

El Rol de los *Data Lakes* y *Data Warehouses* en la Gestión de Datos Masivos

Titulación:

Máster en Big Data y Ciencia de datos

Curso académico

2024 - 2025

Alumno/a: Garro López, Mónica María

D.N.I: 32299176

Director/a de TFM: Gema Pérez Martínez



Universidad
Internacional
de Valencia

3 de marzo de 2025

De:

 Planeta Formación y Universidades

Índice

Índice de ilustraciones	4
Resumen	6
Palabras clave	6
1. Introducción	7
1.1. Motivación del Estudio	7
1.2. Estructura del Trabajo.....	8
2. Objetivos.....	9
3. Estado del Arte y Marco teórico	11
3.1. Fundamentos de <i>Data Lakes</i> y <i>Data Warehouses</i>	11
3.1.1. Tipos de datos y su almacenamiento empresarial	11
3.1.2. Procesos ETL, ELT y ETLT para la gestión de datos.....	13
3.1.3. Datos Empresariales y su gestión en la era del <i>Big Data</i>	14
3.1.4. ¿Qué es un <i>Data Lake</i> ?.....	15
3.1.5. ¿Qué es un <i>Data Warehouse</i> ?	19
3.1.6. ¿Qué son los <i>Data Marts</i>	22
3.2. Evolución y contraste de <i>Data Lakes</i> y <i>Data Warehouses</i>	25
3.2.1. Diferencias Técnicas y Operativas.....	26
3.2.2. Uso de <i>Data Lakes</i> y <i>Data Warehouses</i> en Análisis Avanzado	27
3.2.3. Gobernanza y Calidad de Datos	27
3.2.4. Innovaciones Tecnológicas.....	28
3.3. Aplicabilidad en diferentes industrias	32
3.3.1. Sector Financiero	32
3.3.2. Sector Salud.....	35
3.3.3. Sector <i>Retail</i>	41
3.3.4. Desafíos en las diferentes industrias	45
3.4. Factores clave en la selección de la Infraestructura de Almacenamiento	47
4. Desarrollo del proyecto y resultados	51
4.1. Metodología.....	51
4.2. Planteamiento del problema	51
4.3. Desarrollo del proyecto	52
4.3.1. <i>Data Warehouse</i> en Amazon Redshift.....	52
4.3.2. <i>Data Lake</i> en Amazon S3.....	57



4.4.	Resultados	61
5.	Conclusión recomendaciones y trabajos futuros	62
5.1.	Conclusiones	62
5.2.	Recomendaciones	63
5.3.	Trabajos Futuros	65
6.	Referencias	66

Índice de ilustraciones

Figura 1. Tipos de datos en Big Data (Ortega Candel, 2023)	12
Figura 2. Comparación entre ETL y ELT. (Núria, n.d.)	13
Figura 3. Comparación entre Data Lake y Data Swamp (Torreglosa, 2023).....	16
Figura 4. Comparativa de Plataformas de Data Lakes Hiperescaladores en la Nube y Soluciones Multicloud. (Fis, 2024)	19
Figura 5. Esquema Estrella (¿Qué Es Un Almacén de Datos? IBM, n.d.).....	21
Figura 6. Esquema Copo de Nieve (¿Qué Es Un Almacén de Datos? IBM, n.d.).....	21
Figura 7. Método Kimball (Bottom up). Elaboración propia.....	23
Figura 8. Método Inmon (Top down). Elaboración propia	24
Figura 9. Cronograma de las tareas definidas. Elaboración propia.	51
Figura 10. Página de inicio de Redshift con opción de prueba gratuita sin servidor. Elaboración propia en la plataforma AWS.....	53
Figura 11. Grupo de trabajo creado con los pasos antes indicados. Elaboración propia en la plataforma AWS.	53
Figura 12. Base de datos creada con pasos anteriores. Elaboración propia en la plataforma AWS.....	54
Figura 13. Modelo Entidad-Relación para el Data Warehouse que se creó como práctica. Elaboración propia.....	54
Figura 14. Creación de la tabla clientes desde la opción “Crear tabla”. Elaboración propia en la plataforma AWS.	54
Figura 15. Creación de tabla producto y carga de datos desde S3. Elaboración propia en la plataforma AWS.	55
Figura 16. Creación de tabla ventas_fact y carga de datos desde S3. Elaboración propia en la plataforma AWS.	55
Figura 17. Consulta para calcular los productos más vendidos. Elaboración propia en la plataforma AWS.....	56
Figura 18. Consulta para calcular el promedio de ventas por país. Elaboración propia en la plataforma AWS.	56
Figura 19. Consultas guardadas en Redshift para uso posterior. Elaboración propia en la plataforma AWS.	56
Figura 20. Creación de Bucket en S3. Elaboración propia en la plataforma AWS	57
Figura 21. Creación de capas (carpetas) dentro del bucket. Elaboración propia en la plataforma AWS.....	58
Figura 22. Carga de datos en crudo en la capa bronze. Elaboración propia en la plataforma AWS.....	58
Figura 23. Función creada para limpieza de datos y carga en capa Silver	59
Figura 24. Resultado de la ejecución de la función lambda. Elaboración propia en la plataforma AWS.....	60
Figura 25. Archivo con datos en crudo. Elaboración propia en la plataforma AWS ...	60
Figura 26. Archivo con datos procesados. Elaboración propia en la plataforma AWS	61

Índice de tablas

Tabla 1. Comparación entre Data Lake y Data Warehouse. Elaboración Propia. Fuente: (Azzabi et al. 2024); (Dubey, 2020); (Divya Meena et al.(n.d.); (Harby y Zulkernine, 2025); Nambiar y Mundra 2022).	26
Tabla 2. Comparación entre Data Warehouses, Data Lakes y Data Lakehouses. Fuente: Adaptado de (Nambiar & Mundra, 2022b); (Azzabi et al., 2024); (Mckendrick, 2020).	29

Resumen

La gestión eficiente del almacenamiento de datos es un aspecto crítico en el ciclo de vida de la información dentro del contexto del **Big Data**. La elección de una infraestructura inadecuada puede comprometer la calidad de los datos, afectando la precisión del análisis y la efectividad de modelos avanzados como **Machine Learning** y **Deep Learning**. Este trabajo analiza los principales modelos de almacenamiento de datos masivos, comparando las características y aplicaciones de los **Data Lakes** y los **Data Warehouses**, así como el impacto de los **Data Swamps** y **Delta Lakes** en la integridad y usabilidad de la información.

El estudio examina los procesos de **Extracción, Transformación y Carga (ETL)**, **Extracción, Carga y Transformación (ELT)** y **Extracción, Carga, Transformación y Transferencia (ETLT)**, esenciales en la estructuración y procesamiento de datos en estos modelos. Los **Data Warehouses** emplean **ETL**, priorizando la transformación antes del almacenamiento, mientras que los **Data Lakes** permiten enfoques más flexibles como **ELT**, donde la transformación se realiza después del almacenamiento. Sin una gestión adecuada, esta flexibilidad puede derivar en **Data Swamps**, reduciendo el valor analítico de los datos.

Como alternativa a estas limitaciones, se analiza el modelo híbrido de **Data Lakehouse**, que combina la flexibilidad y escalabilidad de los **Data Lakes** con la gobernanza y optimización de los **Data Warehouses**. Su implementación facilita el acceso estructurado a los datos sin afectar su disponibilidad para análisis avanzados.

Este trabajo se basa en investigaciones previas en el ámbito de los sistemas de almacenamiento y gestión de datos, así como en referencias académicas y estudios de aplicación en diversas industrias. A través de un análisis comparativo, se establecen criterios clave para la selección de infraestructuras de almacenamiento según las necesidades específicas de cada sector, proporcionando un marco de referencia para la toma de decisiones estratégicas en la gestión de datos masivos.

Palabras clave

Big Data, Data Lake, Data Warehouse, Data Lakehouse, Delta Lake, Data Swamp, ETL, ELT.

1. Introducción

El almacenamiento en la nube y el análisis de datos masivos son dos tecnologías que han ganado una popularidad significativa en los últimos años. Con el crecimiento exponencial de la información proveniente de fuentes diversas - como redes sociales, dispositivos *IoT*, registros transaccionales y plataformas en la nube - se ha vuelto esencial contar con infraestructuras que permitan no solo almacenar datos, sino también garantizar su disponibilidad, calidad y accesibilidad para el análisis y la toma de decisiones estratégicas. En este contexto, los **Data Warehouses** y los **Data Lakes** han emergido como **soluciones fundamentales en la administración de datos masivos, ofreciendo enfoques distintos pero complementarios**.

El propósito de este Trabajo de Fin de Máster (TFM) es analizar el papel de estas infraestructuras en la gestión de datos, estableciendo una comparativa entre sus características, ventajas y desafíos. A través de un enfoque analítico, se explorarán los factores clave en la selección de la arquitectura más adecuada para distintos entornos empresariales, así como su impacto en la eficiencia operativa y el desarrollo de modelos avanzados de análisis de datos. Además, se abordará la evolución de estas soluciones hacia **modelos híbridos** como el **Data Lakehouse**, que busca integrar los beneficios de ambos enfoques para optimizar el almacenamiento y procesamiento de información en escenarios de Big Data.

1.1. Motivación del Estudio

Si bien muchos trabajos en el ámbito del **Big Data y la Ciencia de Datos** han centrado su atención en algoritmos de **Machine Learning y Deep Learning**, la infraestructura sobre la cual estos modelos operan es un aspecto igualmente crítico que no ha recibido la misma atención académica. Durante el desarrollo del máster, asignaturas como “**Sistemas de Almacenamiento y Gestión de Big Data**” han permitido comprender la relevancia de las infraestructuras de almacenamiento en el ciclo de vida de los datos, mientras que “**Cloud Computing**” ha proporcionado una visión de los servicios de computación en la nube como alternativa para la escalabilidad y optimización de costes en la gestión de datos masivos.

En este orden de ideas, la motivación de este estudio radica en la necesidad de profundizar en estos aspectos y analizar cómo la elección de una arquitectura de almacenamiento impacta en la calidad del análisis de datos y en la eficiencia de los sistemas de procesamiento.

1.2. Estructura del Trabajo

Para desarrollar el análisis, este estudio se estructura en los siguientes capítulos:

1. **Fundamentos de *Data Lakes* y *Data Warehouses*.** Se introduce el concepto de cada infraestructura, destacando sus características y aplicaciones en la gestión de datos masivos, así como su integración con tecnologías en la nube.
2. **Evolución y contraste de *Data Lakes* y *Data Warehouses*.** Se realiza una comparativa detallada considerando aspectos como evolución, gobernanza, escalabilidad, costes y rendimiento en distintos entornos de negocio.
3. **Aplicabilidad en diferentes industrias.** Se analizan casos de uso en los sectores financiero, salud y *retail*, donde estas infraestructuras han demostrado ser clave para la optimización de procesos y análisis de datos.
4. **Factores clave en la selección de infraestructura.** Se presentan los criterios estratégicos que influyen en la adopción de una u otra solución.
5. **Demostración de un *Data Lake* en Amazon S3 y un *Data Warehouse* en Amazon Redshift** para comprender su configuración básica con datos sintéticos, debido a los altos costos que puede acarrear un caso real.
6. **Conclusiones y recomendaciones:** se sintetizan los hallazgos del estudio y se ofrecen directrices sobre cuándo y cómo implementar estas soluciones para maximizar su impacto en la gestión de datos empresariales.

2. Objetivos

Objetivo General

- Proporcionar criterios técnicos y estratégicos para la selección de la infraestructura más adecuada en función de las necesidades de cada organización, utilizando tanto *Data Lakes* como *Data Warehouses*.

Objetivos Específicos

Para alcanzar el **objetivo general**, se establecen los siguientes objetivos específicos, los cuales permiten analizar y comparar las principales infraestructuras de almacenamiento de datos, evaluando su aplicabilidad en la gestión de datos masivos y su impacto en distintos sectores empresariales.

1. **Definir los conceptos fundamentales** de *Data Lakes*, *Data Swamps*, *Delta Lakes*, *Data Warehouses* y *Data Marts*, describiendo sus características, evolución y diferencias clave en la administración de datos a gran escala y su integración con tecnologías en la nube.
2. **Comparar las infraestructuras de almacenamiento de datos** en términos de estructura, rendimiento, escalabilidad, gobernanza, costos y accesibilidad, identificando sus ventajas y limitaciones en distintos entornos empresariales.
3. **Examinar casos de uso en diversas industrias**, ilustrando cómo estos modelos de almacenamiento contribuyen a la optimización del almacenamiento y análisis de datos en sectores como el financiero, salud y *retail*.
4. **Analizar los factores estratégicos en la selección de infraestructura**, considerando la naturaleza de los datos, los requisitos de procesamiento y análisis, así como las necesidades específicas de cada industria.
5. **Analizar innovaciones tecnológicas en el almacenamiento de datos**, incluyendo modelos emergentes como *Delta Lakes*, evaluando sus beneficios, limitaciones y su impacto en la evolución de los sistemas de almacenamiento. Asimismo, examinar la transición de los modelos tradicionales hacia arquitecturas híbridas como el *Data Lakehouse*, destacando sus ventajas y desafíos en comparación con enfoques convencionales.
6. **Formular recomendaciones sobre la adopción de estas infraestructuras**, proponiendo mejores prácticas para su implementación y gestión, considerando los retos y necesidades específicas de cada organización.



7. **Implementar un *Data Lake* en Amazon S3 y un *Data Warehouse* en Amazon Redshift** para comprender su configuración básica, almacenamiento y consulta de datos, utilizando un conjunto de datos sintéticos.

3. Estado del Arte y Marco teórico

3.1. Fundamentos de *Data Lakes* y *Data Warehouses*

La relevancia de los datos empresariales ha crecido exponencialmente en la última década debido al auge de las redes sociales y aplicaciones en la nube, lo que ha llevado a un aumento masivo en la cantidad de datos disponibles. Aunque muchas organizaciones han recurrido históricamente a los *Data Warehouses* para analizar datos históricos y tomar decisiones estratégicas, el volumen y la diversidad de los datos actuales dificultan el aprovechamiento de su valor completo. Es aquí donde tecnologías como los *Data Lakes* se posicionan como una solución clave (Tomcy & Pankaj, 2017)

La **elección entre un *Data Warehouse* y un *Data Lake* no es excluyente**; en realidad, ambas soluciones se complementan y contribuyen sinérgicamente a la infraestructura de datos de una organización. Si se requieren **decisiones rápidas y reportes predefinidos**, el ***Data Warehouse*** es la opción ideal. Por otro lado, para manejar **datos complejos y no estructurados**, como videos y datos en tiempo real, un ***Data Lake*** resulta ser la elección más adecuada. La decisión final dependerá del tipo de análisis requerido y de la flexibilidad necesaria para trabajar con los datos. En última instancia, independientemente de la herramienta seleccionada, el objetivo primordial es transformar los datos en información valiosa que facilite la toma de decisiones estratégicas.

Mientras que la **selección entre estas infraestructuras** puede parecer inicialmente como una **decisión exclusiva** entre dos tecnologías competitivas, es fundamental reconocer que su complementariedad es esencial para maximizar la eficiencia en el manejo de los variados tipos de datos que las empresas modernas enfrentan. La sinergia entre ambos modelos se manifiesta no solo en su capacidad para almacenar grandes volúmenes de datos, sino también en cómo facilitan el procesamiento y análisis de datos.

3.1.1. Tipos de datos y su almacenamiento empresarial

La **clasificación de los datos** en el contexto de **Big Data** es fundamental para definir estrategias de almacenamiento y procesamiento eficientes. Comprender sus características resulta crucial para obtener resultados exitosos. En la siguiente imagen se pueden apreciar los diferentes grupos en los cuales podemos clasificar los datos:

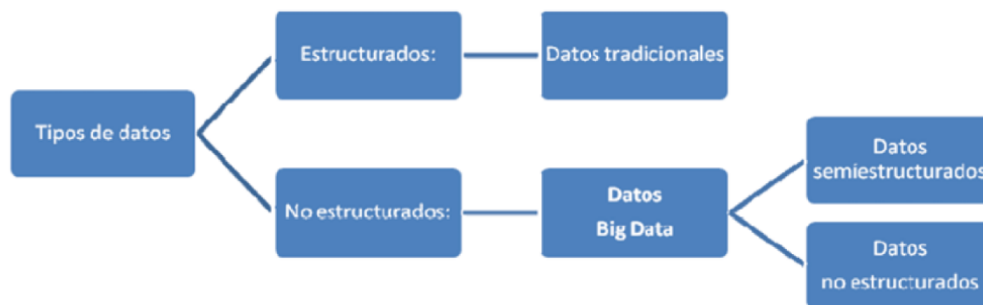


Figura 1. Tipos de datos en Big Data (Ortega Candel, 2023)

Esta **clasificación** es **esencial** para entender cómo se pueden **organizar, almacenar y analizar** los datos en el contexto de Big Data, lo que permite aplicar métodos de análisis más efectivos y precisos según las características específicas de cada tipo:

- **Datos estructurados.** Son un tipo de **información organizada** en un formato predefinido que facilita su almacenamiento, acceso y análisis. Generalmente, se encuentran en bases de datos relacionales (*RDBMS*) y se representan en **tablas con filas y columnas**, donde cada columna define un tipo específico de dato (como números, fechas o texto).
- **Datos no estructurados.** Son aquellos que **carecen de una organización** o jerarquía interna clara, lo que dificulta su clasificación y análisis mediante herramientas tradicionales. Este tipo de datos incluye una **gran variedad de formatos**, como documentos de texto (archivos Word, PDF), archivos multimedia (imágenes, audios, videos), correos electrónicos, mensajes de texto, datos provenientes de redes sociales, dispositivos móviles, y del Internet de las cosas, entre otros.
- **Datos semiestructurados.** Son datos que **poseen un cierto grado de organización interna, pero que no cumplen completamente con** el modelo rígido de las **bases de datos estructuradas**. Aunque contienen elementos que facilitan su clasificación, como etiquetas o marcadores que identifican estructuras y jerarquías, no se organizan en un formato tabular convencional. Este tipo de datos se encuentra comúnmente en archivos web y formatos utilizados para la gestión de información en la *web*, como HTML, XML, OWL, entre otros.

La adecuada clasificación y comprensión de los diferentes tipos de datos no solo es fundamental para un almacenamiento eficiente, sino que también es crucial para implementar estrategias efectivas de gestión de datos.

Cada tipo de dato presenta desafíos y oportunidades únicas que pueden ser explotadas para maximizar el valor que las organizaciones extraen de sus recursos informativos. Con la diversidad y complejidad de datos disponibles hoy en día, las organizaciones deben emplear procesos de integración de datos que no solo manejen eficazmente la variedad y volumen, sino que también apoyen la velocidad y la flexibilidad requeridas para respuestas analíticas en tiempo real.

