**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 | Alumna:  Garro López, Mónica María D.N.I: 32299176  Director/a de TFM: Gema Pérez Martínez | Convocatoria:  Primera |

3 marzo 2025







24 de marzo deÍndice

[Resumen 7](#_Toc193466788)

[1. Introducción 8](#_Toc193466790)

[1.1. Motivación del Estudio 8](#_Toc193466791)

[1.2. Estructura del Trabajo 9](#_Toc193466792)

[2. Objetivos 10](#_Toc193466793)

[3. Estado del Arte y Marco teórico 12](#_Toc193466794)

[3.1. Fundamentos de](#_Toc193466803) *[Data Lakes](#_Toc193466803)* [y](#_Toc193466803) *[Data Warehouses](#_Toc193466803)* [12](#_Toc193466803)

[3.1.1. Clasificación de los tipos de datos y su impacto en el almacenamiento empresarial 12](#_Toc193466805)

[3.1.2. Procesos](#_Toc193466808) *[ETL, ELT](#_Toc193466808)* [y](#_Toc193466808) *[ETLT](#_Toc193466808)* [para la gestión de datos 14](#_Toc193466808)

[3.1.3. Datos Empresariales y su gestión en la era del](#_Toc193466810) *[Big Data](#_Toc193466810)* [15](#_Toc193466810)

[3.1.4. ¿Qué es un](#_Toc193466812) *[Data Lake?](#_Toc193466812)* [16](#_Toc193466812)

[3.1.5. ¿Qué es un](#_Toc193466813) *[Data Warehouse?](#_Toc193466813)* [20](#_Toc193466813)

[3.1.6. ¿Qué son los](#_Toc193466814) *[Data Marts?](#_Toc193466814)* [23](#_Toc193466814)

[3.2. Evolución y contraste de](#_Toc193466815) *[Data Lakes](#_Toc193466815)* [y](#_Toc193466815) *[Data Warehouses](#_Toc193466815)* [26](#_Toc193466815)

[3.2.1. Diferencias Técnicas y Operativas 26](#_Toc193466816)

[3.2.2. Gobernanza y Calidad de Datos 27](#_Toc193466875)

[3.2.3. Innovaciones Tecnológicas 28](#_Toc193466877)

[3.3. Aplicación de arquitecturas de datos en diferentes industrias 31](#_Toc193466879)

[3.3.1. Sector Financiero 31](#_Toc193466881)

[3.3.2. Sector Salud 34](#_Toc193466882)

[3.3.3. Sector](#_Toc193466883) *[Retail](#_Toc193466883)* [40](#_Toc193466883)

[3.3.4. Desafíos en las diferentes industrias 44](#_Toc193466884)

[3.4. Factores clave en la selección de la Infraestructura de Almacenamiento 45](#_Toc193466885)

[4. Desarrollo del proyecto y resultados 49](#_Toc193466886)

[4.1. Metodología 49](#_Toc193466887)

[4.2. Planteamiento del problema 49](#_Toc193466888)

[4.3. Desarrollo del proyecto 50](#_Toc193466889)

[4.3.1.](#_Toc193466890) *[Data Lake](#_Toc193466890)* [en Amazon S3 50](#_Toc193466890)

[4.3.2.](#_Toc193466891) *[Data Lake en Azure](#_Toc193466891)* [54](#_Toc193466891)

[4.3.3.](#_Toc193466892) *[Data Warehouse](#_Toc193466892)* [en Amazon Redshift 58](#_Toc193466892)

[4.3.4.](#_Toc193466893) *[Data](#_Toc193466893)* [Warehouse](#_Toc193466893)[en](#_Toc193466893) *[Azure Synapse Analytics](#_Toc193466893)* [61](#_Toc193466893)

[4.3.5.](#_Toc193466894) *[Delta Lake en Databricks](#_Toc193466894)* [64](#_Toc193466894)

[1.1. Resultados 68](#_Toc193466895)

[5. Conclusión recomendaciones y trabajos futuros 69](#_Toc193466896)

[5.1. Conclusiones 69](#_Toc193466898)

[5.2. Recomendaciones 70](#_Toc193466908)

[5.3. Trabajos Futuros 70](#_Toc193466909)

[6. Referencias 72](#_Toc193466937)

Índice de ilustraciones

**Figura 1**. Tipos de datos en Big Data (Ortega Candel, 2023) 13

**Figura 2**. Comparación entre ETL y ELT. (Núria, n.d.-a) 14

**Figura 3**. Comparación entre Data Lake y Data Swamp (Torreglosa, 2023) 17

**Figura 4**. Comparativa de Plataformas de Data Lakes Hiperescaladores en la Nube y Soluciones Multicloud. (Fis, 2024) 20

**Figura 5**. Esquema Estrella (¿Qué Es Un Almacén de Datos? | IBM, n.d.) 22

**Figura 6**. Esquema Copo de Nieve (¿Qué Es Un Almacén de Datos? | IBM, n.d.) 22

**Figura 7**. Método Kimball (Bottom up). Elaboración propia. 24

**Figura 8**. Método Inmon (Top down). Elaboración propia 25

**Figura 9**. Cronograma de las tareas definidas. Elaboración propia. 49

**Figura 20**. Creación de Bucket en S3. Elaboración propia en la plataforma AWS 51

**Figura 21**. Creación de capas (carpetas) dentro del bucket. Elaboración propia en la plataforma AWS 51

**Figura 22**. Carga de datos en crudo en la capa bronze. Elaboración propia en la plataforma AWS 52

**Figura 23**. Función creada para limpieza de datos y carga en capa Silver 52

**Figura 24.** Configuración de eventos. Elaboración propia en la plataforma AWS. 53

**Figura 25**. Resultado de la ejecución de la función lambda. Elaboración propia en la plataforma AWS 53

**Figura 26**. Archivo con datos en crudo. Elaboración propia en la plataforma AWS 54

**Figura 27**. Archivo con datos procesados. Elaboración propia en la plataforma AWS 54

**Figura 33**. Arquitectura del Data Lake en Azure. Elaboración propia 55

**Figura 34**. Creación del Contenedor ventas en Blob Storage. Elaboración propia en la plataforma Azure 55

**Figura 35**. Creación del Contenedor bronze en Data Lake. Elaboración propia en la plataforma Azure 55

**Figura 36**. Configuración de Linked Services en Azure Data Factory. Elaboración propia en la plataforma Azure 56

**Figura 37**. Creación del Dataset ds\_clientes\_bronze\_bs en Azure Data Factory. Elaboración propia en la plataforma Azure 56

**Figura 38**. Creación del Dataset ds\_clientes\_bronze\_dl en Azure Data Factory. Elaboración propia en la plataforma Azure 57

**Figura 39**. Desarrollo del Pipeline de Ingesta de Datos (pl\_clientes). Elaboración propia en la plataforma Azure 57

**Figura 40**. Verificación del traslado exitoso del archivo a la capa Bronze 58

**Figura 10**. Página de inicio de Redshift con opción de prueba gratuita sin servidor. Elaboración propia en la plataforma AWS. 58

**Figura 11**. Grupo de trabajo creado con los pasos antes indicados. Elaboración propia en la plataforma AWS. 59

**Figura 12**. Base de datos creada con pasos anteriores. Elaboración propia en la plataforma AWS. 59

**Figura 13**. Modelo Entidad-Relación para el Data Warehouse que se creó como práctica. Elaboración propia 59

**Figura 14**. Creación de la tabla clientes desde la opción “Crear tabla”. Elaboración propia en la plataforma AWS. 60

**Figura 15**. Creación de tabla producto y carga de datos desde S3. Elaboración propia en la plataforma AWS. 60

**Figura 16**. Creación de tabla ventas\_fact y carga de datos desde S3. Elaboración propia en la plataforma AWS. 60

**Figura 17**. Consulta para calcular los productos más vendidos. Elaboración propia en la plataforma AWS. 61

**Figura 18**.Consulta para calcular el promedio de ventas por país. Elaboración propia en la plataforma AWS. 61

**Figura 19**. Consultas guardadas en Redshift para uso posterior. Elaboración propia en la plataforma AWS. 61

**Figura 28**. Creación del área de trabajo en Azure Synapse Analytics. Elaboración propia en la plataforma AWS 62

**Figura 29**. Configuración del SQL Pool en Azure Synapse. Elaboración propia en la plataforma AWS 63

**Figura** 30. Creación de tabla de hechos y dimensiones. Elaboración propia en la plataforma AWS 63

**Figura 31**. Carga de datos desde csv. Elaboración propia en la plataforma AWS 63

**Figura 32**. Consulta para análisis de datos 64

**Figura 41**. Arquitectura Medallion.(Núria, n.d.-b) 64

**Figura 42**. Creación de Clúster. Elaboración propia en la plataforma Databricks 65

**Figura 43**. Creación del Workspace. Elaboración propia en la plataforma Databricks 65

**Figura 44**. Directorio principal. Elaboración propia en la plataforma Databricks 65

**Figura 45**. Subdirectorio de Capa Bronze. Elaboración propia en la plataforma Databricks 66

**Figura 46**. Ingesta de datos crudos. Elaboración propia en la plataforma Databricks 66

**Figura 47**. Transformación de datos y almacenamiento en capa Silver. Elaboración propia en la plataforma Databricks 67

**Figura 48**. Validación de resultados en Capa Gold. Elaboración propia en la plataforma Databricks 67

**Figura 49**. Transformaciones en la Capa Gold. Elaboración propia en la plataforma Databricks 68

Índice de tablas

**Tabla 1**. Comparación entre las metodologías Kimball e Inmon. Elaboración propia 25

**Tabla 2**. 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 3**. 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*.

# 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. Además, **“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 Data Lake, Data Warehouse y Delta Lake** para comprender su configuración básica con datos sintéticos, debido a los altos costos que puede acarrear un caso real con grandes volúmenes de datos.
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.**

# 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 ***D*ata 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 y un Data Warehouse en Amazon y Azure, así como la estructura básica de un Delta Lake en Datbricks** para comprender su configuración básica, almacenamiento y consulta de datos, utilizando un conjunto de datos sintéticos.

# Estado del Arte y Marco teórico

(Tomcy & Pankaj, 2017)

## Fundamentos de *Data Lakes* y *Data Warehouses*

El análisis y gestión de datos empresariales requiere infraestructuras robustas que permitan almacenar, procesar y analizar grandes volúmenes de información de manera eficiente. En este contexto, los *Data Lakes* y *Data Warehouses* emergen como soluciones clave, cada una con características y aplicaciones específicas. Esta sección aborda los fundamentos de estas arquitecturas, incluyendo los tipos de datos que gestionan, los procesos de integración de información (ETL, ELT y ETLT), su impacto en la era del *Big Data* y conceptos relacionados, como los *Data Marts*.

Para comprender el funcionamiento de estas infraestructuras, es fundamental analizar los tipos de datos que manejan las organizaciones y cómo su naturaleza influye en las estrategias de almacenamiento y procesamiento.

### Clasificación de los tipos de datos y su impacto en el 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:

Diagrama

Descripción generada automáticamente

**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 datos organizados bajo una estructura fija, generalmente tabular, lo que facilita su almacenamiento, acceso y análisis mediante sistemas de gestión de 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.

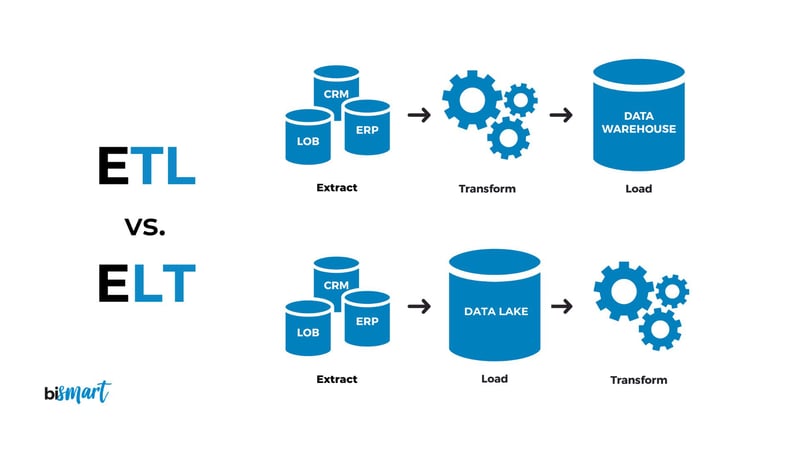
La adecuada clasificación y comprensión de los diferentes tipos de datos no solo es fundamental para un almacenamiento eficiente, sino que también es crucial para implementar estrategias efectivas de gestión de datos.

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 implementar procesos de integración que respondan a los retos de volumen, variedad y velocidad, garantizando respuestas analíticas en tiempo real.

### 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***, este proceso 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.-a)

* **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. En consecuencia, la elección del enfoque adecuado impacta directamente en la calidad de los análisis y en la agilidad de la toma de decisiones.

### 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:

* **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.**
* 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.
* 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.

### ¿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 queestos 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 Lake* 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:

Tabla

Descripción generada automáticamente

**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.

### ¿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 figura 5.

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

**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**. Esto reduce la redundancia de datos y mejora la eficiencia del almacenamiento, aunque puede complicar las consultas debido a la estructura más fragmentada.

Interfaz de usuario gráfica

Descripción generada automáticamente

**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.

* . S(Millalen, 2022)
* . A(Millalen, 2022)
* . L(Millalen, 2022).
* (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 de estos modelos, como los Data Marts, que optimizan la consulta y análisis de información para áreas específicas del negocio.

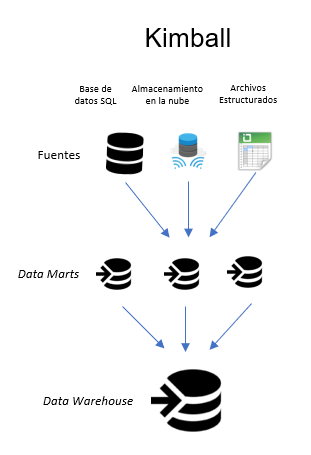
### ¿Qué son los *Data Marts?*

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(Aytas, 2021). Su propósito es optimizar la consulta y análisis de datos en función de las necesidades de distintos usuarios dentro de la organización.

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.

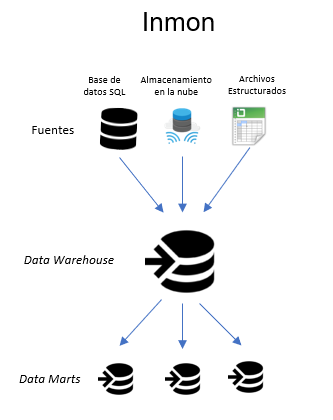
La manera en que se diseñan e implementan los ***Data Marts*** está directamente relacionada con la metodología utilizada en la construcción del Data Warehouse. Dependiendo del enfoque adoptado, los Data Marts pueden ser diseñados desde el inicio como parte integral del almacén de datos o generados posteriormente 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.

1. **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, con base en las siguientes consideraciones:

|  |  |  |
| --- | --- | --- |
| **Criterio** | **Kimball (Bottom-Up)** | **Inmon (Top-Down)** |
| Enfoque inicial | Por Data Marts | Por Data Warehouse |
| Modelo | Dimensional  (Estrella o Copo de Nieve) | Relacional (3FN) |
| Tiempo de implementación | Rápido | Largo |
| Facilidad de mantenimiento | Alta para áreas específicas | Alta consistencia general |
| Flexibilidad | Alta (orientada al negocio) | Baja (orientada a la estructura) |

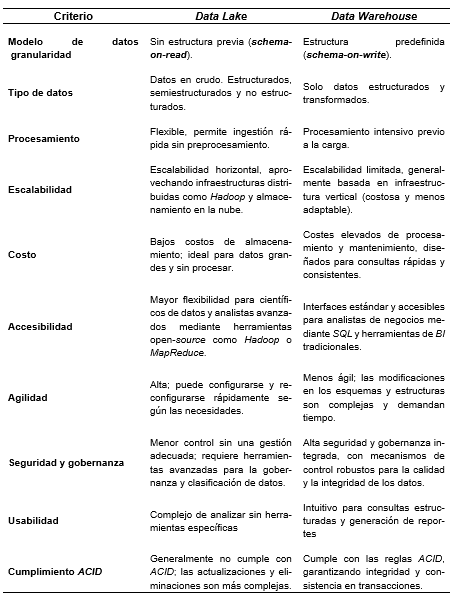
**Tabla 1**. Comparación entre las metodologías Kimball e Inmon. Elaboración propia

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.

## Evolución y contraste de *Data Lakes* y *Data Warehouses*

En la gestión de datos masivos, la elección entre **estos modelos** no solo responden 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). En la sigu, se presenta una comparación detallada entre ambas arquitecturas, enfocándose en sus diferencias técnicas y operativas.

