon Web Services. (2024). *AWS Step Functions Developer Guide*. https://docs.aws.amazon.com/step-functions/latest/dg/welcome.html

Amazon Web Services. (s.f.). *¿Qué es ETL? Extracción, transformación y carga*. https://aws.amazon.com/what-is/etl/

BG&A. (2025, febrero 19). *Aumento (2023–2025) en porcentaje de aporte CCSS*. BG&A Abogados Corporativos. https://bgacorp.com/porcentaje-aporte-ccss/

Cascante Molina, J. A. (2021). *Propuesta de mejora del proceso de gestión del servicio de análisis de datos en la Gerencia de Innovación del Grupo 823*. Trabajo Final de Graduación para optar al grado de Licenciatura en Administración de Tecnología de Información, Instituto Tecnológico de Costa Rica.

Chavarría Sánchez, L. J. (Coord.). (2024). Administración de Tecnología de Información. Instituto Tecnológico de Costa Rica.

Chaves Araya, J. C. (2023). *Propuesta para mejoramiento de los procesos de carga de datos sobre el módulo de servicio al cliente que brinda la plataforma Oracle CX, ofrecido por la empresa Xum Technologies*. Trabajo Final de Graduación para optar al grado de Licenciatura en Administración de Tecnología de Información, Instituto Tecnológico de Costa Rica.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly, 36(4), 1165–1188.

Cordero Pereira, M. de L. A. (2022). *Propuesta de estandarización y automatización para el proceso de integración de datos en la empresa Xumtech*. Trabajo Final de Graduación para optar al grado de Licenciatura en Administración de Tecnología de Información, Instituto Tecnológico de Costa Rica.

DAMA International. (2017). *Data management body of knowledge (DAMA-DMBOK2)* (2nd ed.). Technics Publications.

Deloitte. (s.f.). *Transformación digital y automatización inteligente de procesos*. Deloitte. Recuperado de https://www.deloitte.com/es/es/services/consulting/research/transformacion-digital-y-automatizacion-inteligente-de-procesos.html

Dumas, M., La Rosa, M., Mendling, J., & Reijers, H. A. (2018). *Fundamentals of business process management* (2nd ed.). Springer.

Durairaj, N., & Hertsens, J. (2024). *Setting up a secure and scalable multi-account AWS environment*. AWS Prescriptive Guidance. Amazon Web Services. https://docs.aws.amazon.com/prescriptive-guidance/latest/migration-aws-environment/welcome.html

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill/Interamericana Editores.

Hikmawati, S., Santosa, P. I., & Hidayah, I. (2021). Improving data quality and data governance using master data management: A review. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 5*(3).

Huaire Inacio, E. J. (2019). *Método de investigación*. Acta Académica. Recuperado de https://www.aacademica.org/edson.jorge.huaire.inacio/78

IBM. (s.f.). *What is Business Process Management?* https://www.ibm.com/think/topics/business-process-management

Instituto Tecnológico de Costa Rica. (2024). Reglamento para la presentación del Trabajo Final de Graduación de la Escuela de Administración de Tecnologías de Información (ATI). Última versión: febrero de 2024.

International Organization for Standardization. (2013). *ISO/IEC 19510:2013: Information technology — Object Management Group Business Process Model and Notation (Version 2.0.1)*. ISO.

Maranto Rivera, M., & González Fernández, M. E. (2015). *Fuentes de información*. Universidad Autónoma del Estado de Hidalgo. Recuperado de http://www.uaeh.edu.mx/virtual

Mucci, T., & Stryker, C. (2024, 5 de abril). ¿Qué es la automatización de procesos de negocio? *IBM*. https://www.ibm.com/mx-es/topics/business-process-automation

Nguyen, T., Bonini, M., Langenbahn, J. E., Moser, S., Schneeweis, E. A., Urru, A., & Echelmeyer, W. (2021). Automation? Yes … But Where to Begin? *Proceedings of the 2021 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, 435–441. IEEE. https://doi.org/10.1109/IEEM50564.2021.9673068

Project Management Institute. (2017). *A guide to the project management body of knowledge (PMBOK® guide)* (6th ed.). Project Management Institute.

Skulimowski, A. M. J., & Smętkowski, M. (2016). *Fishbone diagrams for the development of knowledge bases in foresight studies*. In J. Kacprzyk & W. Pedrycz (Eds.), Springer Proceedings in Complexity (pp. 83–92). Springer. https://doi.org/10.1007/978-3-319-45145-9_8

Stachtiari, E., Naskos, A., Ampatzoglou, A., Avgeriou, P., & Chatzigeorgiou, A. (2018). Early validation of system requirements and design through correctness-by-construction. *Journal of Systems and Software*, 143, 50–67. https://doi.org/10.1016/j.jss.2018.05.009

Stencil BPMN. (2013). *BPMN Quick Reference Guide*. BPMN Quick Reference.

Stryker, C., & Belcic, I. (2024, 25 de junio). What is business process modeling and notation (BPMN)? *IBM*. https://www.ibm.com/think/topics/bpmn

Southern Illinois University Carbondale. (2024, 20 de junio). *The Importance of Return on Investment in Business Decision-Making*. Recuperado de https://onlinedegrees.siu.edu/programs/business/mba/finance/importance-of-roi-in-decision-making/

Suranto, A. W. (2015). Software prototypes: Enhancing the quality of requirements engineering process. *Journal of Theoretical and Applied Information Technology, 74*(1), 147–153.

Taulli, T. (2020). The robotic process automation handbook: A guide to implementing RPA systems. Apress.

Visure Solutions. (2021). *Matriz de trazabilidad de requerimientos (RTM): Definición, importancia y fundamentos clave*. https://visuresolutions.com/es/blog/Requerimientos-de-trazabilidad-matriz/

Westerman, G., Bonnet, D., & McAfee, A. (2014). Leading Digital: Turning Technology into Business Transformation. *Harvard Business Review Press*.

Wickramasinghe, S. (2024, abril 15). Prototype testing: Types, benefits, and best practices. *Testsigma*. https://testsigma.com/blog/prototype-testing/

9 Apéndices

9.1 Apéndice A. Cronograma del proyecto

Tarea	Semanas															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ajuste de anteproyecto																
Desarrollo de marco metodológico																
Desarrollo de marco conceptual																
Realización de la Fase 1																
Realización de la Fase 2																
Realización de la Fase 3																
Realización de la Fase 4																
Elaboración de conclusiones y recomendaciones																
Revisión final y ajustes del documento																
Entrega final																

9.2 Apéndice B. Plantilla de minutas de reunión

Reunión No.			**Fecha:**	
Lugar:			Día:	
			Hora de inicio:	
			Hora de finalización:	
Objetivo de la reunión:				
Participantes:	**Presentes:**			
	Ausentes:			
Temas tratados				
No.	Asunto	Comentarios	**Acuerdos**	
1				
2				

3			
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
Firma representante de organización			**Firma estudiante**

9.3 Apéndice C. Plantilla de control de cambios

Hoja de Control de Cambios			
Datos Generales del Cambio			
N° Cambio			
Solicitante		Fecha de solicitud del cambio	
Responsable de la implementación		Fecha de realización del cambio	
Estado	☐ Aprobado ☐ En Revisión ☐ Rechazado		
Detalles del Cambio			
Sección	Introducción / Alcance / Marco Teórico / Metodología /…		
Descripción detallada	Descripción detalla del cambio por realizar.		
Justificación			
Implicaciones de realizar el cambio			
Impacto	Especificar si el cambio genera impacto en otras áreas del proyecto, tales como recursos, cronograma, limitaciones, supuestos, entre otros.		
Comentarios/ Observaciones			
Aprobación			
Revisado por: Nombre tutor Firma		Elaborado por: Nombre estudiante Firma	

(Prof. tutor)	(Estudiante)
Revisado por: **Nombre representante empresa** **Firma** **(Empresa)**	Aprobado por: Nombre Coordinadora TFG Firma (Coordinadora de TFG)

9.4 Apéndice D. Minuta de reunión #1

Reunión No.	**01**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	01 de octubre del 2024
		Hora de inicio:	11:00am
		Hora de finalización:	11:30am
Objetivo de la reunión:	Discutir la viabilidad de realizar un Trabajo Final de Graduación (TFG) en el equipo y la organización, así cómo analizar problemáticas existentes.		
Participantes:	Presentes: Representante de la organización, estudiante		
	Ausentes: Ninguno		

Temas tratados

No.	Asunto	Comentarios	**Acuerdos**
1	Viabilidad del TFG en la organización	Se discutió la pertinencia del proyecto dentro del equipo y cómo puede contribuir a resolver problemáticas actuales.	Acordar la realización del proyecto y definir próximos pasos.
2	Identificación de problemáticas existentes	Se abordaron temas oportunidades de mejora en los procesos actuales.	Priorizar el análisis de estas problemáticas y profundizar en las causas raíz.

Próxima reunión

Temas por tratar	Fecha	**Convocados**
Discutir en profundidad la problemática seleccionada y definir el enfoque específico del TFG.	10 de octubre del 2024	Representante de la organización Estudiante

Firma representante de la organización:		**Firma estudiante:**

9.5 Apéndice E. Minuta de reunión #2

Reunión No.	**02**	**Fecha:**	
Lugar:	Oficinas de la empresa	Día:	**10 de octubre del 2024**
		Hora de inicio:	**01:00pm**

		Hora de finalización:	**02:00pm**
Objetivo de la reunión:	Discutir en profundidad la problemática seleccionada y definir el enfoque específico del TFG.		
Participantes:	Presentes: Representante de la organización, estudiante		
	Ausentes: Ninguno		
Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Contexto del proyecto	Se discutió el enfoque general del proyecto y su importancia para resolver problemáticas existentes.	Confirmar que el proyecto se centrará en la mejora del proceso.
2	Problemática identificada	Se analizaron retos actuales relacionados con la operación y gestión del proceso.	Avanzar con un análisis más estructurado del problema.
3	Causa raíz y posibles impactos	Se identificaron elementos clave que contribuyen a la problemática y los riesgos asociados.	Elaborar herramientas de análisis para estructurar la problemática.
4	Riesgos asociados con la problemática	Se discutió el impacto potencial de no resolver las deficiencias identificadas.	**Definir estrategias iniciales para abordar los riesgos.**
Próxima reunión			
Temas por tratar Discutir en profundidad la problemática seleccionada y definir el enfoque específico del TFG.		Fecha 17 de octubre del 2024	**Convocados** Representante de la organización Estudiante
			Representante de la organización Estudiante
Firma representante de la organización			**Firma estudiante**

9.6 Apéndice F. Minuta de reunión #3

Reunión No.	**03**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	**17 de octubre del 2024**
		Hora de inicio:	**10:00am**
		Hora de finalización:	**10:45am**
Objetivo de la reunión:	Conocer a fondo las causas de la problemática y analizar áreas clave para abordar en el proyecto.		
Participantes:	**Presentes:** Representante de la organización, estudiante		
	Ausentes: Ninguno		

Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Análisis de las causas de la problemática	Se discutieron factores como la falta de claridad en los requerimientos, la ausencia de documentación estructurada, la existencia de tareas manuales repetitivas, y la necesidad de mejorar estándares de gobernanza y control.	Trabajar en la definición de requerimientos, creación de documentación estándar, y propuestas de mejora para automatización y gobernanza.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
Validar secciones específicas del documento del proyecto y profundizar en el entendimiento del proceso.		06 de noviembre del 2024	Representante de la organización Estudiante
Firma representante de la organización			**Firma estudiante**

9.7 Apéndice G. Minuta de reunión #4

Reunión No.	**04**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	**06 de noviembre del 2024**
		Hora de inicio:	**11:00am**
		Hora de finalización:	**12:00pm**
Objetivo de la reunión:	Validar secciones específicas del documento del proyecto y profundizar en el entendimiento del proceso.		
Participantes:	**Presentes:** Representante de la organización, estudiante		
	Ausentes: Ninguno		

Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Proyectos similares	Se discutieron ejemplos de proyectos previos que sirvan como insumo y referencia para el proyecto, asegurando alineación con estándares previos.	Identificar información relevante de proyectos similares y adaptarla al documento.
2	Sección de beneficios	Se definieron beneficios directos y beneficios indirectos	Refinar los beneficios

3	Exclusiones, supuestos y limitaciones	Se acordó definir lo que está fuera del alcance del proyecto, además de establecer supuestos claros y limitaciones.	Completar estas secciones para su inclusión en el documento final.
4	Descripción general del proceso	Se revisó el flujo lógico del proceso seleccionado, para identificar áreas clave que necesitan ser descritas en detalle en el documento.	Estructurar una descripción clara y general del proceso para el documento.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
Analizar el estado actual del proceso y sus principales problemáticas para avanzar en la definición de áreas de mejora.		18 de noviembre del 2024	Representante de la organización Estudiante
Firma representante de la organización			**Firma estudiante**