3.1.2. Procesos ETL, ELT y ETLT para la gestión de datos

Los procesos de *Extract, Transform, Load (ETL)*, *Extract, Load, Transform (ELT)* y *Extract, Transform, Load, Transform (ETLT)* son cruciales para la gestión de datos en entornos modernos, especialmente en la implementación de *Data Lakes* y *Data Warehouses*. Estos procesos describen cómo los datos son preparados y manejados para optimizar tanto el almacenamiento como el análisis posterior.

- **ETL.** Generalmente utilizado en **Data Warehouses**, el ETL implica extraer datos de varias fuentes, transformar estos datos (limpieza, consolidación, reorganización) antes de cargarlos en el almacén. Este proceso es fundamental para asegurar que los datos estén en un formato adecuado y limpio para análisis complejos.
- **ELT.** Más alineado y eficiente con las tecnologías de **Data Lakes**, el ELT permite una mayor flexibilidad al cargar datos directamente en el lago de datos y transformarlos según sea necesario dentro del propio lago. Esto es particularmente útil para manejar grandes volúmenes de datos no estructurados y para escenarios donde la velocidad de carga es crítica.

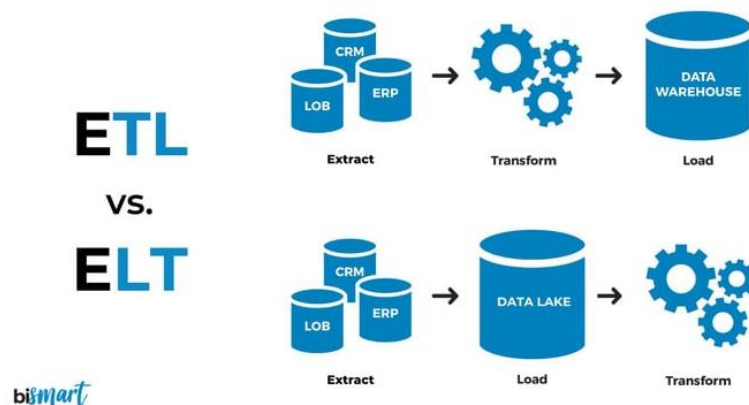


Figura 2. Comparación entre ETL y ELT. (Núria, n.d.)

- **ETLT:** Combina lo mejor de ETL y ELT, proporcionando un **enfoque híbrido** que es útil en situaciones donde diferentes sets de datos requieren distintos tratamientos. Este proceso permite una transformación preliminar, seguida de la carga y transformaciones adicionales más complejas según se requiera para análisis específicos.

Estos enfoques han evolucionado para responder a los desafíos tecnológicos y empresariales actuales. Mientras que el ETL ha sido tradicionalmente el estándar en la gestión de datos estructurados en *Data Warehouses*, el **ELT y el ETLT han ganado relevancia en entornos más dinámicos y con requisitos de flexibilidad**. Esta transición no ocurre en aislamiento, sino que refleja los cambios que han traído el *Big Data* y la adopción de la computación en la nube. Estos avances han transformado las necesidades de las organizaciones, exigiendo mayor velocidad, escalabilidad y adaptabilidad en el procesamiento de datos (Vanga, 2024)

La elección entre ETL, ELT y ETLT depende de varios factores, incluyendo la naturaleza de los datos, los requisitos específicos del análisis y la infraestructura tecnológica existente. La adopción de *cloud computing* ha facilitado la flexibilidad de estos procesos, permitiendo a las organizaciones gestionar eficazmente el volumen creciente y la variedad de datos en la era del *Big Data*.

La eficiencia y efectividad de los procesos ETL, ELT y ETLT no solo influyen en la capacidad de una organización para almacenar y procesar grandes volúmenes de datos, sino que **también desempeñan un papel crucial en la calidad y la utilidad de la información que se utiliza para la toma de decisiones críticas**. A medida que estas técnicas facilitan la integración y el análisis de datos a gran escala, la gestión de los datos empresariales se convierte en una tarea aún más compleja y esencial. Este desafío es particularmente prominente en la era del *Big Data*, donde la velocidad, la variedad y el volumen de los datos generan demandas sin precedentes sobre los sistemas de información empresariales.

3.1.3. Datos Empresariales y su gestión en la era del *Big Data*

Los datos empresariales comprenden toda la información compartida entre los *Stakeholders* internos y externos de una organización, independientemente de su ubicación geográfica. Estos datos **incluyen información clave** como datos financieros, comerciales, de empleados y personales. Su gestión adecuada implica una inversión considerable de tiempo y recursos para garantizar su seguridad y calidad (Tomcy & Pankaj, 2017)

Dentro de una organización, los datos empresariales pueden dividirse en tres categorías principales:

1. **Datos maestros.** Representan las **entidades fundamentales de una empresa**, como clientes, productos o proveedores. Constituyen la base para que las otras categorías de datos tengan un significado útil, y suelen estar gestionados por diferentes departamentos.
2. **Datos transaccionales.** Son aquellos **generados por las aplicaciones internas y externas** al ejecutar procesos empresariales. Incluyen datos relacionados con personas y procesos, proporcionando información valiosa para optimizar operaciones y estrategias comerciales.
3. **Datos analíticos. Derivados de las categorías anteriores**, estos datos ofrecen perspectivas profundas sobre las entidades empresariales y **se combinan con datos transaccionales** para generar recomendaciones que pueden ser implementadas tras la debida diligencia (Tomcy & Pankaj, 2017).

La correcta gestión de estos datos es fundamental para el éxito organizacional. Aquí entra en juego la Gestión de Datos Empresariales (*Enterprise Data Management*, EDM), como estrategia integral para definir, integrar y recuperar datos de una organización y así crear una cultura basada en datos confiables. Este proceso no solo garantiza la calidad de los datos, sino que también establece políticas y responsabilidades claras sobre su manejo, resolviendo conflictos entre departamentos con intereses diversos.

3.1.4. ¿Qué es un *Data Lake*?

Un *Data Lake* (en español, “lagos de datos”) es un **sistema de almacenamiento centralizado diseñado para gestionar grandes volúmenes de datos en su estado bruto**, sin necesidad de estructurarlos previamente. Este sistema puede albergar información proveniente de múltiples fuentes empresariales, incluyendo datos estructurados, semiestructurados y no estructurados, como bases de datos, *logs* de aplicaciones, archivos multimedia, entre otros. Su principal ventaja radica en su enfoque económico y flexible, lo que permite a las organizaciones almacenar y procesar datos de manera eficiente para futuras consultas y análisis.

A diferencia de otros repositorios de datos, como los *Data Warehouses*, los *Data Lakes* se enfocan en almacenar datos en su formato nativo, permitiendo su transformación y análisis posterior según las necesidades específicas. Esto brinda una flexibilidad significativa para implementar casos de uso como el aprendizaje automático, análisis en tiempo real y descubrimientos ad hoc (Mckendrick, 2020).

Desde una perspectiva técnica, este modelo suele **construirse sobre infraestructuras distribuidas**, como *Hadoop* o sistemas en la nube, **permitiendo escalabilidad y**

acceso a grandes volúmenes de datos a un coste relativamente bajo (Divya Meena et al., 2016)

Además, estos fomentan la democratización del acceso a los datos, permitiendo que diferentes áreas de la organización, desde analistas de datos hasta científicos de datos, trabajen directamente con la información en bruto. Sin embargo, la ausencia de gestión de la *metadata*, una mala gestión y clasificación de un gran volumen de los datos puede convertirlos en “*Data Swamp*”.

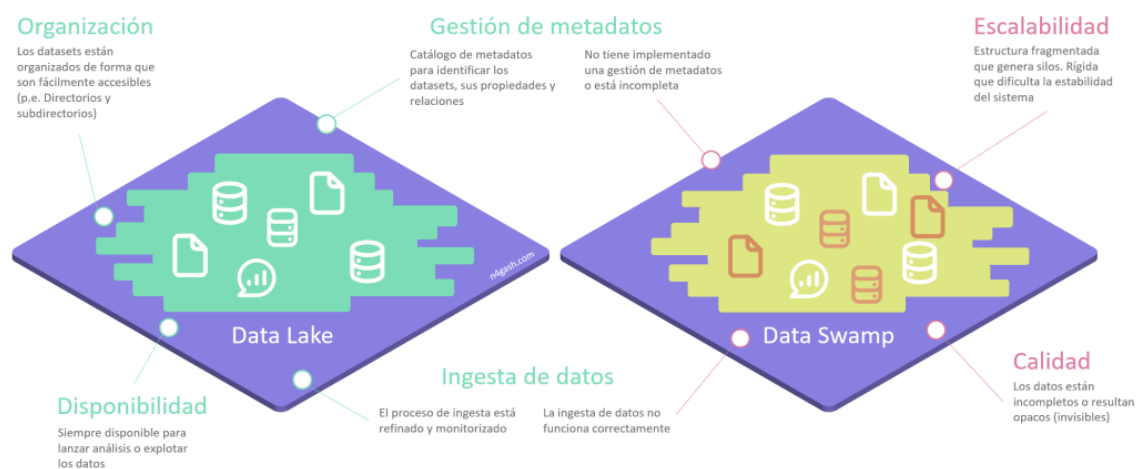


Figura 3. Comparación entre Data Lake y Data Swamp (Torreglosa, 2023)

Para comprender la **diferencia fundamental entre un Data Lake y un Data Swamp**, es esencial observar cómo cada uno gestiona la organización, la calidad y la escalabilidad de los datos. Mientras que un **Data Lake** mantiene una **estructura organizada que facilita el acceso y el análisis** de los datos, un **Data Swamp** carece de una **gestión eficaz**, lo que resulta en **datos desordenados y de difícil acceso**. La imagen mostrada previamente resalta estas diferencias fundamentales en la gestión de datos.

Entre las principales causas de la degradación de un *Data Lake*, encontramos:

- **Falta de gobernanza de datos.** Sin políticas de gestión claras, los datos pueden volverse inconsistentes y difíciles de acceder.
- **Exceso de datos irrelevantes.** Muchas organizaciones recopilan más datos de los que pueden procesar, lo que incrementa el ruido y reduce la utilidad de este modelo (Olavsrud, 2017).
- **Ausencia de metadatos y catalogación.** Sin una estructura adecuada de metadatos, los usuarios no pueden encontrar o interpretar los datos correctamente (Meena & Vidhyameena, 2016).

- **Mala calidad de los datos.** Datos sin limpieza, inconsistentes o con errores pueden invalidar el propósito de este modelo.

Para garantizar que un **Data Lake** cumpla su propósito y no se convierta en un **Data Swamp**, es fundamental **implementar estrategias** que aseguren la calidad, accesibilidad y gobernanza de los datos, como las que se indican a continuación:

- Implementación de un **Marco de Gobernanza** que incluya:
 - **Definición de roles y responsabilidades.** Asignar propietarios de datos y establecer procesos de control.
 - **Políticas de acceso y seguridad.** Uso de controles de acceso basados en roles (RBAC) para proteger la información.
 - **Manejo de la calidad de datos.** Aplicar reglas de validación y limpieza antes de almacenar información en el lago de datos (Divya Meena et al., n.d.).
 - **Recopilar menos datos inicialmente.** Es recomendable comenzar con un conjunto de datos bien definido, asegurando que cada nuevo dato almacenado tenga un valor analítico específico
- Un **sistema de metadatos** bien diseñado que facilite la búsqueda y el uso efectivo de los datos:
 - **Catalogación de datos.** Utilizar herramientas como *Apache Atlas* o *AWS Glue* para documentar datos almacenados.
 - **Etiquetado semántico.** Asociar etiquetas y descripciones a los conjuntos de datos para mejorar su accesibilidad.
 - **Historial y linaje de datos.** Mantener un registro del origen y transformaciones aplicadas a los datos para asegurar su trazabilidad.
- Aplicación de **Machine Learning** para Mantenimiento del **Data Lake** mediante:
 - **Detección de anomalías en los datos.** Identificación de registros duplicados o inconsistentes en tiempo real.
 - **Clasificación automática de datos.** Algoritmos que categorizan datos nuevos según reglas predefinidas.

- **Optimización del acceso.** Sistemas de recomendación que sugieren datos relevantes para los usuarios con base en su historial de consultas (Olavsrud, 2017).

Además de los aspectos fundamentales de los *Data Lakes* y cómo evitar que estos se conviertan en un *Data Swamp*, es crucial considerar las **plataformas en la nube que facilitan su creación y gestión**. En la actualidad, diversas plataformas en la nube ofrecen servicios especializados que permiten a las organizaciones diseñar y administrarlos de manera eficiente y escalable. Según (Fis, 2024), estas soluciones proporcionan las herramientas necesarias para centralizar y procesar grandes volúmenes de datos, apoyando así la implementación de arquitecturas modernas de análisis de datos. Entre las plataformas más destacadas se encuentran:

- **Amazon Web Services (AWS).** Proporciona una variedad de servicios para construir **Data Lakes seguros, flexibles y rentables**. Entre ellos se incluyen *Amazon Simple Storage Service* (S3) para almacenamiento general y *Amazon Elastic MapReduce* (EMR) para procesamiento de datos basado en herramientas de código abierto. Además, *AWS Lake Formation* facilita la configuración y creación de *Data Lakes* en S3.
- **Google Cloud Platform (GCP).** Ofrece un *Data Lakes* que permite la ingesta, **almacenamiento y análisis seguro de grandes volúmenes de datos diversos**. Sus componentes clave incluyen *Google Cloud Storage* (GCS) para almacenamiento general, *Google Dataproc* para procesamiento y análisis de datos a escala, y *Google BigQuery* para consultas nativas en datos almacenados en GCS.
- **Microsoft Azure.** Integrado en la plataforma en la nube de *Microsoft*, proporciona **almacenamiento escalable** que permite realizar **procesamientos y análisis en múltiples plataformas y lenguajes** de programación. Incluye *Azure Data Lake Storage* (ADLS) Gen 2, que combina almacenamiento de sistema de archivos con almacenamiento de objetos para mejorar la escalabilidad y el rendimiento.
- **Soluciones Multicloud.** Cada vez más, las organizaciones están adoptando arquitecturas *multicloud* para **evitar la dependencia de un único proveedor y aprovechar lo mejor de cada plataforma**. Herramientas como **Snowflake** y **Databricks** se destacan en este enfoque, ya que permiten la integración y análisis de datos desde múltiples plataformas, asegurando interoperabilidad y flexibilidad. Estas soluciones son especialmente valiosas en escenarios empresariales complejos donde los datos residen en diferentes entornos de nube o locales.

Para ilustrar más claramente estas diferencias y facilitar la comparación directa entre las opciones disponibles, a continuación, se presenta una imagen comparativa con sus respectivos criterios:

Criterios	LOS HIPERESCALADORES EN LA NUBE			LAS SOLUCIONES MULTICLOUD		
	aws	Google Cloud	Microsoft Azure	CLOUDERA	databricks	snowflake
Servicio de almacenamiento primario	Amazon S3	Google Cloud Storage	ADLS Gen2	Cloudera Data Platform	Lago de datos sobre AWS, GCS o ADLS	Plataforma de datos en la nube Snowflake
Motor de procesamiento	Amazon EMR	Google Dataproc, Dataflow	Azure HDInsight, Azure Synapse	CDP Data Engineering	Delta Engine	Snowflake
Compatibilidad con SQL	Amazon Athena, Redshift, Spectrum	Google BigQuery	Azure Synapse	Servicios de analítica de autoservicio para data warehouses	SQL Analytics Service	Snowflake
Catálogo	AWS Glue	Google Data Catalog	Azure Data Catalog	Cloudera Data Platform	Unity Catalog	Soluciones para partners
Servicio de canales	AWS Glue	Cloud Data Fusion, Dataflow	Azure Data Factory	Cloudera Data Engineering	Delta Live Tables	Snowpark
Compatibilidad con Apache Hive y Apache Spark	✓	✓	✓	✓	Apache Spark	N/A
Compatibilidad con varios lenguajes de programación	✓	✓	✓	✓	✓	✓
Almacenamiento y computación desacoplados	✓	✓	✓	✓	✓	✓
Arquitectura de lakehouse	✓	✓	✓	✓	✓	✓
Multicloud	✗	✗	✗	✓	✓	✓

Figura 4. Comparativa de Plataformas de Data Lakes Hiperescaladores en la Nube y Soluciones Multicloud. (Fis, 2024)

En este orden de ideas, mientras que los **Data Lakes** ofrecen una plataforma versátil y escalable para almacenar y gestionar una **amplia gama de tipos de datos**, desde estructurados hasta no estructurados, para organizaciones que requieren un análisis intensivo de datos con tiempos de respuesta rápidos y consultas complejas basadas en grandes volúmenes de datos históricos, los **Data Warehouses** emergen como una solución indispensable. A continuación, exploraremos cómo este modelo se construye específicamente para soportar operaciones de inteligencia de negocios, reportes y análisis, ofreciendo un entorno altamente optimizado.

3.1.5. ¿Qué es un *Data Warehouse*?

Un *Data Warehouse* (en español, “almacén de datos”) es un sistema **diseñado para centralizar información estructurada de múltiples fuentes** dentro de una organización. Este sistema proporciona acceso a datos históricos y actuales relevantes de la empresa de forma integrada, lo que facilita una toma de decisiones mejor informada.

Principales tipos de datos recopilados incluyen:

- Datos transaccionales provenientes de los sistemas operativos.
- Información de gestión empresarial generada en procesos internos.

- Datos externos relevantes, como tendencias del mercado o información de competidores.

Además de entender los tipos de datos que lo alimentan, es fundamental profundizar en cómo estos datos son organizados internamente para maximizar su utilidad. Identificar y estructurar correctamente los elementos clave, es crucial para capturar y analizar eficientemente los procesos de negocio de una organización.

- **Hechos.** Representan los **procesos de negocio** que una organización desea analizar. Estos son eventos o transacciones que se registran en las tablas de hechos (*fact tables*). Por ejemplo, una venta puede identificarse como un hecho relevante, ya que es un proceso central en la mayoría de las empresas. Cada hecho incluye medidas cuantificables, como el importe de la venta o la cantidad de productos vendidos (Díaz & Caralt, 2015, p. 46).
- **Dimensiones.** Ofrecen un contexto para interpretar los hechos. Estas dimensiones se estructuran como **vistas específicas del proceso** de negocio que se analiza. Por ejemplo, en el caso de una venta, las dimensiones podrían incluir:
 - El cliente que realizó la compra.
 - La fecha en la que se efectuó la transacción.
 - Los productos adquiridos. (Díaz & Caralt, 2015, p. 47)

Las dimensiones permiten un análisis más detallado y granular, facilitando la segmentación y el entendimiento de los datos desde diferentes perspectivas.