### 



**Tabla 2**. 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).

Además, e

* . (Dubey, 2020) Además, (Olavsrud, 2017)
* ***.*** (Mckendrick, 2020)



### Gobernanza y Calidad de Datos

La gobernanza de datos representa un componente esencial en cualquier infraestructura de gestión de datos masivos, ya que influye directamente en la calidad, integridad, trazabilidad y confiabilidad de la información. Su adecuada implementación puede marcar la diferencia entre el éxito o el fracaso de un sistema de almacenamiento y análisis de datos. A continuación, se presentan las principales características de este aspecto en los entornos de ***Data Lakes*** y ***Data Warehouses*.**

* ***Data Lakes*.** Están **expuestos a mayores riesgos de degradación**, transformándose en *data swamps* **cuando no se establecen políticas claras de gobernanza**. Estos riesgos incluyen la pérdida de trazabilidad, la duplicidad de datos, la baja calidad y la inaccesibilidad de los conjuntos de datos almacenados (Olavsrud, 2017).

Para evitar este deterioro, se **recomienda implementar herramientas de gestión de metadatos y catalogación**, como *Apache Atlas* o *AWS Glue*, que permiten estructurar el acceso, describir los datos disponibles y mejorar su calidad, fomentando una cultura de datos más confiable (Nambiar & Mundra, 2022).

* ***Data Warehouses*.** Incorporan, por diseño, **mecanismos de gobernanza integrados** que aseguran una estructura altamente controlada y estandarizada. Estos mecanismos **incluyen el cumplimiento de principios transaccionales** como **ACID** (Atomicidad, Consistencia, Aislamiento y Durabilidad), lo que garantiza la integridad y fiabilidad de los datos almacenados (Harby & Zulkernine, 2025).

Asimismo, este tipo de infraestructura está alineada con marcos regulatorios y de cumplimiento normativo, siendo común su uso en industrias con altos requerimientos de trazabilidad y auditoría, como el sector financiero o el sanitario.

### 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.

En este orden de idea, esta solución intermedia, combina las ventajas de los *Data Warehouses* y *Data Lakes*, al tiempo que minimiza sus debilidades. La tabla 3 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** | *Schema-on-write* | *Schema-on-read* | 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 3**. 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 sin sacrificar rendimiento analítico.
* Integración nativa de datos estructurados y no estructurados.
* Soporte avanzado para IA y machine learning.
* Reducción del riesgo de data swamp mediante gobernanza y metadatos.

Preservación de datos originales para análisis más precisos.

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

* Curva de aprendizaje y adopción tecnológica.
* Complejidad en la gobernanza híbrida.
* Costes de migración desde arquitecturas legacy.

***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***:***

* Soporte para transacciones *ACID*, que garantiza integridad y consistencia.
* Registro detallado de operaciones mediante un *transaction log* en formato *Parquet*.
* Optimización automática del almacenamiento para mejorar el rendimiento.
* Soporte de *Time Travel*, que permite acceder a versiones históricas de los datos.
* Integración con múltiples motores analíticos como *Apache Spark, Presto, Hive* y *Redshift*.
* Posibilidad de realizar actualizaciones (*upserts*) y auditorías de datos.
* Alta escalabilidad para el manejo de grandes volúmenes.
* Compatibilidad con almacenamiento basado en la nube (Armbrust et al., s.f.).

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.

## Aplicación de arquitecturas de datos en diferentes industrias

En la economía digital actual, los datos representan un recurso estratégico de alto valor. Comprender y aplicar adecuadamente las infraestructuras modernas de almacenamiento y análisis se ha convertido en una prioridad para las organizaciones que buscan mantener su competitividad. Estas soluciones no solo permiten gestionar grandes volúmenes de información, sino que también habilitan la toma de decisiones basadas en análisis avanzados, fortaleciendo la innovación y la eficiencia operativa en diversos sectores.

### 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 regulare**s. Además, muchas de estas infraestructuras proporcionan alta disponibilidad, escalabilidad y cumplimiento de propiedades ACID, 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:

* Escalabilidad para responder al crecimiento exponencial de datos financieros (Gupta, 2023).
* Integración ágil de fuentes internas y externas.
* Reducción de costos de preprocesamiento mediante almacenamiento en bruto (Pappil Kothandapani, 2023).
* Integración con fuentes de datos externas para análisis de riesgo crediticio, detección de fraudes o evaluación de tendencias del mercado.
* Soporte a análisis avanzados y en tiempo real para personalización y cumplimiento normativo

***Data Warehouses* en el Sector Financiero**

Este modelo constituye una pieza esencial en la infraestructura de datos del sector financiero, ya que permiten **almacenar, organizar y analizar grandes volúmenes de datos estructurados** de forma eficiente. Su diseño, optimizado para **consultas rápidas y análisis complejos**, los convierte en herramientas estratégicas para la **toma de decisiones empresariales**, la **gestión de riesgos** y el **cumplimiento normativo** (Romero-Chuquital & Melendres-Velasco, 2023).

Estos sistemas proporcionan información con **alto nivel de gobernanza**, datos **limpios y consolidados**, y estructuras listas para el análisis de tendencias financieras, proyecciones y generación de informes para entes reguladores.

En la práctica, instituciones líderes han demostrado su efectividad. En el caso de **JPMorgan Chase**, la integración de **algoritmos de aprendizaje automático** en su infraestructura de *Data Warehouse* ha fortalecido la **detección de fraudes y la evaluación del riesgo crediticio**, reduciendo significativamente las tasas de incumplimiento y mejorando los protocolos de seguridad. Por su parte, **Bank of America** ha logrado avances similares mediante la implementación de **modelos de detección de anomalías**, que han reforzado la **seguridad operativa** y la **confianza del cliente** en sus servicios (Seethala, 2025).

En este orden de ideas, entre las principales ventajas de los *Data Warehouses* en el ámbito Financiero se encuentran:

* Soporte para decisiones financieras basadas en datos
* Gestión de riesgos y cumplimiento normativo.
* Soporte a la inteligencia de negocios.
* Infraestructura para innovación financiera.

***Data Lakehouses* en el Sector Financiero:**

En el sector financiero, donde la **agilidad operativa**, la **gobernanza de los datos** y el **cumplimiento normativo** son aspectos críticos, los ***Data Lakehouses*** se posicionan como una solución integral para impulsar la innovación, optimizar la toma de decisiones y mantener la competitividad en un entorno altamente regulado y dinámico. Lo anterior, dado que, esta arquitectura permite gestionar de manera simultánea tanto **datos estructurados** (como transacciones, balances o métricas de riesgo) como **datos no estructurados** (como documentos legales, correos electrónicos o registros de voz), lo cual resulta esencial para entidades financieras que requieren consolidar fuentes de datos heterogéneas para realizar análisis más completos y contextualizados.

Además, estos sistemas habilitan el procesamiento tanto en **tiempo real** como por **lotes (batch)**, lo que los convierte en un soporte eficaz para modelos de analítica avanzada, especialmente en tareas de **detección de fraudes** y **personalización de servicios** mediante técnicas de *machine learning*, cada vez más extendidas en el sector.

Otro aspecto destacable es su **estructura de gobernanza**, que facilita el cumplimiento de exigencias regulatorias al asegurar la **calidad, trazabilidad y auditabilidad** de la información utilizada en reportes financieros y procesos de auditoría.

A partir de lo anterior, se pueden destacar las siguientes ventajas clave:

* **Optimización del procesamiento de datos financieros**, integrando múltiples formatos y fuentes en una plataforma unificada.
* **Soporte para análisis avanzados**, incluyendo modelos predictivos e inteligencia artificial.
* **Adaptabilidad a regulaciones**, garantizando el cumplimiento normativo mediante políticas robustas de gobernanza de datos.

***Delta Lakes* en el Sector Financiero:**

Este tipo de infraestructuras se ha consolidado como una herramienta especialmente valiosa para el sector financiero, ya que permite a las instituciones gestionar grandes volúmenes de datos de forma eficiente y segura. Una de sus principales fortalezas es la capacidad de soportar transacciones ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad), lo que proporciona un marco robusto que garantiza la integridad de los datos, aspecto fundamental en una industria donde la precisión y la fiabilidad son imprescindibles.

Gracias a su capacidad para realizar análisis en tiempo real, los Delta Lakes facilitan la toma de decisiones oportunas, permitiendo ejecutar evaluaciones de riesgo, garantizar el cumplimiento normativo y optimizar procesos de auditoría con mayor precisión y agilidad, al disponer de datos actualizados y confiables (Pagidi et al., 2022).

Asimismo, al integrar datos estructurados y no estructurados en un entorno unificado, esta tecnología contribuye a eliminar los silos de información que dificultan una visión global y eficiente de los datos. Esto es especialmente beneficioso en un sector donde la información proviene de múltiples fuentes, tales como transacciones bancarias, mercados de valores y evaluaciones de crédito (Armbrust et al., n.d.).

Finalmente, la funcionalidad de time travel de Delta Lake permite a las instituciones consultar versiones históricas de sus conjuntos de datos, lo cual facilita la trazabilidad de decisiones y transacciones y contribuye al cumplimiento efectivo y transparente de las normativas vigentes (Pagidi et al., 2022).

En resumen, los principales beneficios que los Delta Lakes aportan al sector financiero son:

* Adaptabilidad a condiciones económicas cambiantes.
* Optimización de la gestión de datos mediante la consolidación de múltiples fuentes.
* Acceso a información histórica que fortalece los procesos de auditoría y cumplimiento normativo.

### 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 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. Además, 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..
* **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.
* **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).

***Data Warehouses* en el Sector Salud**

En el ámbito de la salud, los almacenes de datos desempeñan un papel fundamental en la gestión y análisis de información estructurada, ya que facilitan la generación de informes esenciales para la toma de decisiones clínicas informadas, la gestión de riesgos y el cumplimiento normativo. Su implementación no solo optimiza los procesos operativos dentro de las instituciones sanitarias, sino que también mejora la calidad de la atención al paciente al proporcionar un acceso más eficiente y seguro a los datos clínicos.

Estos sistemas están específicamente diseñados para maximizar la velocidad de consulta, permitiendo a los profesionales de la salud acceder rápidamente a información crítica y fundamentar sus decisiones en datos clínicos precisos, actualizados y consolidados.

Además, ofrecen un entorno altamente estructurado y seguro para el almacenamiento de datos, lo cual es especialmente relevante en un sector donde la privacidad y la seguridad de la información del paciente son prioridades ineludibles. Por otro lado, estos sistemas permiten analizar relaciones complejas entre factores clínicos y desenlaces en salud, lo que aporta información clave para la optimización de tratamientos y la mejora continua de las prácticas médicas.

Asimismo, incorporan protocolos avanzados de protección de datos, garantizando la confidencialidad de la información sensible, el cumplimiento de normativas internacionales de privacidad y una reducción significativa de los riesgos asociados a accesos no autorizados o filtraciones de datos personales (Alaa Khalaf Hamoud et al., 2018).

Teniendo en cuenta lo anterior, los principales beneficios de los Data Warehouses en el sector salud pueden resumirse en:

* Acceso rápido a datos estructurados.
* Gobernanza de datos clínicos en entornos seguros.
* Identificación de patrones y correlaciones relevantes para la toma de decisiones médicas.
* Seguridad y privacidad de la información del paciente.

***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.

***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.

### 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. **Gracias a su flexibilidad y escalabilidad, 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.

Entre las ventajas de este tipo de almacenamiento en el entorno 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 realizar, por ejemplo, el análisis de sentimientos en redes sociales en correlación con las cifras de ventas y otros indicadores clave del negocio.
* **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 resaltan:

* **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**. Consolida 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.

En conjunto, estas capacidades hacen que *Delta Lake* represente una solución altamente eficaz para responder a los retos del entorno retail moderno, caracterizado por la volatilidad del comportamiento del consumidor y la necesidad de tomar decisiones basadas en datos confiables y actualizados.

Esta eficacia queda evidenciada en el estudio citado por (Pagidi et al., 2022), que menciona la experiencia de un **retailer analizado por Lee et al. (2020)**. En este caso, la adopción de Delta Lake permitió **mejoras significativas en la eficiencia del procesamiento de datos**, optimizando el análisis de la cadena de suministro y habilitando análisis en tiempo real sobre el comportamiento de compra de los clientes. Como resultado, la empresa pudo ajustar sus estrategias de inventario y marketing de forma más ágil y efectiva.

Esto nos permite concluir que, **este tipo de modelos no solo contribuye a mejorar la eficiencia operativa en el sector *retail***, sino que también fortalece la toma de decisiones estratégicas, **aumentando la satisfacción del cliente y generando un impacto positivo en las ventas**.

***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. **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.

### 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**. Por ejemplo, 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.

En relación con la implementación de *Data Lakehouses,* a pesar de sus múltiples beneficios, estos también enfrentan 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.

Finalmente, entre los desafíos en la adopción de los ***Delta Lakes***, 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.

## Factores clave en la selección de la Infraestructura de Almacenamiento

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 (Mckendrick, 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**. Esto también constituye 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, 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).

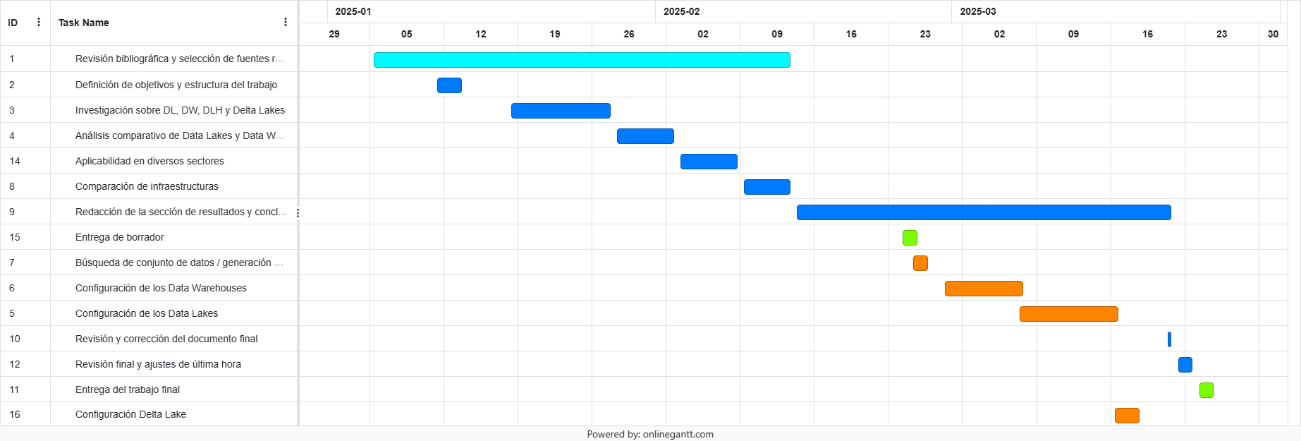
* **Seguridad y gobernanza de datos**. La seguridad y la gobernanza de datos son cruciales para el cumplimiento normativo y la protección de información sensible. Los *Data Warehouses* ofrecen un control más estricto sobre la calidad y la integridad de los datos debido a su estructura organizada. En 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.

# Desarrollo del proyecto y resultados

## 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.

## 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*.

El objetivo principal de este trabajo fue implementar y comparar las arquitecturas de ***Data Lake*** y ***Data Warehouse*** en entornos *cloud* (*Amazon y Azure*), utilizando un conjunto de datos sintéticos. Asimismo, se implementó un **Delta Lake en Databricks**, con el fin de explorar su modelo transaccional y estructura de tres capas (*bronze, silver, gold*). Esta comparativa permite evaluar el desempeño, escalabilidad, facilidad de uso y aplicabilidad de cada tecnología.

El problema central identificado radica en la escasa claridad sobre cómo estas infraestructuras pueden **coexistir, complementarse o diferenciarse en entornos empresariales reales**, especialmente en lo relativo a su configuración y aprovechamiento dentro de plataformas *cloud*. Aunque estas tecnologías son ampliamente conocidas, existe una brecha entre el conocimiento teórico y su aplicación práctica.

Mediante esta implementación guiada, se pretende proporcionar una referencia clara y aplicable para profesionales que deseen optimizar su gestión de datos en la nube, estableciendo **recomendaciones concretas** sobre cuándo y cómo utilizar cada arquitectura según las necesidades de la organización.

## Desarrollo del proyecto

A continuación, se describen las implementaciones prácticas realizadas para evaluar el comportamiento, la facilidad de uso, el rendimiento y la escalabilidad de distintas arquitecturas de almacenamiento de datos en la nube. Con base en el planteamiento del problema y los objetivos definidos, se llevaron a cabo pruebas en cinco plataformas y tecnologías: *Amazon Redshift* y *Azure Synapse Analytics* (como soluciones de *Data Warehouse), Amazon S3 y Azure Data Lake Storage Gen2* (como soluciones de Data Lake), y Databricks con *Delta Lake* (como enfoque híbrido transaccional). Cada implementación se desarrolló utilizando datos sintéticos, aplicando modelos de carga, transformación y consulta, a fin de comparar su aplicabilidad en escenarios empresariales reales.