9.8 Apéndice H. Minuta de reunión #5

Reunión No.	**05**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	**18 de noviembre del 2024**
		Hora de inicio:	**11:00am**
		Hora de finalización:	**12:00pm**
Objetivo de la reunión:	Analizar el estado actual del proceso y sus principales problemáticas para avanzar en la definición de áreas de mejora.		
Participantes:	**Presentes:** Representante de la organización, estudiante		
	Ausentes: Ninguno		

Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Estado actual del proceso.	Se revisó el flujo actual del proceso	Documentar el proceso actual en detalle para su análisis.
2	Roles y responsabilidades del equipo	Se definieron los roles clave dentro del equipo, alineados a un marco ágil, y su impacto en la gestión y operación del proceso.	Incluir los roles definidos en la documentación del proyecto.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**

		Representante de la organización Estudiante
Firma representante de la organización		**Firma estudiante**

9.9 Apéndice I. Minuta de Reunión #6

Reunión No.	**06**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	4 de marzo de 2025
		Hora de inicio:	08:00 pm
		Hora de finalización:	09:00 pm
Objetivo de la reunión:	Definir el alcance del Trabajo Final de Graduación, revisar los objetivos específicos y establecer la estructura metodológica del proyecto.		
Participantes:	**Presentes:** Estudiante y profesor tutor		
	Ausentes: Ninguno		

Temas tratados

No.	**Asunto**	**Comentarios**	**Acuerdos**
1	Definición del alcance del proyecto	Se discutieron los niveles de alcance del proyecto, diferenciando entre diseño, desarrollo e implementación. Se recomendó que el objetivo sea desarrollar una propuesta de solución, sin comprometer la implementación en la organización.	Ajustar el alcance del proyecto en el documento del TFG para enfocarse en una propuesta viable sin riesgos de implementación inmediata.
2	Revisión de objetivos específicos	Se identificó la necesidad de separar el diseño del proceso y la incorporación de herramientas tecnológicas. Se debaten posibles ajustes en la redacción.	Modificar los objetivos específicos para reflejar claramente los enfoques de análisis del proceso actual, definición del nuevo proceso y propuesta de automatización.
3	Medición y evaluación de eficiencia	Se estableció la importancia de definir métricas para evaluar el impacto del nuevo proceso en tiempo y errores operativos. Se debatió si incluir costos financieros.	Definir indicadores clave de desempeño (KPIs) para medir la mejora del proceso, sin incluir costos financieros en los objetivos.

4	Beneficios directos del proyecto	Se identificaron mejoras en eficiencia operativa, reducción de errores y confiabilidad de datos como beneficios clave. Se ajustó el lenguaje en los objetivos para reflejar estos beneficios.	Redactar con precisión los beneficios en la documentación del TFG y evitar afirmaciones sobre implementación futura.
5	Metodología y marcos de referencia	Se discutieron *frameworks* relevantes como CMMI y DMBOK, así como herramientas de recolección de datos (entrevistas, análisis documental y FODA).	Determinar los *frameworks* a utilizar y documentarlos en la metodología del proyecto.
6	Plan de trabajo y cronograma	Se aclaró que el TFG consta de 13 semanas efectivas y las primeras 6 son clave. Se sugirió realizar reuniones adicionales y mantener una bitácora de trabajo.	Definir un plan de trabajo detallado y registrar el tiempo dedicado al TFG en una bitácora.
7	Próximos pasos	Se estableció la necesidad de modificar los objetivos específicos, actualizar el alcance y preparar una reunión con la empresa.	Enviar los cambios en los objetivos y beneficios antes del sábado 8 de marzo. Gestionar la coordinación de la próxima reunión con la empresa.

Próxima reunión		
Temas por tratar	Fecha	**Convocados**
Explicación de los roles en el TFG, incluyendo las funciones del tutor académico, el estudiante y la empresa en el desarrollo del proyecto.	11 de marzo del 2025	Profesor tutor Estudiante Representante de la empresa
Firma de profesor tutor		**Firma estudiante**

9.10 Apéndice J. Minuta de Reunión #7

Reunión No.	**07**	**Fecha:**	
Lugar:	Zoom Meetings	Día:	11 de marzo de 2025
		Hora de inicio:	01:30 pm
		Hora de finalización:	02:00 pm
Objetivo de la reunión:	Establecer la relación entre la empresa y el Instituto Tecnológico de Costa Rica en el marco del Trabajo Final de Graduación (TFG), clarificando los roles de los participantes en el proceso.		
Participantes:	**Presentes:** Estudiante, profesor tutor y representante de la empresa		
	Ausentes: Ninguno		

Temas tratados			
No.	**Asunto**	**Comentarios**	**Acuerdos**

1	Explicación de los roles en el TFG	El profesor tutor explicó la estructura del TFG y el papel de cada participante.	Se confirmó la comprensión de los roles y sus responsabilidades.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
Revisión de ajustes en la documentación del TFG, refinamiento de los objetivos y beneficios directos, y definición del enfoque para la validación del prototipo de solución.		12 de marzo del 2025	Profesor tutor Estudiante
Firma de profesor tutor			**Firma estudiante**
Firma representante de la empresa			

9.11 Apéndice K. Minuta de Reunión #8

Reunión No.	**08**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	12 de marzo de 2025
		Hora de inicio:	08:00 pm
		Hora de finalización:	09:00 pm
Objetivo de la reunión:	Revisar los ajustes en la documentación del TFG, definir mejoras en los objetivos y beneficios directos, y establecer la estructura para la validación del prototipo de solución.		
Participantes:	**Presentes:** Estudiante y profesor tutor		
	Ausentes: Ninguno		
Temas tratados			

No.	**Asunto**	**Comentarios**	**Acuerdos**
1	Correcciones en la documentación	Se revisaron cambios en la numeración, estructura y redacción de la documentación del TFG.	Se acordó mejorar la organización de viñetas y numeración, así como la coherencia en la redacción.
2	Definición de terminología	Se discutió la nomenclatura de licencias de datos y acceso a información en la empresa.	Se definieron términos precisos para describir la gestión de licencias de acceso a datos financieros y de otros tipos.
3	Ajustes en los objetivos del TFG	Se modificó el segundo objetivo para enfocarse en un marco conceptual e integración de herramientas de automatización.	Se estableció la creación de un cuarto objetivo para la validación del prototipo de solución.
4	Prototipo de solución	Se definió que el prototipo debe incluir los elementos clave del proceso, roles, herramientas y estructura del flujo de datos.	Se mantendrá el término "prototipo de solución" por considerarse más formal que "propuesta de solución".

5	Validación del prototipo	Se analizó la importancia de realizar pruebas y validaciones del prototipo sin implementación real.	Se agregó un objetivo específico para la validación del prototipo.
6	Beneficios directos	Se ajustaron los beneficios directos del TFG para alinearlos con los nuevos objetivos.	Se revisará el impacto de la validación en los beneficios definidos.
7	Plan de trabajo	Se estableció la necesidad de cerrar el primer capítulo del TFG para avanzar al marco conceptual.	Se enviará la versión final del capítulo 1 antes del viernes 14 de marzo para revisión.
8	Minutas de reuniones	Se recordó la importancia de documentar todas las reuniones y obtener firmas a tiempo.	El estudiante se comprometió a actualizar y completar las minutas pendientes.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
Seguimiento al avance del Trabajo Final de Graduación (TFG) con la empresa.		8 de abril, 2025	Profesor tutor Estudiante Representante de la organización
Firma de profesor tutor			**Firma estudiante**

9.12 Apéndice L. Minuta de reunión #9

Reunión No.	**09**	**Fecha:**	
Lugar:	Zoom Meetings	Día:	8 de abril, 2025
		Hora de inicio:	01:30pm
		Hora de finalización:	02:00pm
Objetivo de la reunión:	Dar seguimiento al avance del Trabajo Final de Graduación (TFG) y explicar al representante de la empresa el procedimiento para realizar la primera evaluación organizacional.		
Participantes:	**Presentes:** Profesor tutor, representante de la organización, estudiante		
	Ausentes: Ninguno		
Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Seguimiento del TFG con la empresa	Se presentaron los avances actuales del proyecto a la empresa.	La empresa ratificó su compromiso de seguir apoyando el desarrollo del proyecto.
2	Explicación del proceso de evaluación	El tutor académico explicó el procedimiento que debe seguir la empresa para realizar la	Se acordó que el representante de la empresa completará la evaluación según el formato oficial provisto por el ITCR.

		primera evaluación del estudiante.	
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
N/A		N/A	N/A
Firma de profesor tutor			**Firma estudiante**
Firma representante de la empresa			

9.13 Apéndice M. Plantilla de entrevista semiestructurada

Entrevista No. #			
Entrevistador:		**Fecha:**	
		Hora inicio:	
		Hora fin:	
Entrevistado:		**Rol:**	
		Equipo:	
Propósito:			
Preguntas:			
Observaciones:			

Evidencia:

9.14 Apéndice N. Entrevista técnica sobre el proceso actual de carga de datos

Entrevista No. 1			
Entrevistador:	Kemuel Chavarría (estudiante)	**Fecha:**	23 de abril, 2025
		Hora inicio:	10:00 am
		Hora fin:	11:00 am
Entrevistado:	Diego Quirós	**Rol:**	*Senior Data Engineer*
		Equipo:	*Data Operations*
Propósito:	Obtener información técnica y operativa clave sobre el proceso actual de carga de datos en la plataforma de Gestión de Datos Maestros, con el fin de identificar deficiencias, validar métricas operativas y evaluar oportunidades de mejora alineadas a los objetivos de la Fase 1 del Trabajo Final de Graduación.		
Preguntas:			

1. ¿Cuáles son las etapas principales del proceso de carga de datos desde tu perspectiva técnica?

Desde la perspectiva técnica del Senior Data Engineer, el proceso actual de carga de datos se compone de diversas etapas manuales y fragmentadas. Inicialmente, los archivos son recibidos mediante canales como correo electrónico o descarga desde protocolos de red (FTP). Luego, se almacenan de forma temporal en un repositorio compartido (SharePoint). Posteriormente, los datos se procesan localmente en hojas de Excel, donde se realiza la consolidación manual de registros. Finalmente, la información es cargada a la base de datos de la plataforma MDM mediante consultas estructuradas escritas en SQL.

2. ¿Qué herramientas se utilizan en cada etapa?

Las herramientas involucradas en el proceso incluyen:

- Correo electrónico y FTP, como canales de entrada de datos.
- Microsoft SharePoint, utilizado para el almacenamiento intermedio de archivos.
- Microsoft Excel, empleado para el procesamiento y consolidación de la información.
- PostgreSQL, como motor de base de datos para la ejecución de sentencias de carga.
- Finalmente, la plataforma MDM de destino a la cual se realiza la carga final de los datos estructurados.

Entrevista No. 1

3. ¿Cuánto tiempo toma, en promedio, cada etapa del proceso (descarga, procesamiento, carga en MDM)?

El tiempo promedio estimado para ejecutar el ciclo completo de carga es de aproximadamente 29 horas por analista (equivalentes a 3.6 días hábiles por ciclo). La descarga inicial toma cerca de 3 horas, mientras que las tareas de procesamiento, validación y revisión manual representan el mayor consumo de tiempo, abarcando alrededor del 80 % del total estimado. Finalmente, la carga en la plataforma de Gestión de Datos Maestros requiere aproximadamente 2 horas, aunque este valor puede variar según la calidad y consistencia de los datos consolidados.

4. ¿Cuántos archivos se procesan por ciclo? ¿Qué volumen tienen estos archivos (en filas promedio)?

La cantidad y volumen de archivos procesados por ciclo varía significativamente. En promedio, se manejan entre 50.000 y varios millones de registros por archivo. En algunos casos, los archivos más grandes deben dividirse en lotes debido a limitaciones técnicas en su procesamiento. Esta variabilidad representa un reto para mantener una ejecución uniforme del proceso.

5. ¿Qué tareas realizas manualmente en cada etapa?

Las tareas manuales realizadas en el procesamiento incluyen la validación de duplicados, detección de valores faltantes (NA) en columnas críticas, y revisión general de consistencia. Adicionalmente, se lleva a cabo una estandarización de datos, asegurando que las filas cumplan con los identificadores requeridos, aunque estos no siempre sean llaves primarias formales. Estas acciones requieren un criterio técnico por parte del analista y se ejecutan fuera de un sistema automatizado.

6. ¿Cuáles de esas tareas consideras críticas por su complejidad o riesgo de error?

El Senior Data Engineer identificó como tarea crítica la validación de calidad de los datos, particularmente la detección de duplicaciones y valores vacíos en columnas específicas. Estas validaciones son esenciales, ya que un error en esta etapa puede comprometer la integridad del proceso completo. Al no existir una verificación automatizada previa a la carga, el riesgo de omisiones o inconsistencias recae totalmente en la revisión manual del analista, lo cual eleva la probabilidad de error humano.

7. ¿Existe alguna guía o procedimiento formal para estas tareas?

