Arquitectura Cloud desde la Ciencia de Datos: Cómo el Conocimiento en Infraestructura de Almacenamiento Mejora el Análisis de Datos para la Toma de Decisiones

Juan Manuel Sánchez Ortega
Catherine Balbuena Franco

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
Maria Bernarda Salazar Sánchez, Doctora (Ph.D.) en Ingeniería Electrónica

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2025

Cita	(Balbuena Franco & Sánchez Ortega, 2025)
Referencia **Estilo APA 7 (2020)**	Balbuena Franco, C & Sánchez Ortega, J.M. (2025). *Arquitectura Cloud desde la Ciencia de Datos: Cómo el Conocimiento en Infraestructura de Almacenamiento Mejora el Análisis de Datos para la Toma de Decisiones* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Danny Alexandro Múnera Ramírez
Coordinadora del Programa: Maria Bernarda Salazar Sánchez

Dedicatoria

Juan Manuel Sánchez Ortega

Dedico este artículo y esta especialización en general a mi familia, en quienes siempre encuentro motivación para dar lo mejor de mí.

Catherine Balbuena Franco

Dedico esta especialización a mi abuela, por su amor incondicional; a JuanG, por acompañarme y guiarme, y a Iris, por mirarme siempre con confianza y orgullo.

Tabla de contenido

Lista de Figuras

Resumen

En los últimos años, la generación de datos ha crecido de forma exponencial. Sectores como la salud, la economía, el medio ambiente y prácticamente todos los ámbitos del conocimiento producen información valiosa que, si se analiza correctamente, puede influir significativamente en la toma de decisiones de empresas, gobiernos y otras organizaciones. En este contexto, la ciencia de datos cumple un rol fundamental al convertir estos grandes volúmenes de datos en información clara, comprensible y útil para los tomadores de decisiones. Sin embargo, entre las fuentes de datos y los análisis que permiten tomar decisiones estratégicas, hay un componente clave: la forma en que los datos son almacenados, organizados y preparados para su uso. La eficiencia y escalabilidad de los sistemas de almacenamiento influyen directamente en la calidad del análisis final. Aunque estas tareas son desarrolladas principalmente por los ingenieros de datos, comprenderlas es esencial para cualquier científico de datos, ya que impactan su trabajo diario. Este artículo, aborda el contenido del curso "Data Management and Storage in the Cloud" de Google, profundizando en conceptos como almacenamiento eficiente, modelos de datos, normalización y desnormalización, gobernanza de datos, y buenas prácticas que optimizan los proyectos de ciencia de datos. En este sentido un buen diseño en la estructuración del almacenamiento no solo mejora el rendimiento de los sistemas, sino que permite a los científicos de datos generar análisis más precisos, relevantes y confiables, fortaleciendo así la toma de decisiones en cualquier campo de aplicación.

Palabras clave— Data Lake, Data Warehouse, Data Lakehouse, Almacenamiento.

1. Introducción

La cuarta revolución industrial que se está desarrollando actualmente implica la generación de una gran cantidad de datos que, muchas veces, no alcanzamos a dimensionar. Las interacciones digitales, las diferentes transacciones financieras, monitoreo de sistemas (médicos, ambientales, urbanos, etc), el denominado internet de las cosas (IoT, por sus siglas en inglés Internet of Things) generan una cantidad significativa de información que crece de forma exponencial. Bajo estas circunstancias, los científicos de datos enfrentamos un reto contundente en términos de cómo almacenar, estructurar, acceder y procesar datos para encontrar y extraer información útil que permita la toma de decisiones a partir de ese océano de datos. Para esto no sólo es necesario contar con conocimientos en estadística o la puesta en producción de modelos de aprendizaje automático (Ml, del inglés machine learning), sino también, que los profesionales en datos tengan una comprensión profunda de la infraestructura de datos moderna. En este sentido el curso "*Data Management and Storage in the Cloud*" nos introduce a los profesionales de datos en el mundo de la infraestructura de almacenamiento en servicios Cloud, abordando desde conceptos fundamentales hasta casos prácticos que permiten generar experiencia en la construcción de estas arquitecturas (Google LLC, 2024).

Una de las arquitecturas más relevantes en el almacenamiento moderno de datos es el uso de técnicas avanzadas como el modelo de Data Lakehouse el cual combina la flexibilidad y escalabilidad de los data lakes con las ventajas analíticas de los *Data Warehouses*. Este modelo permite trabajar con datos estructurados y no estructurados en un mismo entorno, sin necesidad de duplicar o transformar constantemente la información. En este sentido y dada la importancia del uso de Data *Lakehouse* en la actual demanda de servicios de almacenamiento de datos de diferentes fuentes, el uso de herramientas Cloud como Google Cloud Storage, BigQuery y Dataproc que permiten procesar grandes volúmenes de datos con diferentes estructuras, representan una oportunidad para que los científicos de datos centremos nuestros esfuerzos en el análisis, modelado y toma de decisiones, en lugar de dedicar tiempo a gestionar servidores o recursos locales.