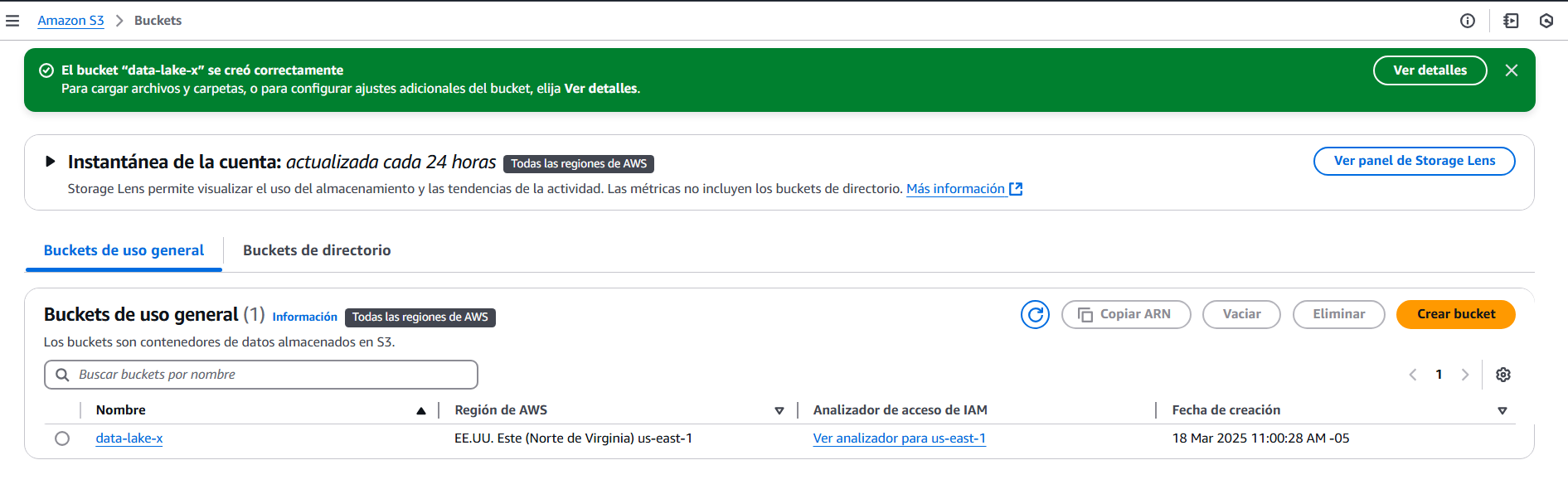
### *Data Lake* en Amazon S3

**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, además se hará un proceso de limpieza normalización previa carga en silver mediante AWS Lambda.

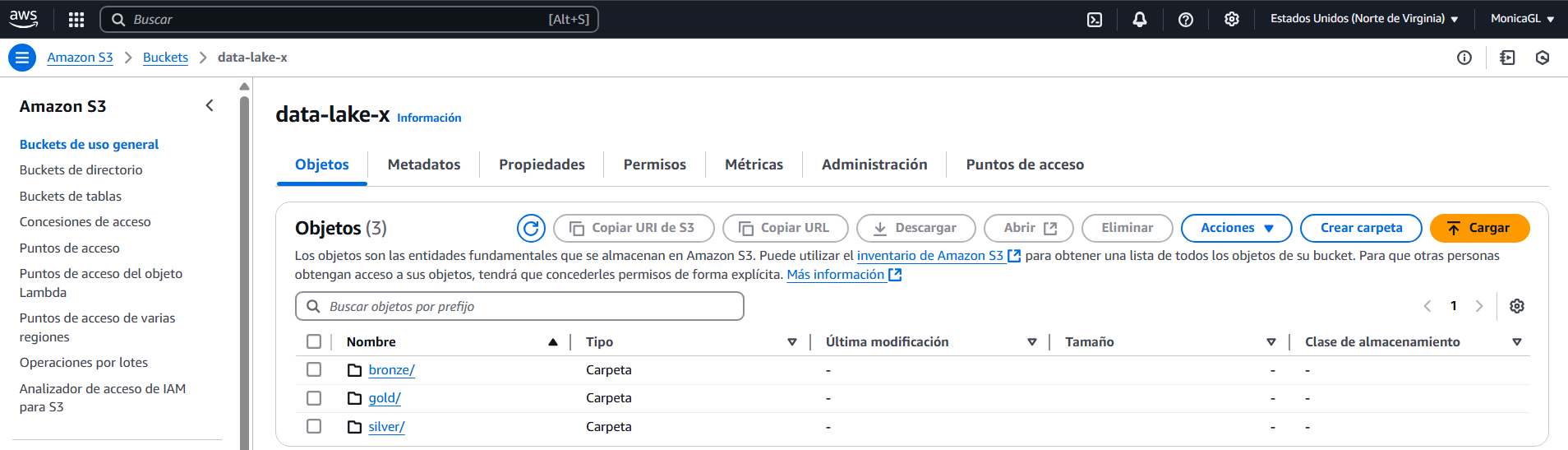
A continuación, se detallan los pasos para la implementación del Data Lake en Amazon S3:

* **Creación del Bucket en Amazon S3.** Para la creación del *bucket* en Amazon S3, que funcionará como el **repositorio principal del Data Lake**, se debe iniciar sesión en AWS Console y acceder al servicio S3. Una vez dentro, se selecciona la opción "**Crear bucket**" e ingresa un nombre único, como data-lake-x. Posteriormente, se elige la región más cercana a la ubicación del usuario y se configuran las opciones de acceso y permisos de acuerdo con los requisitos de seguridad. Para optimizar costos, es recomendable desactivar la opción de versionado. Finalmente, tras revisar la configuración, se procede a la creación del bucket haciendo clic en la opción correspondiente.



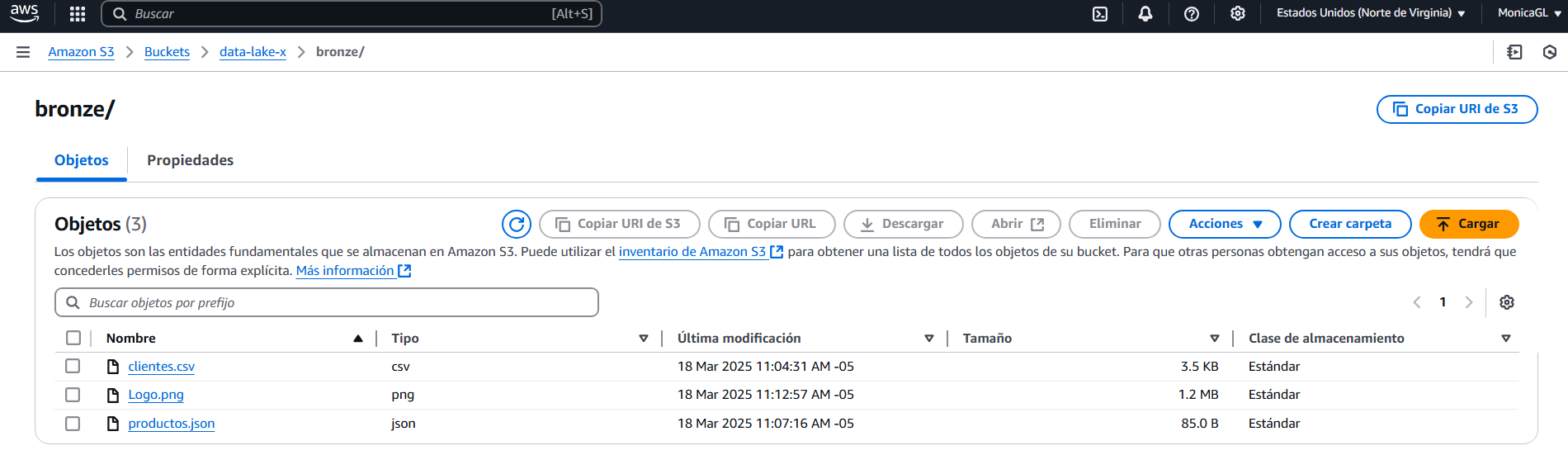
**Figura 20**. Creación de Bucket en S3. Elaboración propia en la plataforma AWS

* **Definición de la estructura del Data Lake.** Es tipo de almacenamiento se organiza en tres capas lógicas: Bronze (datos crudos), Silver (datos procesados) y Gold (datos agregados para analítica avanzada). Para su implementación en Amazon S3, se creó un bucket llamado "data-lake-x" con las carpetas correspondientes a cada capa.



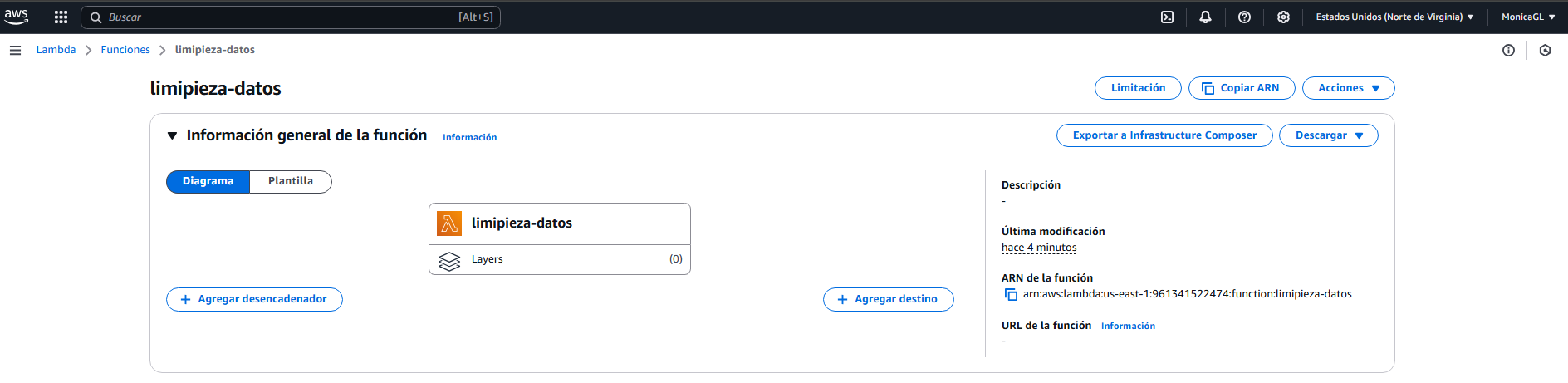
**Figura 21**. Creación de capas (carpetas) dentro del bucket. Elaboración propia en la plataforma AWS

* **Carga de datos en la capa Bronze. Para la carga de datos en la capa Bronze y la simulación de datos reales, se utilizará un archivo clientes.csv que contiene valores nulos y datos que requieren modificación. Para ello, se accede a AWS S3 y se navega hasta el *bucket data-lake-x*, donde se ingresa a la carpeta bronze/. Luego, se selecciona la opción "Cargar" y se suben los archivos clientes.csv, logo.png y *productos.json*, que representan datos estructurados, no estructurados y semiestructurados, respectivamente. Con este proceso, los datos quedan almacenados en la capa *Bronze* del *Data Lake* para su posterior procesamiento.**



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

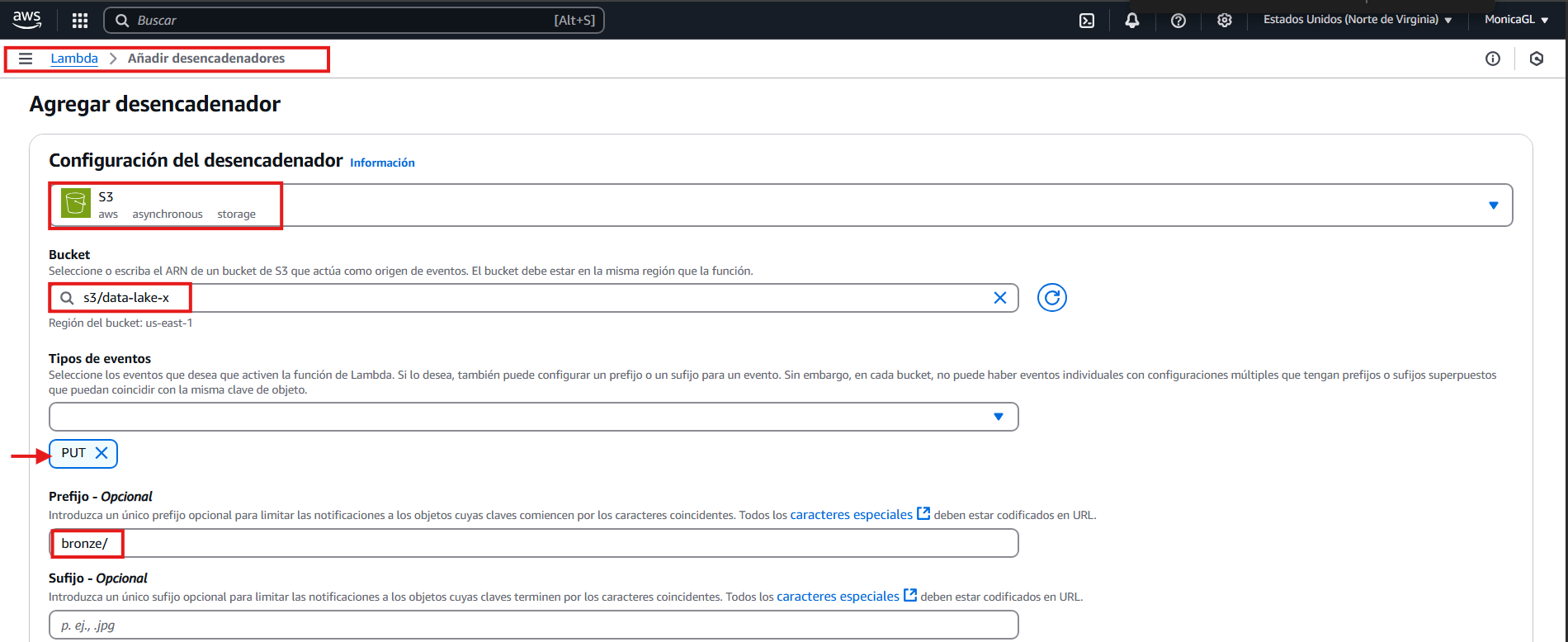
* **Limpieza de datos con AWS Lambda.** La función procesará automáticamente los archivos almacenados en la capa bronze para eliminar inconsistencias y mejorar la calidad de los datos. Para la creación de la función Lambda, primero se accede a AWS Lambda y se selecciona la opción "Crear función". A continuación, se elige la opción "Crear desde cero" y se asigna el nombre *limpiar\_clientes\_s3*. Como entorno de ejecución, se configura Python 3.13 y se ajustan los permisos IAM para permitir el acceso a S3, incluyendo la política *Amazon S3 Full Access* desde IAM. Finalmente, tras revisar la configuración, se completa el proceso haciendo clic en "Crear función".



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

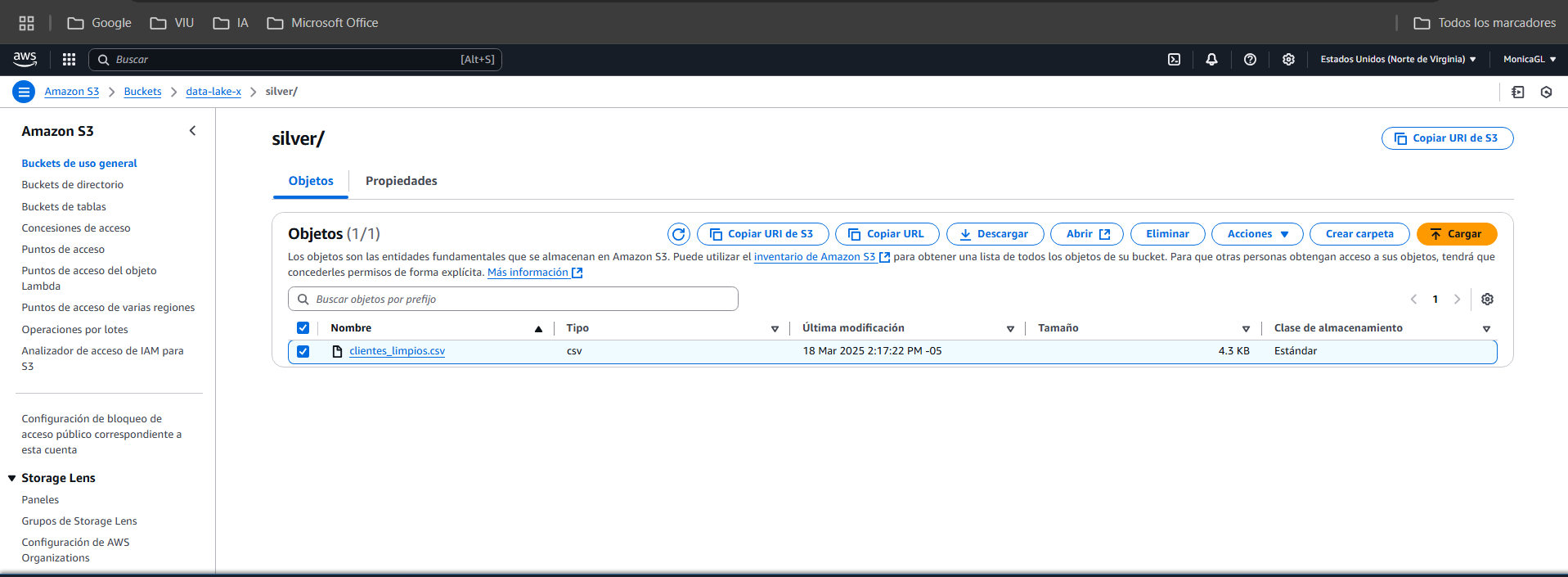
* **Desarrollo del código en Lambda. El código de Lambda eliminará filas nulas y cambiará género F por Femenino y M por Masculino (ver anexo X).**
* **Automatización del proceso con notificaciones S3.** Para que la limpieza de datos sea automática, se configura un evento en S3 que dispare la función Lambda cuando se suba un archivo a bronze/.