Existe una guía básica no exhaustiva que sirve como referencia para los analistas. Este documento describe los aspectos clave a revisar dentro de cada archivo, como la ausencia de duplicados o la presencia de datos faltantes en columnas obligatorias. Sin embargo, dicha guía no estandariza completamente los procedimientos ni asegura uniformidad en los criterios aplicados por cada colaborador.

8. ¿Qué pasos consideras ineficientes o repetitivos? ¿Por qué?

Desde la perspectiva del entrevistado, todos los pasos que dependen de la intervención humana deben considerarse ineficientes. Actividades como la descarga, validación, revisión manual,

Entrevista No. 1

consolidación y carga requieren una inversión considerable de tiempo y están sujetas a errores. La repetitividad y la ausencia de mecanismos automatizados generan una dependencia excesiva del conocimiento empírico del colaborador, dificultando la escalabilidad del proceso.

9. ¿Qué errores específicos suelen ocurrir durante la carga de datos en la base de datos de la plataforma MDM?

Los errores más frecuentes durante la carga en la plataforma MDM están relacionados con registros duplicados y campos incompletos (NA). Además, existen dificultades en la identificación de entidades debido a inconsistencias entre los identificadores únicos de diferentes fuentes de datos. Por ejemplo, una misma entidad puede estar nombrada de forma distinta en herramientas como Corporate Intelligence Database o NewsEdge, lo que complica el mapeo adecuado entre registros y afecta la calidad de la carga.

10. ¿Estos errores son detectados de inmediato o después de la carga?

Generalmente, los errores se detectan después de la carga, cuando el analista realiza una revisión manual de los datos ya almacenados en la base. No existen mecanismos automatizados ni reglas de validación internas en MDM que permitan anticipar todos los errores durante la carga. Como resultado, los errores se identifican de forma reactiva, lo cual implica un retraso en el ciclo de revisión y reprocesamiento.

11. ¿Cuáles son las causas más comunes de estos errores?

Según el entrevistado, las causas más comunes de los errores en la carga de datos derivan de la falta de estandarización entre las distintas fuentes de información. Cada proveedor utiliza su propio formato y convenciones de identificación, lo que dificulta el mapeo entre entidades equivalentes. Por ejemplo, un mismo dato maestro puede tener diferentes identificadores y nombres entre sistemas como Corporate Intelligence Database y NewsEdge. Estas diferencias semánticas generan conflictos al consolidar los datos en la plataforma MDM.

12. ¿Con qué frecuencia se presentan errores en la carga de datos (por semana o por ciclo)?

Los errores se presentan de manera constante en cada ciclo de carga. No existe un ciclo libre de errores, ya que siempre se detectan inconsistencias, duplicaciones o campos faltantes que requieren revisión y validación por parte del analista. Esta recurrencia representa un reto persistente que retrasa la disponibilidad operativa de los datos.

13. ¿Cómo se corrigen los errores de carga? ¿Se repite todo el proceso o solo una parte?

La corrección de errores no implica repetir todo el proceso completo, sino únicamente la parte afectada. En caso de que no se logre resolver la inconsistencia localmente, se emite un ticket al proveedor correspondiente para que verifique si el error proviene de un identificador incorrecto o de un registro mal ingresado. Este flujo de corrección requiere coordinación externa y añade complejidad al proceso.

14. ¿Existen mecanismos automatizados o manuales para verificar la calidad de los datos una vez cargados?

Entrevista No. 1

No existen mecanismos automatizados robustos para verificar la calidad de los datos una vez que estos han sido cargados en MDM. Más allá de algunas reglas básicas que la estructura del sistema permite configurar, la verificación depende exclusivamente de la revisión manual del analista. Esto limita la capacidad de respuesta inmediata ante errores y retrasa la validación final del ciclo de carga.

15. ¿Cuánto tiempo adicional toma corregir errores o realizar reprocesos?

La mayor parte del tiempo del ciclo de carga se consume en actividades de revisión y corrección de errores. De los tres días estimados para completar el proceso, entre dos y dos días y medio se dedican exclusivamente a validar manualmente los datos, identificar inconsistencias y ajustar los archivos. Esta proporción refleja un uso ineficiente del tiempo operativo debido a la falta de automatización.

16. ¿Cómo ha evolucionado el volumen de datos en los últimos meses?

El entrevistado señaló que el volumen de datos ha experimentado un crecimiento constante. Aunque dicho crecimiento no es abrupto, sí representa un incremento sostenido en cada ciclo, con la incorporación regular de entre mil y cinco mil nuevas entidades. Este aumento continuo, aunque progresivo, ejerce presión sobre la capacidad operativa del proceso actual, especialmente considerando su carácter manual.

17. ¿Crees que el proceso puede mantenerse igual si el volumen aumenta? ¿Hasta qué punto?

El proceso, en su estado actual, es funcional pero presenta una escalabilidad limitada. Según el entrevistado, un incremento adicional del volumen en un 10% o 20% afectaría directamente el tiempo de procesamiento, alargando potencialmente el ciclo de tres a cuatro días. Al estar basado en la ejecución manual y depender del criterio individual del analista, el proceso no se adapta fácilmente a un entorno de crecimiento acelerado, lo que refuerza la necesidad de automatización.

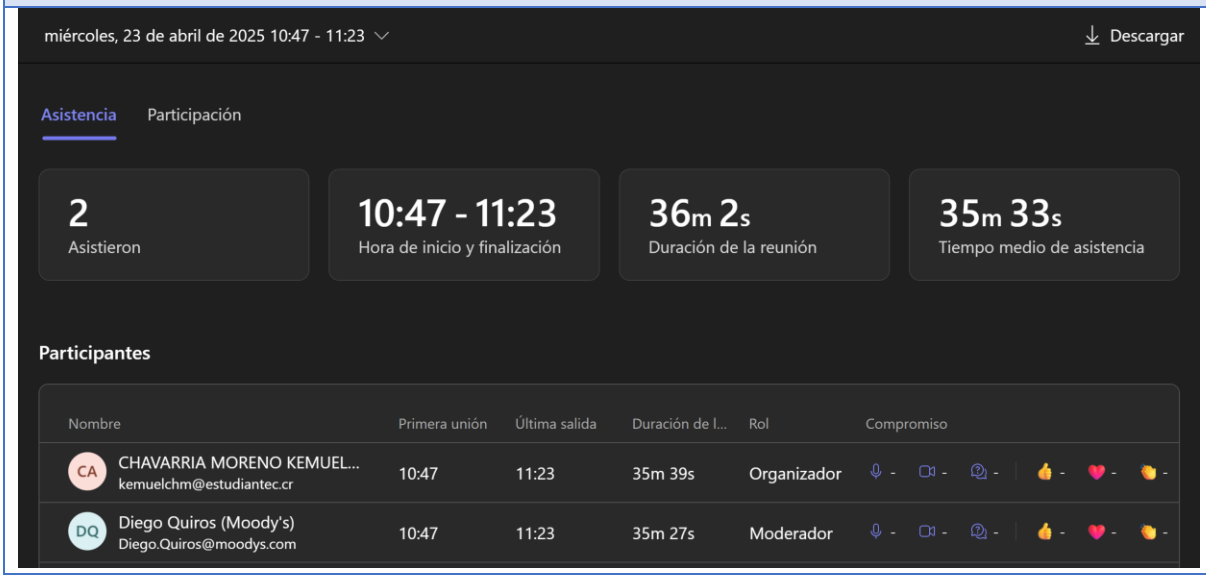
18. ¿Qué tareas crees que deberían automatizarse primero para mejorar la eficiencia?

El entrevistado fue claro en señalar que las tareas relacionadas con la validación de datos deben ser las primeras en ser automatizadas. Esto incluye verificaciones de calidad y consistencia, como la detección de duplicaciones, campos vacíos y errores estructurales. Al automatizar estas validaciones, se reduciría significativamente la carga de trabajo manual y se minimizaría el margen de error, lo que permitiría un proceso más eficiente y escalable.

Observaciones:

- El proceso actual no se sostiene ante aumentos de volumen mayores al 20%.
- La consolidación de datos representa el mayor cuello de botella.
- No existen métricas automatizadas ni alertas para errores.
- La guía actual utilizada es informal y carece de estandarización.

Entrevista No. 1
Evidencia:



9.15 Apéndice O. Plantilla de Revisión Documental

Ficha No. #	
Tipo de documento:	Libro, artículo científico, sitio web, ensayo, reporte técnico, documentación organizacional.
Autor:	
Título:	
Fuente en APA:	
Palabras clave:	
Descripción general del documento:	
Observaciones:	
Contiene definiciones clave, teorías, datos relevantes, resultados, ejemplos aplicados, herramientas de medición, hipótesis validadas, y enfoques del tema estudiado.	

9.16 Apéndice P. Revisión documental de plantilla de validación de datos

Ficha No. #	
Tipo de documento:	Documentación organizacional en Confluence (guía interna no estandarizada)
Autor:	*Data Operations Team*
Título:	*Data Review Checklist for Master Data Loading Process*
Fuente en APA:	Data Operations Team. (2025). *Data review checklist for master data loading process*. Internal document.

Palabras clave:	*Manual validation, data quality, data loading, MDM, SharePoint, Excel processing.*
Descripción general del documento:	
Este documento constituye una guía operativa utilizada por el equipo de *Data Operations* para validar manualmente los archivos antes de proceder con la carga en la plataforma de Gestión de Datos Maestros (MDM). La guía establece una lista de verificación básica que incluye aspectos como la recepción de archivos CSV, la integridad de su estructura, el formato de datos y la consolidación en Excel. La revisión incluye además controles previos a la carga mediante herramientas manuales, sin que exista un protocolo formal o automatizado. La guía no ha sido sometida a procesos de revisión formal ni cuenta con control de versiones.	
Observaciones:	
El documento evidencia la informalidad en los procesos actuales de validación, destacando la ausencia de estandarización y el alto grado de dependencia del criterio individual. No incluye métricas formales ni herramientas de medición, aunque recoge ejemplos prácticos sobre errores recurrentes y enfoques empíricos de control. Refuerza las debilidades identificadas en la Fase 1 del análisis, vinculadas a la falta de procesos sistematizados y a la necesidad de implementar mecanismos de validación automatizados para garantizar la calidad de los datos cargados.	

9.17 Apéndice Q. Revisión documental de plantilla de documentación de errores

Ficha No. #	
Tipo de documento:	Documentación organizacional en Confluence (formato interno para control de errores)
Autor:	*Data Operations Team*
Título:	*Data Load Error Documentation Template*
Fuente en APA:	Data Operations Team. (2025). *Data load error documentation template*. Internal document.
Palabras clave:	*Error tracking, data load process, MDM, manual correction, data quality, operational control.*
Descripción general del documento:	
La plantilla revisada corresponde a un formato interno utilizado por el equipo de *Data Operations* para documentar los errores detectados durante el proceso manual de carga de datos en la plataforma MDM. La estructura de la plantilla incluye secciones específicas para registrar información detallada sobre el tipo de error, su ubicación, método de detección, acciones correctivas aplicadas y análisis preliminar de causa raíz. Además, permite adjuntar evidencias complementarias como archivos, capturas de pantalla o registros SQL. Aunque cumple una función operativa esencial, la plantilla no se encuentra integrada dentro de un sistema formal de gestión de calidad y tampoco es parte de un procedimiento normado.	
Observaciones:	

Este documento refleja el intento del equipo por sistematizar la documentación de errores, aunque bajo un esquema manual y no estandarizado. Proporciona ejemplos prácticos de cómo se gestionan las fallas en el flujo actual. La existencia de esta plantilla respalda las deficiencias señaladas en la Fase 1, especialmente la falta de automatización y control formal sobre los errores, así como la necesidad de establecer mecanismos más trazables de seguimiento de incidentes operativos.

9.18 Apéndice R. Entrevista técnica sobre rediseño del proceso de carga de datos y Automatización

Entrevista No. 2			
Entrevistador:	Kemuel Chavarría (estudiante)	**Fecha:**	24 de abril, 2025
		Hora inicio:	02:00 pm
		Hora fin:	02:50 pm
Entrevistado:	Diego Quirós	**Rol:**	*Senior Data Engineer*
		Equipo:	*Data Operations*
Propósito:	Obtener insumos técnicos, operativos y estratégicos para el rediseño del proceso automatizado de carga de datos. La entrevista se centró en explorar oportunidades de mejora, herramientas tecnológicas viables, puntos críticos del flujo actual, criterios de selección tecnológica, y métricas clave para evaluar la nueva solución.		
Preguntas:			

Oportunidades de Mejora y Reestructuración

¿Cuáles etapas del proceso actual deben automatizarse o eliminarse?

El entrevistado indicó que el primer paso lógico hacia la automatización consiste en eliminar la gestión manual de la obtención de archivos. Esto implica sustituir la descarga manual desde SharePoint, correos electrónicos o FTP por mecanismos automáticos que ubiquen y transfieran los archivos hacia un repositorio común en la nube, definido como una zona de aterrizaje (*Landing Zone*) en Amazon Web Services (AWS). Esta automatización inicial permitiría que el analista trabaje directamente con los datos ya centralizados, mejorando de inmediato la eficiencia del flujo operativo.