Sin embargo, la creciente generación de datos exige consigo no sólo garantizar el almacenamiento de estos, sino también una correcta estructuración del modelado de datos. En este sentido, definir con claridad la necesidad del esquema de bases de datos, como los modelos normalizados o no normalizados, y el tipo de modelo en estrella o copo de nieve, cobran gran importancia en los proyectos de análisis y ciencia de datos. Lo anterior debido a que calidad de los modelos predictivos y la coherencia de los resultados dependen fuertemente de cómo se organizan y relacionan los datos, dado que esquemas bien diseñados ayudan a reducir la redundancia, mejorar la integridad de los datos y optimizar las consultas complejas (Batini, Cappiello, Francalanci, & Maurino, 2009)

Por otro lado, los profesionales en datos enfrentamos otro desafío cuando trabajamos con grandes volúmenes de datos. La capacidad de extraer datos desde su fuente origen y moverlos a un entorno analítico en la nube o procesarlos en servidores locales, son habilidades necesarias en análisis y ciencia de datos. El curso "Data Management and Storage in the Cloud" no es solo una introducción conceptual al almacenamiento en la nube, sino que proporciona herramientas como laboratorios con ejercicios reales para llevar a la práctica lo abordado a lo largo del mismo, permitiendo desarrollar competencias y conocimientos necesarios para cualquier profesional de datos.

2. Justificación académica

En un mundo digitalizado, donde cada interacción humana, comercial o tecnológica genera enormes volúmenes de datos con estructuras diversas (Lee, 2017), surge una preocupación esencial: ¿dónde y cómo almacenar esta información para aprovecharla eficazmente? La respuesta a esta interrogante no se limita al resguardo de datos, sino que abarca su organización, análisis y transformación para agregar valor a los negocios (Harby & Farhana , 2025).

El conocimiento sobre infraestructura de almacenamiento se ha vuelto un elemento clave, dado que permite preservar la información de manera segura, optimizar su análisis y hacer un uso estratégico de ella. En este contexto, plataformas como Google Cloud ofrecen oportunidades relevantes, como el curso *"Data Management and Storage in the Cloud"*, que proporciona una comprensión integral de las principales operaciones relacionadas con la gestión y almacenamiento de datos en entornos cloud (Google LLC, 2024).

Este curso está diseñado para abordar desafíos empresariales actuales mediante la elección adecuada de soluciones de almacenamiento, analizando sus capacidades y limitaciones. A través del estudio de los *Data Warehouses* y *Data Lakes*, e introduciendo el enfoque híbrido del *Data Lakehouse*; se enfoca en aprenden a combinar lo mejor de ambos modelos: la capacidad de ingesta rápida de datos no estructurados y el procesamiento eficiente para análisis posteriores (Janssen, Ilayperuma, Jayasinghe, Bukhsh, & Daneva, 2024). Google Cloud al ser una plataforma versátil, facilita la organización de la información dentro de estas arquitecturas, permitiendo trabajar tanto con esquemas normalizados como no normalizados. Esta flexibilidad operativa permite adaptarse a distintos objetivos analíticos, priorizando la optimización y reduciendo la carga computacional al evitar procesos complejos como múltiples uniones entre tablas, incluso si eso implica relajar ciertos principios de integridad de los datos (datos duplicados, inconsistencias, entre otros) (Kimball & Ross, 2013).

Un componente fundamental que también se tiene en cuenta en este ecosistema, es la gobernanza de datos, cada vez más relevante para asegurar que la información pueda ser gestionada, procesada y analizada de manera segura, confiable y exacta (Yébenes Serrano, 2022). Esta estructuración de políticas y estándares de datos, fortalece la confianza en los resultados, facilita el cumplimiento normativo y fomenta la transparencia en la toma de decisiones basada en datos (Harby & Farhana , 2025). Ahora bien, ¿cómo llevar los datos a la infraestructura cloud? Por medio del diseño e implementación de procesos eficientes de Extracción, Transformación y Carga (ETL), los cuales actúan como el pilar técnico que permite trasladar información desde múltiples fuentes hacia estos entornos analíticos (Azzabi, Alfughi, & Ouda, 2024).