Para configurar la notificación de eventos en **AWS S3**, se accede al bucket data-lake-x y se navega a la sección de Propiedades. Dentro del apartado Eventos, se selecciona la opción "Crear notificación" y se asigna un nombre. Luego, se configura el evento **PUT** para que la notificación se active cuando se suban archivos, definiendo el prefijo de monitoreo como bronze/. A continuación, se elige **AWS Lambda** como destino y se asigna la función limpieza-datos. Finalmente, se guardan los cambios para completar la configuración.



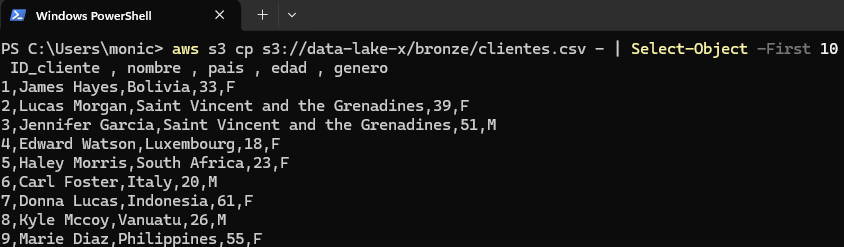
**Figura 24.** Configuración de eventos. Elaboración propia en la plataforma AWS.

* **Verificación del proceso de limpieza.** Para verificar que el proceso de limpieza funciona correctamente, se sube un archivo con datos en bruto a la carpeta bronze/ y, tras unos segundos, se comprueba que el archivo procesado ha sido trasladado automáticamente a la carpeta silver/. Adicionalmente, se utiliza la opción de prueba de la función Lambda para validar su ejecución. Finalmente, se revisa el contenido del archivo procesado para confirmar que no contiene filas con valores vacíos y que la columna de género ha sido normalizada, transformando "F" en Femenino y "M" en Masculino.

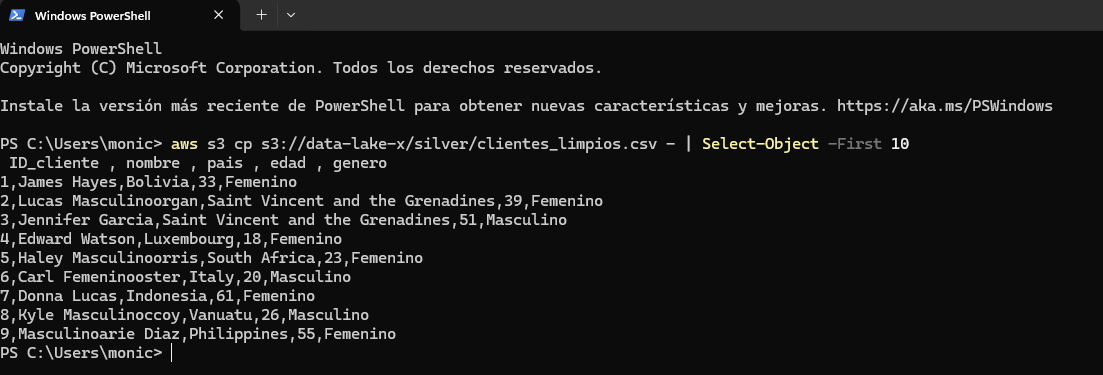


**Figura 25**. Resultado de la ejecución de la función lambda. Elaboración propia en la plataforma AWS

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



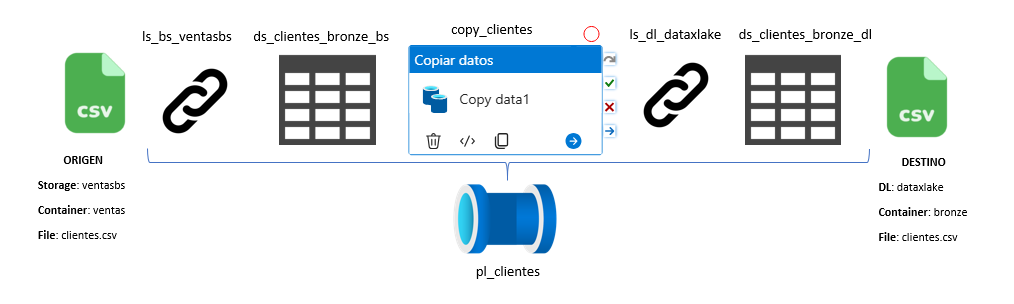
**Figura 26**. Archivo con datos en crudo. Elaboración propia en la plataforma AWS



**Figura 27**. Archivo con datos procesados. Elaboración propia en la plataforma AWS

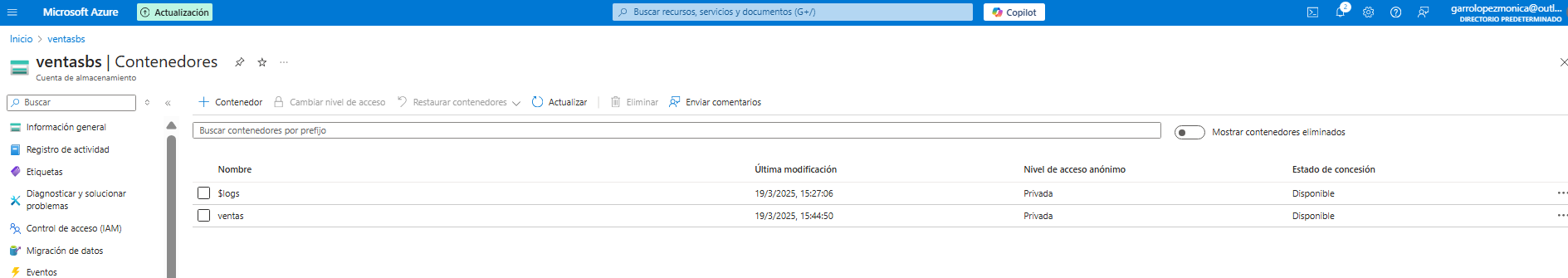
### *Data Lake en Azure*

En esta práctica se implementa un proceso de **ingesta de datos** en un **Data Lake en Azure**, utilizando diversos servicios de **Microsoft Azure** para gestionar y transformar la información de manera eficiente. **Azure Blob Storage** actúa como el repositorio inicial donde se almacena el archivo clientes.csv en el contenedor ventas, mientras que **Azure Data Lake Storage Gen2** se configura como destino final en la capa Bronze. Para automatizar la transferencia de datos entre estos servicios, se emplea **Azure Data Factory**, donde se crea un **Pipeline (pl\_clientes)** que orquesta la ejecución del proceso de copiado de datos (copy\_clientes). En este flujo, se establecen **Linked Services *(ls\_bs\_ventasbs y ls\_dl\_datalake)*** para conectar los repositorios de origen y destino, junto con **Datasets *(ds\_clientes\_bronze\_bs y ds\_clientes\_bronze\_dl)*** que definen la estructura del archivo en ambas ubicaciones. La siguiente imagen ilustra este proceso de ingesta, destacando cada componente clave involucrado en la transferencia de datos y su integración en el **Data Lake**.



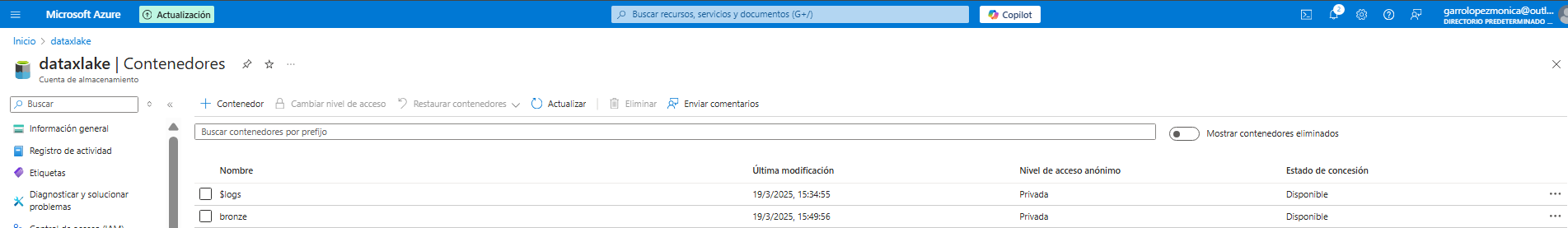
**Figura 33**. Arquitectura del Data Lake en Azure. Elaboración propia

* **Configuración del Entorno.** Se crearon los recursos necesarios en **Azure** para el almacenamiento de datos en bruto. En **Azure Blob Storage**, se estableció una cuenta de almacenamiento denominada ventasbs, con un contenedor llamado ventas, en el que se almacenó el archivo clientes.csv. De manera similar, en **Azure Data Lake Storage Gen2**, se configuró una cuenta de almacenamiento llamada dataxlake, donde se creó el contenedor bronze para recibir el archivo después de su procesamiento
* **Preparación de los Contenedores de Almacenamiento.** Se crearon dos espacios para organizar los datos de manera estructurada. En primer lugar, el contenedor ventas dentro de ventasbs (ver figura 34), que sirve como punto de ingesta inicial del archivo.



**Figura 34**. Creación del Contenedor ventas en Blob Storage. Elaboración propia en la plataforma Azure

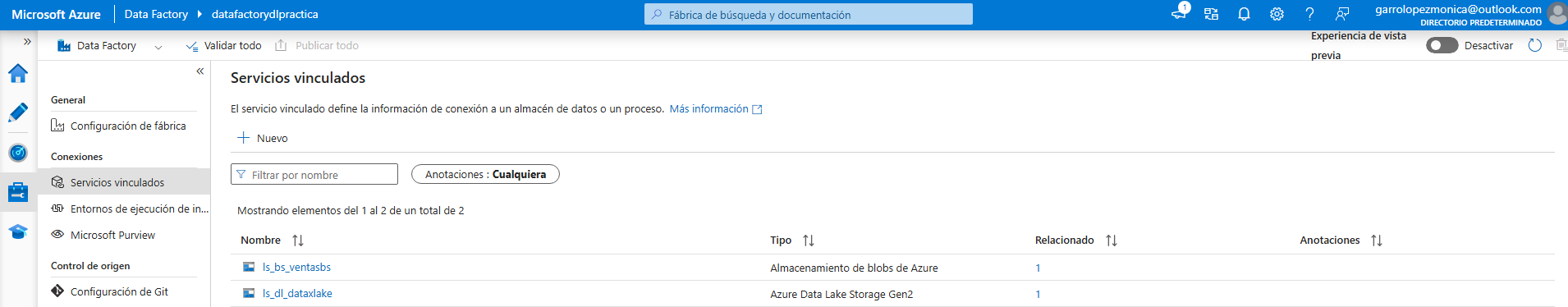
Posteriormente, en **Data Lake Storage**, se creó el contenedor bronze, donde se almacenarán los datos una vez transferidos desde Blob Storage (ver figura 35).



**Figura 35**. Creación del Contenedor bronze en Data Lake. Elaboración propia en la plataforma Azure

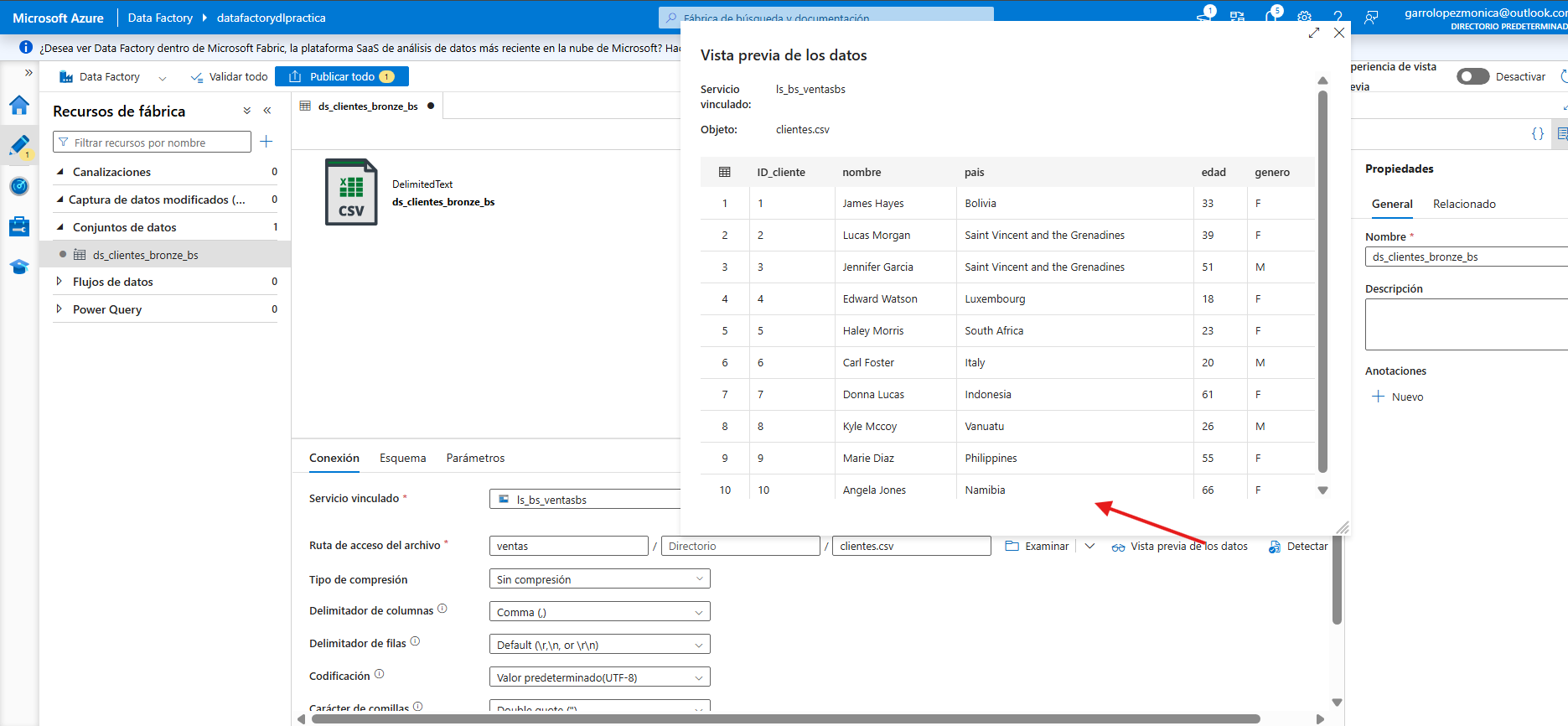
Para la carga de datos, el archivo clientes.csv fue cargado en el contenedor ventas dentro de **Blob Storage**, asegurando su disponibilidad para el proceso de ingesta hacia **Data Lake Storage**.