- **Métricas.** Son los **indicadores cuantitativos** asociados a los hechos, y permiten medir el rendimiento de un proceso de negocio. Estas métricas están directamente relacionadas con las tablas de hechos. Por ejemplo, en una venta, las métricas pueden incluir el monto total de la transacción o el número de unidades vendidas, lo que proporciona datos objetivos para análisis de desempeño (Curto Díaz & Conesa Caralt, 2015, p. 47).
- **Esquemas de modelado.** Para **estructurar los datos** en un *Data Warehouse*, se utilizan principalmente dos tipos de esquemas:
 - **Esquema en estrella.** Estructura más utilizada, en la cual una tabla central, conocida como tabla de hechos, se conecta a varias tablas de dimensiones desnormalizadas a través de claves foráneas. Esto facilita consultas rápidas y eficientes al concentrar las métricas clave en la tabla central y las cualidades descriptivas en las dimensiones, tal y como se muestra en la siguiente figura.

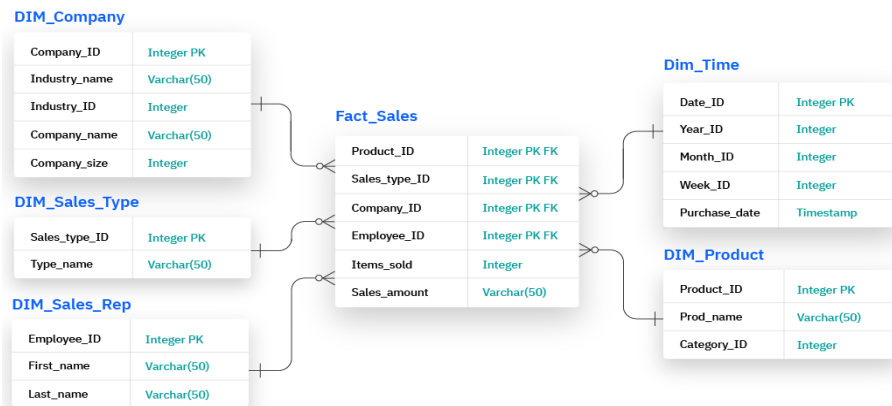


Figura 5. Esquema Estrella (¿Qué Es Un Almacén de Datos? | IBM, n.d.)

- **Esquema en copo de nieve** Variante del modelo estrella en la que se mantiene la misma estructura central, pero las tablas de dimensiones se normalizan, dividiéndose en tablas adicionales, como se muestra en la siguiente figura. Esto reduce la redundancia de datos y mejora la eficiencia del almacenamiento, aunque puede complicar las consultas debido a la estructura más fragmentada.

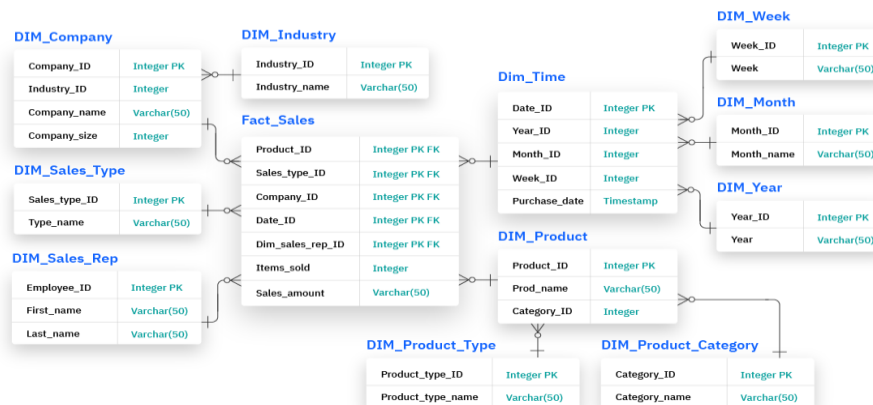


Figura 6. Esquema Copo de Nieve (¿Qué Es Un Almacén de Datos? | IBM, n.d.)

Sin embargo, a medida que las organizaciones manejan volúmenes crecientes de información, los esquemas tradicionales de modelado deben complementarse con soluciones tecnológicas que permitan mayor escalabilidad y eficiencia. En este sentido, el uso de **plataformas en la nube** ha ganado relevancia, ofreciendo almacenamiento flexible, procesamiento distribuido y una mejor integración con herramientas analíticas. Entre las plataformas más destacadas se encuentran:

- **Amazon Redshift:** parte de *Amazon Web Services (AWS)*, *Redshift* es una solución de **almacenamiento de datos en la nube que permite a las organizaciones analizar datos a gran escala de manera eficiente**. Ofrece integración con otras herramientas de *AWS*, facilitando un ecosistema cohesionado para el análisis de datos (Millalen, 2022).
- **Azure Synapse Analytics:** anteriormente conocido como *Azure SQL Data Warehouse*, este servicio de Microsoft **combina capacidades de almacenamiento de datos y análisis de Big Data**. Permite consultas tanto de datos estructurados como no estructurados, integrándose con herramientas como *Power BI* y *Azure Machine Learning* para análisis avanzados (Millalen, 2022).
- **Google BigQuery:** la propuesta de *Google Cloud Platform* para almacenamiento de datos en la nube, *BigQuery*, es un **almacén de datos sin servidor y altamente escalable**. Facilita el análisis de grandes volúmenes de datos mediante consultas SQL estándar y se integra con otras soluciones de *Google* para análisis de datos y aprendizaje automático (Millalen, 2022).
- **Oracle Autonomous Data Warehouse:** este servicio de Oracle ofrece un **almacén de datos en la nube totalmente gestionado que automatiza tareas** como la configuración, seguridad y escalado. Está diseñado para simplificar la administración de datos y mejorar el rendimiento en consultas analíticas (Millalen, 2022).

La adopción de plataformas en la nube no solo ha transformado la infraestructura de los **Data Warehouses**, sino que también ha optimizado la manera en que las organizaciones acceden y organizan subconjuntos de información. En este contexto, surge la necesidad de estructuras más focalizadas dentro del **Data Warehouse**, como los *Data Marts*, que optimizan la consulta y análisis de información para áreas específicas del negocio.

3.1.6. ¿Qué son los *Data Marts*

Los *Data Marts*, son un subconjunto especializado de un *Data Warehouse* diseñado para atender los requerimientos de un área específica del negocio, como ventas, clientes o proveedores. Su propósito es optimizar la consulta y análisis de datos en función de las necesidades de distintos usuarios dentro de la organización. A diferencia de los *Data Warehouses*, que almacena datos de toda la empresa, los ***Data Marts* contienen solo la información relevante para un departamento en particular, lo que mejora el rendimiento y la accesibilidad de las consultas.**(Aytas, 2021)

Un ***Data Mart*** se alimenta de datos almacenados en un ***Data Warehouse*** o directamente de fuentes operacionales, dependiendo de su tipo y modelo de

implementación. En el caso de los **Data Marts dependientes**, los datos provienen del *Data Warehouse*, donde han sido previamente integrados y transformados, lo que permite a las empresas disponer de información más precisa y focalizada sin afectar el rendimiento del almacén de datos central.

Un caso común de uso es la creación de **tableros de control y reportes especializados** para áreas como ventas o marketing, facilitando el acceso a información relevante y optimizando la toma de decisiones de manera más ágil y eficaz.

La manera en que se diseñan e implementan está directamente relacionada con la metodología utilizada en la construcción del *Data Warehouse*. Dependiendo del enfoque adoptado, los *Data Marts* pueden ser diseñados desde el inicio como parte integral del almacén de datos o generados posteriormente en función de las necesidades de cada departamento. En este contexto, existen dos metodologías ampliamente utilizadas: *Kimball* e *Inmon*, las cuales se detallan a continuación:

1. Metodología de Kimball (Modelo Descendente o *Bottom-Up*)

- Se basa en la construcción de *Data Marts* específicos para diferentes áreas de negocio, como ventas o marketing, y luego se integran en un *Data Warehouse* global.
- Utiliza un modelo dimensional con esquemas en *estrella* o *copo de nieve* para optimizar el rendimiento de las consultas.
- Está orientado al análisis de negocio y es más flexible para la generación de reportes y dashboards.

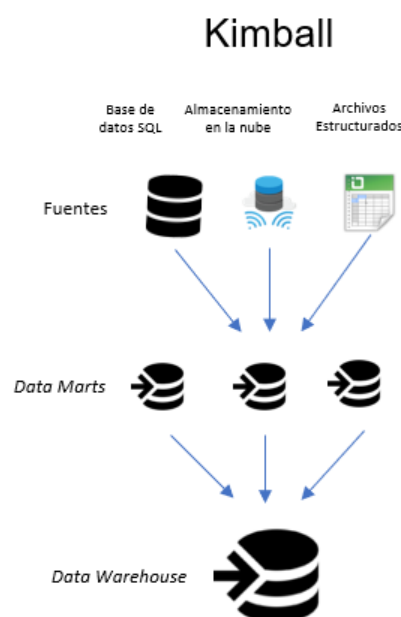


Figura 7. Método Kimball (*Bottom up*). Elaboración propia.

2. Metodología de Inmon (Modelo Ascendente o *Top-Down*)

- Propone un *Data Warehouse* centralizado y normalizado en tercera forma normal (3NF), del cual se derivan los *Data Marts* según las necesidades del negocio.
- Su enfoque estructurado facilita la gobernanza y la calidad de los datos, asegurando la consistencia en toda la organización.
- Se considera más robusto para grandes volúmenes de datos, pero puede ser más complejo de implementar y mantener.

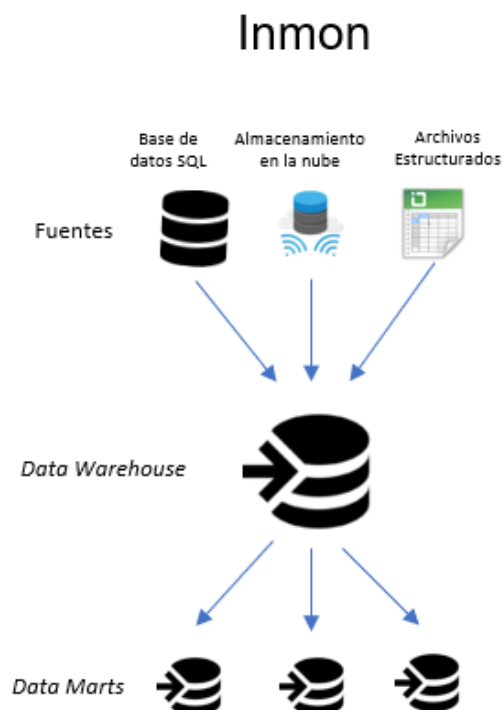


Figura 8. Método Inmon (*Top down*). *Elaboración propia*

La elección entre las metodologías de **Kimball** e **Inmon** dependerá de las necesidades específicas de cada organización, el volumen de datos manejado y los objetivos estratégicos de análisis. Mientras que el enfoque **bottom-up** de Kimball ofrece mayor flexibilidad y rapidez en la implementación de **Data Marts**, el modelo **top-down** de Inmon prioriza la integridad y gobernanza de los datos desde un **Data Warehouse** centralizado. Sin embargo, en la actualidad, las organizaciones no solo deben decidir entre estas arquitecturas tradicionales, sino que también enfrentan la evolución de los modelos de almacenamiento de datos a gran escala.

En este contexto, es fundamental comprender la evolución y el contraste entre **Data Lakes** y **Data Warehouses**, dado que las necesidades de almacenamiento han cambiado con el crecimiento exponencial de los datos. Las empresas buscan soluciones

que permitan almacenar y procesar datos estructurados y no estructurados de manera eficiente, impulsando la adopción de arquitecturas híbridas y nuevas tecnologías de gestión de datos.

3.2. Evolución y contraste de *Data Lakes* y *Data Warehouses*

En la gestión de datos masivos, la elección entre ambos modelos no solo responde a criterios tecnológicos, sino también a las necesidades analíticas y estratégicas de las organizaciones. Mientras que los *Data Warehouses* han sido el estándar en la gestión de datos estructurados para inteligencia de negocios, los *Data Lakes* han surgido como una solución flexible para almacenar y procesar grandes volúmenes de datos en formatos diversos (Nambiar & Mundra, 2022).

A continuación, se presenta una comparación detallada entre ambas arquitecturas, enfocándose en sus diferencias técnicas y operativas.

3.2.1. Diferencias Técnicas y Operativas

Criterio		Data Lake	Data Warehouse
Modelo de granularidad	datos/	Sin estructura previa (<i>schema-on-read</i>).	Estructura predefinida (<i>schema-on-write</i>).
Tipo de datos		Datos en crudo. Estructurados, semiestructurados y no estructurados.	Solo datos estructurados y transformados.
Procesamiento		Flexible, permite ingestión rápida sin preprocesamiento.	Procesamiento intensivo previo a la carga.
Escalabilidad		Escalabilidad horizontal, aprovechando infraestructuras distribuidas como <i>Hadoop</i> y almacenamiento en la nube.	Escalabilidad limitada, generalmente basada en infraestructura vertical (costosa y menos adaptable).
Costo		Bajos costos de almacenamiento; ideal para datos grandes y sin procesar.	Costes elevados de procesamiento y mantenimiento, diseñados para consultas rápidas y consistentes.
Accesibilidad		Mayor flexibilidad para científicos de datos y analistas avanzados mediante herramientas <i>open-source</i> como <i>Hadoop</i> o <i>MapReduce</i> .	Interfaces estándar y accesibles para analistas de negocios mediante <i>SQL</i> y herramientas de <i>BI</i> tradicionales.
Agilidad		Alta; puede configurarse y reconfigurarse rápidamente según las necesidades.	Menos ágil; las modificaciones en los esquemas y estructuras son complejas y demandan tiempo.
Seguridad y gobernanza		Menor control sin una gestión adecuada; requiere herramientas avanzadas para la gobernanza y clasificación de datos.	Alta seguridad y gobernanza integrada, con mecanismos de control robustos para la calidad y la integridad de los datos.
Usabilidad		Complejo de analizar sin herramientas específicas	Intuitivo para consultas estructuradas y generación de reportes
Cumplimiento ACID		Generalmente no cumple con ACID; las actualizaciones y eliminaciones son más complejas.	Cumple con las reglas ACID, garantizando integridad y consistencia en transacciones.

Tabla 1. Comparación entre Data Lake y Data Warehouse. *Elaboración Propia. Fuente: (Azzabi et al. 2024); (Dubey, 2020); (Divya Meena et al.(n.d.); (Harby y Zulkernine, 2025); Nambiar y Mundra 2022).*

3.2.2. Uso de *Data Lakes* y *Data Warehouses* en Análisis Avanzado

El **crecimiento de disciplinas** como la inteligencia artificial (IA) y el aprendizaje automático **ha redefinido los requisitos de las infraestructuras de datos**. En este caso los modelos estudiados juegan roles diferentes en este panorama:

- ***Data Lakes*:**
 - Ideales para entrenar modelos de *Machine Learning* debido a su capacidad para almacenar datos en bruto y sin preprocesamiento (Dubey, 2020).
 - Soportan análisis en tiempo real y exploratorios gracias a tecnologías como *Apache Spark* y *Hadoop* (Olavsrud, 2017).

- ***Data Warehouses*:**

Optimizados para reportes de negocio y análisis histórico. Aunque menos flexibles, proporcionan datos preprocesados y consolidados, útiles para análisis descriptivos y reportes estructurados (Mckendrick, 2020) .

3.2.3. Gobernanza y Calidad de Datos

La gobernanza de datos es un área crítica que puede determinar el éxito o fracaso de una infraestructura:

- ***Data Lakes*:**
 - Presentan mayores riesgos de convertirse en *Data Swamps* si no se implementan estrategias adecuadas de gobernanza (Olavsrud, 2017) , para evitarlo se han presentado previamente diversas técnicas.
 - La calidad y accesibilidad dependen de herramientas de catalogación y metadatos, como *Apache Atlas* o *AWS Glue*, las cuales permiten el acceso a datos relevantes y mejoran su calidad (Nambiar & Mundra, 2022).
- ***Data Warehouses*:**
 - Incorporan mecanismos de gobernanza integrados que garantizan la consistencia, calidad y cumplimiento de normativas como *ACID*. (Harby & Zulkernine, 2025)

3.2.4. Innovaciones Tecnológicas

En el ámbito del almacenamiento y gestión de datos, la rápida evolución de la tecnología ha impulsado el desarrollo de arquitecturas más eficientes, flexibles y escalables. A medida que las organizaciones enfrentan volúmenes crecientes de información estructurada y no estructurada, surgen nuevas soluciones diseñadas para optimizar el almacenamiento, procesamiento y análisis de datos. En este contexto, han emergido innovaciones tecnológicas como **Data Lakehouse** y **Delta Lake**, que buscan superar las limitaciones de los modelos tradicionales de **Data Warehouses** y **Data Lakes**, proporcionando mejoras en gobernanza, transaccionalidad y rendimiento analítico.

Estas innovaciones responden a la necesidad de contar con infraestructuras de datos más robustas, capaces de manejar grandes volúmenes de información sin comprometer la calidad ni la accesibilidad. A lo largo de este apartado, se explorarán estas tecnologías y su impacto en la modernización de la gestión de datos empresariales, analizando sus beneficios, desafíos y diferencias clave respecto a los modelos previos.

Data Lakehouse

La **convergencia entre** las capacidades de los **Data Lakes** y **Data Warehouses** ha **llevado al desarrollo de arquitecturas híbridas como el Data Lakehouse**, que combina lo mejor de ambos mundos.

El auge de este modelo está directamente relacionado con la creciente necesidad de una infraestructura de datos que pueda manejar:

- Diversidad de formatos de datos.
- Flexibilidad y escalabilidad.
- Procesamiento de datos en tiempo real e integración de analítica avanzada y *Machine Learning*.
- Eficiencia operativa garantizando reducción del costo y la complejidad en la gestión de datos.

De acuerdo con (Mckendrick, 2020), la tendencia de las empresas hacia la adopción de arquitecturas híbridas ha llevado a la convergencia de los entornos de datos en soluciones como los *Data Lakehouses*, impulsadas por la necesidad de análisis más ágiles y democratización del acceso a los datos.

Este modelo ha surgido como una solución intermedia que combina las ventajas de los *Data Warehouses* y *Data Lakes*, al tiempo que minimiza sus debilidades. La tabla 2

sintetiza las principales diferencias entre estas arquitecturas, de acuerdo a las principales características de estas infraestructuras.