¿Qué tareas son más repetitivas o críticas y podrían reestructurarse para mejorar eficiencia?

Las tareas más repetitivas señaladas por el ingeniero son la unificación y estandarización de los datos, especialmente durante la fase de carga. Desde su perspectiva, el inicio debe centrarse en SharePoint, por su menor complejidad, seguido de correos electrónicos y, finalmente, APIs, las cuales requieren mayor nivel de personalización. Estas tareas, al estar distribuidas entre varias fuentes con estructuras disímiles, representan un cuello de botella operativo que debería resolverse mediante funciones Lambda y scripts en Python.

¿Qué errores frecuentes podrían evitarse con automatización?

Entre los errores más comunes que podrían eliminarse mediante automatización, el entrevistado mencionó los campos vacíos y las discrepancias entre registros. Al no depender del análisis manual, un proceso automatizado tendría la capacidad de revisar grandes

Entrevista No. 2

volúmenes de datos de forma uniforme y constante, detectando errores desde el inicio. Además, si se llegara a generar una excepción, esta podría registrarse en un archivo de *logs* para su análisis posterior por parte del equipo.

¿Cuáles son los puntos críticos que deberían tener prioridad en la mejora?

Se identificaron dos puntos críticos que requieren atención prioritaria. El primero es la extracción automatizada de archivos, ya que actualmente se realiza de manera manual. El segundo, aún más relevante en términos de consumo de tiempo, es el procesamiento de datos. En esta fase, los analistas realizan múltiples tareas manuales que podrían ser replicadas y reemplazadas por funciones Lambda dentro del ecosistema de AWS. Este cambio reduciría significativamente el tiempo de ejecución y aumentaría la consistencia operativa.

¿Qué impacto tendría en el equipo si esas tareas fueran automatizadas?

Según el entrevistado, la automatización de estas tareas liberaría una parte considerable del tiempo operativo del equipo, permitiendo que se enfoquen en actividades de mayor valor agregado, como revisiones de calidad de datos. Actualmente, muchas de estas tareas se encuentran postergadas debido a la prioridad que exige el proceso de carga manual. Automatizarlo permitiría restaurar ese equilibrio y fortalecer la gobernanza de datos.

Herramientas y Tecnologías Compatibles

¿Qué herramientas o tecnologías sugieres que sean compatibles con nuestra infraestructura?

El entrevistado recomendó priorizar herramientas que ya cuenten con soporte organizacional y licencias activas, como es el caso de Amazon Web Services (AWS). Mencionó que la empresa mantiene un convenio de colaboración (*partnership*) con esta plataforma, lo cual facilita el acceso a recursos técnicos, materiales de capacitación y soporte organizacional. Si bien reconoció que existen alternativas viables como Google Cloud o Microsoft Azure, destacó que el uso de AWS representa una ventaja estratégica debido a su disponibilidad inmediata y madurez tecnológica.

¿Qué limitaciones técnicas debemos considerar al seleccionar nuevas herramientas?

Desde su experiencia, no se identifican limitaciones técnicas significativas para implementar herramientas modernas como las que ofrece AWS. No obstante, señaló que es fundamental validar que cada componente requerido por el proceso esté disponible o sea factible de implementar dentro del ecosistema seleccionado. Dado el alto grado de madurez de estas plataformas, es razonable esperar que muchas de las funcionalidades requeridas; como extracción, validación y procesamiento ya se encuentren incorporadas o sean fácilmente desarrollables.

¿Existen herramientas ya disponibles en la organización que podríamos aprovechar?

Sí. El entrevistado afirmó que actualmente se dispone de acceso organizacional a diversas soluciones dentro de AWS, incluyendo ambientes de prueba, recursos de cómputo, y herramientas para la automatización y almacenamiento. Además, se cuenta con licencias activas y entornos de capacitación formal que pueden ser aprovechados para la implementación de una solución sostenible y alineada con la infraestructura corporativa.

¿Qué integraciones deben mantenerse con sistemas actuales como SharePoint o PostgreSQL?

En cuanto a las integraciones existentes, se resaltó que algunas herramientas, como SharePoint o FTP, no serán reemplazadas, ya que actúan como fuentes originales de los datos. Lo que

Entrevista No. 2

cambiará es el mecanismo mediante el cual se extrae y procesa dicha información. En el caso de PostgreSQL, su integración debe mantenerse, dado que actualmente alberga el modelo de datos implementado en MDM. Esto asegura la compatibilidad con la plataforma MDM utilizada por la organización y garantiza la continuidad operacional del flujo actual.

Requerimientos Técnicos y Funcionales Clave

¿Qué características técnicas debe cumplir la solución para manejar el volumen y seguridad de datos esperados?

El entrevistado subrayó que la solución debe contar con capacidades de escalabilidad robustas, como las que ofrece AWS, para manejar volúmenes elevados de datos sin degradar el rendimiento. Señaló que servicios como AWS Glue y Lambda permiten ejecutar procesos de extracción, transformación y carga de manera distribuida y eficiente, incluso con archivos que contienen cientos de miles o millones de registros. Adicionalmente, destacó que la solución debe ofrecer mecanismos integrados para la gestión de seguridad, incluyendo control de accesos, encriptación de datos y trazabilidad de las operaciones ejecutadas.

¿Qué validaciones automáticas serían esenciales para asegurar la calidad de los datos cargados?

El entrevistado indicó que las funciones Lambda deberán incorporar validaciones automatizadas equivalentes a las que actualmente realiza el equipo de analistas. Estas incluyen la detección de campos vacíos, verificación de formatos incorrectos, y validación de valores atípicos en columnas clave. Automatizar estos controles dentro del flujo garantizaría que los errores sean identificados tempranamente, sin necesidad de revisión manual posterior.

¿Qué tareas deberían ejecutarse sin intervención manual en el nuevo flujo?

Idealmente, todas las tareas del flujo deberían ser automatizadas. No obstante, el Senior Data Engineer sugirió iniciar con la automatización de la extracción de archivos y la consolidación de datos, ya que representan aproximadamente el 60% del trabajo operativo actual. Estas tareas son repetitivas, intensivas en tiempo y altamente susceptibles a errores humanos, lo que justifica su priorización en el rediseño del proceso.

¿Qué tipos de reportes o alertas debería generar la solución para facilitar el monitoreo?

Se propuso que el sistema genere *logs* estructurados que documenten cada ejecución, incluyendo errores detectados y estatus del proceso. Estos registros podrían almacenarse en una base de datos y visualizarse mediante reportes en Power BI. La revisión de estos reportes podría realizarse con una frecuencia diaria, semanal o mensual, según la criticidad del proceso y la periodicidad de su ejecución.

¿Cuál sería un tiempo de procesamiento aceptable por carga dentro del nuevo flujo?

Actualmente, el ciclo completo toma aproximadamente tres días. El entrevistado indicó que un objetivo razonable para la nueva solución sería reducir este tiempo a menos de un día, sin sacrificar la calidad del proceso. Esta mejora sería viable gracias a la ejecución paralela de funciones y al tratamiento automatizado de grandes volúmenes de datos en la nube.

¿Qué restricciones operativas debemos tener en cuenta al automatizar el proceso?

No se identificaron restricciones operativas críticas dentro del equipo actual que limiten la automatización. El entrevistado confirmó que existe libertad técnica para implementar la solución propuesta, lo cual elimina barreras organizativas que pudieran retrasar su ejecución.

Métricas Clave del Nuevo Proceso

¿Qué métricas consideras clave para evaluar el desempeño del nuevo proceso?

Entrevista No. 2

El Senior Data Engineer identificó como métricas clave la cantidad de errores reportados a los proveedores de datos y el tiempo total requerido para completar el ciclo de carga. En el proceso actual, dicho ciclo puede extenderse hasta tres días. La expectativa para el nuevo diseño es que el tiempo se reduzca a menos de un día, lo cual implicaría una mejora significativa en desempeño. La reducción de errores y la disminución del tiempo de ejecución representan, en conjunto, los principales indicadores de éxito.

¿Qué indicadores podríamos utilizar para medir la eficiencia tras la automatización?

El Senior Data Engineer sugirió utilizar indicadores relacionados con la reducción de horas-hombre empleadas en tareas manuales, así como un análisis cruzado entre el tiempo de ejecución y la cantidad de errores registrados. Propuso incluso emplear una visualización tipo cuadrante (inspirada en modelos como los cuadrantes de Gartner), que permita representar simultáneamente la eficiencia temporal y la mejora en calidad de datos.

¿Cómo deberíamos monitorear la calidad de los datos cargados (ej. porcentaje de errores, tiempos de carga)?

La calidad de los datos cargados puede monitorearse mediante el porcentaje de registros con errores, la frecuencia de ocurrencia de valores nulos o duplicados, y el tiempo requerido para procesar cada lote de datos. Estas métricas deben capturarse de forma automática y almacenarse en registros que permitan generar reportes periódicos para su análisis por parte del equipo.

¿Qué frecuencia de revisión de métricas sería adecuada para asegurar el control continuo?

El Senior Data Engineer recomendó revisar las métricas cada vez que se ejecute el proceso completo, lo cual sucede de forma mensual. Esta revisión permitiría identificar patrones, evaluar la estabilidad del flujo automatizado y tomar decisiones de mejora continua a partir de datos concretos.

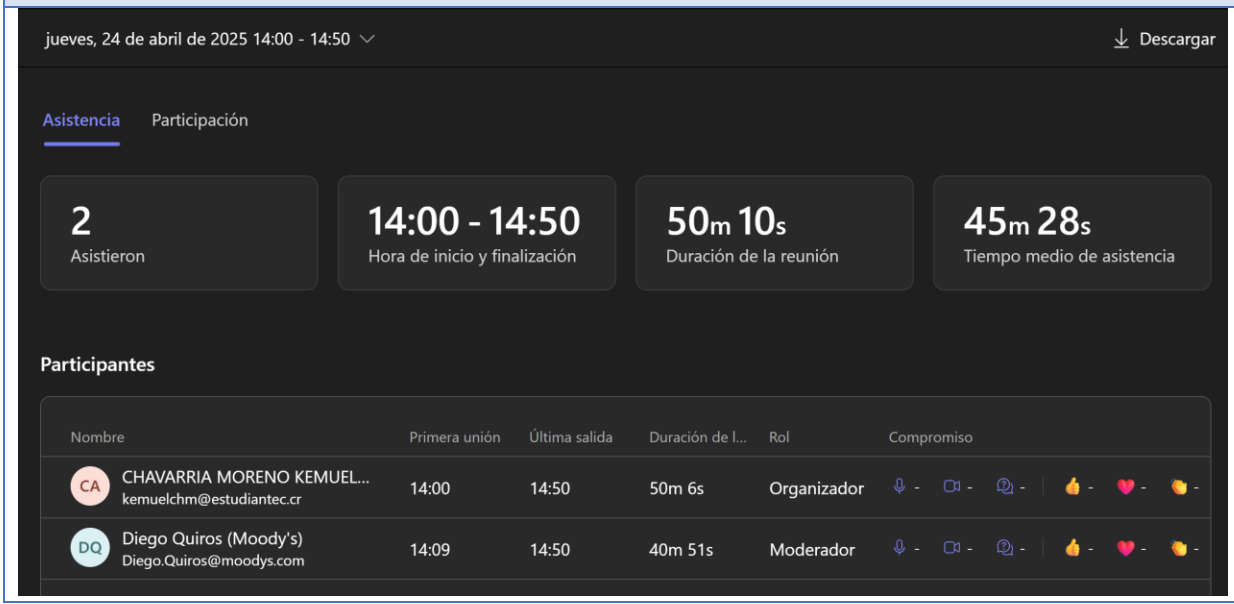
¿Existen métricas operativas que el equipo ya utilice y deban mantenerse o adaptarse en el nuevo flujo?

Actualmente, se realiza un seguimiento no documentado del tiempo de ejecución del proceso. Aunque no se formaliza como una métrica operativa, el equipo es consciente de esta duración y la utiliza como referencia interna. Se sugiere incorporar esta métrica dentro del nuevo diseño para asegurar su trazabilidad y facilitar la comparación histórica.

Observaciones:

- Los entrevistados manifestaron una posición estratégica favorable hacia el uso de AWS, respaldada por el *partnership* organizacional de la empresa con Amazon Web Services.
- Se recomendó iniciar la automatización por las etapas más críticas y repetitivas.
- Se identificaron errores frecuentes que podrían eliminarse con validaciones automáticas: campos vacíos, duplicados y valores erróneos (por ejemplo, identificadores incorrectos).
- La automatización liberaría tiempo a los analistas para enfocarse en tareas de mayor valor como el aseguramiento de calidad de datos.
- Se enfatizó que el proceso actual carece de trazabilidad sistematizada.