* **Transferencia de datos.** Se realizó mediante **Azure Data Factory**, un servicio que permite la integración y el movimiento eficiente de datos entre distintos entornos. Para ello, se configuraron **Linked Services**, que establecen conexiones entre los servicios de almacenamiento de **Azure**, asegurando una comunicación fluida entre **Blob Storage** y **Data Lake Storage Gen2**

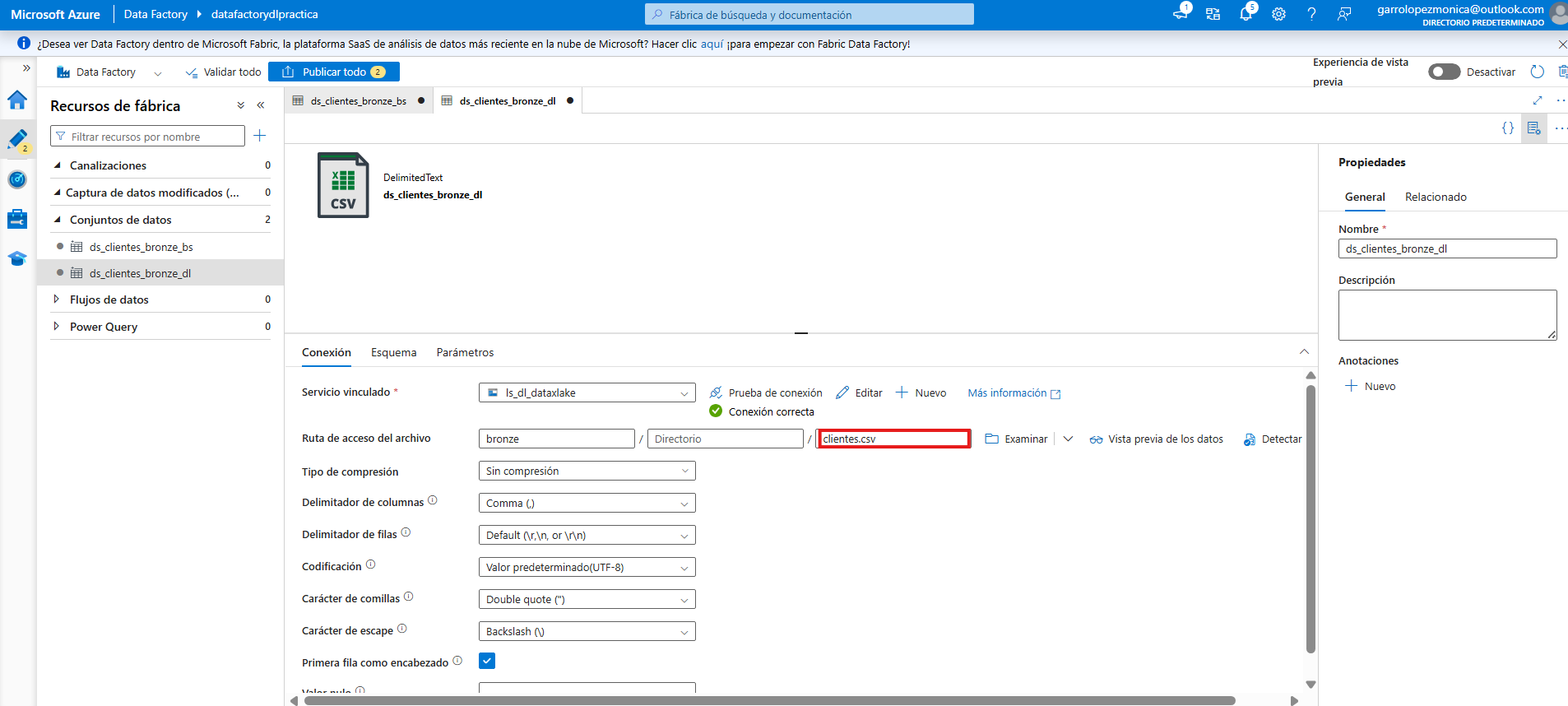


**Figura 36**. Configuración de Linked Services en Azure Data Factory. Elaboración propia en la plataforma Azure

En esta fase también se definieron los **Datasets** que especifican los formatos de datos de entrada y salida. A continuación, se diseñó un **Pipeline** que gestiona el flujo de trabajo y permite mover los datos desde **Blob Storage** hasta **Data Lake**, asegurando una ingesta estructurada y eficiente.

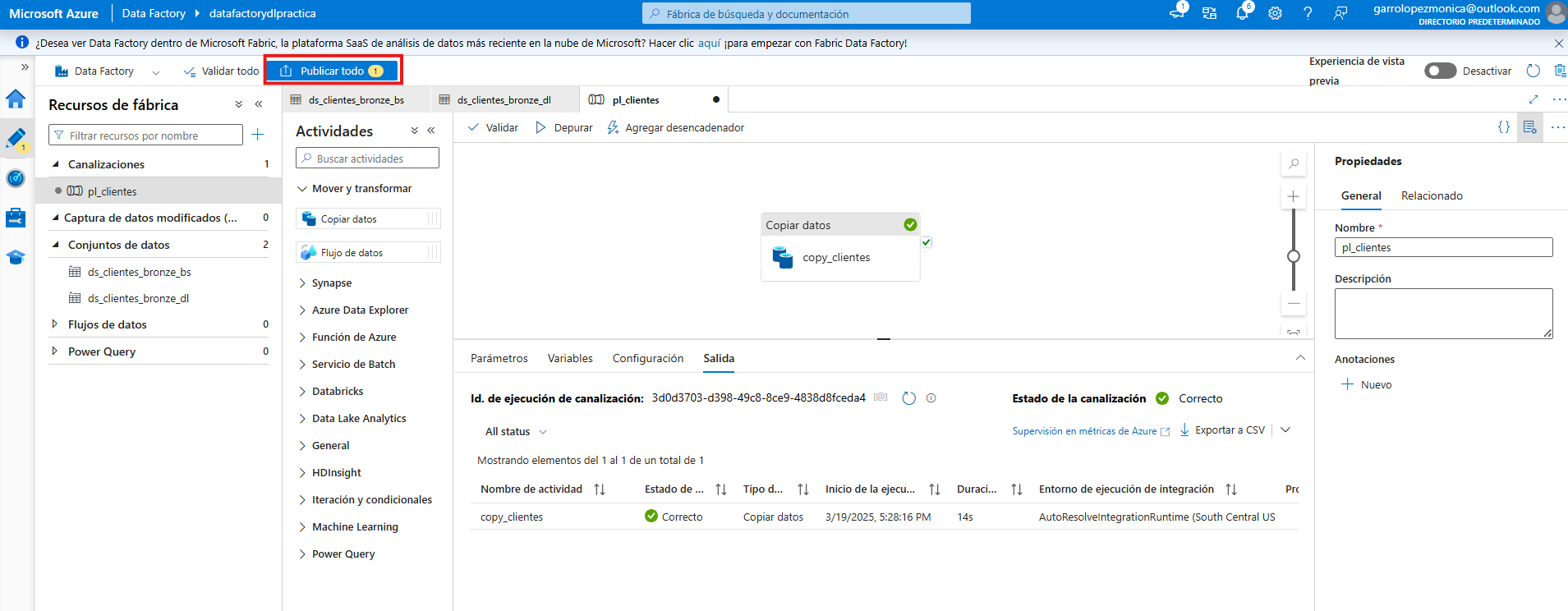


**Figura 37**. Creación del Dataset ds\_clientes\_bronze\_bs en Azure Data Factory. Elaboración propia en la plataforma Azure



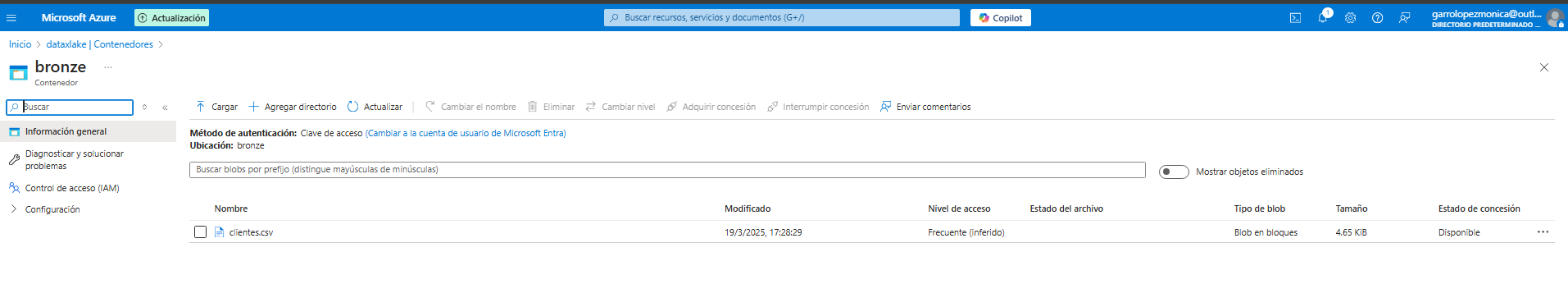
**Figura 38**. Creación del Dataset ds\_clientes\_bronze\_dl en Azure Data Factory. Elaboración propia en la plataforma Azure

* **Ejecución del Pipeline.** Una vez configurado el **Pipeline**, se ejecutó el proceso de ingesta, trasladando el archivo clientes.csv desde **Blob Storage** hasta la capa Bronze del **Data Lake**. La ejecución se verificó a través del monitor de **Azure Data Factory**, asegurando que el proceso se completara sin errores.



**Figura 39**. Desarrollo del Pipeline de Ingesta de Datos (pl\_clientes). Elaboración propia en la plataforma Azure

Finalmente, se validó la presencia del archivo en el contenedor bronze dentro de dataxlake, confirmando que la transferencia se llevó a cabo exitosamente.



**Figura 40**. Verificación del traslado exitoso del archivo a la capa Bronze

### *Data Warehouse* en Amazon Redshift

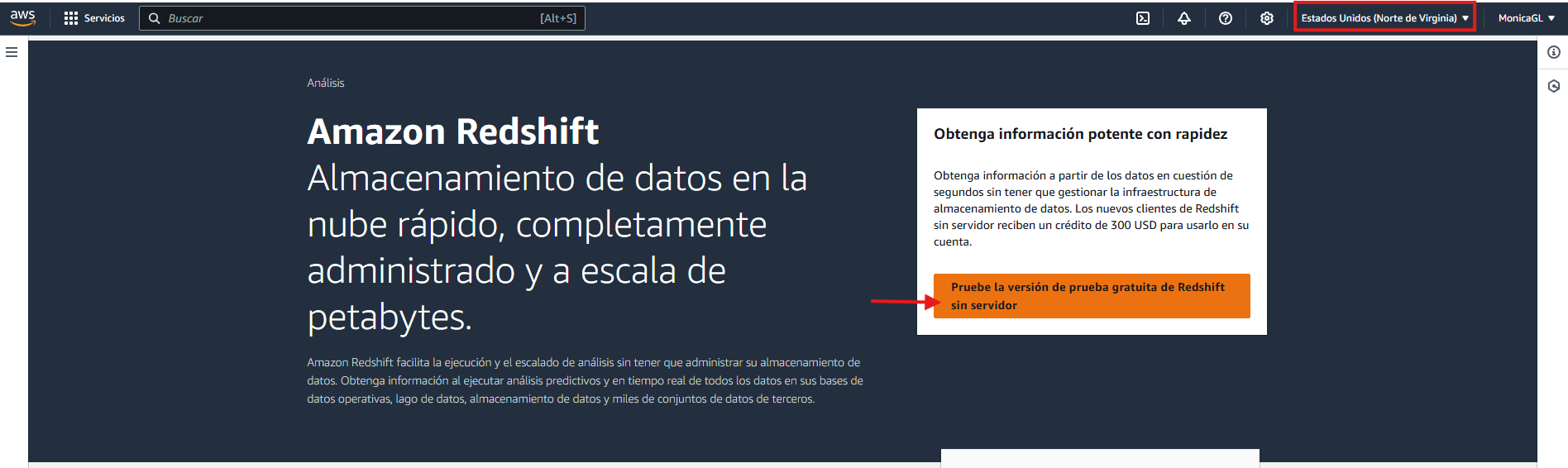
Amazon Redshift es un servicio de *Data Warehouse* en la nube de AWS **que permite gestionar grandes volúmenes de datos de manera eficiente mediante SQL**. Su capacidad de escalabilidad y procesamiento masivo lo hace ideal para análisis de datos a gran escala.

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.

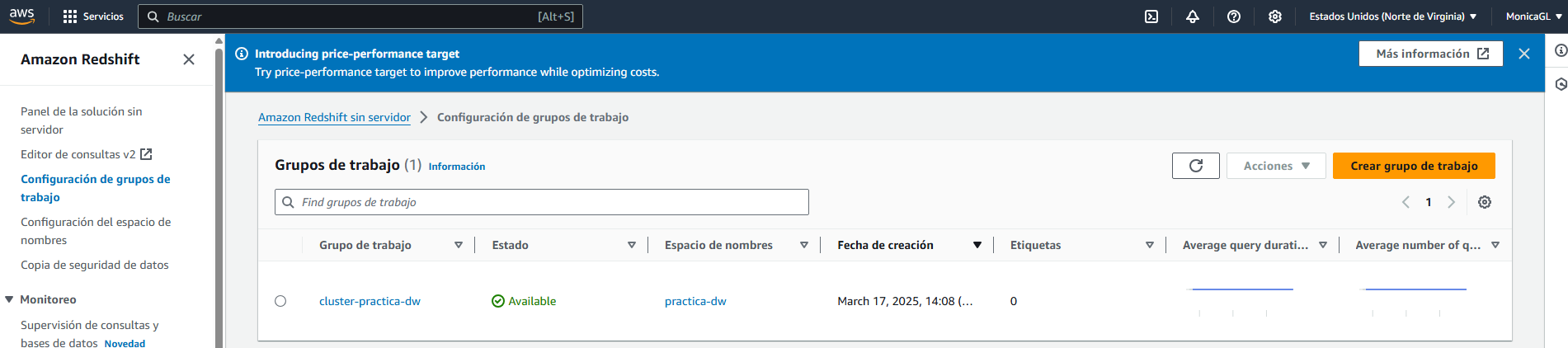
A continuación, se detallan los pasos para la realización de la práctica:

* + **Configuración del clúster en Amazon Redshift.** Acceder a la consola de **AWS** y buscar **Amazon Redshift**.



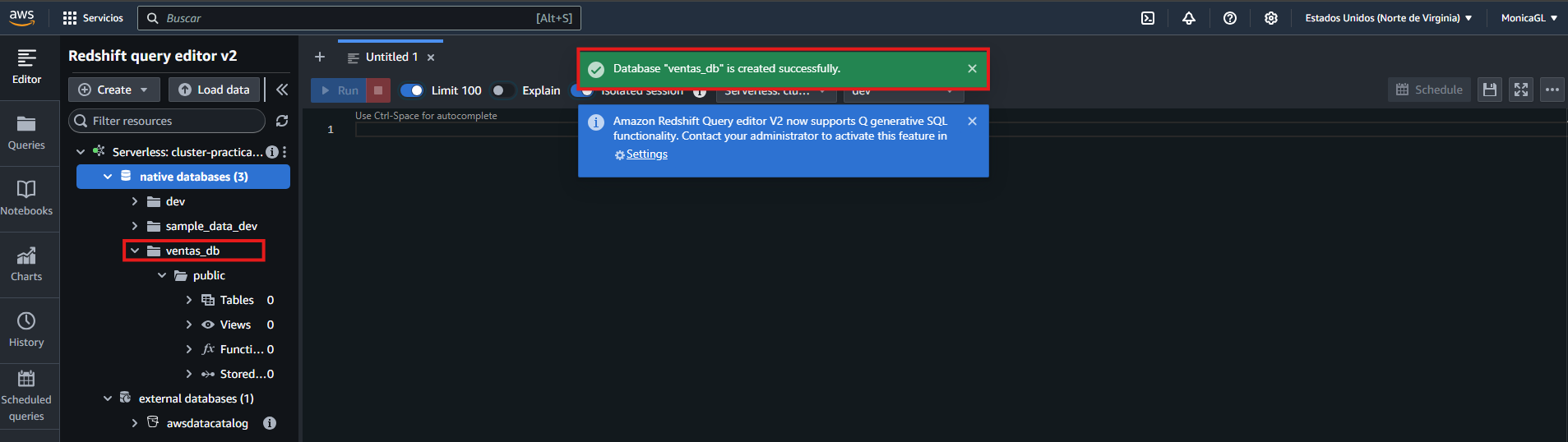
**Figura 10**. Página de inicio de Redshift con opción de prueba gratuita sin servidor. Elaboración propia en la plataforma AWS.

* + Posteriormente, se busca **Amazon Redshift** y se selecciona la opción "**Crear Clúster**". Posteriormente, se define el nombre del clúster (practica-dw), los parámetros de configuración y la base de datos predeterminada dev. Una vez configuradas las credenciales IAM para el acceso seguro, se guarda la configuración y se completa la creación del clúster.



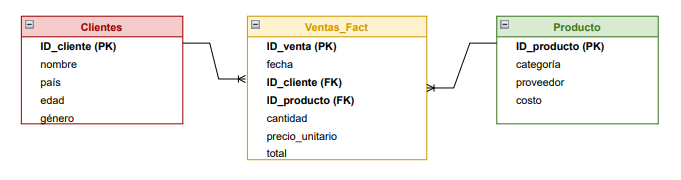
**Figura 11**. Grupo de trabajo creado con los pasos antes indicados. Elaboración propia en la plataforma AWS.

* **Creación de la base de datos.** En el **Query Editor** de **Redshift**, se selecciona el clúster creado, se elige la opción "**Create Database**", se especifica el nombre de la base de datos (ventas\_db) y se verifica su correcta configuración.