Característica	Data Warehouse	Data Lake	Data Lakehouse
Estructura de datos	Altamente estructurados	No estructurados o semiestructurados	Estructurados y no estructurados
Esquema	<i>Schema-on-write</i>	<i>Schema-on-read</i>	Híbrido
Procesamiento de datos	ETL	ELT	ELT con optimización transaccional
Costo de almacenamiento	Elevado debido a procesamiento previo	Bajo debido a almacenamiento en bruto	Optimizado mediante estructuras indexadas
Flexibilidad	Baja, optimizado para reportes y BI	Alta, permite consultas ad hoc	Equilibrado, optimizado para múltiples casos de uso
Tiempo de respuesta	Rápido para consultas estructuradas	Lento sin preprocesamiento	Eficiente con indexación y metadatos optimizados
Uso en Machine Learning	Limitado	Ideal para entrenamiento, pero sin optimización para consultas	Optimizado para machine learning e inteligencia artificial
Gobernanza y seguridad	Altamente gobernado y seguro	Débil en control y calidad de datos	Gobernanza avanzada con escalabilidad

Tabla 2. Comparación entre Data Warehouses, Data Lakes y Data Lakehouses. Fuente: Adaptado de (Nambiar & Mundra, 2022b); (Azzabi et al., 2024); (Mckendrick, 2020).

La evolución del *Data Lakehouse* ha sido impulsada por la necesidad de abordar problemas comunes en los enfoques tradicionales de gestión de datos. Sus principales beneficios incluyen:

- **Optimización de costes.** Almacenamiento económico sin comprometer el rendimiento analítico.
- **Mayor eficiencia en la integración de datos.** Permite el uso simultáneo de datos estructurados y no estructurados sin necesidad de transformación previa.
- **Soporte avanzado para inteligencia artificial y *Machine Learning*.** Proporciona una base de datos más accesible para modelos predictivos.
- **Reducción del riesgo de “Data Swamp”.** Esto gracias a la implementación de gobernanza, *metadata* y control de calidad de datos.
- **Fidelidad de los Datos:** Los *data lakehouses* preservan los datos originales, evitando la pérdida de información que puede ocurrir durante la preprocesamiento y transformación, lo que permite la realización de análisis más precisos y completos.

A pesar de sus beneficios, la adopción del *Data Lakehouse* también implica **desafíos y consideraciones en la implementación**, como:

- **Curva de aprendizaje tecnológica.** Las organizaciones deben capacitar a sus equipos en nuevas herramientas y arquitecturas.
- **Gobernanza de datos compleja.** Combinar estructuras rígidas y flexibles requiere políticas avanzadas de seguridad y control de acceso.
- **Costes de migración.** Aunque reduce costes a largo plazo, la transición desde sistemas *legacy* puede ser costosa, además de las pérdidas económicas en las que se debe incurrir durante la transición por interrupciones que afecten la operativa normal de la empresa.

Delta Lake

Si bien los **Data Lakes** ofrecen flexibilidad para almacenar grandes volúmenes de datos en diversos formatos, también presentan desafíos como la falta de transaccionalidad, esquemas inconsistentes y problemas de calidad de datos. Los **Delta Lakes** surgen como una **tecnología que se integra con arquitecturas de datos modernas**, específicamente en el contexto de los *Data Lakes*. Se trata de un almacenamiento de datos de código abierto que mejora las capacidades de los *Data Lakes* tradicionales mediante la implementación de características avanzadas como transacciones *ACID*

(Atomicidad, Consistencia, Aislamiento, Durabilidad), manejo escalable de metadatos y evolución y aplicación de esquemas (Pagidi et al., 2022).

Delta Lake fue desarrollado por Databricks como un formato de almacenamiento open-source basado en Apache Parquet, que **permite que un Data Lake tradicional opere con características de un sistema transaccional**. Esto se logra a través de un **log de transacciones**, que registra cada cambio en los datos y permite revertir operaciones erróneas o consultar versiones históricas.

De acuerdo con (Pagidi et al., 2022) las **características** clave de esta tecnología le permiten superar las limitaciones de los métodos *ETL* tradicionales. las características más notables de este tipo de gestor y almacenamiento de datos es:

1. **Transacciones ACID.** Garantiza que todas las operaciones son atómicas, consistentes, aisladas y duraderas, lo que permite una gestión de datos confiable.
2. **Registro de transacciones.** Utiliza un registro de transacciones (*transaction log*) que se compone y almacena en formato Parquet, lo que permite una rápida búsqueda y acceso a los metadatos de las tablas.
3. **Optimización de la disposición de datos.** Ofrece características avanzadas como la optimización automática de la disposición de datos, permitiendo consultas más rápidas y eficientes.
4. **Soporte para tiempo de viaje (Time Travel).** La capacidad de ver versiones anteriores de los datos facilita la recuperación de información y la comparación de algoritmos en procesos como el aprendizaje automático.
5. **Acceso a múltiples aplicaciones.** Se integra con diversas herramientas y sistemas como *Apache Spark*, *Hive*, *Presto*, *Redshift* y otros, proporcionando versatilidad en el acceso y manejo de datos.
6. **Facilidad en la gestión de datos.** Permite acciones como actualizaciones (*upserts*) y mantenimiento de auditorías, lo cual simplifica los procesos de manipulación de datos.
7. **Escalabilidad y rendimiento.** Puede gestionar grandes volúmenes de datos y ofrece un rendimiento mayor, con mejoras significativas en comparación con arquitecturas más complejas.
8. **Compatibilidad con almacenamiento basado en la nube.** Está diseñado para funcionar efectivamente con almacenes de objetos en la nube, lo que lo convierte en una opción atractiva para la construcción de lagos de datos y almacenes de datos económicos. (Armbrust et al., n.d.)

En este orden de ideas, al comparar los *Delta Lakes* con los *Data Lakes* tradicionales, la diferencia más significativa radica en la capacidad de manejo de datos estructurados y no estructurados mediante el uso de transacciones *ACID*, algo que los *Data Lakes* convencionales no proporcionan. Esta capacidad no solo mejora la calidad y la consistencia de los datos, sino que también crea una base sólida para la aplicación de técnicas de análisis de datos avanzadas como la inteligencia artificial y el aprendizaje automático.

3.3. Aplicabilidad en diferentes industrias

Dado que los datos se consideran el nuevo petróleo, comprender y aplicar correctamente las infraestructuras de almacenamiento y análisis de datos se ha convertido en una prioridad para las organizaciones que buscan mantenerse competitivas. Estas soluciones no solo ofrecen capacidades avanzadas de almacenamiento y gestión de datos, sino que también permiten la explotación inteligente de grandes volúmenes de información, facilitando la toma de decisiones estratégicas basadas en análisis avanzados.

3.3.1. Sector Financiero

En el sector financiero, donde la seguridad, el cumplimiento normativo y la gestión eficiente de datos son prioritarios, las infraestructuras de almacenamiento modernas ofrecen soluciones específicas para manejar la complejidad y el volumen creciente de datos. Estas tecnologías no solo facilitan la toma de decisiones estratégicas y el cumplimiento normativo, sino que también **garantizan la protección de datos sensibles mediante estándares rigurosos de seguridad, como la encriptación, el control de acceso y auditorías regulares**. Además, muchas de estas infraestructuras proporcionan alta disponibilidad, escalabilidad y cumplimiento de propiedades ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad), lo que asegura la continuidad operativa en un sector donde el tiempo de inactividad puede resultar en pérdidas significativas.

En un contexto de transformación digital acelerada, las instituciones financieras enfrentan el desafío de modernizar sus infraestructuras de almacenamiento para aprovechar al máximo el potencial de los grandes volúmenes de datos. Según (Eshghi, 2022), la **adopción de plataformas de datos modernas es clave para mejorar el análisis predictivo, la gestión de riesgos, la personalización de servicios y la detección de fraudes**. Estas tecnologías permiten a las instituciones financieras **mantenerse competitivas en un mercado en constante evolución**, al tiempo que les brindan la capacidad de extraer información valiosa de sus datos, impulsando la innovación y la creación de nuevos productos y servicios.

A continuación, se analizan en detalle la aplicabilidad de cada una de estas infraestructuras en el ámbito financiero:

Data Lakes en el Sector Financiero:

Este tipo de almacenamiento es vital para el sector financiero, no solo por su capacidad de almacenar y gestionar grandes volúmenes de datos, sino también por su rol en habilitar **análisis avanzados y en tiempo real que soportan decisiones críticas de negocio**, cumplimiento normativo, y la personalización de la experiencia del cliente. Asimismo, estos sistemas ayudan a las instituciones financieras a mantenerse competitivas y a innovar en un entorno de rápido cambio, ofreciendo los siguientes beneficios:

- **Mejora en la gestión de datos.** Facilita la gestión de grandes volúmenes de datos de diferentes tipos, lo que es crucial para enfrentar los desafíos del *Big Data* en el sector financiero.
- **Flexibilidad y Escalabilidad.** Su arquitectura permite escalar según las necesidades de almacenamiento y procesamiento de datos, lo que es vital en un entorno financiero en el cual se presenta un rápido crecimiento de datos (Gupta, 2023).
- **Integración con fuentes de datos externas.** Permiten integrar datos de fuentes como redes sociales, mercados financieros o proveedores, lo que es crucial para análisis de riesgo crediticio, detección de fraudes o evaluación de tendencias del mercado.
- **Coste-efectividad para datos a gran escala.** Al almacenar datos en bruto y sin procesar, reducen costes asociados con el preprocesamiento, lo que es especialmente útil para instituciones financieras (Pappil Kothandapani, 2023).

Delta Lakes en el Sector Financiero:

Este tipo de infraestructuras se ha convertido en una herramienta especialmente valiosa para el sector financiero, ya que permite a las instituciones **gestionar grandes volúmenes de datos de manera eficiente y segura**. Su capacidad para soportar transacciones ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad) proporciona un marco sólido que garantiza la integridad de los datos, un aspecto fundamental en una industria donde la precisión y la fiabilidad son esenciales. Entre sus principales beneficios se destacan:

- **Adaptabilidad a condiciones económicas cambiantes.** Debido a su capacidad para realizar análisis en tiempo real, facilita la toma de decisiones oportunas. Esto permite a las organizaciones ejecutar evaluaciones de riesgo,

garantizar el cumplimiento normativo y optimizar auditorías con mayor precisión y rapidez, al contar con datos actualizados y confiables (Pagidi et al., 2022).

- **Optimización de la gestión de datos.** Al integrar datos estructurados y no estructurados en un solo entorno, ayuda a eliminar los silos de información que dificultan un análisis eficiente. Esto es especialmente beneficioso en un sector donde los datos provienen de múltiples fuentes, como transacciones bancarias, mercados de valores y evaluaciones de crédito (Pagidi et al., 2022; Armbrust et al., 2020).
- **Acceso a información histórica para auditorías y cumplimiento normativo.** Las capacidades de *time travel* de *Delta Lake* permiten a las instituciones revisar el historial de cambios en sus conjuntos de datos, facilitando la trazabilidad de decisiones y transacciones. Esto asegura el cumplimiento de las normativas vigentes de manera efectiva y transparente (Pagidi et al., 2022).

Data Warehouses en el Sector Financiero:

Esta es una pieza fundamental en la infraestructura de datos de las instituciones financieras, ya que permiten **almacenar, organizar y analizar grandes volúmenes de datos estructurados de manera eficiente**. Su diseño **optimizado para consultas rápidas y análisis complejos** los convierte en una herramienta indispensable para la toma de decisiones estratégicas, la gestión de riesgos y el cumplimiento normativo financiero. Además, su capacidad para integrar datos históricos y actuales facilita la innovación financiera y el desarrollo de nuevos productos y servicios, manteniendo a las instituciones competitivas en un mercado en constante evolución. Entre sus principales ventajas se encuentran:

- **Soporte para decisiones basadas en datos.** Este modelo está optimizado para realizar análisis rápidos y eficientes, lo que apoya las decisiones empresariales basadas en datos (Romero-Chuquital & Melendres-Velasco, 2023).
- **Gestión de riesgos y cumplimiento normativo.** Facilitan una gestión de riesgos efectiva integrando y analizando datos de múltiples sistemas. Esto también permite a las instituciones financieras identificar patrones y predecir posibles fraudes o incumplimientos normativos, tal como lo demuestran los casos de instituciones como JPMorgan Chase, en la cual han integrado algoritmos de aprendizaje automático en su infraestructura de almacén de datos, lo que ha mejorado significativamente sus capacidades de detección de fraude y evaluación del riesgo crediticio, además de generar una reducción notable en las tasas de incumplimiento y fortalecer los protocolos generales de seguridad del banco (Seethala, 2020).

- **Soporte a la inteligencia de negocios.** Apoyan fuertemente las actividades de inteligencia de negocios al proporcionar datos incluso históricos con alto nivel de gobernanza, limpios, consolidados y listos para el análisis de tendencias financieras, proyecciones y presentación de informes a entes de control.
- **Infraestructura para innovación financiera.** La capacidad para integrar y analizar datos históricos y actuales hace que este modelo de almacenamiento sea una pieza clave en la innovación financiera, soportando el desarrollo de nuevos productos financieros y servicios. En el caso de Bank of America, han mejorado la detección de transacciones fraudulentas mediante la implementación de modelos de detección de anomalías, lo que refuerza la seguridad y la confianza del cliente en sus servicios.(Seethala, 2020)

Data Lakehouses en el Sector Financiero:

En el sector financiero, donde la agilidad, la gobernanza de datos y el cumplimiento normativo son críticos, esta infraestructura emerge como una herramienta esencial para **impulsar la innovación, optimizar la toma de decisiones y mantener la competitividad en un mercado dinámico y regulado**, ofreciendo una solución integral que permite:

- **Optimización del procesamiento de datos financieros.** Este tipo de modelos permiten manejar tanto datos estructurados (transacciones, balances, riesgos) como datos no estructurados (documentación legal, correos electrónicos, registros de voz). Esto es crucial para bancos y aseguradoras que necesitan combinar diversas fuentes de datos para análisis más completos.
- **Soporte para análisis avanzados.** Permiten realizar análisis en tiempo real y *batch*, apoyando modelos predictivos y de *Machine Learning*, que son cada vez más utilizados en el sector financiero para la detección de fraudes y personalización de servicios.
- **Adaptabilidad a regulaciones.** Su estructura de gobernanza de datos facilita el cumplimiento de requisitos regulatorios, asegurando la calidad y trazabilidad de la información utilizada en auditorías y reportes financieros.

3.3.2. Sector Salud

El uso de *Big Data* en la atención médica está revolucionando la forma en que los profesionales de la salud diagnostican y tratan a los pacientes. La recopilación y el análisis de grandes volúmenes de datos a través de sistemas unificados permiten

identificar enfermedades en etapas tempranas, facilitando tratamientos más eficaces y reduciendo los costes asociados a la atención médica.

Sin embargo, Panwar et al. (2022) advierten que la implementación de *Big Data* en la salud también presenta desafíos significativos, como la privacidad de la información, la seguridad de los datos y la interoperabilidad entre diferentes sistemas tecnológicos. Por su parte, Ristevski & Chen (2018) señalan que, aunque la adopción de *Big Data* en medicina supone un proceso complejo, pero su correcta utilización puede generar un impacto positivo considerable en la calidad de la atención y el desarrollo del conocimiento biomédico. Entre los principales beneficios subrayan:

- **Mejora de la atención médica.** A pesar de los desafíos mencionados, la implementación efectiva de *Big Data* puede traducirse en mejoras sustanciales en la calidad del servicio médico. La detección de patrones y tendencias en grandes volúmenes de datos contribuye a diagnósticos más precisos y al diseño de tratamientos personalizados, optimizando la atención a los pacientes.
- **Desarrollo de modelos predictivos.** La analítica avanzada facilita la creación de modelos que pueden prever brotes de enfermedades, virus o complicaciones en pacientes, permitiendo a los profesionales de la salud adoptar un enfoque preventivo y mejorar la eficiencia de la atención médica.
- **Nuevas perspectivas de investigación.** La integración de volúmenes masivos de datos provenientes de distintas disciplinas posibilita la exploración de nuevas hipótesis y la identificación de patrones ocultos en la salud pública y en investigaciones relacionada con el sector salud, acelerando el descubrimiento de nuevos tratamientos y estrategias médicas.
- **Optimización de procesos administrativos.** Además de los beneficios clínicos, el uso de *Big Data* en salud también mejora la eficiencia operativa de las instituciones sanitarias, optimizando la gestión de recursos y reduciendo costes administrativos.

En este contexto, las arquitecturas de almacenamiento y gestión de datos han emergido como soluciones claves para abordar estos desafíos. Estas tecnologías no solo permiten almacenar grandes volúmenes de información médica, sino que también integran y analizan datos de manera eficiente, facilitando la investigación biomédica, la toma de decisiones clínicas y la optimización de los recursos sanitarios.

En relación con los **casos de uso de *Big Data* en el Sector Salud**, Raghupathi & Raghupathi (2014) documentan múltiples aplicaciones del *Big Data* en el ámbito de la salud, destacando su contribución en la **mejora de la eficiencia y los resultados clínicos**. Algunos ejemplos clave incluyen:

- **Detección de infecciones.** El *Hospital para Niños Enfermos (Sick Kids)* en Toronto ha implementado sistemas avanzados de análisis de datos que permiten identificar signos tempranos de infecciones en bebés con riesgo de infecciones nosocomiales. Esta tecnología ha logrado detectar posibles infecciones hasta 24 horas antes que los métodos convencionales.
- **Análisis de datos clínicos.** En el *Instituto Rizzoli de Ortopedia* en Italia, el uso de análisis de datos ha permitido comprender mejor las variaciones clínicas dentro de familias, lo que ha resultado en una reducción del 30% en hospitalizaciones anuales y una disminución del 60% en pruebas de imagen.
- **Mejora en cirugías de reemplazo de rodilla.** Cirujanos ortopédicos en el *Hospital Brigham y Mujeres* en Boston han estandarizado su enfoque para las cirugías de reemplazo de rodilla utilizando datos analíticos. Esta metodología ha optimizado los resultados clínicos y reducido los costes operativos.
- **Uso de datos para prevenir problemas de salud.** El sistema de salud de la *Universidad de Michigan* ha implementado analíticas avanzadas para optimizar la administración de transfusiones de sangre, lo que ha resultado en una disminución del 31% en transfusiones.
- **Modelos Predictivos para enfermedades.** En el caso de la diabetes, se han desarrollado aplicaciones que utilizan datos de pacientes para predecir resultados clínicos y segmentar grupos de riesgo, facilitando estrategias de monitoreo y prevención en salud pública.
- **Detección de fraude en salud.** Aplicaciones avanzadas de análisis de datos han sido empleadas para detectar y minimizar fraudes en el sector salud, mejorando la precisión en la gestión de reclamaciones y garantizando la transparencia en los procesos administrativos.
- **Predicción de brotes epidémicos.** Datos provenientes de *Google Flu Trends* (servicio de Google que estimaba la actividad de la gripe en varios países) y actualizaciones en *Twitter* han sido utilizados para prever aumentos en visitas a salas de emergencia por gripe, así como para rastrear la propagación del cólera en Haití, permitiendo a los sistemas de salud anticipar y responder de manera más eficaz a emergencias sanitarias.