Entrevista No. 2
Evidencia:



9.19 Apéndice S. Minuta de reunión #10

Reunión No.	**10**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	26 de abril, 2025
		Hora de inicio:	11:11am
		Hora de finalización:	11:38am
Objetivo de la reunión:	Discutir con el tutor los riesgos metodológicos y técnicos en el diseño del TFG, en especial sobre la validez de proyecciones de eficiencia en procesos teóricos, la estructura del análisis de resultados de la Fase 1, y aclarar el concepto y documentación del prototipo funcional requerido en la Fase 3 y 4.		
Participantes:	**Presentes**: Profesor tutor y estudiante		
	Ausentes: Ninguno		

Temas tratados

No.	**Asunto**	**Comentarios**	**Acuerdos**
1	Proyección de eficiencia en procesos teóricos	Se discutió la variable "eficiencia" en el objetivo específico 2. Michael señaló que no es válido proyectar eficiencia basándose únicamente en el diseño BPMN sin respaldo numérico ni estadísticas.	Se acordó reformular la variable hacia un enfoque más cualitativo o justificar con métricas proyectadas únicamente si se dispone de suficiente evidencia.
2	Calidad del respaldo de métricas actuales	Kemuel presentó como base una entrevista con el *Senior Data Engineer,* donde se mencionan duraciones estimadas de	Se recomienda buscar literatura adicional, tesis comparables o registros históricos de tiempos para mejorar la solidez de la proyección.

		hasta 3 días por ciclo. Michael lo consideró insuficiente para respaldo estadístico confiable.	
3	Mejora del análisis de resultados	Kemuel explicó que la Fase 1 incluye descripción del proceso actual, hallazgos técnicos, discusión crítica, métricas, y análisis FODA. Michael consideró que la narrativa está bien construida.	Michael sugiere incluir un diagrama de Ishikawa (causa-efecto) antes del FODA para fortalecer el diagnóstico de origen de los problemas.
4	Concepto y alcance del prototipo funcional	Se debatió qué significa "prototipo" en el contexto del TFG. Michael explicó que un prototipo debe representar una versión tangible y funcional parcial de la solución, aunque no esté en producción.	Se acordó que Kemuel debe documentar claramente qué partes del prototipo están implementadas y cuáles son proyectadas, usando evidencia técnica y capturas.
5	Forma de documentar el prototipo	Se propuso usar diagramas, capturas de pantallas, scripts en Python o código comentado que explique cómo se realizaría cada paso del nuevo proceso automatizado.	Michael avala que esto es válido como documentación del prototipo funcional siempre que esté bien descrito y alineado al flujo TO-BE propuesto.
6	Evaluación del prototipo en la Fase 4	Se habló de cómo documentar la efectividad del prototipo, midiendo reducción de tareas manuales, tiempos de ejecución y validación de estructuras de datos.	Se confirma que la evaluación de la Fase 4 será el insumo principal para mostrar mejoras concretas y justificar los resultados del prototipo.
Próxima reunión			
Temas por tratar N/A		**Fecha** **N/A**	**Convocados** N/A

9.20 Apéndice T. Minuta de Reunión #11

Reunión No.	**11**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	28 de abril, 2025
		Hora de inicio:	03:23pm
		Hora de finalización:	03:42pm

Objetivo de la reunión:	Validar el diseño técnico del proceso automatizado de carga de datos en AWS, establecer una arquitectura viable para el prototipo, definir herramientas esenciales.		
Participantes:	**Presentes**: Product Owner, Senior Data Engineer, Project Manager y Estudiante		
	Ausentes: Ninguno		
Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Validación del rediseño técnico del proceso de carga de datos	El estudiante presentó la propuesta de diseño a miembros del equipo *Data Operations*	Se aprueba oficialmente el diseño técnico y se autoriza la creación del prototipo en AWS.
Próxima reunión			
Temas por tratar	**Fecha**	**Convocados**	
N/A	N/A	N/A	

9.21 Apéndice U. Minuta de reunión #12

Reunión No.	**12**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	30 de abril, 2025
		Hora de inicio:	02:00pm
		Hora de finalización:	02:33pm
Objetivo de la reunión:	Revisar y validar la versión final del diseño funcional del flujo automatizado Además, se analizó el modelo de costos para la estimación del análisis costo-beneficio del prototipo.		
Participantes:	**Presentes**: Product Owner, Senior Data Engineer, Project Manager y Estudiante		
	Ausentes: Ninguno		
Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Revisión del diseño funcional completo	Se repasaron todas las etapas del flujo de datos, desde la extracción hasta la carga en Aurora. Se confirmó el uso correcto de Step Functions, Lambda, entre otros componentes	Se aprueba la lógica del diseño como base oficial del prototipo a documentar en el TFG.
2	Estimación de costos y análisis financiero	Se discutió la estimación del costo operativo de AWS. Se comparó contra el costo del proceso actual.	Se autoriza el uso de estos datos como base para el análisis costo-beneficio en el TFG.
Próxima reunión			
Temas por tratar	Fecha	**Convocados**	
N/A	N/A	N/A	

9.22 Apéndice V. Minuta de reunión #13

Reunión No.	**13**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	01 de mayo, 2025
		Hora de inicio:	06:38pm
		Hora de finalización:	06:58pm
Objetivo de la reunión:	Validar con el tutor el diseño del prototipo automatizado implementado con AWS Free Tier, discutir el riesgo académico de no usar datos reales, así cómo entornos internos de la organización, y evaluar alternativas de justificación ante el comité académico y el lector externo.		
Participantes:	**Presentes**: Profesor tutor y estudiante		
	Ausentes: Ninguno		

Temas tratados

No.	Asunto	Comentarios	**Acuerdos**
1	Prototipo técnico con AWS Free Tier	Kemuel explicó que, por restricciones de Compliance, no puede usar los entornos de Dev, QA ni producción de la organización, por lo que construyó el prototipo completo en una cuenta de AWS Free Tier con datos genéricos.	Michael valida el diseño técnico y confirma que es sólido, pero advierte que hay un riesgo académico si no se justifica correctamente en el TFG.
2	Estrategia de documentación y evidencia	Se discutió cómo presentar la documentación del prototipo: uso de diagramas, explicación técnica paso a paso y evidencia en formato de screenshots.	Se aprueba utilizar capturas, scripts y explicaciones como prueba funcional.
3	Validación del diseño técnico completo	Se repasó todo el flujo automatizado con arquitectura medallion (Bronze, Silver, Gold), funciones Lambda, orquestación con Step Functions y carga en Aurora PostgreSQL. Michael reconoció que la solución técnica está correctamente implementada.	Se aprueba el diseño técnico como funcional y aplicable en entornos reales

Próxima reunión

Temas por tratar	**Fecha**	**Convocados**
N/A	**N/A**	N/A

9.23 Apéndice W. Minuta de reunión #14

Reunión No.	**14**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	03 de mayo, 2025
		Hora de inicio:	11:30am
		Hora de finalización:	11:52am
Objetivo de la reunión:	Analizar los riesgos y alcances de continuar el desarrollo del prototipo en un entorno AWS Free Tier, definir la estrategia de validación y evaluación de resultados bajo este esquema. Establecer cómo documentar técnicamente el prototipo sin comprometer lineamientos de seguridad y confidencialidad de la organización		
Participantes:	**Presentes**: Profesor tutor y estudiante		
	Ausentes: Ninguno		

Temas tratados

No.	Asunto	Comentarios	**Acuerdos**
1	Uso del entorno AWS Free Tier	Debido a restricciones de acceso a entornos organizacionales, se propone continuar el desarrollo del prototipo en AWS Free Tier utilizando datos genéricos.	Michael aprueba esta estrategia como válida.
2	Validación funcional del prototipo	El prototipo debe demostrar que la lógica del flujo automatizado funciona correctamente, incluso con datos simulados, y que reduce intervención humana y mejora los tiempos de ejecución respecto al proceso actual.	Se considerará exitoso si se comprueba su funcionamiento y se valida con el equipo de *Data Operations*.
3	Documentación técnica del prototipo	Kemuel explicó que está detallando qué hace cada función Lambda y Step Function, y adjuntando scripts como evidencia técnica.	Se recomienda mantener esa estructura. Si se accede al entorno de la organización, se evitarán capturas directas y se incluirá una nota por razones de confidencialidad.
4	Estrategia de transición si se aprueba acceso	Si se logra acceso a un entorno corporativo (QA o desarrollo), se haría un cambio inmediato, actualizando únicamente la redacción del documento.	En caso contrario, se formalizará la ejecución sobre Free Tier como solución definitiva.

5	Evaluación del prototipo sin entorno real	Se acuerda que en la Fase 4 se usarán métricas proyectadas si no es posible ejecutar el proceso completo con carga real. Se podría estimar consumo, duración y efectividad según pruebas simuladas y documentación técnica.	Michael valida que este enfoque es correcto para el TFG siempre que se justifique bien en la redacción.
6	Generación de datos *dummy* y marco conceptual	Se debe incluir un apartado en el marco conceptual explicando cómo se generan los datos genéricos (*Dummy Data*).	Se acordó incorporar este detalle en la documentación del TFG como parte de la preparación del prototipo.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
N/A		N/A	N/A

9.24 Apéndice X. Minuta de reunión #15

Reunión No.	**15**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	05 de mayo, 2025
		Hora de inicio:	02:00pm
		Hora de finalización:	02:24pm
Objetivo de la reunión:	Presentar y validar el prototipo funcional del flujo automatizado de carga de datos implementado en AWS, confirmar su replicabilidad en el entorno organizacional, así como el modelo de costos estimado para el análisis costo-beneficio del TFG.		
Participantes:	**Presentes**: Senior Data Engineer y Estudiante		
	Ausentes: Ninguno		

Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Validación del prototipo funcional	Se mostró el flujo completo en AWS con sus componentes: Step Functions, Lambda Functions, S3 (Bronze, Silver, Gold), Aurora y CloudWatch. Se confirmó la lógica funcional.	El equipo de *Data Operations* valida el funcionamiento general del prototipo y da su aprobación para su inclusión en el TFG.
2	Replicabilidad en la organización	Se consultó si el diseño actual puede ser implementado en el	Senior Data Engineer confirma que la lógica es completamente replicable y funcional dentro del entorno corporativo de la empresa.

		entorno organizacional real.	
3	Plan de implementación estimado	Se consultó el tiempo necesario para una implementación completa en producción.	Se estimó un plazo de entre 1 a 2 meses, considerando ajustes técnicos, pruebas y despliegue progresivo.
4	Modelo de costos estimado	Se discutió un costo operativo estimado, considerando procesamiento.	Se autorizó utilizar este dato como base para el análisis costo-beneficio en el capítulo respectivo del TFG.

Próxima reunión		
Temas por tratar	Fecha	**Convocados**
N/A	N/A	N/A

9.25 Apéndice Y. Minuta de reunión #16

Reunión No.	**16**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	06 de mayo, 2025
		Hora de inicio:	01:30pm
		Hora de finalización:	01:43pm
Objetivo de la reunión:	Revisar el estado del prototipo automatizado desarrollado por el estudiante en AWS Free Tier, validar su replicabilidad en el entorno organizacional, establecer restricciones de confidencialidad y analizar la forma en que la evidencia técnica puede ser documentada en el TFG sin comprometer la seguridad corporativa.		
Participantes:	**Presentes**: Profesor tutor, *Senior Data Engineer* y Estudiante		
	Ausentes: Ninguno		

Temas tratados			
No.	Asunto	Comentarios	**Acuerdos**
1	Estado del entorno corporativo (DEV/QA/PROD)	Diego indicó que existen los entornos DEV, QA y PROD en AWS, pero por política interna no es posible exponer información contenida allí fuera de la organización. Esto se debe a la naturaleza comercial de los datos maestros utilizados.	Se confirma que no es viable utilizar estos entornos con datos reales en el desarrollo del TFG.
2	Desarrollo del prototipo en Free Tier	Kemuel explicó que replicó exitosamente todo el flujo automatizado en AWS Free Tier utilizando datos ficticios. El prototipo fue validado	Se aprueba esta implementación como base funcional válida para el TFG.

		por Diego y otros miembros del equipo.	
3	Valor académico del prototipo	Michael señaló que el enfoque actual del prototipo cumple con los objetivos del TFG. La evidencia técnica debe centrarse en la lógica, arquitectura y funcionamiento, sin mostrar datos sensibles.	Se aprueba el enfoque actual siempre que se documente la replicabilidad y se justifique la ausencia de datos reales como una restricción externa.
4	Evidencia técnica permitida	Diego aclaró que no es factible mostrar pantallazos extraídos desde la infraestructura de la empresa. Sin embargo, se puede describir detalladamente la lógica de cada función, las bibliotecas utilizadas, tipos de datos simulados y procesos implicados.	Se acepta incluir descripciones técnicas, scripts comentados y lógica de funcionamiento como evidencia del prototipo.
5	Evaluación y segunda revisión	Michael solicitó que Diego complete la evaluación correspondiente a la semana 11 del cronograma, tomando en cuenta el análisis de resultados y la validación funcional realizada.	Diego confirma que enviará su evaluación posterior a la revisión del capítulo entregado por el estudiante.
6	Posibilidad de replicación en la organización	El equipo confirmó que el prototipo puede ser replicado en el entorno de desarrollo del la empresa sin inconvenientes técnicos. Esto refuerza el valor de la solución planteada en el TFG.	Se recomienda dejar constancia de esta validación técnica en la redacción del TFG y en un apéndice que documente esta reunión como respaldo.
Próxima reunión			
Temas por tratar		Fecha	**Convocados**
N/A		N/A	N/A

9.26 Apéndice Z. Minuta de reunión #17

Reunión No.	**17**	**Fecha:**	
Lugar:	Microsoft Teams	Día:	03 de junio, 2025
		Hora de inicio:	01:59pm
		Hora de finalización:	02:12pm
Objetivo de la reunión:	Evaluar el impacto final del proyecto dentro de la organización y coordinar el cierre formal del proceso académico.		
Participantes:	**Presentes**: Profesor tutor, *Senior Data Engineer* y Estudiante		
	Ausentes: Ninguno		

Temas tratados

No.	Asunto	Comentarios	**Acuerdos**
1	Impacto del proyecto	El representante de la empresa valoró positivamente el prototipo, destacando su utilidad para evitar roles operativos en nuevas migraciones.	La empresa reiteró su interés en replicar el prototipo a nivel organizacional, una vez se apruebe el presupuesto.
2	Evaluación organizacional	Se recordó que la tercera evaluación organizacional debe completarse esta semana.	El representante se comprometió a completarla en cuanto reciba nuevamente el enlace correspondiente.
3	Preparación para la defensa	El tutor recomendó realizar al menos tres simulacros de defensa, idealmente cinco, para ajustar tiempos y fortalecer el discurso.	El estudiante aceptó la recomendación y considerará incluir personas de confianza y del entorno profesional como público de práctica.