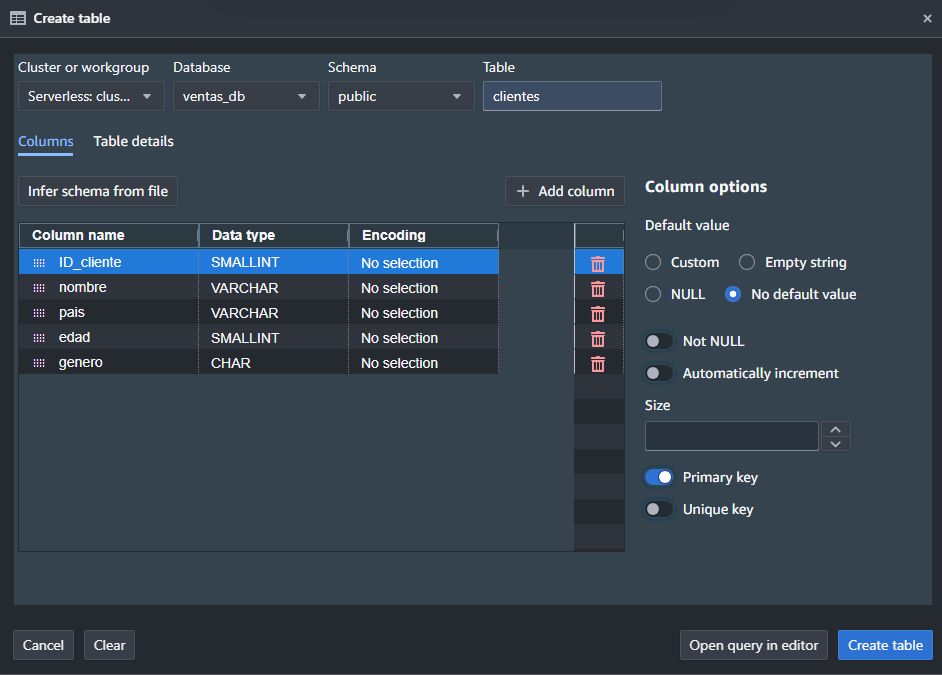
**Figura 12**. Base de datos creada con pasos anteriores. Elaboración propia en la plataforma AWS.

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



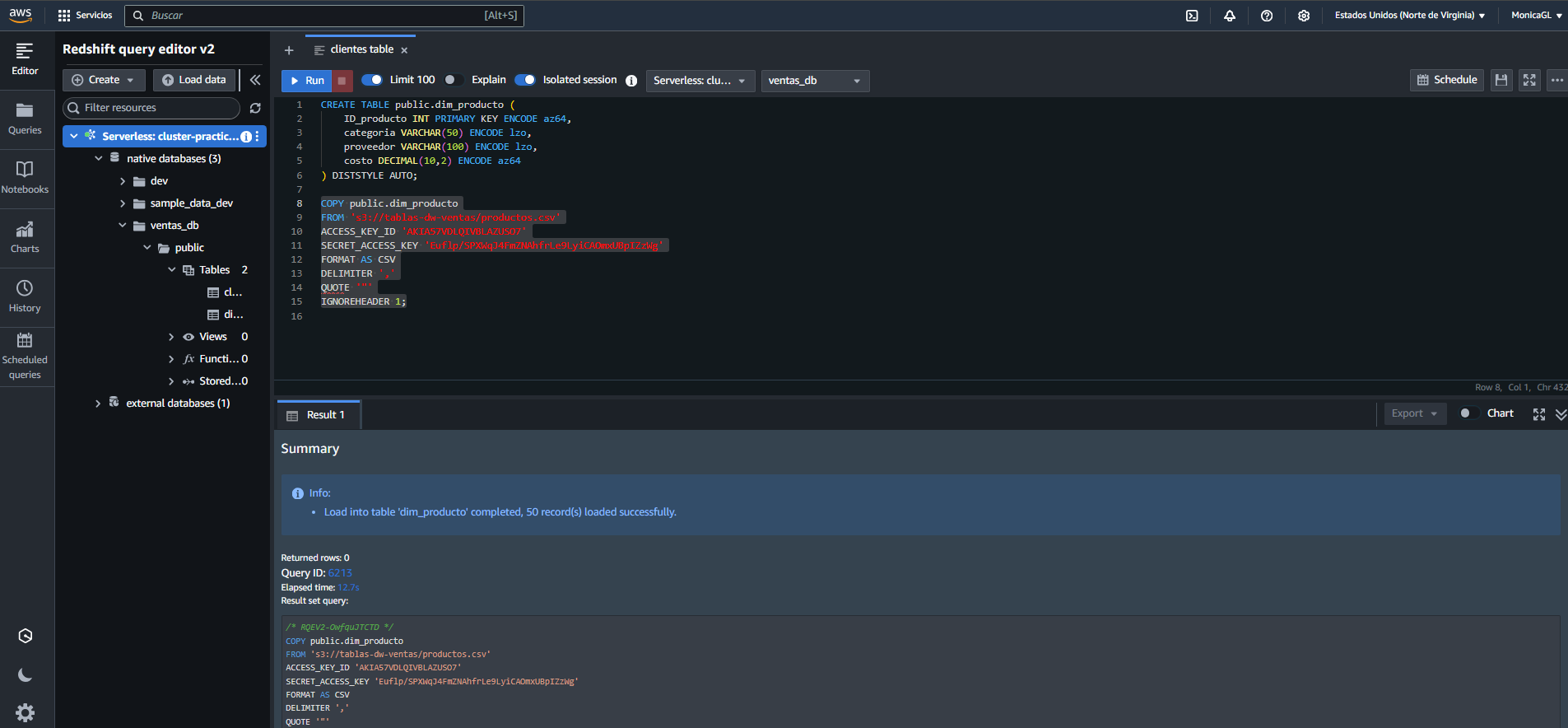
**Figura 13**. Modelo Entidad-Relación para el Data Warehouse que se creó como práctica. Elaboración propia

* **Creación de tablas en Redshift**. Las tablas se crean en *Redshift* mediante el *Query Editor v2*. Para las tablas de dimensiones (clientes, dim\_producto), se utilizó la opción gráfica "*Create Table*", especificando manualmente los nombres de las columnas y sus tipos de datos (ver figura 14).

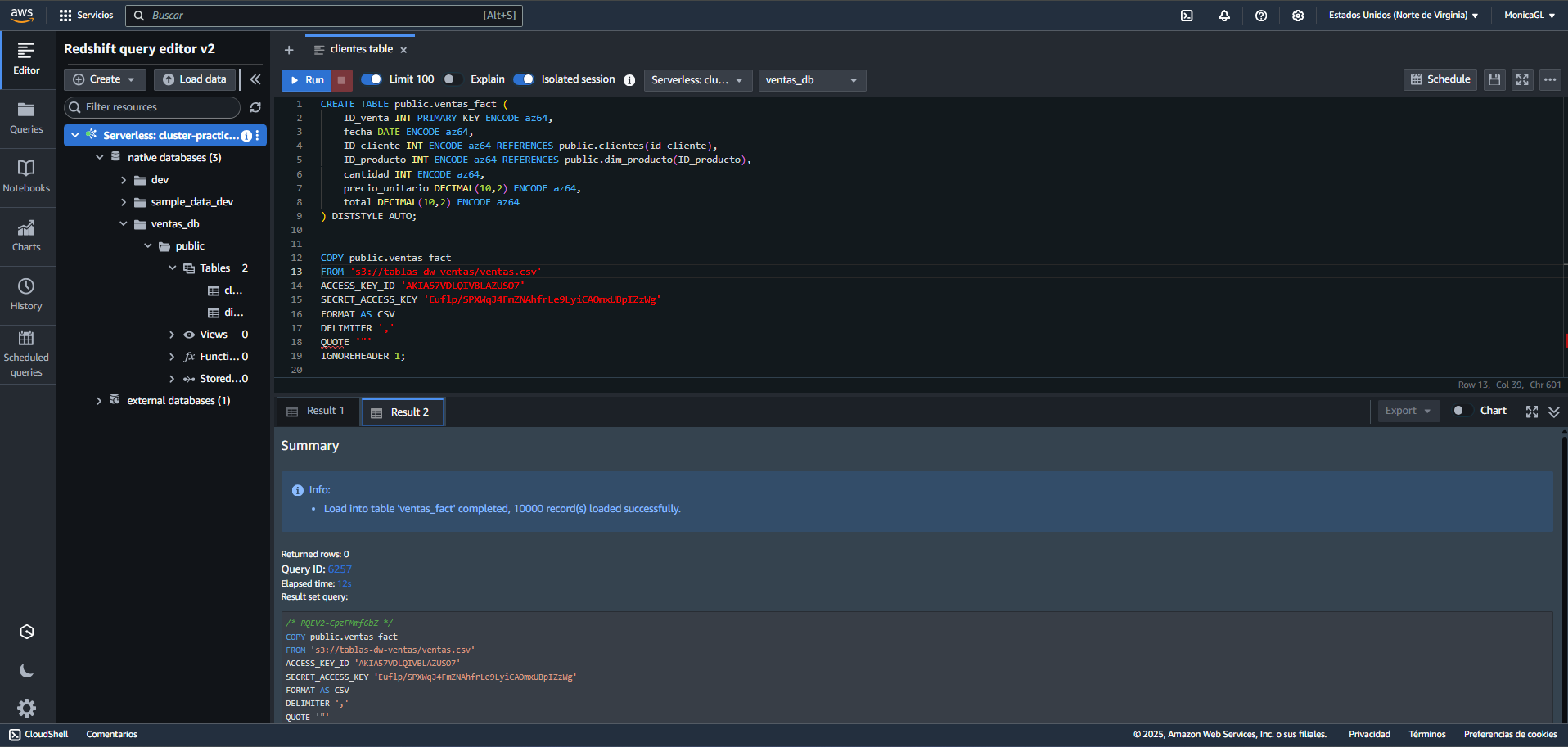


**Figura 14**. Creación de la tabla clientes desde la opción “Crear tabla”. Elaboración propia en la plataforma AWS.

En cambio, las tablas relacionadas con transacciones (ventas\_fact) fueron creadas mediante consultas SQL y cargadas desde S3 con la instrucción *COPY* (ver figura 15)

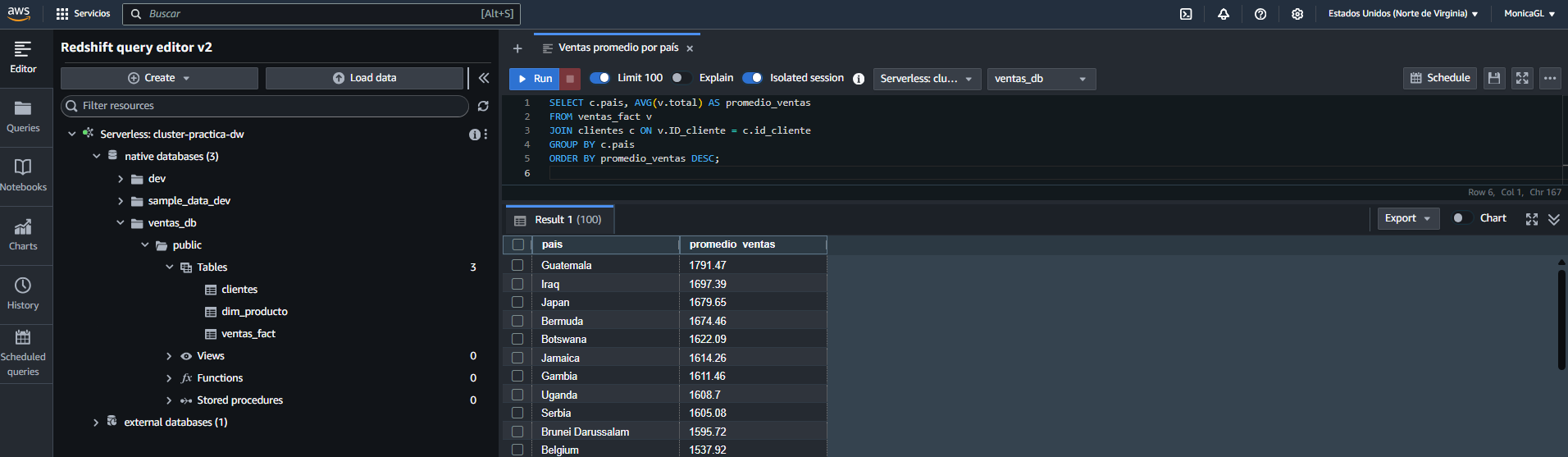


**Figura 15**. Creación de tabla producto y carga de datos desde S3. Elaboración propia en la plataforma AWS.



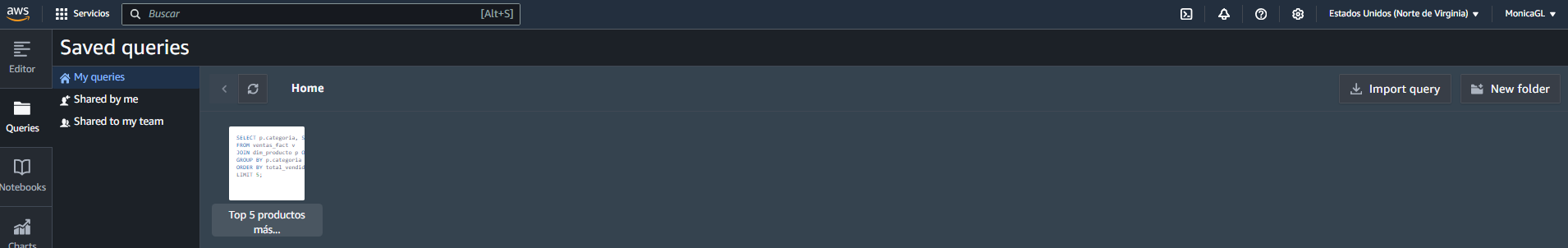
**Figura 16**. Creación de tabla ventas\_fact y carga de datos desde S3. Elaboración propia en la plataforma AWS.

* **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:**



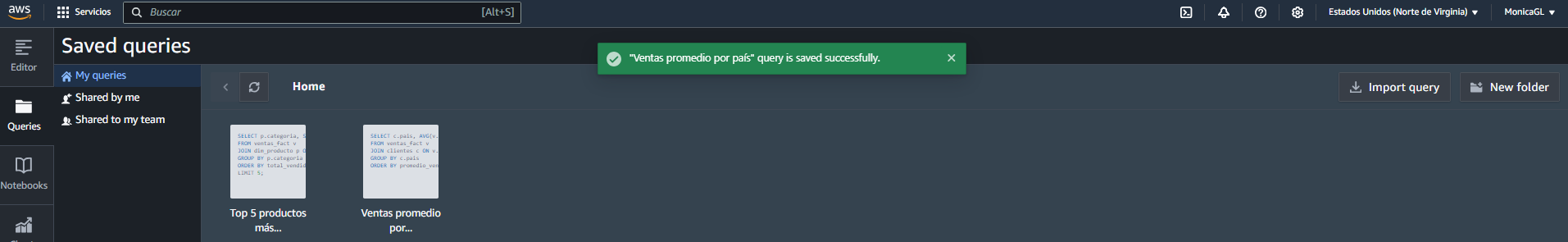
**Figura 17**. Consulta para calcular los productos más vendidos. Elaboración propia en la plataforma AWS.

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



**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.

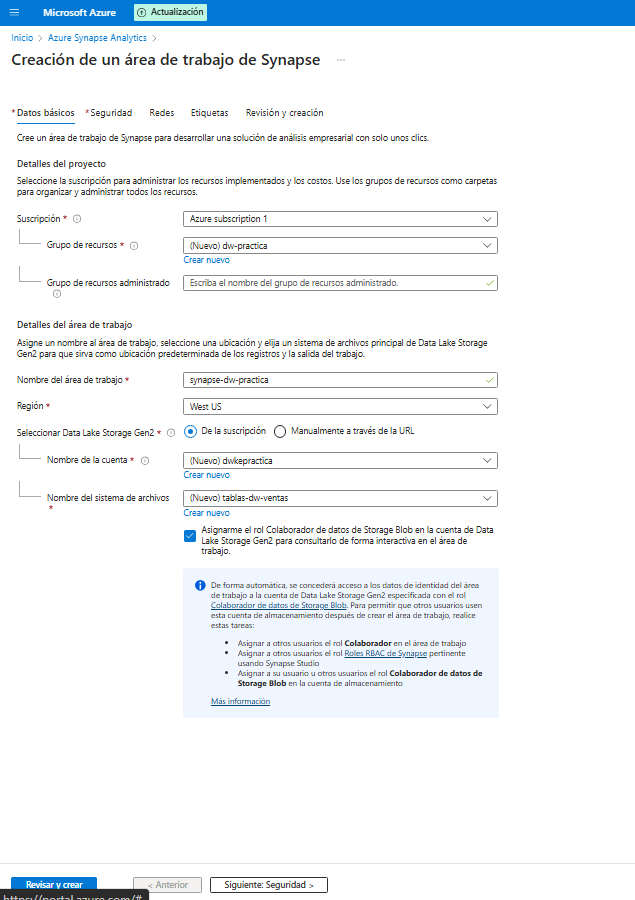
### *Data* Warehouseen *Azure Synapse Analytics*

**Azure Synapse Analytics** es un servicio de análisis en la nube de Microsoft que combina capacidades de almacenamiento y procesamiento de grandes volúmenes de datos con herramientas avanzadas de análisis.

En esta práctica, se implementó un **Data Warehouse** utilizando **Azure Synapse Analytics**, con **Azure Data Lake Storage Gen2** como repositorio para la ingesta de datos. Al igual que en la práctica realizada en Redshift, el objetivo principal es diseñar un modelo de datos basado en la arquitectura estrella, permitiendo almacenar información de ventas, clientes y productos para ejecutar consultas analíticas optimizadas.