Data Lakes en el Sector Salud

La implementación de estos modelos de almacenamiento en el sector salud ha transformado la gestión y el análisis de grandes volúmenes de datos, permitiendo almacenar información de manera eficiente. Con el crecimiento exponencial de datos

provenientes de diversas fuentes, como historiales clínicos electrónicos (*EHR*), dispositivos de monitoreo y estudios de investigación, estos repositorios facilitan la integración y el procesamiento de información valiosa para **mejorar la toma de decisiones clínicas y optimizar los recursos sanitarios por parte de diferentes actores del ecosistema de salud**, como hospitales, investigadores y aseguradoras, lo que mejora significativamente la calidad de la atención médica y permite desarrollar estrategias de cuidado más personalizadas (Gentner et al., 2023). Entre sus beneficios se resaltan los siguientes:

- **Flexibilidad y escalabilidad.** La capacidad de combinar datos clínicos, de laboratorio y de pacientes permite detectar patrones que podrían pasar desapercibidos en entornos de bases de datos tradicionales. El uso de herramientas analíticas avanzadas facilita la identificación de tendencias en enfermedades, la evaluación de la efectividad de tratamientos y la personalización de la atención médica. En enfermedades crónicas como la diabetes o la hipertensión, esta integración posibilita la correlación de factores como el estilo de vida, la genética y los tratamientos, lo que contribuye a estrategias de intervención más eficaces.
- **Acceso rápido a datos.** Proporciona información en tiempo real, permitiendo que las decisiones clínicas y operativas se tomen con mayor rapidez y precisión.
- **Catálogo de datos.** Facilita la creación de índices que funcionan como catálogos de registros de salud, garantizando una consulta eficiente. Estos catálogos incluyen identificaciones únicas de los usuarios, enlaces encriptados a la información y marcas temporales de las transacciones, reforzando tanto la seguridad como la privacidad de los datos almacenados (Panwar et al., 2022).
- **Intercambio seguro de datos.** Implementa avanzados métodos de autenticación y seguridad, asegurando el cumplimiento de normativas de protección de datos y la confidencialidad de la información del paciente.
- **Eficiencia operativa.** Los hospitales pueden utilizar *Data Lakes* para mejorar la gestión logística y el análisis de datos sociales a gran escala, optimizando la planificación operativa (Gentner et al., 2023).
- **Investigación médica.** Facilita el acceso de investigadores a conjuntos de datos de hospitales, clínicas y biobancos, impulsando descubrimientos en áreas como la epidemiología y el desarrollo de nuevos fármacos.
- **Colaboración entre instituciones.** Permite la interconexión de diversas entidades con sistemas de datos independientes, facilitando estudios más completos sin comprometer la confidencialidad de los datos.

- **Investigación biomédica.** Se utilizan para almacenar y procesar grandes volúmenes de datos clínicos y biológicos, facilitando estudios de correlación entre muestras biológicas y características clínicas.
- **Desarrollo de nuevos medicamentos.** La recopilación y análisis de información relevante de diversas fuentes permite acelerar la investigación y producción de nuevos tratamientos.
- **Estudios epidemiológicos.** Al integrar datos de múltiples biobancos, posibilitan el análisis de patrones de enfermedades y la evaluación de tratamientos en diferentes poblaciones.
- **Genómica y medicina personalizada.** Son clave para el almacenamiento y análisis de datos genómicos junto con datos clínicos, impulsando el desarrollo de tratamientos personalizados (Eder & Shekhovtsov, 2021).
- **Prevención de emergencias.** Brindan acceso a datos clínicos en tiempo real, permitiendo a los gestores de atención anticipar problemas y prevenir visitas innecesarias a salas de emergencia (Tom, 2022).

Delta Lakes en el Sector Salud

En el ámbito de la salud, la gestión eficiente de datos es esencial para mejorar la calidad de la atención y optimizar la toma de decisiones clínicas y operativas. Este tipo de mejora tecnológica se ha convertido en una solución clave para consolidar datos clínicos, administrativos e investigativos en un entorno confiable y estructurado. Su capacidad para manejar grandes volúmenes de información y garantizar la trazabilidad de los datos resulta fundamental para hospitales, centros de investigación y organismos de salud pública. Entre sus principales beneficios en este sector se incluyen:

- **Unificación de datos clínicos y administrativos.** Permite integrar información de registros electrónicos de salud, historial médico, tratamientos y procesos hospitalarios en una única plataforma, facilitando la interoperabilidad entre diferentes sistemas de salud.
- **Mantenimiento de precisión y seguridad en los datos.** Al garantizar transacciones ACID, se evita la corrupción o pérdida de información crítica, asegurando que los registros médicos sean confiables y consistentes.
- **Impulso a la investigación médica y epidemiológica.** La funcionalidad *time travel* posibilita el análisis retrospectivo de datos clínicos, permitiendo evaluar la evolución de enfermedades, la eficacia de tratamientos y la identificación de tendencias epidemiológicas.

- **Respuesta ágil ante emergencias sanitarias.** Durante crisis de salud pública, el acceso en tiempo real a datos actualizados facilita la asignación eficiente de recursos, la toma de decisiones informadas y el monitoreo de la evolución de brotes o pandemias.

Data Warehouses en el Sector Salud

En el ámbito de la salud, este tipo de infraestructura desempeña un papel clave en la gestión y análisis de información estructurada, facilitando la generación de informes esenciales para la toma de decisiones informadas, la gestión de riesgos y el cumplimiento normativo. Su implementación no solo optimiza los procesos operativos dentro de las instituciones, sino que también **mejora la calidad de la atención al paciente al proporcionar un acceso más eficiente y seguro a los datos clínicos**. Entre sus principales ventajas destacan:

- **Acceso rápido a datos estructurados.** Estos sistemas están diseñados para optimizar la velocidad de consulta, permitiendo a los profesionales de la salud acceder a información clínica de manera ágil y fundamentar sus decisiones en datos precisos y actualizados.
- **Gobernanza de datos.** Proporcionan un entorno altamente organizado y seguro para el almacenamiento de información, lo que resulta crucial en un sector donde la privacidad y la seguridad de los datos son prioritarias.
- **Identificación de patrones y correlaciones.** Facilitan el análisis de relaciones entre diferentes factores clínicos y desenlaces de salud, proporcionando información clave para la optimización de tratamientos y la mejora de las prácticas médicas.
- **Seguridad y privacidad de los datos.** Incorporan protocolos avanzados para la protección de información sensible de los pacientes, asegurando el cumplimiento de normativas de privacidad y reduciendo los riesgos asociados a la exposición de datos personales (Alaa Khalaf Hamoud et al., 2018).

Data Lakehouses en el Sector Salud

La adopción de esta infraestructura en el sector salud ha revolucionado la gestión y el análisis de datos biomédicos, permitiendo a instituciones y empresas extraer información relevante para la investigación y la atención clínica. Esta arquitectura se ha consolidado como una solución integral para almacenar, integrar y analizar grandes volúmenes de datos heterogéneos, **facilitando la toma de decisiones informadas y mejorando la eficiencia operativa en el ámbito sanitario**. Entre sus principales ventajas (Gentner et al., 2023; Ristevski & Chen, 2018) se detallan las siguientes:

- **Escalabilidad y flexibilidad en la integración de datos.** Permiten almacenar y gestionar grandes volúmenes de información, lo cual es crucial en un sector donde la generación de datos está en constante crecimiento. Su capacidad para integrar información de diversas fuentes, como registros médicos electrónicos y datos genómicos, facilita un análisis más completo y una visión completa de la salud del paciente. Esta integración favorece la toma de decisiones clínicas basadas en datos precisos y actualizados.
- **Análisis en tiempo real.** Posibilitan el procesamiento inmediato de la información, lo que resulta fundamental en escenarios donde la rapidez en la respuesta es crucial, como en el tratamiento de enfermedades críticas o emergentes.
- **Simplicidad en el análisis y mejor interpretación de resultados.** Sus interfaces intuitivas y métodos interactivos permiten a los investigadores analizar datos biomédicos sin necesidad de conocimientos técnicos avanzados, democratizando el acceso a herramientas de análisis. Además, la estructura de los *Data Lakehouses* mejora la interpretación de los resultados, facilitando su aplicación en el desarrollo de diagnósticos y tratamientos.
- **Reducción de tiempos de espera y mayor eficiencia operativa.** Al permitir la carga de datos personalizados sin necesidad de aprobación de un servidor centralizado, plataformas como *BioLake* eliminan demoras asociadas con la administración de datos, optimizando los flujos de trabajo y acelerando el acceso a información crítica.
- **Optimización de la investigación en salud.** La consolidación de múltiples fuentes de información en un solo entorno facilita la identificación de factores que afectan la salud pública, potenciando los estudios epidemiológicos y las investigaciones clínicas. Esto permite un mejor entendimiento de las enfermedades y la implementación más eficaz de medidas preventivas.

3.3.3. Sector *Retail*

El sector *retail* enfrenta una transformación digital acelerada, impulsada por mercados altamente dinámicos y competitivos. **Para mantenerse a la vanguardia, los *retailers* invierten en nuevas tecnologías para digitalizar y automatizar sus operaciones,** con el objetivo de aumentar la productividad, reducir costes y fortalecer su ventaja competitiva. **Esto ha multiplicado las fuentes de información,** generando grandes volúmenes de datos y una gestión más compleja. En este contexto, los datos se han convertido en un activo estratégico, ya que ofrecer una experiencia de compra

omnicanal fluida y personalizada requiere el uso eficiente de información en tiempo real de múltiples canales.

Una estrategia de datos bien diseñada permite a los *retailers* transformar significativamente las prácticas de marketing, permitiendo una personalización más efectiva y una mejora en la interacción con los clientes (Johnson et al., 2024). **La capacidad de gestionar datos de manera efectiva** no solo mejora la eficiencia operativa, sino que **también se convierte en un motor clave de la transformación digital, impulsando la competitividad y el crecimiento en el sector.**

Data Lakes en el Sector Retail:

Este modelo de almacenamiento permite gestionar la diversidad y el volumen creciente de información generada, que abarca desde transacciones en tiendas y plataformas de *e-commerce* hasta interacciones en redes sociales, datos de sensores *IoT* y vídeos de tiendas inteligentes. Su flexibilidad y escalabilidad lo convierten en una solución ideal para consolidar grandes volúmenes de datos heterogéneos sin necesidad de estructurarlos previamente. **Gracias a este enfoque, los *retailers* pueden maximizar el valor de la información,** facilitando tanto el análisis retrospectivo como la aplicación de modelos predictivos avanzados.

Además, **los *Data Lakes* desempeñan un papel clave en la comprensión del comportamiento del cliente de manera integral,** ya que permiten integrar datos de múltiples fuentes **para obtener una comprensión completa y coherente del cliente y de su comportamiento a lo largo de toda su experiencia de compra.** Esto ayuda a identificar patrones de compra, preferencias y puntos de fricción, lo que posibilita estrategias como la personalización de ofertas o la optimización del surtido en cada tienda. En un entorno donde constantemente emergen nuevos tipos de datos y casos de uso analítico, este tipo de almacenamiento se han consolidado como un pilar fundamental de la transformación digital en el *retail*, impulsando la toma de decisiones basada en datos y fortaleciendo la competitividad del negocio.

Entre las ventajas de este tipo de almacenamiento en el sector retail, encontramos:

- **Escalabilidad y ahorro de costes.** Almacenar datos crudos permite a los *retailers* conservar grandes volúmenes de información a lo largo del tiempo, como todos los tickets de venta de la última década o registros de clics en su web desde su lanzamiento, sin incurrir en costes elevados.
- **Consolidación total de datos.** Este tipo de infraestructura elimina la necesidad de mantener bases de datos separadas, permitiendo correlacionar información diversa, como el análisis de sentimientos en redes sociales con las cifras de ventas.

- **Soporte para analítica avanzada e inteligencia artificial.** Proporciona un entorno ideal para uso de modelos de recomendación, segmentación de clientes, detección de fraudes transaccionales y análisis de textos de reseñas para análisis de satisfacción. Su capacidad para preservar la granularidad y riqueza de los datos originales permite aplicar algoritmos avanzados y descubrir patrones ocultos, impulsando la innovación en *retail*.

Delta Lake en el sector retail

Esta tecnología se presenta como una solución poderosa que permite a las empresas gestionar grandes volúmenes de datos para optimizar sus operaciones y ofrecer una **experiencia más personalizada a los consumidores**. Esta ofrece características fundamentales como la gestión de datos en tiempo real, la fiabilidad en la integridad de los datos y la capacidad de realizar **análisis profundos sobre patrones de consumo**. Entre los principales beneficios en el sector *retail* se destacan:

- **Gestión en tiempo real de inventarios y ventas.** La actualización en tiempo real de datos sobre stock y ventas permite a los minoristas reaccionar rápidamente a cambios en la demanda, evitando sobreabastecimiento o falta de productos.
- **Personalización de la experiencia del cliente.** A través del análisis de datos históricos y comportamentales, se pueden generar recomendaciones de productos y estrategias de marketing más precisas, incrementando la satisfacción del consumidor.
- **Integración de múltiples fuentes de datos.** Cconsolida información proveniente de diferentes canales, como tiendas físicas, plataformas de comercio electrónico y campañas de marketing digital, proporcionando una visión holística del negocio y facilitando el desarrollo de estrategias más precisas.
- **Análisis histórico y predictivo de tendencias.** La función *time travel* permite examinar patrones de consumo a lo largo del tiempo, evaluar el impacto de promociones y ajustar estrategias de ventas basadas en datos concretos.
- **Garantía de calidad e integridad de los datos** La implementación de reglas de gobernanza y control de datos asegura que la información utilizada para la toma de decisiones sea precisa, coherente y confiable.

(Pagidi et al., 2022) mencionan el estudio realizado por Lee et al. (2020), en el cual un *retailer* realiza la adopción de *Delta Lake*, lo cual lo llevó a mejoras significativas en la

eficiencia del procesamiento de datos, lo que optimizó el análisis de la cadena de suministro. Esto se tradujo en una mayor capacidad para realizar análisis en tiempo real sobre el comportamiento de compra de los clientes, permitiendo a los minoristas ajustar sus estrategias de inventario y marketing de manera más ágil. A través de estas capacidades, *Delta Lake* no solo mejora la eficiencia operativa, sino que también fortalece la toma de decisiones, resultando en un aumento de la satisfacción del cliente y, potencialmente, en un incremento de las ventas. En este contexto, *Delta Lake* se posiciona como una solución clave para los minoristas que buscan maximizar su rendimiento en un entorno competitivo impulsado por datos.

***Data Warehouses* en el Sector *Retail*:**

Tradicionalmente, los ***retailers*** han utilizado ***Data Warehouses*** para consolidar información clave sobre ventas, inventarios, tiendas y marketing, con el objetivo de realizar análisis históricos y generar reportes gerenciales. La información se extrae de las fuentes operacionales, se transforma y limpia mediante procesos ETL, y luego se carga en el *Data Warehouse* bajo un esquema estructurado alineado con las necesidades del negocio (productos, tiendas, clientes, fechas, etc.). **Esto garantiza datos confiables y estandarizados, permitiendo a analistas y directivos explorar métricas de desempeño con precisión.**

En este contexto, los *Data Warehouses* aportan inteligencia de negocio confiable al *retail*, facilitando reportes detallados sobre ventas, niveles de stock y tendencias de mercado. Además, mejoran la eficiencia operativa al optimizar la toma de decisiones en áreas clave como la gestión de tiendas, la planificación de promociones y la optimización de la cadena de suministro, gracias a la integración y el análisis estructurado de los datos. Entre sus principales ventajas se destacan:

- **Integridad y confiabilidad de los datos.** Las propiedades ACID aseguran que, una vez realizada una compra, el sistema actualice correctamente el inventario, procese el pago y confirme la transacción de manera segura. Esto garantiza que la información se almacene de forma confiable y consistente, incluso en caso de fallos del sistema, evitando errores o inconsistencias en los registros.
- **Seguridad y gobernanza integradas.** Ofrecen medidas avanzadas de seguridad, como cifrado de extremo a extremo, controles de acceso robustos y *backups* automatizados. Esto garantiza la protección de datos sensibles de clientes y facilita el cumplimiento de regulaciones de privacidad y seguridad de la información de los clientes y proveedores.
- **Compatibilidad con herramientas de BI.** Brindan soporte nativo a herramientas de inteligencia de negocios (BI) ampliamente utilizadas, lo que permite la generación de informes para gerencia y entes reguladores,

reduciendo errores manuales y asegurando la integridad de los datos en auditorías y análisis financieros.

Data Lakehouses en el Sector Retail:

La adopción de esta arquitectura de almacenamiento simplifica la gestión de datos al unificar el almacenamiento y reducir la dependencia de múltiples sistemas, lo que disminuye la cantidad de conjuntos de datos que deben ser administrados. Como resultado, se optimizan los costes de desarrollo y operación. Además, al consolidar la información en una única plataforma, se elimina la necesidad de replicar datos entre distintos sistemas, garantizando una fuente única de verdad y proporcionando un acceso unificado para analistas y científicos de datos (Schneider et al., 2024).

Este enfoque permite a las empresas minoristas aprovechar datos estructurados y no estructurados para mejorar su eficiencia operativa, optimizar la experiencia del cliente y tomar decisiones basadas en información en tiempo real. Según (*Lakehouse for Retail Overview* | Databricks, n.d.), los principales beneficios de los *Lakehouses* en *retail* incluyen:

- **Análisis en tiempo real para decisiones ágiles.** Permite la ingesta y procesamiento de datos en tiempo real, facilitando la toma de decisiones inmediatas en áreas clave como precios dinámicos, optimización de inventarios y recomendaciones personalizadas en e-commerce.
- **Colaboración abierta y rentable.** La colaboración en el uso de datos y análisis es clave para fomentar la innovación y la interacción entre todos los socios de la cadena de valor. Mejorar la colaboración acelera las operaciones, permite análisis más completos y reduce los costes de alineación en toda la organización.
- **Aprovechamiento de datos multimodales.** Actualmente, solo entre el 5 % y el 10 % de los datos empresariales están estructurados. Explorar y analizar el 90 % restante (datos no estructurados y semiestructurados) permite a las empresas comprender mejor su entorno y tomar decisiones más informadas.