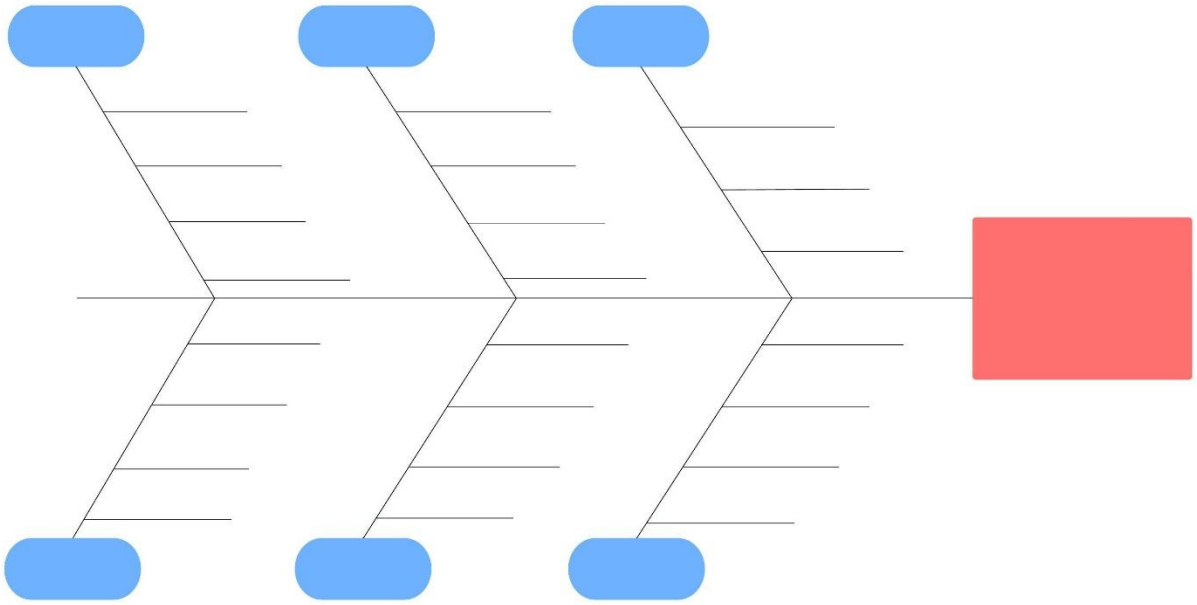
Próxima reunión

Temas por tratar	Fecha	**Convocados**
N/A	N/A	N/A

9.27 Apéndice AA. Plantilla de Análisis FODA.

Análisis FODA	
Fortalezas	**Oportunidades**
Debilidades	**Amenazas**

9.28 Apéndice AB. Plantilla de Diagrama de Ishikawa (*Fishbone*)



9.29 Apéndice AC. Matriz de trazabilidad de requerimientos

Matriz requerimientos vs diseño		
Área	**Especificación del requerimiento**	**Solución implementada en el diseño *To-Be***

10 Anexos

10.1 Anexo I. *Data Review Checklist for Master Data Loading Process*

Data Review Checklist for Master Data Loading Process

Purpose

This document provides basic guidance for analysts on key elements to verify before proceeding with the manual data load into the Master Data Management (MDM) platform. It is intended as a practical reference based on team experience and does not constitute an official or exhaustive procedure. The analyst must apply professional judgment to address issues not covered here.

1. File Reception and Source Verification

- Confirm reception of all expected CSV files from the following sources: **Credit Edge**, **Ratings**, **News Edge**, **ESG**, **Orbis**.
- Validate the file transfer method used (**FTP**, **Email**, **Interface**) and confirm successful receipt.
- Check file names for correct naming conventions (e.g., `credit_edge_YYYYMMDD.csv`).
- Ensure no files are duplicated or missing.
- Confirm files are placed in the correct **SharePoint** directories.

2. File Integrity and Structure Validation

- Open each CSV file and verify:
  - Correct and consistent column headers.
  - Presence of mandatory columns: **Entity**, **Parent Entity**, **Tax ID**, **Exchange**, **LOB**, **Date**, **Amount**.
  - No unexpected columns.
- Check for:
  - Blank rows or columns.
  - Encoding issues or special characters.

3. Data Format and Field-Level Checks

- Validate field types:
  - **Numeric Fields** contain valid numbers.
  - **Dates** follow **YYYY-MM-DD** format.
  - **Booleans** are consistently True/False.
- Identify:
  - Leading/trailing spaces.
  - Case inconsistencies in categorical fields.

4. Duplicates and Cross-File Consistency

- Manually check for duplicates:
  - **Tax ID**, **Entity Names**, **Exchange Codes**.
- Compare with previous cycles to spot:
  - Deviations in volume.
  - Repeated anomalies.

5. Consolidation Verification

- After correction:
  - Merge files into one Excel workbook.
  - Check column order consistency.
  - Ensure no data loss during consolidation.

6. Pre-Load Quality Assurance

- Validate final Excel:
  - Use Excel validation tools.
  - Perform visual review for outliers.
- Conduct a **test SQL load** sample.
- Document manual corrections and escalate issues.

2. File Integrity and Structure Validation

- Open each CSV file and verify:
  - Correct and consistent column headers.
  - Presence of mandatory columns: **Entity**, **Parent Entity**, **Tax ID**, **Exchange**, **LOB**, **Date**, **Amount**.
  - No unexpected columns.
- Check for:
  - Blank rows or columns.
  - Encoding issues or special characters.

3. Data Format and Field-Level Checks

- Validate field types:
  - **Numeric Fields** contain valid numbers.
  - **Dates** follow **YYYY-MM-DD** format.
  - **Booleans** are consistently True/False.
- Identify:
  - Leading/trailing spaces.
  - Case inconsistencies in categorical fields.

10.2 Anexo II. *Data Load Error Documentation Template*

Data Load Error Documentation Template

1. General Information

- **Error Reported By**:
  (Name of the analyst reporting the issue)
- **Date and Time of Detection**:
  (Timestamp of when the error was identified)
- **Cycle Reference**:
  (Indicate the cycle or batch the error belongs to, e.g., April 2025 Load Cycle)
- **Data Source Involved**:
  (e.g., Credit Edge, Ratings, ESG, Orbis, etc.)

2. Error Description

- **Error Type**:
  (Select one or more)
  - Missing Data (NAs in critical fields)
  - Duplicate Records
  - Incorrect Data Format (e.g., date, numeric, boolean)
  - Mismatched Identifiers (Parent/Child entities)
  - Structural Errors in CSV (missing columns, wrong headers)
  - Validation Failures (PostgreSQL load rejection)
  - Encoding Issues (special characters, symbols)
  - Other: _____
- **Detailed Description of the Error**:
  (Explain what was found, where, and its potential impact on data integrity or processing.)
- **Location in File**:
  (Specify the file name, row numbers, and columns where the error appears)

3. Error Detection Method

- Manual Visual Review
- Excel Validation Tool
- SQL Test Load
- Peer Review / Team Validation
- Comparison with Previous Cycle
- Other: _____

4. Immediate Actions Taken

- **Correction Applied**:
  (Describe any manual correction or workaround performed to proceed with the process.)
- **Files or Records Affected**:
  (List all files or specific data entries that were impacted or corrected.)
- **Responsible Analyst for Correction**:
  (Name of the person who applied the correction.)

5. Root Cause Analysis (Optional for Informal Use)

- **Suspected Cause**:
  (Human error, source data inconsistency, tool limitation, missing validation, etc.)
- **Frequency**:
  (Has this error occurred before? If yes, how often?)
- **Escalation**:
  (Was this issue escalated? If so, to whom and when?)

6. Recommendations for Future Prevention

(Optional, but valuable for proposing automation or process improvements based on recurring issues.)

7. Attachments

- Affected File(s)
- Screenshots or Logs
- SQL Queries (if applicable)

8. Sign-Off

- **Reviewed By**: _____
- **Date**: _____

10.3 Anexo III. Salarios mínimos mensuales del sector privado (Año 2025)

MINISTERIO DE TRABAJO Y SEGURIDAD SOCIAL | GOBIERNO DE COSTA RICA | DEPARTAMENTO DE SALARIOS MÍNIMOS

LISTA DE SALARIOS MÍNIMOS
SECTOR PRIVADO
AÑO 2025

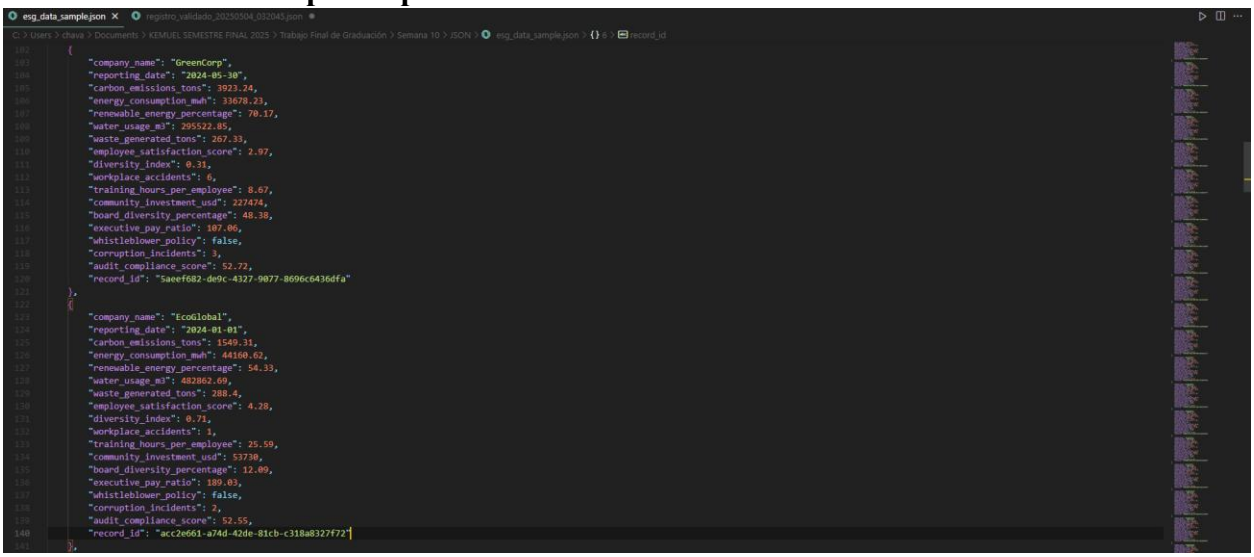
Según Decreto N°44756-MTSS, publicado en La Gaceta N°232, del 10 de diciembre del 2024
Rige a partir del 01 de enero del 2025

SIGLAS Y SALARIOS MÍNIMOS

TONC	Trabajador en Ocupación No Calificada	¢ 12.236,95
TOSC	Trabajador en Ocupación Semicalificada	¢ 13.306,79
TOC	Trabajador en Ocupación Calificada	¢ 13.767,45
TOE	Trabajador en Ocupación Especializada	¢ 15.983,96
TES	Trabajador de Especialización Superior	¢ 24.805,47
TONCG	Trabajador en Ocupación No Calificada (Genérico)	¢ 367.108,55
TOSCG	Trabajador en Ocupación Semicalificada (Genérico)	¢ 399.203,69
TOCG	Trabajador en Ocupación Calificada (Genérico)	¢ 413.023,64
TMED	Técnico Medio en Educación Diversificada	¢ 432.819,25
TOEG	Trabajador en Ocupación Especializada (Genérico)	¢ 476.866,07
TEdS	Técnico de Educación Superior	¢ 533.402,13
DES	Diplomado de Educación Superior	¢ 576.094,24
Bach.	Bachiller Universitario	¢ 653.427,21
Lic.	Licenciado Universitario	¢ 784.139,53