A lo largo del proceso, se configuró un espacio de trabajo en Synapse, se estableció un **SQL Pool Serverless** para la ejecución de consultas, y se implementó el modelo de datos con carga de información desde archivos CSV almacenados en **Azure Data Lake Storage Gen2**. Finalmente, se realizaron consultas SQL para evaluar la efectividad del modelo y extraer información relevante para el análisis de datos. Las acciones realizadas para cumplir el objetivo son las siguientes:

* **Creación del Workspace en Azure Synapse Analytics**. Para iniciar la implementación, se creó un área de trabajo en **Azure Synapse Analytics**, permitiendo así la administración y análisis eficiente de los datos. Para ello, se accedió a la barra de búsqueda en el portal de **Azure**, se seleccionó la opción correspondiente a **Azure Synapse Analytics** y se procedió a la creación de un nuevo Workspace, configurándolo con los nombres y parámetros especificados en la Figura 28.



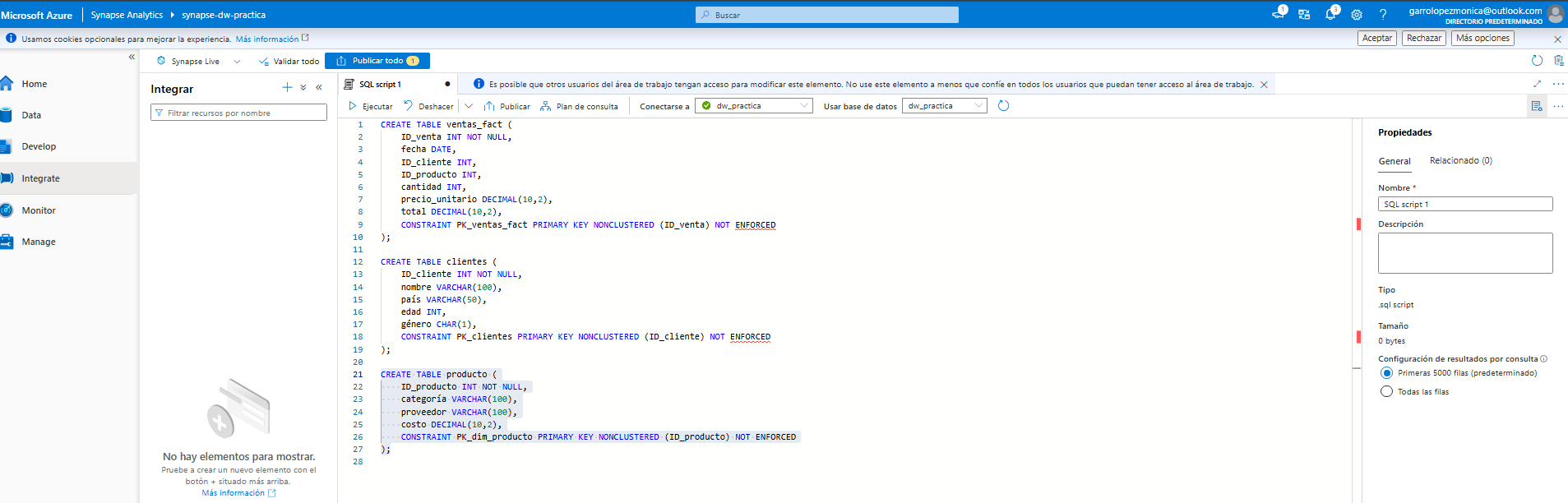
**Figura 28**. Creación del área de trabajo en Azure Synapse Analytics. Elaboración propia en la plataforma AWS

* **Configuración del SQL Pool en Synapse.** Se hizo uso de **Synapse Studio** con el propósito de ejecutar consultas y administrar bases de datos de manera eficiente. Para ello, se accedió a la pestaña Data, donde se creó un SQL Pool Serverless denominado dw\_practica. Finalmente, se estableció la conexión entre el SQL Pool y Azure Data Lake Storage Gen2, permitiendo la integración de los datos almacenados para su posterior procesamiento y análisis.



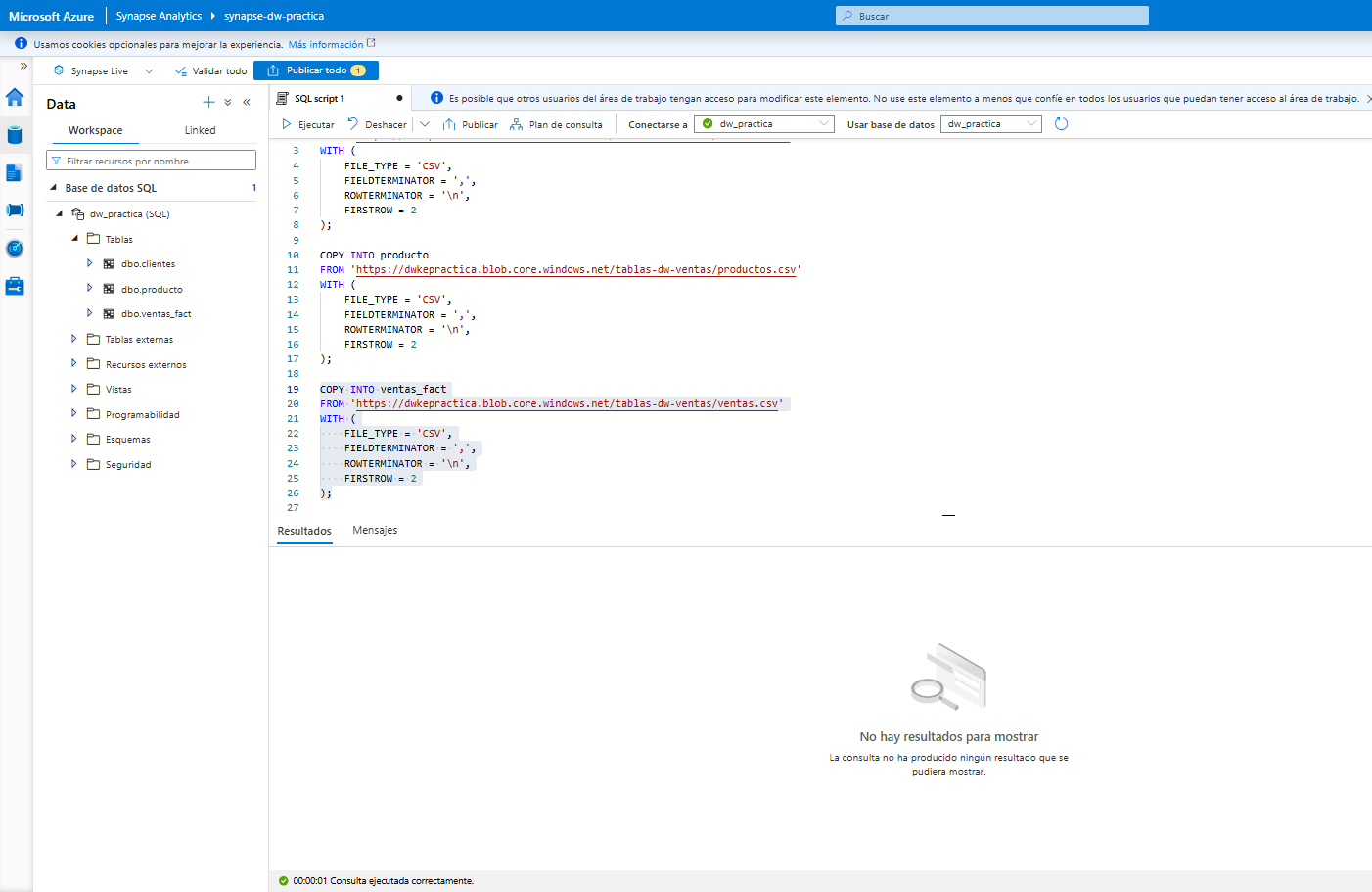
**Figura 29**. Configuración del SQL Pool en Azure Synapse. Elaboración propia en la plataforma AWS

* **Creación del modelo de datos (Modelo Estrella).** Se diseñó el **modelo estrella**, compuesto por una tabla de hechos y dos tablas de dimensiones.



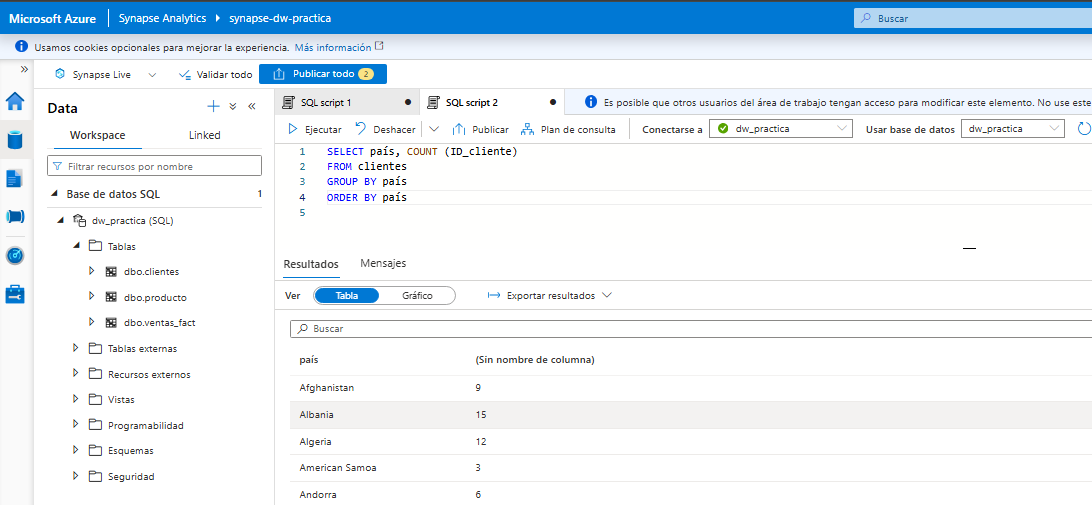
**Figura** 30. Creación de tabla de hechos y dimensiones. Elaboración propia en la plataforma AWS

* **Carga de datos en el Data Warehouse.** Para poblar las tablas, se utilizaron archivos *csv* almacenados en **Azure Data Lake Storage Gen2** y el comando *COPY INTO*.



**Figura 31**. Carga de datos desde csv. Elaboración propia en la plataforma AWS

* **Consultas SQL para análisis de datos.** Como prueba final se ejecuta consulta que permite identificar cantidad de clientes por país.

****

**Figura 32**. Consulta para análisis de datos

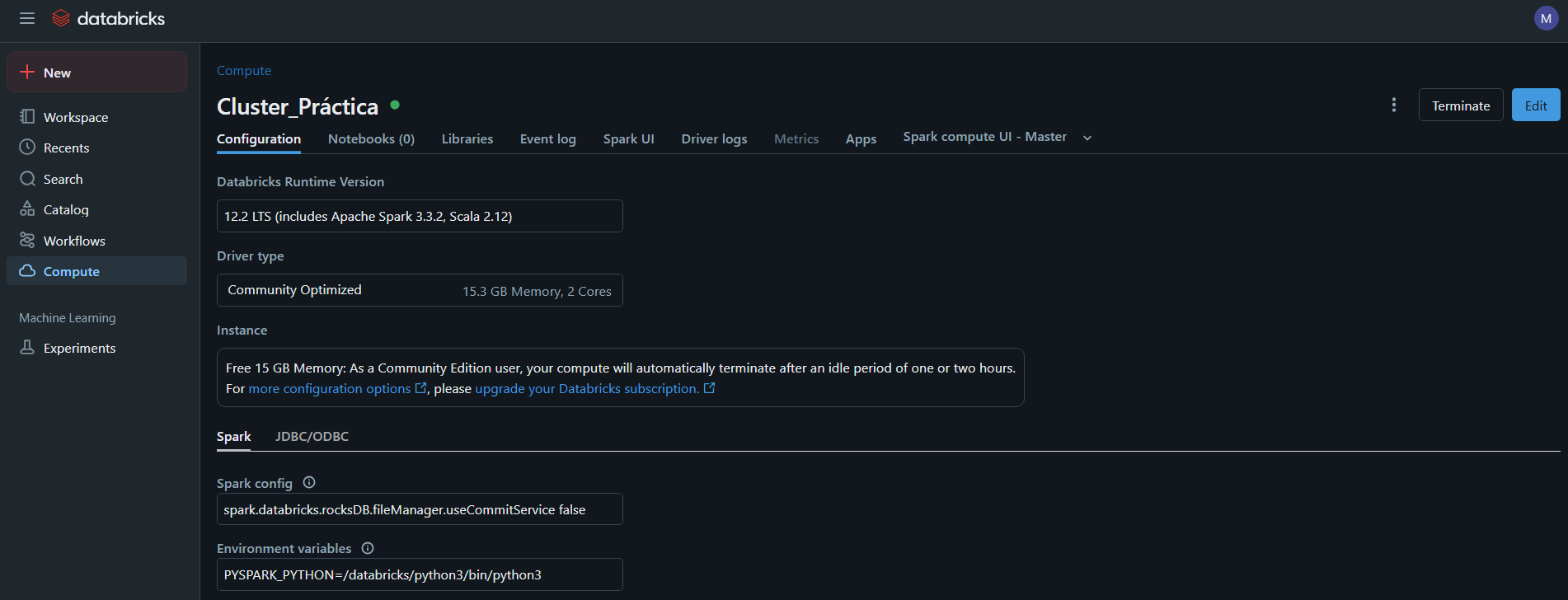
### *Delta Lake en Databricks*

En esta práctica se implementa una arquitectura de datos utilizando **Delta Lake sobre la plataforma Databricks**, con el objetivo de demostrar cómo se gestiona el ciclo de vida de los datos a través de tres capas estructuradas en una estructura *Medallion*: **Bronze** (procesamiento de datos en bruto), **Silver** (transformaciones y limpiezas) y **Gold** (información disponible para ser consumida por el negocio). Se utilizará un archivo de clientes (clientes.csv) como fuente inicial de datos.



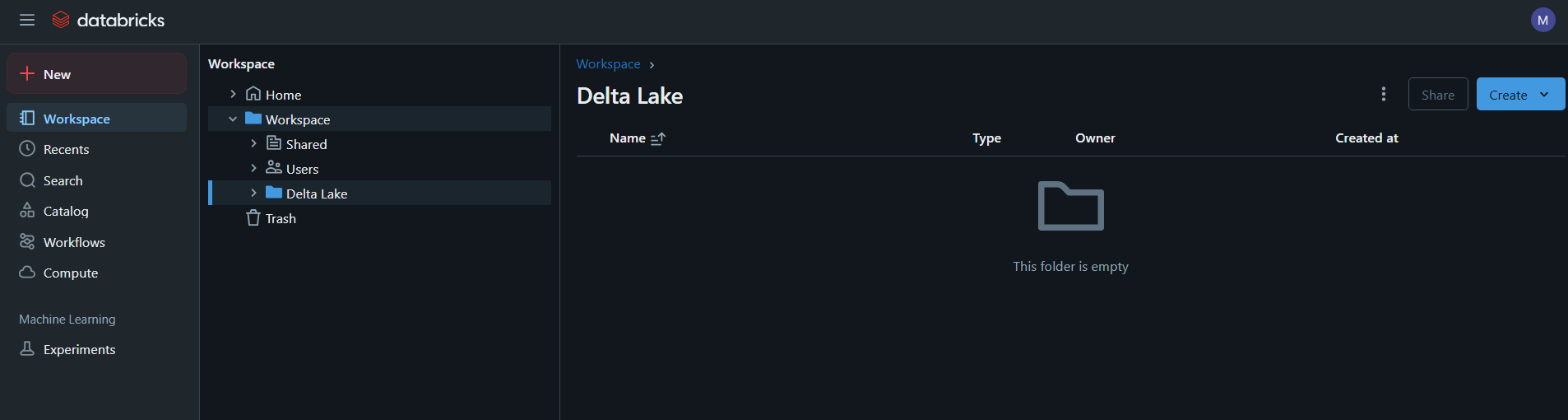
**Figura 41**. Arquitectura Medallion.(Núria, n.d.-b)

El primer paso del proceso consiste en crear un clúster en *Databricks*, configurado para ejecutar tareas con bajo consumo de recursos, utilizando un solo nodo con una versión estable de runtime como se evidencia en la figura 41.