3.3.4. Desafíos en las diferentes industrias

La implementación de *Data Lakes* en los tres sectores antes analizados enfrenta diversos desafíos, como los siguientes:

- **Calidad de los datos:** La falta de estandarización y la presencia de datos incompletos o inexactos pueden afectar la utilidad de la información almacenada, lo que dificulta su aplicación en estudios y análisis clínicos.
- **Privacidad y seguridad:** El manejo de datos sensibles de pacientes exige el cumplimiento de normativas estrictas, como la *Health Insurance Portability and Accountability Act (HIPAA)*, que establece estándares para la protección de información personal de salud y la gestión de la privacidad en el ámbito sanitario (Panwar et al., 2022; Eder & Shekhovtsov, 2021). Estos mismos principios de protección de datos se extienden también al manejo de información sensible en los sectores financiero y *retail*.

En relación con la implementación de *Data Warehouses* también se enfrentan desafíos que deben ser abordados para maximizar su eficacia:

- **Manejo de datos no estructurados:** estos sistemas están optimizados para gestionar datos estructurados, lo cual limita significativamente su capacidad para procesar información no estructurada. En el sector salud, esta limitación afecta el manejo de imágenes médicas, registros de sensores y notas clínicas detalladas. En el sector *retail*, los sistemas encuentran dificultades al procesar facturas, contratos e imágenes de productos. En el sector financiero, esta restricción se extiende a la gestión de documentos tales como extractos bancarios en formatos libres, comunicaciones por correo electrónico y transcripciones de llamadas de servicio al cliente.
- **Altos costes de mantenimiento y escalabilidad:** la administración y expansión de un *Data Warehouse* puede representar una inversión significativa, especialmente en entornos con grandes volúmenes de datos como los que se han analizado, donde la infraestructura debe ser constantemente optimizada para mantener un rendimiento eficiente.

A pesar de sus múltiples beneficios, la implementación de *Data Lakehouses* también enfrenta algunos retos claves como los siguientes:

- **Privacidad y seguridad:** al igual que en los *Data Lakes*, la protección de datos sensibles representa un desafío crítico, ya que su correcta gestión debe cumplir con estrictas normativas de seguridad y privacidad para evitar accesos no autorizados.
- **Capacitación del personal:** la adopción de *Data Lakehouses* requiere profesionales con conocimientos especializados en tecnologías avanzadas de análisis de datos, lo que implica la necesidad de formación continua para su correcta implementación y uso.

A pesar de sus ventajas, la adopción de **Delta Lake** también presenta desafíos significativos, entre los que se destacan:

- **Complejidad en la integración con sistemas existentes:** la implementación de *Delta Lake* puede requerir modificaciones sustanciales en la infraestructura de datos, lo que implica evaluar cuidadosamente la compatibilidad con arquitecturas preexistentes.
- **Capacitación del personal:** la adopción de *Delta Lake* también exige que los equipos de datos adquieran nuevas competencias en el manejo de transacciones ACID, control de versiones y optimización de rendimiento en grandes volúmenes de información.
- **Gestión del almacenamiento y rendimiento:** aunque esta infraestructura mejora la eficiencia en la gestión de datos, es necesario aplicar estrategias adecuadas para evitar el crecimiento descontrolado de archivos pequeños y garantizar un rendimiento óptimo en consultas analíticas.

3.4. Factores clave en la selección de la Infraestructura de Almacenamiento

El almacenamiento de grandes volúmenes de datos se ha convertido en un componente estratégico para las organizaciones. La selección de la infraestructura depende de múltiples factores, incluyendo la estructura de los datos, los casos de uso y los requisitos de procesamiento, la escalabilidad y el costo. A continuación, se presenta un análisis de los factores clave en la selección de una infraestructura de almacenamiento óptima para diferentes escenarios empresariales y tecnológicos.

- **Estructura y naturaleza de los datos**

Uno de los criterios fundamentales para elegir entre los diferentes modelos de almacenamiento es la estructura de los datos. Los **Data Lakes**, al permitir almacenar datos en su formato bruto y sin procesar, son ideales para datos no estructurados o semiestructurados. En contraste, los **Data Warehouses** requieren datos estructurados con un esquema predefinido.

Por otro lado, los **Data Lakehouses** combinan las ventajas de ambos, permitiendo el almacenamiento de datos en bruto y ofreciendo capacidades de análisis estructurado, lo que los convierte en una opción versátil para organizaciones que buscan agilidad en el manejo de datos. Los **Delta Lakes**, por su parte, mejoran la confiabilidad y consistencia de los **Data Lakes** mediante la introducción de transacciones ACID y versiones de datos, asegurando que los

datos sean confiables y rastreables sin perder la flexibilidad del almacenamiento en bruto.

- **Propósito del análisis y casos de uso**

El tipo de análisis que se pretende realizar es un factor clave en la selección de la infraestructura de almacenamiento de datos. **Los Data Lakes** son especialmente adecuados para aplicaciones de *machine learning* y análisis exploratorio, ya que almacenan grandes volúmenes de datos en su formato original sin necesidad de transformación previa, lo que facilita la experimentación y el descubrimiento de patrones ocultos en los datos (Giebler et al., 2019). A diferencia de los **Data Warehouses**, que son más eficientes para analítica de datos e inteligencia empresarial, donde la integridad, consistencia y calidad de los datos son fundamentales para la toma de decisiones basada en información confiable (Mcken-drick, 2020).

Los **Data Lakehouses** proporcionan una plataforma unificada que permite a las organizaciones realizar tanto análisis descriptivos como predictivos dentro de un mismo entorno. Por otro lado, **Delta Lake** se posiciona como una solución clave en entornos que requieren procesamiento avanzado y en tiempo real, acceso concurrente, rapidez en las consultas y consistencia en los datos (*Delta Lake vs. Data Lake: Diferencias Clave* | Airbyte, n.d.).

- **Escalabilidad y Rendimiento**

Los **Data Lakes** ofrecen escalabilidad horizontal masiva a menor costo, debido a que pueden almacenar datos sin necesidad de procesamiento previo. Sin embargo, el acceso a los datos puede ser más lento debido a la falta de estructura. En cambio, los **Data Warehouses** están optimizados para consultas rápidas y eficientes, pero requieren un mayor esfuerzo de preparación y transformación de datos.

Los **Data Lakehouses** buscan equilibrar estas limitaciones al ofrecer escalabilidad similar a la de los **Data Lakes** con un rendimiento mejorado para consultas estructuradas. Por su parte, los **Delta Lakes** optimizan la lectura y escritura de datos dentro de los **Data Lakes**, permitiendo almacenamiento incremental y mejorando significativamente la eficiencia de procesamiento sin perder la flexibilidad del modelo Data Lake (Avril, 2024).

- **Costo de implementación y mantenimiento**

El costo es un factor determinante en la elección de la infraestructura de almacenamiento. Los **Data Lakes** suelen ser más económicos, rápidos y adaptables en términos de almacenamiento, ya que utilizan tecnologías de bajo costo como *Hadoop* o almacenamiento en la nube, aunque pueden incurrir en gastos adicionales debido a la necesidad de limpieza de datos, gestión y costes de almacenamiento a medida que la complejidad aumenta (Harby & Zulkernine, 2025). Por otro lado, los **Data Warehouses** suele residir en almacenamiento especializado (discos y hardware de base de datos) o servicios *cloud* de alto rendimiento, por lo que el costo por volumen de datos es elevado. Por su parte, los **Data Lakehouses** adoptan también almacenamiento de bajo costo, por lo que mantiene la eficiencia de costes y escalabilidad de un *Data Lake* (Cherradi & Haddadi, 2024b).

En términos de **costes operativos**, el mantenimiento de un **Data Warehouse** implica un esfuerzo significativo en modelado y administración para garantizar su rendimiento óptimo. Por otro lado, un **Data Lake** requiere inversiones en gobernanza para evitar la degradación de los datos y el riesgo de convertirse en un *Data Swamp*. Los **Data Lakehouses** buscan un equilibrio entre ambos modelos, reduciendo la complejidad operativa mediante mecanismos avanzados de gestión de datos.

A su vez, **Delta Lake** ofrece una optimización adicional al mejorar la gobernanza y la calidad de los datos dentro de un **Data Lake**, lo que impacta directamente en la reducción de costes operativos. Su capacidad para almacenar las rutas de los archivos **Parquet** en un registro de transacciones independiente elimina la necesidad de realizar costosas operaciones de enumeración de archivos en la nube. Esto resulta particularmente beneficioso en entornos donde el volumen de archivos es elevado, ya que acelera el acceso a los datos y minimiza los tiempos de procesamiento en comparación con los archivos **Parquet** tradicionales (Avril, 2024).

- **Seguridad y gobernanza de datos**

La seguridad y la gobernanza de datos son cruciales para el cumplimiento normativo y la protección de información sensible. Los **Data Warehouses** ofrecen un control más estricto sobre la calidad y la integridad de los datos debido a su estructura organizada. En contraste, los **Data Lakes** pueden presentar riesgos de calidad y seguridad si no se gestionan adecuadamente (Derakhshannia et al., 2019). Los **Data Lakehouses** implementa políticas de gobernanza robustas que garantizan un control riguroso sobre las capas donde se almacenan y procesan los datos, lo que permite una gestión coherente y alineada con las mejores prácticas de la industria (Cherradi & Haddadi, 2024a).



Por otro lado, los **Delta Lakes** proporcionan capacidades avanzadas de control de versiones, auditoría y acceso basado en permisos dentro de los **Data Lakes**, alineándose con las mejores prácticas de la industria y fortaleciendo la seguridad y gobernanza del almacenamiento de datos.

4. Desarrollo del proyecto y resultados

4.1. Metodología

El proyecto se ha desarrollado siguiendo una metodología estructurada, basada en la planificación y gestión de tareas para cumplir con los objetivos propuestos. A continuación, se detalla el proceso seguido:

1. **Evaluación inicial de tareas:** Se identificaron las tareas necesarias para completar el trabajo, desde la revisión bibliográfica hasta la implementación práctica y la redacción final.
2. **Planificación y cronograma:** Se estableció un cronograma ajustado al tiempo disponible, considerando las fechas clave y la priorización de tareas.
3. **Seguimiento y gestión:** Se realizó un seguimiento continuo del progreso, ajustando las tareas según fuera necesario para cumplir con los plazos.

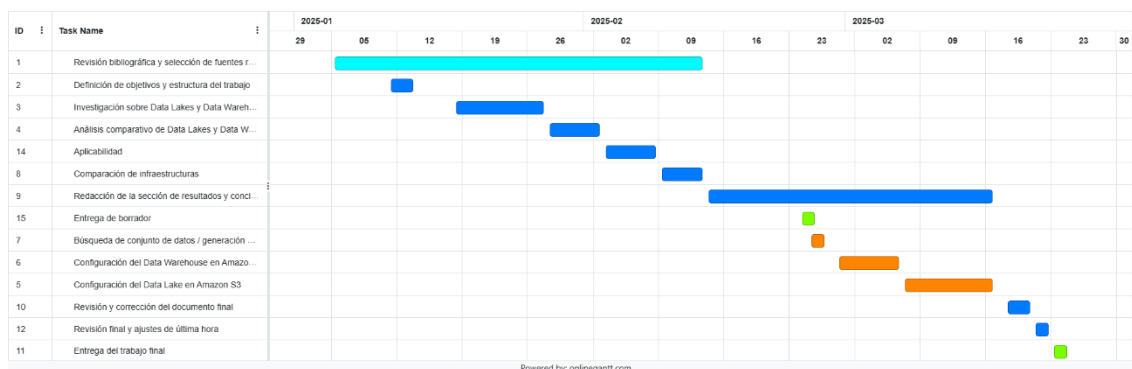


Figura 9. Cronograma de las tareas definidas. Elaboración propia.

4.2. Planteamiento del problema

En la actualidad, la gestión eficiente de grandes volúmenes de datos se ha convertido en un desafío crítico para las organizaciones. La elección de la infraestructura de almacenamiento adecuada no solo impacta en la capacidad de almacenar y procesar datos, sino también en la calidad del análisis y la toma de decisiones. En este contexto, los *Data Lakes* y los *Data Warehouses* han surgido como dos de las soluciones más prominentes, cada una con sus propias ventajas y limitaciones. Sin embargo, la implementación y configuración de estas infraestructuras requieren un entendimiento

profundo de sus características técnicas y operativas, especialmente en entornos cloud como Amazon Web Services (AWS).

El objetivo de esta práctica es implementar y comparar un *Data Lake* en *Amazon S3* y un *Data Warehouse* en *Amazon Redshift*, utilizando un conjunto de datos sintéticos. A través de esta implementación, se busca comprender la configuración básica, el almacenamiento y la consulta de datos en ambas infraestructuras, así como identificar las diferencias clave en términos de rendimiento, escalabilidad y facilidad de uso. Este ejercicio permitirá evaluar cuándo y cómo cada una de estas soluciones puede ser más adecuada según las necesidades específicas de una organización.

El problema central radica en la falta de claridad sobre cómo estas dos tecnologías pueden coexistir o complementarse en un entorno empresarial, especialmente en términos de su implementación práctica en la nube. Aunque ambas soluciones son ampliamente utilizadas, existe una brecha en la comprensión de cómo configurarlas de manera eficiente y cómo aprovechar sus capacidades para maximizar el valor de los datos.

A través de esta implementación práctica, se espera proporcionar una guía clara sobre cómo configurar y utilizar estas tecnologías en AWS, así como ofrecer recomendaciones sobre cuándo y cómo implementarlas en función de las necesidades específicas de una organización. Este trabajo contribuirá a cerrar la brecha entre la teoría y la práctica, ofreciendo información de valor para profesionales y organizaciones que buscan optimizar su gestión de datos en la nube.

4.3. Desarrollo del proyecto

4.3.1. *Data Warehouse* en Amazon Redshift

Amazon Redshift es un servicio de *Data Warehouse* en la nube de AWS **que permite gestionar grandes volúmenes de datos de manera eficiente mediante SQL**. Su capacidad de escalabilidad y procesamiento masivo lo hace ideal para análisis de datos a gran escala.

El objetivo de esta práctica es implementar un **Data Warehouse** utilizando **Amazon Redshift** para gestionar datos de ventas. Se ha optado por un modelo **estrella**, compuesto por una **tabla de hechos** (ventas_fact) y **tablas de dimensiones** (clientes y producto). Este modelo facilita el análisis de grandes volúmenes de datos de manera eficiente, permitiendo consultas optimizadas para reportes de negocio.

Para la carga de datos, se utilizan archivos **CSV almacenados en Amazon S3**, los cuales se importan a Redshift mediante la instrucción COPY. Posteriormente, se realizan consultas SQL para analizar las ventas y generar reportes de productos más vendidos.

A continuación, se detallan los pasos para la realización de la práctica:

- **Configuración del clúster en Amazon Redshift**

- Acceder a la consola de **AWS** y buscar **Amazon Redshift**.



Figura 10. Página de inicio de Redshift con opción de prueba gratuita sin servidor. Elaboración propia en la plataforma AWS.

- Seleccionar la opción **Crear Clúster**.
- Ingresar el nombre del clúster (**practica-dw**) y definir los parámetros de configuración.
- Seleccionar la opción de generación automática de la base de datos llamada **dev**.
- Configurar las credenciales IAM para habilitar el acceso seguro.
- Guardar los cambios y proceder con la creación del clúster.

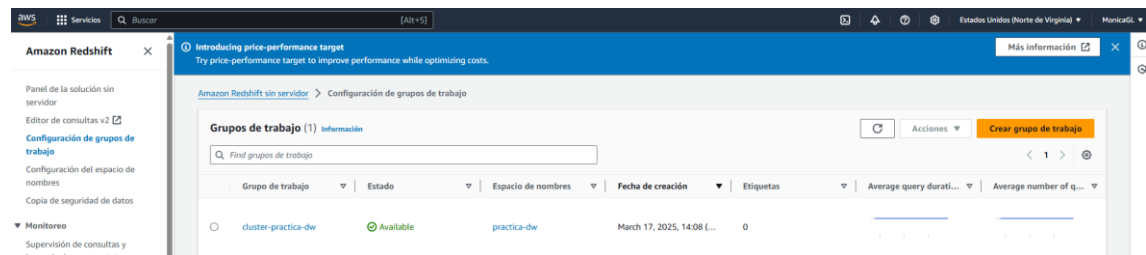


Figura 11. Grupo de trabajo creado con los pasos antes indicados. Elaboración propia en la plataforma AWS.

- **Creación de la base de datos**

- Ingresar al **Query Editor** dentro de la consola de Redshift.
- Seleccionar el clúster previamente creado.
- Presionar el botón **“Create”** y elegir la opción **“Database”**.
- Especificar el nombre de la base de datos (**ventas_db**).
- Confirmar la creación y verificar su correcta configuración.

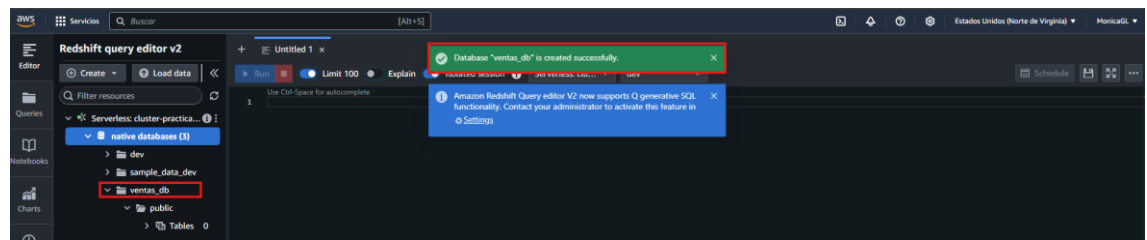


Figura 12. Base de datos creada con pasos anteriores. Elaboración propia en la plataforma AWS.

- **Definición de la estructura del Data Warehouse.** El modelo de datos se estructura de la siguiente manera:



Figura 13. Modelo Entidad-Relación para el Data Warehouse que se creó como práctica. Elaboración propia

- **Creación de tablas en Redshift**
 - Acceder al **Query Editor v2** en la consola de Redshift.
 - Presionar el botón **“Create”** y seleccionar **“Table”**.
 - Especificar el nombre de la tabla (clientes, dim_producto, ventas_fact).
 - Definir manualmente las columnas y asignar los respectivos tipos de datos.
 - Confirmar la creación de la tabla presionando **“Create Table”**.