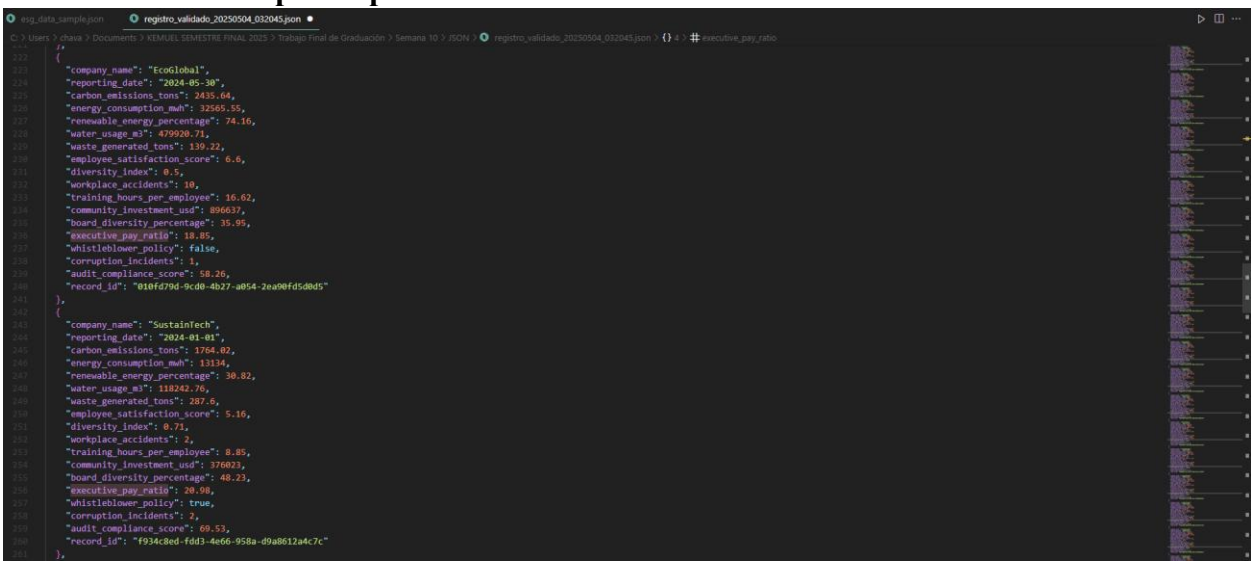
***Salario Mínimo Mensual.**
El Salario Mínimo que no tiene ninguna indicación (*), está por jornada ordinaria

10.4 Anexo IV. Muestra de archivo JSON (Dataset Original) con datos demostrativos para el desarrollo del prototipo



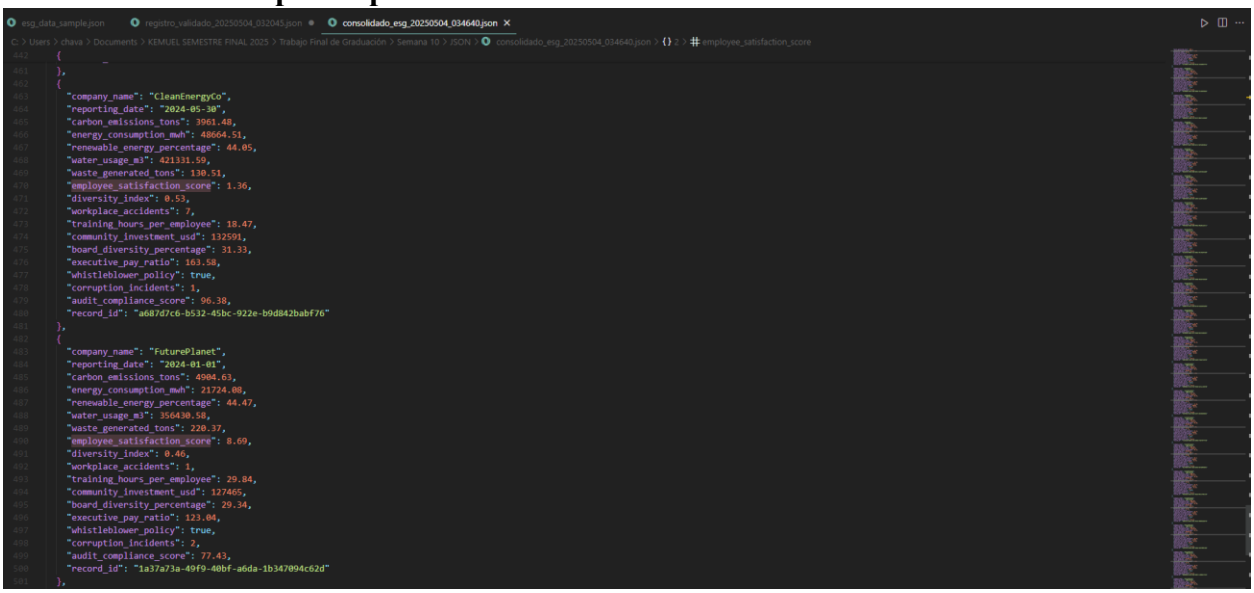

```
esg_data_sample.json ×  registro_validado_20250504_032045.json
C: > Users > chava > Documents > KEMUEL SEMESTRE FINAL 2025 > Trabajo Final de Graduación > Semana 10 > JSON > esg_data_sample.json > {} 6 > record_id
    {
        "company_name": "GreenCorp",
        "reporting_date": "2024-05-30",
        "carbon_emissions_tons": 3923.24,
        "energy_consumption_mwh": 33678.23,
        "renewable_energy_percentage": 70.17,
        "water_usage_m3": 295522.85,
        "waste_generated_tons": 267.33,
        "employee_satisfaction_score": 2.97,
        "diversity_index": 0.31,
        "workplace_accidents": 6,
        "training_hours_per_employee": 8.67,
        "community_investment_usd": 227474,
        "board_diversity_percentage": 48.38,
        "executive_pay_ratio": 107.06,
        "whistleblower_policy": false,
        "corruption_incidents": 3,
        "audit_compliance_score": 52.72,
        "record_id": "5aeef682-de9c-4327-9077-8696c6436dfa"
    },
    {
        "company_name": "EcoGlobal",
        "reporting_date": "2024-01-01",
        "carbon_emissions_tons": 1549.31,
        "energy_consumption_mwh": 44160.62,
        "renewable_energy_percentage": 54.33,
        "water_usage_m3": 482862.69,
        "waste_generated_tons": 288.4,
        "employee_satisfaction_score": 4.28,
        "diversity_index": 0.71,
        "workplace_accidents": 1,
        "training_hours_per_employee": 25.59,
        "community_investment_usd": 53730,
        "board_diversity_percentage": 12.09,
        "executive_pay_ratio": 189.03,
        "whistleblower_policy": false,
        "corruption_incidents": 2,
        "audit_compliance_score": 52.55,
        "record_id": "acc2e661-a74d-42de-81cb-c318a8327f72"
    },
```


10.5 Anexo V. Muestra de archivo JSON (Archivo Silver) con datos demostrativos para el desarrollo del prototipo




```
esg_data_sample.json  registro_validado_20250504_032045.json ●
C: > Users > chava > Documents > KEMUEL SEMESTRE FINAL 2025 > Trabajo Final de Graduación > Semana 10 > JSON > registro_validado_20250504_032045.json > {} 4 > # executive_pay_ratio
  /*
  {
    "company_name": "EcoGlobal",
    "reporting_date": "2024-05-30",
    "carbon_emissions_tons": 2435.64,
    "energy_consumption_mwh": 32565.55,
    "renewable_energy_percentage": 74.16,
    "water_usage_m3": 479920.71,
    "waste_generated_tons": 139.22,
    "employee_satisfaction_score": 6.6,
    "diversity_index": 0.5,
    "workplace_accidents": 10,
    "training_hours_per_employee": 16.62,
    "community_investment_usd": 896637,
    "board_diversity_percentage": 35.95,
    "executive_pay_ratio": 18.85,
    "whistleblower_policy": false,
    "corruption_incidents": 1,
    "audit_compliance_score": 58.26,
    "record_id": "010fd79d-9cd0-4b27-a054-2ea90fd5d0d5"
  },
  {
    "company_name": "SustainTech",
    "reporting_date": "2024-01-01",
    "carbon_emissions_tons": 1764.02,
    "energy_consumption_mwh": 13134,
    "renewable_energy_percentage": 30.82,
    "water_usage_m3": 118242.76,
    "waste_generated_tons": 287.6,
    "employee_satisfaction_score": 5.16,
    "diversity_index": 0.71,
    "workplace_accidents": 2,
    "training_hours_per_employee": 8.85,
    "community_investment_usd": 376023,
    "board_diversity_percentage": 48.23,
    "executive_pay_ratio": 20.98,
    "whistleblower_policy": true,
    "corruption_incidents": 2,
    "audit_compliance_score": 69.53,
    "record_id": "f934c8ed-fdd3-4e66-958a-d9a8612a4c7c"
  },
```

10.6 Anexo VI. Muestra de archivo JSON (Archivo Gold) con datos demostrativos para el desarrollo del prototipo



10.7 Anexo VII. Conversión de JSON a CSV en el entorno AWS Lambda

En el desarrollo del prototipo, los archivos manipulados durante la simulación se estructuraron en formato .json por su facilidad de manejo dentro del entorno AWS Lambda, así como por su compatibilidad con el servicio Amazon CloudWatch y la necesidad de utilizar datos *dummy* con estructura jerárquica. No obstante, la implementación real dentro de la organización se realizará con archivos en formato .json y .csv, según los requerimientos funcionales establecidos en la **Fase 2** del proyecto.

Con el fin de asegurar la viabilidad técnica de dicha transición, se documenta a continuación una función Lambda escrita en Python que permite convertir archivos .json almacenados en Amazon S3 a formato .csv, también alojado en S3. Este procedimiento demuestra la lógica de transformación adaptada a formatos estructurados planos sin comprometer la secuencia técnica del proceso de carga automatizado.

La función realiza los siguiente pasos:

- Lee el archivo .json ubicado en el folder Gold/ del *bucket* S3.
- Interpreta su contenido como un objeto o lista de objetos de datos maestros.
- Escribe los datos en memoria como un archivo .csv utilizando csv.DictWriter.
- Guarda el resultado en el mismo *bucket*, bajo una ruta Gold/esg/ con timestamp.

Este tipo de conversión permite mantener la trazabilidad y continuidad del proceso incluso cuando el formato de entrada cambia, fortaleciendo la flexibilidad técnica del prototipo desarrollado.

json_to_csv_lambda.py > ...

```python
import json
import csv
import boto3
import io
from datetime import datetime

s3 = boto3.client('s3')

def lambda_handler(event, context):
    bucket = 'prototipo-mdm-kemuel'
    source_key = 'Gold/esg/registro_consolidado_20250504_XXXX.json'

    try:
        # Leer archivo JSON desde S3
        response = s3.get_object(Bucket=bucket, Key=source_key)
        json_content = response['Body'].read().decode('utf-8')
        data = json.loads(json_content)

        # Convertir JSON a CSV en memoria
        csv_buffer = io.StringIO()
        fieldnames = list(data.keys())
        writer = csv.DictWriter(csv_buffer, fieldnames=fieldnames)
        writer.writeheader()
        writer.writerow(data)  # Si es un solo registro JSON

        # Si el JSON es una lista, usar esto en lugar de .writerow():
        # for item in data:
        #     writer.writerow(item)

        # Generar nombre destino .csv
        timestamp = datetime.utcnow().strftime("%Y%m%d_%H%M%S")
        destination_key = f'Gold/esg/registro_csv_{timestamp}.csv'

        # Subir archivo CSV a S3
        s3.put_object(
            Bucket=bucket,
            Key=destination_key,
            Body=csv_buffer.getvalue()
        )

        return {
            'statusCode': 200,
            'body': json.dumps({
                'mensaje': 'Archivo convertido exitosamente a CSV',
                'csv_key': destination_key
            })
        }

    except Exception as e:
        return {
            'statusCode': 500,
            'body': json.dumps({'error': str(e)})
        }
```

10.8 Anexo VIII. Carta de revisión filológica

San José, 7 de junio de 2025

Señores (as)
Tecnológico de Costa Rica
Escuela de Administración de Tecnología de Información

Estimados señores (as)

Por este medio notifico formalmente que en mi calidad de Filóloga he revisado minuciosamente el trabajo final de graduación del estudiante Kemuel Abiel Chavarría Moreno cédula 604610204, denominado ***"Propuesta de mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros para una empresa del sector financiero en Costa Rica durante el I Semestre del 2025"***, para optar por el grado de Licenciatura en Administración de Tecnología de Información.