**Figura 42**. Creación de Clúster. Elaboración propia en la plataforma Databricks

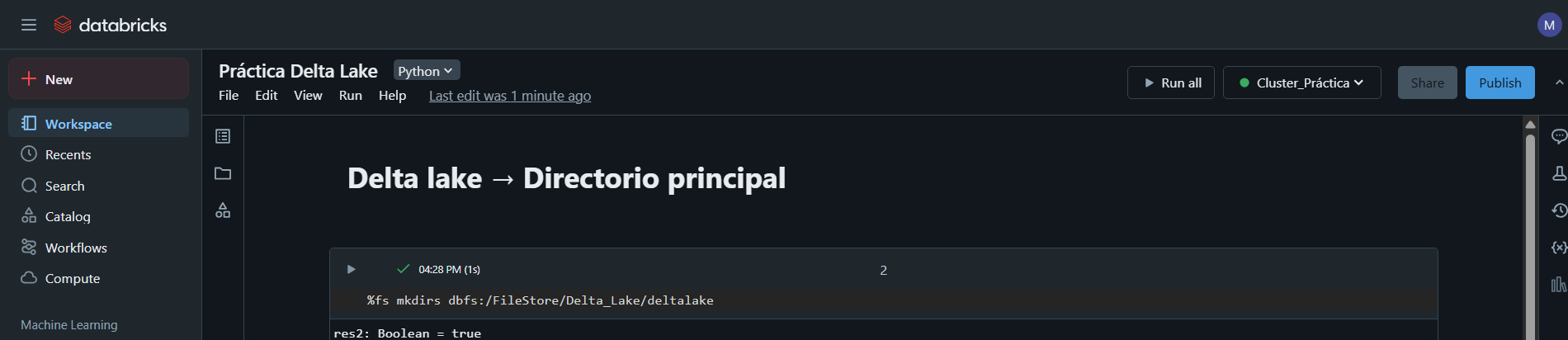
Además, se realiza la **creación del Workspace,** teniendo en cuenta que,*Databricks* organiza el almacenamiento mediante el sistema de archivos **DBFS (Databricks File System)**, que permite gestionar archivos y directorios dentro de la plataforma.



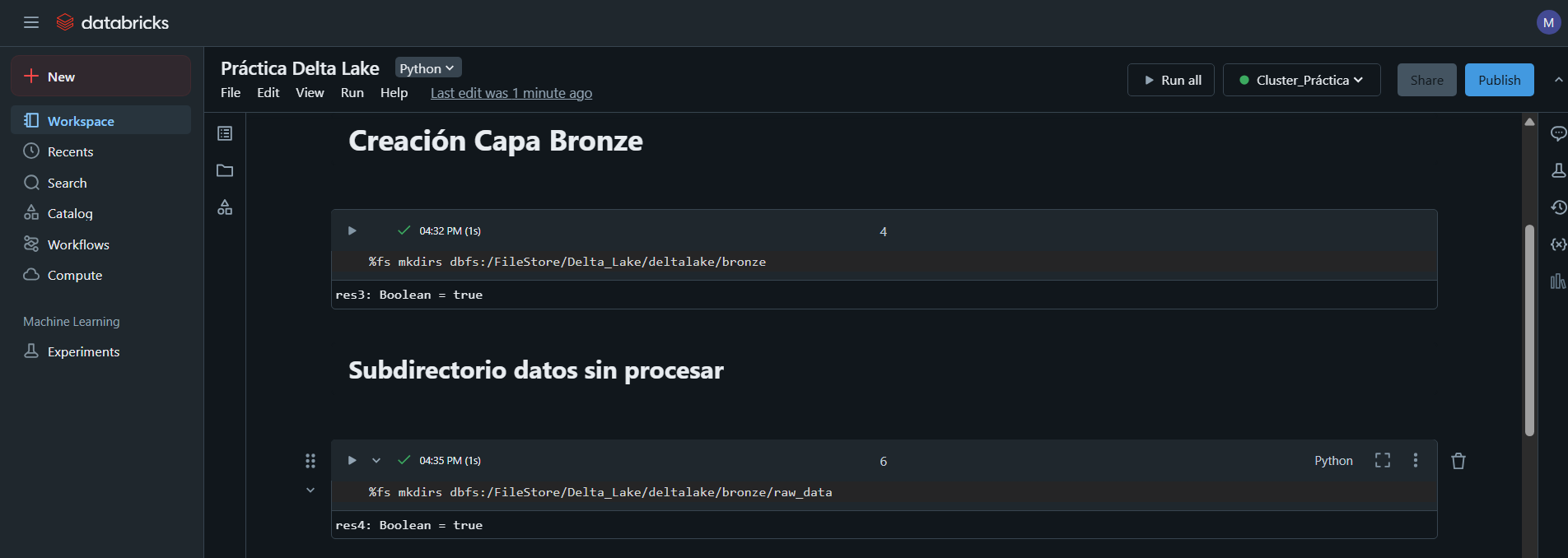
**Figura 43**. Creación del Workspace. Elaboración propia en la plataforma Databricks

Para la implementación de la arquitectura se realiza el siguiente procedimiento:

* **Creación del Notebook**. Se crea un notebook llamado **"Práctica Delta Lake"** para alojar el código de la práctica, se define la ruta principal del Delta Lake en DBFS y posteriormente se crean subdirectorios para cada capa como se muestra en la figura 45.



**Figura 44**. Directorio principal. Elaboración propia en la plataforma Databricks



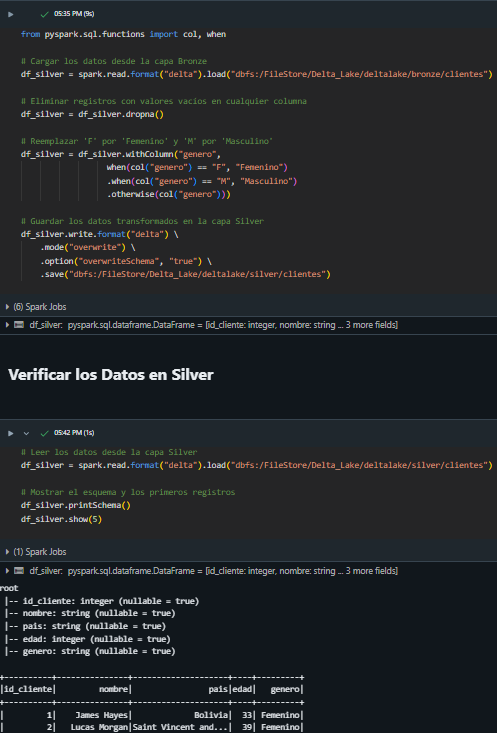
**Figura 45**. Subdirectorio de Capa Bronze. Elaboración propia en la plataforma Databricks

* **Capa bronce - Ingesta de datos crudos**. Se carga el archivo clientes.csv a la ruta *dbfs:/FileStore/Delta\_Lake/deltalake/bronze/raw\_data/.* Además se valida que los datos registren cargados correctamente



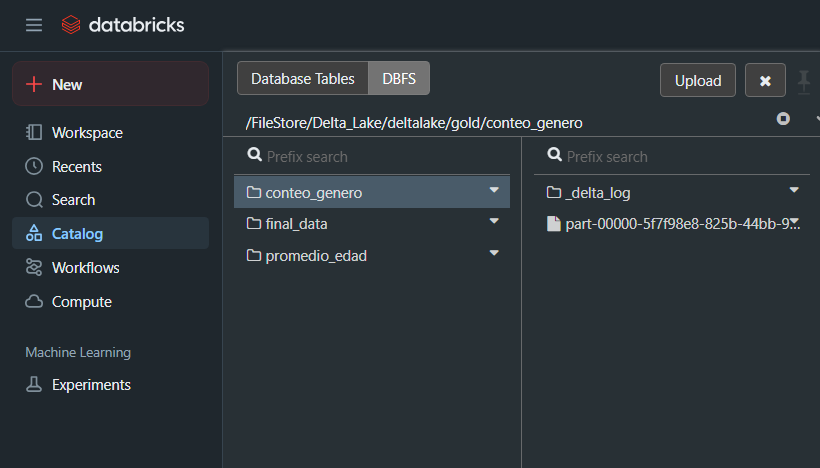
**Figura 46**. Ingesta de datos crudos. Elaboración propia en la plataforma Databricks

* **Capa Silver**: Transformación de datos que consiste en eliminar registros con datos incompletos y transformar el campo género: "F" por "Femenino", "M" por "Masculino".

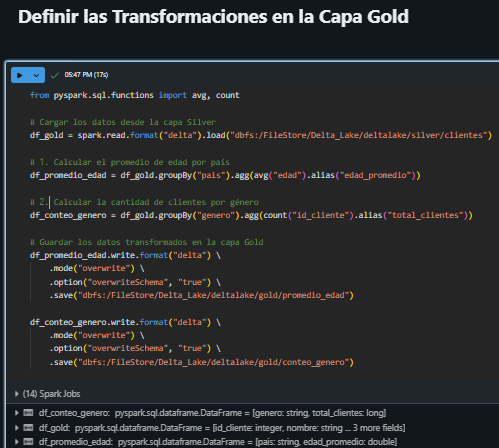


**Figura 47**. Transformación de datos y almacenamiento en capa Silver. Elaboración propia en la plataforma Databricks

* **Capa Gold**: En esta capa se generan datos transformados listos para análisis. Por ejemplo: Cálculo del promedio de edad por país y Conteo de clientes por género los cuales se almacenan en esta última capa.



**Figura 48**. Validación de resultados en Capa Gold. Elaboración propia en la plataforma Databricks



**Figura 49**. Transformaciones en la Capa Gold. Elaboración propia en la plataforma Databricks

## Resultados

Los resultados obtenidos a partir de las implementaciones permiten identificar diferencias significativas entre las arquitecturas evaluadas:

* ***Amazon Redshift (Data Warehouse*)** destacó por su alto rendimiento en consultas SQL estructuradas y su integración eficiente con *S3* mediante *COPY*.
* ***Amazon S3 (Data Lake)*** mostró gran flexibilidad para almacenar datos estructurados, no estructurados y semiestructurados. La implementación de capas permitió separar datos en bruto de datos limpiados, mejorando la calidad y gobernanza de la información.
* ***Azure Synapse Analytics (Data Warehouse*)** proporcionó una experiencia de desarrollo similar a *Redshift*, con ventajas en la integración nativa con *Azure Data Lake Storage* y capacidades analíticas para entornos *Microsoft*.
* ***Azure Data Lake + Data Factory*** destacó por la automatización del flujo de datos entre sistemas y su compatibilidad con arquitecturas escalables. La gestión del *pipeline* y los *linked services* facilitó la ingesta ordenada de datos.
* ***Databricks + Delta Lake*** sobresalió por su modelo transaccional y su estructura *medallion*, que aporta robustez en la trazabilidad de datos y facilidad para aplicar transformaciones avanzadas. Las consultas sobre la capa gold fueron eficientes y orientadas al consumo del negocio.

En conjunto, los resultados evidencian que **no existe una solución única ideal**, sino que la elección de arquitectura depende del **caso de uso específico**, los **requisitos de gobernanza**, la **necesidad de integración** y el **presupuesto de la organización**. Este trabajo permite establecer una base comparativa para futuras decisiones tecnológicas en proyectos de *Big Data* en la nube.

# Conclusión recomendaciones y trabajos futuros

## Conclusiones

La evolución del almacenamiento y gestión de datos masivos ha dado lugar a distintas arquitecturas, cada una con características particulares que responden a diferentes necesidades organizacionales. A través de este trabajo se ha realizado una revisión conceptual exhaustiva de los modelos más relevantes complementada con su implementación práctica en entornos *cloud* como *Amazon Web Services* (AWS), *Microsoft Azure* y *Databricks*.

* **Infraestructuras tradicionales y su evolución.** La implementación del modelo estrella en *Amazon Redshift* y *Azure Synapse* permitió constatar el alto rendimiento de los *Data Warehouses* en consultas analíticas estructuradas y su valor en entornos con requisitos de informes normativos.
* **El auge de los *Data Lakes* y el riesgo de los *Data Swamps***. La práctica en *Amazon S3* y *Azure Data Lake Storage* evidenció la flexibilidad de los *Data Lakes* para almacenar datos en bruto de distintos formatos. Sin embargo, se observó que sin una capa de transformación (como la plata y oro) y sin gobernanza adecuada, la calidad y trazabilidad de los datos se ve comprometida.
* **Delta Lake como solución para la gobernanza y calidad de datos**. La arquitectura *Medallion* implementada en *Databricks* con *Delta Lake* también demostró cómo la estructuración en capas (*bronze, silver, gold*) mejora la trazabilidad, calidad y el análisis de los datos, mitigando los riesgos asociados a los Data Lakes tradicionales.
* **Casos de uso y aplicabilidad en distintas industrias.** A través del análisis comparativo, se identificaron fortalezas específicas: *Data Warehouses* para reportes y cumplimiento normativo; *Data Lakes* para analítica exploratoria y almacenamiento escalable; y *Delta Lake* para entornos mixtos donde se requiere trazabilidad, eficiencia y análisis avanzado.
* **Data *Lakehouses* como 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.

## Recomendaciones

A partir de la experiencia práctica y los hallazgos obtenidos durante el desarrollo de este proyecto, se presentan las siguientes recomendaciones para una implementación efectiva de arquitecturas de almacenamiento de datos en entornos empresariales:

* **Realizar implementaciones piloto**: Antes de adoptar una arquitectura de forma definitiva, se recomienda ejecutar pruebas en entornos controlados en la nube para validar aspectos como rendimiento, escalabilidad y facilidad de uso con datos propios.
* **Aplicar una estructura en capas en los *Data Lakes (bronze, silver, gold*)**: Esta organización, evidenciada en las prácticas, permite mejorar la calidad, trazabilidad y gobernanza de los datos a lo largo de su ciclo de vida.
* **Utilizar plataformas integradas como Databricks**: Herramientas que combinan almacenamiento, procesamiento y analítica en un solo entorno facilitan el trabajo colaborativo, reducen la complejidad técnica y mejoran la trazabilidad.
* **Evaluar las necesidades específicas del negocio** antes de seleccionar una solución de almacenamiento, considerando factores como tipos de datos, volumen, necesidades analíticas y cumplimiento normativo.
* **Diseñar una estrategia de integración clara** con otras plataformas de la organización, especialmente en escenarios *multicloud* o híbridos.
* **Capacitar al equipo de datos** en las herramientas de gestión, transformación y análisis de datos adoptadas, para maximizar el valor de la infraestructura tecnológica implementada.

## Trabajos Futuros

A pesar del avance significativo en la adopción e implementación de arquitecturas de almacenamiento como ***Data Warehouses, Data Lakes, Delta Lakes* y *Data Lakehouses***, aún persisten diversos retos técnicos y conceptuales que abren nuevas líneas de investigación. A partir de los hallazgos obtenidos en este trabajo y su aplicación práctica, se proponen las siguientes áreas de desarrollo futuro:

* **Gobernanza y calidad de datos en entornos de *Data Lake***: Es necesario profundizar en estrategias automatizadas de gestión de metadatos, catalogación, linaje y control de acceso para evitar la degeneración de los Data Lakes en *Data Swamps* y maximizar su valor analítico.
* **Optimización del rendimiento en arquitecturas híbridas (*Data Lakehouse*)**: El creciente interés por combinar lo mejor de los *Data Lakes* y los *Data Warehouses* plantea desafíos en la integración, consistencia y rendimiento de sistemas híbridos. Se requieren investigaciones sobre estructuras eficientes y modelos de orquestación de datos en entornos *cloud* y *multicloud*.
* **Automatización inteligente de procesos ETL/ELT**: La incorporación de técnicas de inteligencia artificial y aprendizaje automático para automatizar tareas de ingestión, transformación, limpieza y validación puede mejorar notablemente la eficiencia operativa en proyectos de *Big Data*, especialmente en entornos como *Databricks*, *Azure Data Factory* o *AWS Glue*.
* **Seguridad y cumplimiento normativo en arquitecturas distribuidas**: A medida que las organizaciones migran hacia arquitecturas *multicloud* o híbridas, se vuelve esencial desarrollar nuevos esquemas de cifrado, autenticación federada y auditoría de datos sensibles, que permitan cumplir normativas como GDPR, HIPAA o ISO 27001 sin comprometer el rendimiento del sistema.
* **Exploración del impacto de la computación cuántica en la gestión de datos masivos**: Si bien aún en etapa temprana, se proyecta que la computación cuántica transformará la velocidad de procesamiento de grandes volúmenes de datos. Investigar cómo podrían integrarse estos avances con las infraestructuras actuales de almacenamiento representa un campo emergente de gran potencial.
* **Comparativas de rendimiento y coste entre plataformas *cloud***: Futuras investigaciones podrían analizar en detalle las diferencias de rendimiento, coste-beneficio y facilidad de mantenimiento entre tecnologías como *Amazon Redshift*, *Azure Synapse* y *Databricks*, considerando distintos volúmenes de datos y casos de uso.

# Referencias

Alaa Khalaf Hamoud, A., Salah Hashim, A., & Akeel Awadh, W. (2018). CLINICAL DATA WAREHOUSE: A REVIEW. *Iraqi Journal for Computers and Informatics*. https://doi.org/10.25195/2017/4424

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. 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=TQrCDMeO9scTALIhjECVytfoXaEQbHRqvcI2sRnCITA%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.-a). *¿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

Núria, E. (n.d.-b). *¿Qué es la arquitectura Medallion en un contexto Data Lakehouse?* Retrieved March 19, 2025, from https://blog.bismart.com/arquitectura-medallion-data-lakehouse

Olavsrud, T. (2017). *3 keys to keep your data lake from becoming a data swamp - ProQuest*. https://www.proquest.com/docview/1933320250?parentSessionId=QCe%2FARhihFDauFqaihPMug7vtjq3fJ3QEyG647oYNiA%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

Seethala, S. C. (2025). The Role of AI in Revolutionizing Finance Data Warehouses for Predictive Financial Modeling&amp;nbsp; *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 Torregrosa*. 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