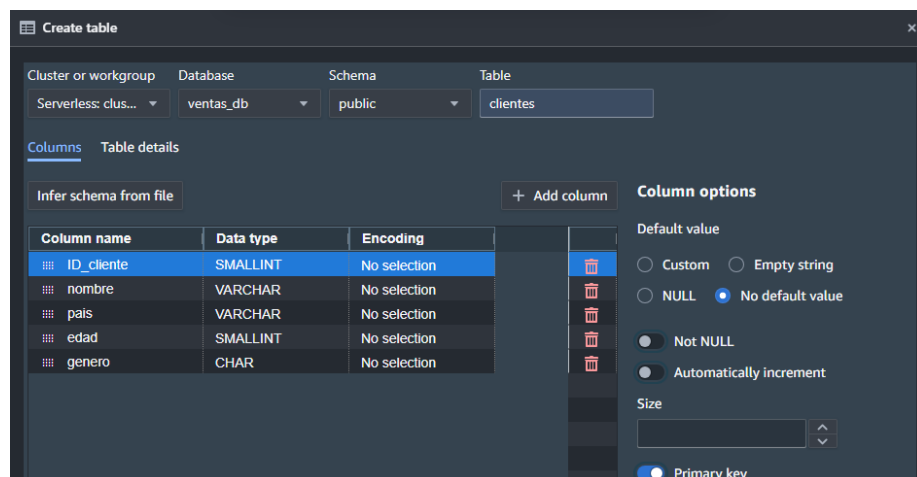


Figura 14. Creación de la tabla clientes desde la opción “Crear tabla”. Elaboración propia en la plataforma AWS.

La creación de las tablas productos y ventas_fact se realizó a través de queries como se evidencia en las siguientes imágenes:

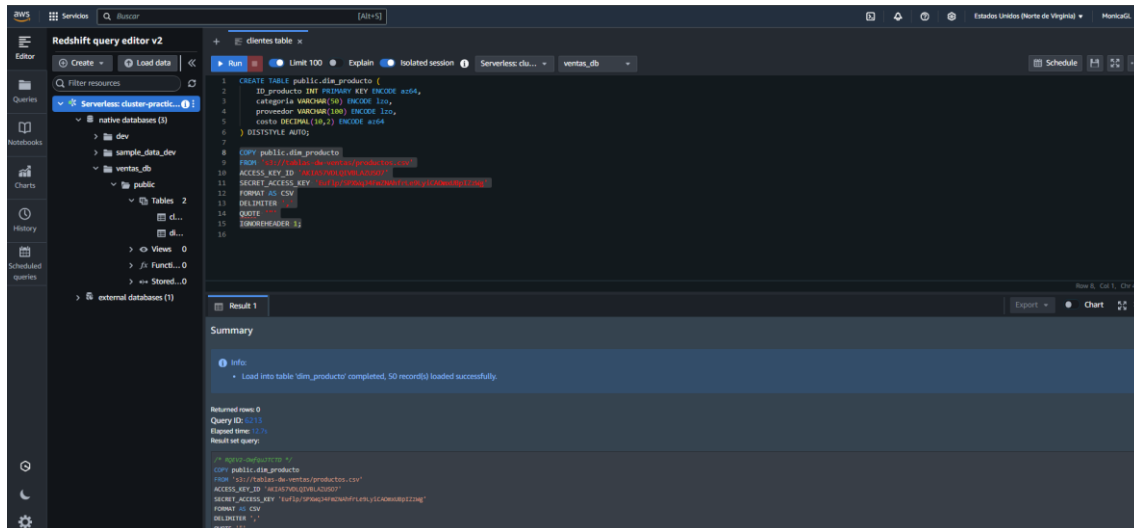


Figura 15. Creación de tabla producto y carga de datos desde S3. Elaboración propia en la plataforma AWS.

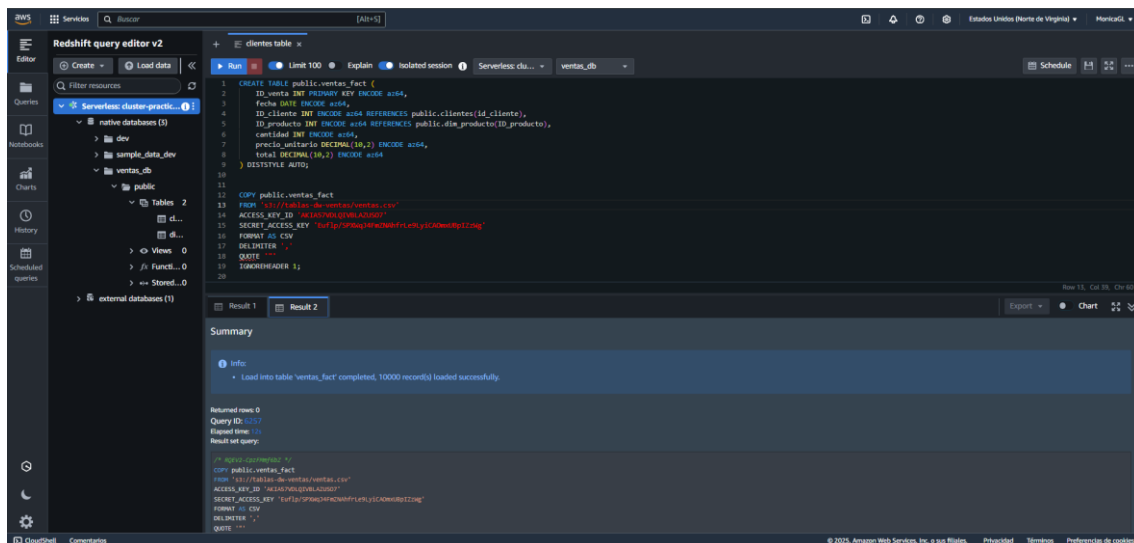


Figura 16. Creación de tabla ventas_fact y carga de datos desde S3. Elaboración propia en la plataforma AWS.

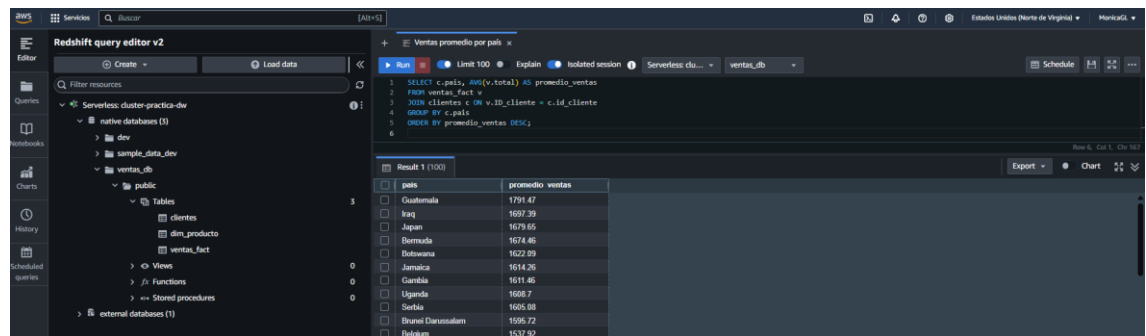
- **Carga de datos en las tablas.** Los datos a cargar se encuentran en archivos CSV almacenados en Amazon S3. Para importar la información en Redshift, se ejecuta el siguiente comando COPY:

```
COPY clientes
FROM 's3://tablas-dw-ventas/clientes.csv'
ACCESS_KEY_ID 'xxxxx'
SECRET_ACCESS_KEY xxxxxxxxxxxxxx'
FORMAT AS CSV
DELIMITER ','
```

```
QUOTE ' '
IGNOREHEADER 1;
```

El mismo procedimiento se aplica para la carga de las tablas `ventas_fact` y `dim_producto`, reemplazando el nombre del archivo y la tabla destino.

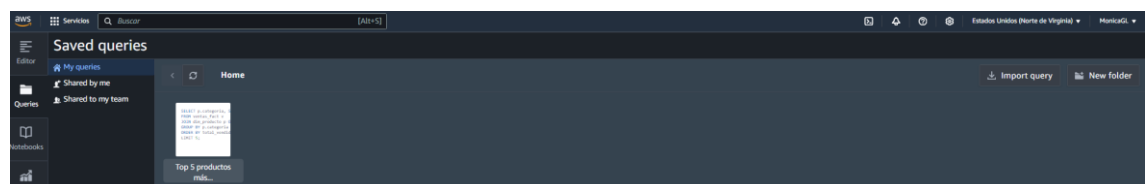
- **Análisis de datos con SQL.** Una vez cargados los datos en Redshift, se pueden ejecutar consultas SQL para analizar la información. A continuación, se presentan algunos ejemplos:
 - **Obtener los 5 productos más vendidos:**



pais	promedio ventas
Guatemala	1791.47
Iraq	1697.39
Japan	1679.65
Bermuda	1674.46
Indonesia	1622.39
Jamaica	1614.36
Guadalupe	1611.46
Uganda	1608.7
Serbia	1605.08
Brunei Darussalam	1595.72
Bahamas	1537.42

Figura 17. Consulta para calcular los productos más vendidos. Elaboración propia en la plataforma AWS.

- **Calcular el promedio de ventas por país:**



pais	promedio ventas
Guatemala	1791.47
Iraq	1697.39
Japan	1679.65
Bermuda	1674.46
Indonesia	1622.39
Jamaica	1614.36
Guadalupe	1611.46
Uganda	1608.7
Serbia	1605.08
Brunei Darussalam	1595.72
Bahamas	1537.42

Figura 18. Consulta para calcular el promedio de ventas por país. Elaboración propia en la plataforma AWS.

Las consultas realizadas se pueden guardar en la opción de Queries que tiene Redshift para tal fin

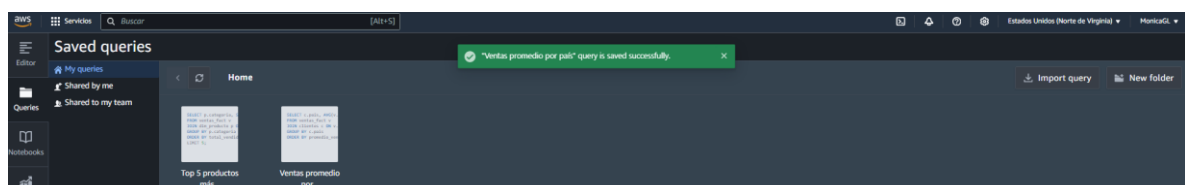


Figura 19. Consultas guardadas en Redshift para uso posterior. Elaboración propia en la plataforma AWS.

4.3.2. *Data Lake* en Amazon S3

Amazon S3 (*Simple Storage Service*) es un **servicio de almacenamiento en la nube altamente escalable que permite gestionar grandes volúmenes de datos en distintos formatos**. Su flexibilidad y capacidad de integración con otros servicios de AWS lo hacen ideal para la implementación de *Data Lakes*.

El objetivo de esta práctica es implementar un *Data Lake* en Amazon S3 utilizando una estructura de tres capas (*bronze*, *silver* y *gold*), permitiendo la ingestión, transformación y análisis de datos. En este caso, se procesarán datos sintéticos de clientes, los cuales serán almacenados inicialmente en la capa *bronze*, luego serán limpiados y normalizados en *silver* mediante AWS Lambda, y finalmente estarán listos para su consumo en *gold*.

A continuación, se detallan los pasos para la implementación del *Data Lake* en Amazon S3:

- **Creación del Bucket en Amazon S3.** Amazon S3 será el repositorio principal del *Data Lake*, donde se almacenarán los datos en diferentes niveles de procesamiento.
 - Iniciar sesión en AWS Console y buscar el servicio S3.
 - Hacer clic en “**Crear bucket**”.
 - Ingresar un nombre único para el bucket, por ejemplo: `data-lake-x`.
 - Seleccionar la región más cercana a la ubicación del usuario.
 - Configurar las opciones de acceso y permisos según los requisitos de seguridad.
 - Mantener la opción de versionado desactivada para evitar costos innecesarios.
 - Hacer clic en “**Crear bucket**” para completar la configuración.

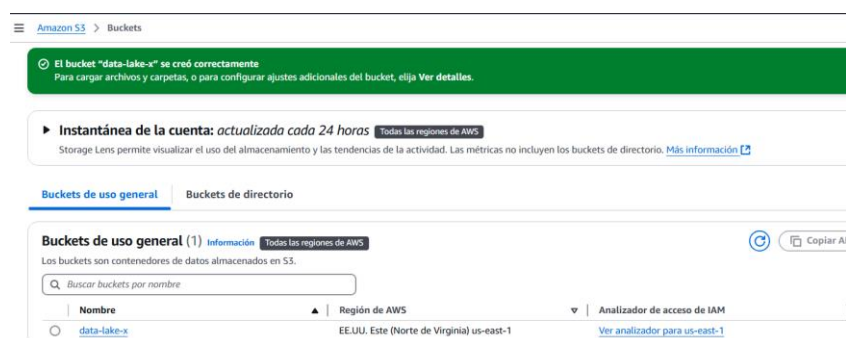


Figura 20. Creación de Bucket en S3. Elaboración propia en la plataforma AWS

- **Definición de la estructura del Data Lake.** Este modelo se organiza en tres capas lógicas para garantizar un procesamiento estructurado de la información:
 - **Capa Bronze:** Almacena los datos crudos sin modificaciones.
 - **Capa Silver:** Contiene datos limpios y procesados listos para análisis.
 - **Capa Gold:** Aloja datos agregados y optimizados para consumo en analítica avanzada y modelos de *machine learning*.

Para crear la estructura en S3:

- Ingresar al *bucket* data-lake-x en S3.
- Crear las siguientes las carpetas *bronze*, *silver* y *gold* dentro del *bucket*
- Guardar los cambios.

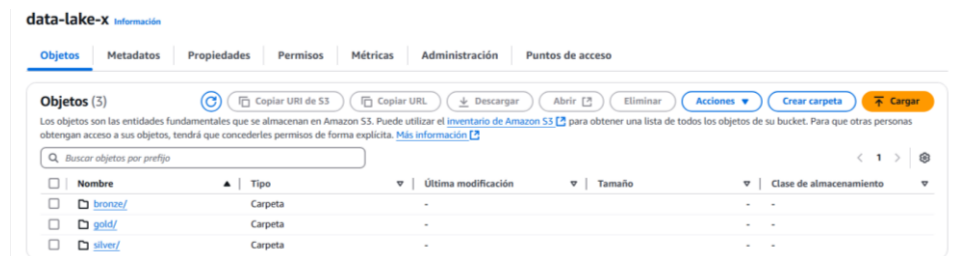


Figura 21. Creación de capas (carpetas) dentro del bucket. Elaboración propia en la plataforma AWS

- **Carga de datos en la capa Bronze.** Para simular datos reales, se utilizará un archivo clientes.csv con valores nulos y datos que requieren modificación. Este archivo se subirá a la capa bronze.
 - Acceder a **AWS S3** y navegar hasta el *bucket* data-lake-x.
 - Ingresar a la carpeta bronze/.
 - Hacer clic en **“Cargar”**, seleccionar clientes.csv, logo.png y productos.json (Archivos estructurados, no estructurados y semiestructurados) y subirlos al *Data Lake*.

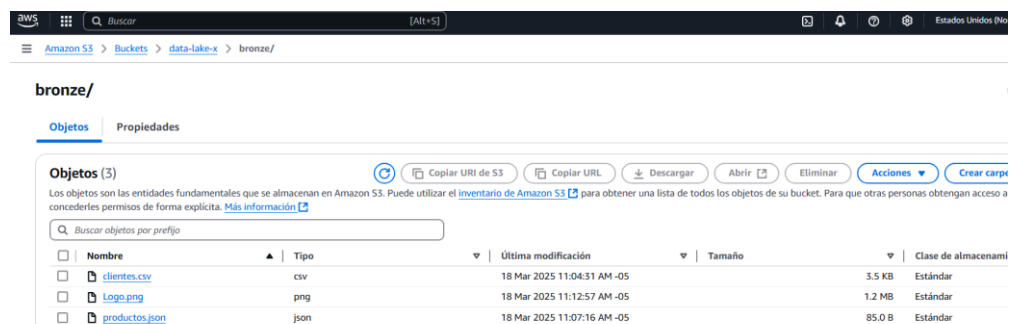


Figura 22. Carga de datos en crudo en la capa bronze. Elaboración propia en la plataforma AWS

- **Limpieza de datos con AWS Lambda.** La función procesará automáticamente los archivos almacenados en la capa *bronze* para eliminar inconsistencias y mejorar la calidad de los datos. Para la **creación de la Función Lambda seguimos los siguientes pasos**
 - Iniciar sesión en **AWS Lambda** y hacer clic en “**Crear función**”.
 - Seleccionar “**Crear desde cero**”.
 - Ingresar el nombre: `limpiar_clientes_s3`.
 - Seleccionar **Python 3.13** como entorno de ejecución.
 - Configurar permisos *IAM* para permitir acceso a S3. Posteriormente incluir *Amazon s3 full Access* desde *IAM*.
 - Hacer clic en “**Crear función**”.

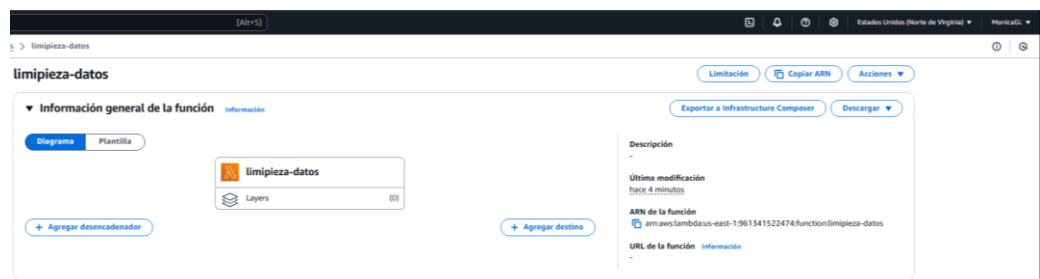


Figura 23. Función creada para limpieza de datos y carga en capa Silver

- **Desarrollo del código en Lambda.** El código de Lambda eliminará filas nulas y cambiará género F por Femenino y M por Masculino (ver anexo X).
- **Automatización del proceso con notificaciones S3.** Para que la limpieza de datos sea automática, se configura un evento en S3 que dispare la función Lambda cuando se suba un archivo a *bronze/*.
 - Ir a **AWS S3** y seleccionar el *bucket* data-lake-x.
 - Navegar a **Propiedades** → **Eventos** → **Crear notificación**.
 - Asignar un nombre
 - Seleccionar el evento **PUT** (cuando se suban archivos).
 - Definir el prefijo de monitoreo: *bronze/*.
 - Seleccionar **AWS Lambda** como destino y asignar la función *limpieza-datos*.
 - Guardar los cambios.