Corregí el trabajo en aspectos tales como: construcción de párrafos, vicios del lenguaje que se trasladan a lo escrito , ortografía, puntuación y otros relacionados con el campo filológico y según lo establecido por APA 7. Desde ese punto de vista considero que está listo para ser presentado como Trabajo Final de Graduación; por cuanto cumple con los requisitos establecidos por la Universidad.

Suscribe de Ustedes cordialmente,

LISBETH FIORELLA ALVAREZ RAMIREZ
Firmado digitalmente por LISBETH FIORELLA ALVAREZ RAMIREZ
Fecha: 2025.06.07 11:53:08 -06'00'

Lic. Fiorella Alvarez Ramírez
Cédula 401890154
Código Colypro 43535

10.9 Anexo IX. Firma de minutas del representante de la organización

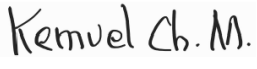
Validación de participación del representante de la organización

Por este medio,

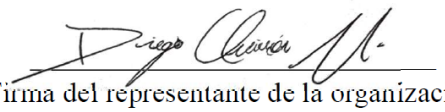
Se agrupan y se firman las minutas de reuniones realizadas a lo largo del desarrollo del Trabajo Final de Graduación titulado "*Propuesta de mejora y automatización del proceso de carga de datos en una plataforma de Gestión de Datos Maestros para una empresa del sector financiero en Costa Rica durante el I Semestre del 2025*", elaborado por el estudiante Kemuel Abiel Chavarría Moreno, carné 2020035032.

El Senior Data Engineer del equipo de Data Operations de la empresa, Diego Quirós Morales, valida su participación en las siguientes sesiones:

- Minuta de reunión #1
- Minuta de reunión #2
- Minuta de reunión #3
- Minuta de reunión #4
- Minuta de reunión #5
- Minuta de reunión #7
- Minuta de reunión #9
- Minuta de reunión #12
- Minuta de reunión #15
- Minuta de reunión #16
- Minuta de reunión #17
- Entrevista #1
- Entrevista #2



Firma del estudiante
Kemuel Abiel Chavarría Moreno



Firma del representante de la organización
Diego Quirós Morales

11 Glosario

Enterprise Data Management & Governance: Área de negocio responsable de supervisar y optimizar la gestión de datos dentro de la organización. Sus funciones incluyen la gobernanza de datos, estandarización de procesos, documentación de flujos de trabajo, administración de la calidad de los datos, y soporte para la toma de decisiones estratégicas.

Data Operations: Equipo especializado dentro del área de *Enterprise Data Management & Governance*, responsable de la gestión operativa de los datos maestros y de referencia dentro de la organización. *Data Operations* administra el proceso de intake de datos, ejecuta tareas asociadas a la gestión horizontal de procesos de Enterprise Data Management (EDM), coordina proyectos, gestiona liberaciones y pruebas, documenta procesos técnicos y lidera la estandarización interna de actividades relacionadas con los datos.

FTP (File Transfer Protocol): Protocolo de red que permite la transferencia de archivos entre sistemas a través de una conexión TCP/IP. En el contexto del TFG, es uno de los medios utilizados por algunas fuentes de datos para entregar los archivos que luego son procesados por el equipo de Data Operations.

Archivo CSV: Formato de archivo de texto plano estructurado en columnas separadas por comas (Comma-Separated Values). Es uno de los formatos más comunes para el intercambio de datos tabulares. En este proyecto, varios de los archivos entregados por las fuentes de datos se presentan en formato CSV.

API (Application Programming Interface): Conjunto de reglas y protocolos que permiten la comunicación entre aplicaciones. Algunas fuentes de datos del proceso actual utilizan APIs para entregar información al equipo de Data Operations, aunque su integración aún requiere validaciones manuales.

SharePoint: Plataforma de colaboración y almacenamiento utilizada por la organización. En el proceso actual, los archivos enviados por las fuentes se almacenan en carpetas de SharePoint antes de ser procesados por el equipo *Data Operations*.

Microsoft Excel: Herramienta de hoja de cálculo utilizada en el proceso actual para transformar, validar y preparar los datos antes de su carga en el sistema maestro. En este TFG, se identifica como una herramienta operativamente útil pero no escalable ni automatizable en su forma actual de uso.

MDM (Master Data Management): Estrategia de gestión que busca garantizar que los datos clave de la organización; como productos, clientes, proveedores o entidades financieras, sean únicos, consistentes y accesibles desde un repositorio central. En este TFG, hace referencia a la plataforma empresarial de Gestión de Datos Maestros, utilizada para consolidar la información proveniente de distintas fuentes externas mediante un proceso de carga estructurado. Esta plataforma constituye el destino final de los archivos procesados y el eje central del flujo de automatización propuesto.

ATI: Administración de Tecnologías de Información. Se refiere a la carrera universitaria y corresponde al programa académico bajo el cual se desarrolla el trabajo final de graduación.

PostgreSQL: Sistema de gestión de bases de datos relacional de código abierto, ampliamente utilizado en entornos empresariales para almacenar, consultar y administrar grandes volúmenes de datos de forma estructurada. En el proceso manual, PostgreSQL sirve como repositorio final donde se cargan los datos procesados por el equipo de *Data Operations*. En el prototipo desarrollado, esta base de datos es replicada en Amazon Aurora PostgreSQL, funcionando como destino de los datos procesados a través del flujo implementado en AWS.

Diagrama As-Is: Representación gráfica del flujo actual del proceso de carga de datos en la plataforma de Gestión de Datos Maestros. En este TFG, ilustra las tareas manuales realizadas por el equipo de Data Operations.

Diagrama To-Be: Modelo gráfico del proceso automatizado propuesto, construido con notación BPMN 2.0. Refleja la nueva arquitectura basada en servicios de AWS, con tareas orquestadas mediante Step Functions, procesamiento distribuido por funciones Lambda, almacenamiento segmentado en zonas Bronze, Silver y Gold, y monitoreo con CloudWatch. Este diagrama representa la solución ideal y sirve como guía para la implementación del flujo automatizado.

AWS (Amazon Web Services): Plataforma de servicios en la nube ofrecida por Amazon, que proporciona infraestructura escalable, almacenamiento, procesamiento y servicios especializados bajo demanda. En este TFG, AWS fue el entorno seleccionado para desarrollar el prototipo de automatización del proceso de carga de datos.

Credit Risk Data: Fuente externa de datos que proporciona información relacionada con riesgos crediticios, utilizada como insumo para alimentar la plataforma de Gestión de Datos Maestros. Su integración al flujo automatizado se simula mediante funciones Lambda que extraen, almacenan y procesan los archivos correspondientes

News Edge: Fuente de datos basada en noticias financieras y corporativas. En el contexto del TFG, se trata de una fuente adicional prevista en el entorno organizacional real que aporta información crítica para alimentar datos maestros con eventos relevantes, como adquisiciones, cambios directivos o incidentes reputacionales.

ESG: Siglas de *Environmental, Social and Governance*. Representa una de las fuentes de datos principales utilizadas en el prototipo. Incluye indicadores como emisiones de carbono, diversidad laboral o gobernanza empresarial. En este TFG, los datos ESG se utilizaron para validar la funcionalidad del flujo automatizado, desde su extracción hasta la carga en la base de datos Aurora PostgreSQL

Corporate Intelligence Database: Fuente de datos orientada a consolidar información estratégica sobre entidades corporativas, como estructuras organizativas, jerarquías legales o afiliaciones. En el flujo propuesto, se prevé su inclusión como una fuente relevante para enriquecer los datos maestros de la organización.

Bronze, Silver y Gold (arquitectura Medallion): Modelo de arquitectura de almacenamiento por capas utilizado en *data lakes*. En este TFG, Bronze representa la zona de ingreso de datos *raw* desde fuentes externas; Silver almacena los datos validados y transformados; Gold consolida los datos finales, listos para ser cargados en la base de datos de la plataforma MDM (Aurora PostgreSQL). Esta segmentación mejora la trazabilidad, calidad y control sobre el ciclo de vida de los datos procesados.

DBeaver: Herramienta de código abierto utilizada para la administración y consulta de bases de datos relacionales. En este TFG, se empleó como entorno de validación técnica durante la fase de evaluación del prototipo, permitiendo verificar que los datos cargados automáticamente en Aurora PostgreSQL cumplieran con la estructura esperada, la unicidad de registros y la integridad semántica de los valores procesados.

Código dummy: Fragmento de código o conjunto de datos ficticios diseñados para simular condiciones reales de operación. En este TFG, se utilizó código dummy para probar las funciones del prototipo sin comprometer información confidencial, permitiendo validar la lógica de extracción, transformación y carga dentro del flujo automatizado.

Bucket S3: Contenedor lógico utilizado en Amazon Web Services para almacenar objetos como archivos, carpetas o flujos de datos. En este proyecto, cada bucket representa una zona del flujo automatizado (Bronze, Silver o Gold), donde se almacenan los archivos en diferentes etapas del procesamiento, siguiendo el modelo de arquitectura por capas.

Python: Lenguaje de programación interpretado, de alto nivel, ampliamente utilizado para automatización, procesamiento de datos y desarrollo de scripts en la nube. Todas las funciones Lambda desarrolladas en el prototipo de este TFG fueron escritas en Python, permitiendo interacción con AWS mediante el SDK Boto3 y ejecución de transformaciones estructuradas sobre archivos JSON.

JSON (JavaScript Object Notation): Formato de intercambio de datos estructurado y jerárquico, ideal para representar objetos con atributos anidados. En el TFG, fue el formato elegido para la simulación de datos ESG y el procesamiento automatizado en Lambda, debido a su compatibilidad con servicios AWS y su capacidad para conservar la semántica de los registros durante las etapas de transformación y carga.

Data Lake: Repositorio centralizado que permite almacenar grandes volúmenes de datos en su formato original o transformado, ya sea estructurado, semiestructurado o no estructurado. En el prototipo de este TFG, se simula un *data lake* mediante la estructura de almacenamiento en capas en S3 (Bronze, Silver y Gold), lo cual facilita la trazabilidad y el control de calidad de los datos procesados.

Dataset: Conjunto de datos estructurados que pueden ser procesados, transformados o analizados. En el contexto del prototipo, los *datasets* corresponden a archivos JSON agrupados por fuente y período, que son validados y consolidados para su posterior carga en la base de datos Aurora PostgreSQL.

UUID (Universally Unique Identifier): Identificador único generado automáticamente, diseñado para evitar colisiones incluso entre sistemas distribuidos. En el TFG, se utilizó para generar claves primarias (*record_id*) dentro de cada registro cargado en la base de datos Aurora, garantizando unicidad e integridad referencial.

Timestamp: Marca temporal que registra la fecha y hora exactas de un evento. En este proyecto, los archivos generados por las funciones Lambda incluyen un *timestamp* en su nombre para asegurar su unicidad, permitir trazabilidad y facilitar el orden cronológico de ejecución del flujo automatizado.

Logs: Registros automáticos generados durante la ejecución de funciones o procesos técnicos. En este TFG, los logs de ejecución fueron capturados mediante Amazon CloudWatch para verificar el comportamiento del flujo automatizado, identificar errores potenciales y confirmar que cada etapa se completó de forma satisfactoria.

psycopg2: Biblioteca de Python utilizada para conectarse y operar con bases de datos PostgreSQL. En el prototipo del TFG, se implementó dentro de una función Lambda (*insert_to_aurora_v2*) para establecer la conexión con Amazon Aurora PostgreSQL y ejecutar sentencias SQL parametrizadas durante la carga final de datos.

SDK Boto3: Biblioteca oficial de Amazon Web Services (AWS) para Python que permite interactuar programáticamente con los servicios de AWS, como S3, Lambda, CloudWatch y Aurora. En este TFG, se utilizó dentro de las funciones Lambda para realizar operaciones automatizadas como la lectura y escritura de archivos en los buckets S3, el manejo de logs, y la conexión con la base de datos Aurora PostgreSQL, facilitando así la ejecución integral del flujo automatizado desde Python.

Parsing: Proceso de análisis sintáctico mediante el cual un sistema interpreta la estructura interna de un archivo o conjunto de datos para extraer su contenido de forma organizada. En este TFG, el término se refiere a la capacidad de las funciones Lambda de AWS para interpretar directamente archivos en formato JSON sin necesidad de transformaciones intermedias. Gracias al motor de *parsing* nativo de AWS, los registros fueron procesados de forma estructurada, permitiendo acceder a campos jerárquicos con precisión dentro del flujo automatizado.

Amazon States Language (ASL): Lenguaje de definición de estados utilizado para construir flujos de trabajo dentro del servicio AWS Step Functions. En este TFG, ASL permitió estructurar la lógica de ejecución del flujo automatizado, definiendo transiciones entre tareas, condiciones, manejos de errores y puntos de espera entre funciones Lambda.