Configuración del desencadenador

Bucket: s3/data-lake-x

Tipos de eventos: PUT

Prefijo - Opcional: bronze/

Sufijo - Opcional: p-*.jpg

- **Verificación del proceso de limpieza.** Para comprobar que el proceso de limpieza funciona correctamente:
 - Subir un archivo con datos erróneos a *bronze/*.
 - Esperar unos segundos y verificar que el archivo limpio ha sido movido a *silver/*. Para esto también se ha usado la opción de prueba de la función lambda.
 - Revisar el contenido del archivo para confirmar que no contiene filas con valores vacíos y que la columna de género ha sido normalizada (F → Femenino, M → Masculino).

Objetos (1/1)

Nombre	Tipo	Última modificación	Tamaño	Clase de almacenamiento
clientes_limpios.csv	csv	18 Mar 2025 2:17:22 PM -05	4.3 KB	Estándar

Figura 24. Resultado de la ejecución de la función lambda. Elaboración propia en la plataforma AWS

Finalmente, se consultan las primeras 10 filas del archivo previo y el archivo que ha sido transformado, a través de AWS CLI

```
PS C:\Users\monic> aws s3 cp s3://data-lake-x/bronze/clientes.csv - | Select-Object -First 10 ID_cliente , nombre , pais , edad , genero
1,James Hayes,Bolivia,33,F
2,Lucas Morgan,Saint Vincent and the Grenadines,39,F
3,Jennifer Garcia,Saint Vincent and the Grenadines,51,M
4,Edward Watson,Luxembourg,18,F
5,Haley Morris,South Africa,23,F
6,Carl Foster,Italy,20,M
7,Donna Lucas,Indonesia,61,F
8,Kyle McCoy,Vanuatu,26,M
9,Marie Diaz,Philippines,55,F
```

Figura 25. Archivo con datos en crudo. Elaboración propia en la plataforma AWS

```
Windows PowerShell
Copyright (C) Microsoft Corporation. Todos los derechos reservados.

Instale la versión más reciente de PowerShell para obtener nuevas características y mejoras. https://aka.ms/PSWindows

PS C:\Users\monic> aws s3 cp s3://data-lake-x/silver/clientes_limpios.csv - | Select-Object -First 10
ID_cliente , nombre , pais , edad , genero
1,James Hayes,Bolivia,33,Femenino
2,Lucas Masculinoorgan,Saint Vincent and the Grenadines,39,Femenino
3,Jennifer Garcia,Saint Vincent and the Grenadines,51,Masculino
4,Edward Watson,Luxembourg,18,Femenino
5,Haley Masculinoorris,South Africa,23,Femenino
6,Carl Femeninooster,Italy,20,Masculino
7,Donna Lucas,Indonesia,61,Femenino
8,Kyle Masculinoccoy,Vanuatu,26,Masculino
9,Masculinoarie Diaz,Philippines,55,Femenino
PS C:\Users\monic> |
```

Figura 26. Archivo con datos procesados. Elaboración propia en la plataforma AWS

4.4. Resultados

5. Conclusión recomendaciones y trabajos futuros

5.1. Conclusiones

La evolución del almacenamiento y gestión de datos masivos ha llevado al desarrollo de diversas infraestructuras, cada una con características, ventajas y desafíos particulares. A lo largo de este trabajo, se han examinado en profundidad los modelos de Data Warehouses, Data Lakes, Data Swamps, Delta Lakes y la convergencia de estos en arquitecturas híbridas como los Data Lakehouses.

- **Infraestructuras tradicionales y su evolución.** Los *Data Warehouses* continúan siendo esenciales para la inteligencia de negocios y el análisis estructurado de datos históricos, proporcionando seguridad, gobernanza y optimización en consultas analíticas. Sin embargo, presentan limitaciones en escalabilidad y flexibilidad frente a los volúmenes crecientes de datos no estructurados.
- **El auge de los *Data Lakes* y el desafío de los *Data Swamps*.** Los *Data Lakes* surgieron como una alternativa flexible y escalable, permitiendo el almacenamiento en bruto de datos de distintos formatos. No obstante, sin una estrategia de gobernanza adecuada, pueden evolucionar a *Data Swamps*, donde los datos se vuelven inservibles y de difícil acceso.
- ***Delta Lake* como solución para la calidad y gobernanza de datos.** En respuesta a los problemas de los *Data Lakes*, *Delta Lake* ha demostrado ser una solución eficiente para mejorar la gobernanza de datos, incorporando transacciones *ACID*, control de versiones y optimización de consultas. Su capacidad para garantizar integridad y trazabilidad lo convierte en una alternativa valiosa en sectores con altos requerimientos normativos, como el financiero y el de salud.
- **Casos de uso y aplicabilidad en distintas industrias.** Se ha identificado que cada modelo de almacenamiento tiene aplicaciones óptimas según el contexto empresarial. Mientras los *Data Warehouses* siguen siendo fundamentales en sectores con necesidades estrictas de informes estructurados, los *Data Lakes* han potenciado la analítica avanzada en industrias con datos heterogéneos. Por su parte, *Delta Lake* ha permitido optimizar procesos en tiempo real y mejorar la calidad de los datos, facilitando su aprovechamiento en sectores dinámicos como *retail* y financiero.

- **Data Lakehouses como una solución híbrida emergente.** La combinación de las ventajas de los *Data Lakes* y los Data Warehouses ha dado lugar a los *Data Lakehouses*, que permiten un almacenamiento escalable sin sacrificar la gobernanza y calidad de los datos. Su adopción está en crecimiento, especialmente en empresas que buscan una solución integral para la gestión de datos masivos.

5.2. Recomendaciones

Con base en los hallazgos del estudio, se presentan las siguientes directrices para la implementación de estas tecnologías en entornos empresariales:

1. ¿Cuándo implementar un *Data Lake*?:

- Si la organización trabaja con grandes volúmenes de datos en formatos heterogéneos (estructurados y no estructurados).
- Para aplicaciones de aprendizaje automático, análisis en tiempo real y descubrimiento de patrones complejos.
- Cuando la flexibilidad y la exploración de datos en crudo son más importantes que la inmediatez en el acceso a datos estructurados.
- Si se prevé la necesidad de escalar el almacenamiento de datos sin incurrir en costes elevados de preprocesamiento.

2. Mejores prácticas para evitar un *Data Swamp*:

- Implementar estrategias de gobernanza de datos: estandarizar metadatos, establecer políticas de acceso y asegurar calidad en la ingesta de información.
- Utilizar herramientas de catalogación y linaje de datos para facilitar su identificación y recuperación.

3. ¿Cuándo implementar *Delta Lake*?

- Cuando se requiere mayor confiabilidad y rendimiento en un *Data Lake* sin perder flexibilidad en la ingesta de datos.
- Para gestionar datos en múltiples formatos y fuentes, facilitando su integración en entornos analíticos.
- En casos donde se necesitan transacciones ACID para garantizar la consistencia de los datos y evitar corrupciones o duplicidades.
- Si es necesario optimizar la velocidad de las consultas gracias a la indexación avanzada y la abstracción de metadatos en un registro de transacciones.
- Cuando diferentes equipos necesitan acceder simultáneamente a los mismos datos sin comprometer la calidad ni la seguridad de la información.

- Para ejecutar operaciones de datos complejas como actualizaciones, eliminación de registros o cambios en esquemas sin afectar la estabilidad del sistema.
- En organizaciones que buscan reducir su dependencia de formatos de datos propietarios y garantizar compatibilidad con herramientas de análisis y ciencia de datos.

4. ¿Cuándo utilizar un *Data Warehouse*?:

- Cuando la empresa necesita informes estructurados y reportes regulares.
- Para análisis históricos y toma de decisiones basadas en datos preprocesados.
- Si el negocio requiere cumplimiento normativo y calidad de datos asegurada.
- Cuando las consultas deben ejecutarse con alto rendimiento y baja latencia.

5. ¿Cuándo optar por una solución híbrida (*Data Lakehouse*)?:

- Si la empresa necesita un equilibrio entre gobernanza y flexibilidad, evitando la duplicación de datos en múltiples sistemas.
- Cuando se requiere un único entorno para gestionar tanto datos operativos como datos analíticos.
- Para optimizar los costes de almacenamiento y procesamiento mediante tecnologías que permitan trabajar con datos en crudo y datos procesados en un mismo ecosistema.

6. Consideraciones finales sobre la adopción de estas tecnologías:

- Evaluar las necesidades específicas del negocio antes de elegir una solución de almacenamiento de datos.
- Definir una estrategia clara de integración con otras plataformas empresariales, especialmente en entornos *multicloud*.
- Asegurar la formación del equipo de datos en herramientas de gestión, procesamiento y análisis para maximizar el valor obtenido de estas infraestructuras.

5.3. Trabajos Futuros

A pesar del amplio desarrollo y adopción de estas infraestructuras en la industria, aún existen desafíos y oportunidades de mejora en la gestión de datos masivos. Algunas líneas de investigación y desarrollo futuras incluyen:

- **Gobernanza y calidad de datos en *Data Lakes*.** Se requiere una mayor investigación en estrategias avanzadas de *metadata*, catalogación automática y control de acceso para prevenir la conversión de los *Data Lakes* en *Data Swamps* y garantizar su valor analítico.
- **Optimización del rendimiento en arquitecturas híbridas.** El auge de los *Data Lakehouses* presenta nuevos retos en la integración eficiente de almacenamiento estructurado y no estructurado.
- **Automatización de procesos ETL y ELT.** La incorporación de técnicas de inteligencia artificial y aprendizaje automático en la automatización de la ingesta, limpieza y transformación de datos podría mejorar la eficiencia operativa y reducir los tiempos de procesamiento en entornos de Big Data.
- **Seguridad y cumplimiento normativo en entornos distribuidos.** Con la creciente adopción de arquitecturas *multicloud*, es esencial investigar nuevas estrategias de seguridad, encriptación y control de acceso para garantizar la protección de datos sensibles en infraestructuras descentralizadas.
- **Impacto de la computación cuántica en el procesamiento de datos:** a medida que la computación cuántica avanza, se espera que tenga un impacto significativo en la velocidad y eficiencia de procesamiento de grandes volúmenes de datos. Explorar cómo esta tecnología puede integrarse con las diferentes infraestructuras de gestión y almacenamiento de datos será un área de investigación crucial en los próximos años.

6. Referencias

- Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., Torres, J., van Hovell, H., Ionescu, A., Łuszczak, A., Szafrá nski, M., Li, X., Ueshin, T., Mokhtar, M., Boncz, P., Ghodsi, A., Paranjpye, S., Senster, P., Xin, R., ... Berkeley, U. (n.d.). *Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores*. <https://doi.org/10.14778/3415478.3415560>
- Avril, A. (2024, May 15). *Delta Lake vs Data Lake: ¿cuál es la diferencia?* | Delta Lake. <https://delta.io/blog/delta-lake-vs-data-lake/>
- Aytas, Y. (2021). *Designing Big Data Platforms : How to Use, Deploy, and Maintain Big Data Systems*. John Wiley & Sons, Incorporated. <https://ebookcentral.proquest.com/lib/universidadviu/detail.action?docID=6659001#>
- Cherradi, M., & Haddadi, A. El. (2024a). Data Lakehouse: Next Generation Information System. *Seminars in Medical Writing and Education*, 3, 67–67. <https://doi.org/10.56294/MW202467>
- Cherradi, M., & Haddadi, A. El. (2024b). View of Data Lakehouse: Next Generation Information System. *Seminars in Medical Writing and Education*, 67–67. <https://mw.ageditor.ar/index.php/mw/article/view/48/55>
- Delta Lake vs. Data Lake: diferencias clave* | Airbyte. (n.d.). Retrieved March 6, 2025, from <https://airbyte.com/data-engineering-resources/delta-lake-vs-data-lake>
- Derakhshannia, M., Gervet, C., Hajj-Hassan, H., Laurent, A., & Martin, A. (2019). Life and Death of Data in Data Lakes: Preserving Data Usability and Responsible Governance. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11938 LNCS, 302–309. https://doi.org/10.1007/978-3-030-34770-3_24
- Díaz, J. C., & Caralt, J. C. (2015). *¿Cómo crear un data warehouse?* 106. <http://elibro.bibliotecabuap.elogim.com/es/lc/bibliotecasbuap/titulos/114035>
- Divya Meena, M. S., Vidhya, M. S., & Be-Cse, M. (n.d.). DATA LAKES-A NEW DATA REPOSITORY FOR BIG DATA ANALYTICS WORKLOADS. *International Journal of Advanced Research in Computer Science*, 7(5). Retrieved January 31, 2025, from <http://www.clir.org/pubs/reports/pub160/pub160.pdf>
- Divya Meena, M. S., Vidhya, M. S., & Be-Cse, M. (2016). DATA LAKES-A NEW DATA REPOSITORY FOR BIG DATA ANALYTICS WORKLOADS. *International Journal of Advanced Research in Computer Science*, 7(5). <http://www.clir.org/pubs/reports/pub160/pub160.pdf>

- Dubey, A. (2020). *Data Lakes vs. Data Warehouses – common arguments* - ProQuest. <https://www.proquest.com/docview/2434752874/fulltext/51925C872D9F43EBPQ/1?accountid=198016&sourcetype=Trade%20Journals>
- Eshghi, K. (2022, July 21). *How Financial Services Can Enable Modern Data Platforms For Digital Transformation*. Forbes Technology Council. <https://www.forbes.com/councils/forbestechcouncil/2022/07/21/how-financial-services-can-enable-modern-data-platforms-for-digital-transformation/>
- Fis, E. (2024). *Guía comparativa de los mejores data lakes en la nube 2024* - Data IQ. <https://dataiq.com.ar/blog/guia-mejores-data-lakes-2024/>
- Gupta, P. (2023). Beyond Banking: The Trailblazing Impact of Data Lakes on Financial Landscape Sivakumar Ponnusamy. *International Journal of Computer Applications*, 185(47), 975–8887.
- Harby, A. A., & Zulkernine, F. (2025). Data Lakehouse: A survey and experimental study. *Information Systems*, 127, 102460. <https://doi.org/10.1016/J.IS.2024.102460>
- Johnson, O., Brown, W., & Wilson, G. (2024). *Examining the Impact of Technology Adoption on Marketing Strategies in Retail*. <https://doi.org/10.20944/PREPRINTS202407.1215.V1>
- Lakehouse for Retail Overview | Databricks*. (n.d.). Retrieved February 26, 2025, from <https://www.databricks.com/glossary/lakehouse-for-retail>
- Mckendrick, J. (2020). *The future of Analytics: Leveraging Data Lakes and Data Warehouses* - ProQuest. <https://www.proquest.com/docview/2463167769?parentSessionId=TQrCDMeO9scTALihjECVytfoXaEQbHRqvcl2sRnCITA%3D&pq-origsite=summon&accountid=198016&sourcetype=Trade%20Journals>
- Millalen, A. (2022, August 22). *AWS vs Azure vs Google vs Snowflake, cual es el mejor Data Warehouse en la nube | El Blog de Ale*. <https://alejandromillalen.com/aws-vs-azure-vs-google-vs-snowflake-cual-es-el-mejor-data-warehouse-en-la-nube/>
- Nambiar, A., & Mundra, D. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data and Cognitive Computing*, 6(4). <https://doi.org/10.3390/BDCC6040132>
- Núria, E. (n.d.). *¿ETL o ELT? Diferencias y casos de uso*. Retrieved February 1, 2025, from <https://blog.bismart.com/etl-o-elt-diferencias-y-casos-de-uso>
- Olavsrud, T. (2017). *3 keys to keep your data lake from becoming a data swamp* - ProQuest. <https://www.proquest.com/docview/1933320250?parentSessionId=QCe%2FARhiehFDauFqaihPMug7vtjq3fJ3QEYg647oYNiA%3D&pq-origsite=summon&accountid=198016&sourcetype=Trade%20Journals>

- Ortega Candel, J. Manuel. (2023). *Big data, machine learning y data science en Python*. RA-MA Editorial.
- Pagidi, R. K., Kolli, R. K., Mokkapati, C., Goel, O., Khan, Dr. S., & Jain, Prof. (Dr.) A. (2022). Enhancing ETL Performance Using Delta Lake in Data Analytics Solutions. *Universal Research Reports*, 9(4), 473–495. <https://doi.org/10.36676/URR.V9.I4.1381>
- Pappil Kothandapani, H. (2023, June 29). (PDF) *EMERGING TRENDS AND TECHNOLOGICAL ADVANCEMENTS IN DATA LAKES FOR THE FINANCIAL SECTOR: AN IN-DEPTH ANALYSIS OF DATA PROCESSING, ANALYTICS, AND INFRASTRUCTURE INNOVATIONS*. https://www.researchgate.net/publication/386275841_EMERGING_TRENDS_AND_TECHNOLOGICAL_ADVANCEMENTS_IN_DATA_LAKES_FOR_THE_FINANCIAL_SECTOR_AN_IN-DEPTH_ANALYSIS_OF_DATA_PROCESSING_ANALYTICS_AND_INFRASTRUCTURE_INNOVATIONS
- ¿Qué es un almacén de datos? | IBM. (n.d.). Retrieved February 1, 2025, from <https://www.ibm.com/es-es/topics/data-warehouse>
- Romero-Chuquital, A., & Melendres-Velasco, J. J. (2023). Uso de data Warehouse para la toma de decisiones empresariales: una revisión literaria. *Revista Científica de Sistemas e Informática*, 3(2), e543. <https://doi.org/10.51252/RCSI.V3I2.543>
- Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2024). The Lakehouse: State of the Art on Concepts and Technologies. *SN Computer Science*, 5(5), 1–39. <https://doi.org/10.1007/S42979-024-02737-0/TABLES/8>
- Seethala, S. C. (2020). The Role of AI in Revolutionizing Finance Data Warehouses for Predictive Financial Modeling. *SSRN Electronic Journal*. <https://doi.org/10.2139/SSRN.5113359>
- Tomcy, J., & Pankaj, M. (2017). *Data Lake for Enterprises*. Packt Publishing Ltd.
- Torreglosa, M. (2023, January 31). *Data Swamp: ¿Qué es y cómo evitarlo? - Marcos Torreglosa*. <https://n4gash.com/data-swamp-que-es-y-como-evitarlo/>
- Vanga, R. R. (2024). *ETL vs ELT: Evolving Approaches to Data Integration*. 6(5). <https://www.ijfmr.com/papers/2024/5/29481.pdf>