En definitiva, el aprovechamiento estratégico de los datos en la era digital exige una comprensión profunda de su infraestructura de almacenamiento, lo que incluye la aplicación de arquitecturas (Janssen, Ilayperuma, Jayasinghe, Bukhsh, & Daneva, 2024); la elección entre normalización o no normalización, según los objetivos analíticos; la implementación de procesos ETL eficientes y escalables (Azzabi, Alfughi, & Ouda, 2024), así como el cumplimiento de principios sólidos de gobernanza de datos. Todo ello garantiza la integridad, seguridad y utilidad de la información obtenida por el profesional en datos y, se consolidan los datos como un activo valioso de la organización. Así, la integración coherente entre arquitectura, procesos y gobernanza no solo fortalece los proyectos de ciencia de datos, sino que impulsa la innovación y la competitividad en las organizaciones contemporáneas (Koçyiğit, Kayabay , Eren, Gökalp, & Gökalp, 2021).

3. Contenido del curso

Basados en el curso "*Data Management and Storage in the Cloud*" se abordan tres conceptos fundamentales que relacionan la ingeniería de datos con la ciencia de datos, mostrando que, si bien son ciencias diferentes, tienen, en realidad, una fuerte conexión entre sí. Este análisis permite evidenciar que el rol del científico de datos no es ajeno al conocimiento de cómo se estructuran y almacenan los datos y como dicho conocimiento conduce a optimizar los tiempos de preprocesamiento de datos, obtener datos de mejorar calidad y obtener resultados de alta confiabilidad. En la primera parte de esta sección se abordan las 3V del BigData (Volumen, Variedad y Velocidad), y la arquitectura óptima para el manejo de grandes y complejos volúmenes de datos. Luego se analiza como una correcta elección del modelo de datos, técnicas de normalización y gobernanza de datos impactan en el quehacer diario del científico de datos y finalmente, se estudia como la capacidad para extraer, mover y consultar datos de manera eficiente optimiza la gestión de la información.

3.1. Almacenamiento eficiente en el manejo de datos

La era digital ha revolucionado la forma en que se generan, se obtienen y se almacenan los datos. Estos datos provienen desde diferentes fuentes, campos de acción y contextos: bases de datos tradicionales, registros de interacciones en páginas web, redes sociales, datos geoespaciales, archivos de audio y video que muestran que la generación de datos diaria tiene un crecimiento exponencial. Este fenómeno se ha denominado Big Data y comprende: Volumen, relacionado con la gran cantidad de datos que se generan y como trabajar con este gran conjunto de información; Variedad, ligado con el volumen dado que existen diferentes formas de representarlos, ya sea estructurados, semiestructurados y no estructurados; y finalmente Velocidad, la cual se refiere a la velocidad con que se crean los datos (Camargo Vega, Camargo Ortega,, & Joyanes Aguilar, 2015).

Con base en lo anterior, se hace necesario utilizar arquitecturas que sean capaces de dar la flexibilidad y escalabilidad que el Big Data requiere. Las arquitecturas de almacenamiento tradicionales como el Data Warehouse funciona como el repositorio central de una organización que contiene datos que pasaron por procesos de ETL desde diferentes fuentes (Ver Figura 1a). Mientras que el Data Lake, es una arquitectura propuesta para solucionar algunas limitaciones del Data Warehouse, almacena grandes cantidades de datos sin procesar en su formato original hasta que sean necesarios (Ver Figura 1b). Sin embargo, estas arquitecturas han presentado limitaciones tanto estructurales como operacionales surgiendo así el modelo de Data Lakehouse que toma las mejores características de los Data Warehouse y data Lake para abordar y/o superar las limitaciones de estas metodologías tradicionales (Ver Figura 1c). Este denominado Data Lakehouse, es una arquitectura híbrida que combina las capacidades de análisis y la arquitectura estructurada de los Data Warehouse con la flexibilidad y bajo costo de almacenamiento de los Data Lake. Permite almacenar datos estructurados, como bases de datos relacionales, pero también datos no estructurados como imágenes o vídeos en un mismo entorno, con una única capa para administrar y controlar el acceso a dichos datos. Esto permite así realizar análisis directamente sobre los datos crudos o semiprocesados según sean requeridos por el usuario.

El entendimiento de estas nuevas arquitecturas de almacenamiento impacta significativamente el quehacer diario del científico de datos, pues ayudan a reducir el tiempo y esfuerzo que se dedica a la preparación, limpieza y transferencia de datos entre diferentes fuentes. Además, ya que el Data Lake permite el almacenamiento en bruto de los datos, elimina la necesidad de realizar transformaciones costosas computacionalmente permitiendo así acceder y analizar los datos en su formato original, lo

cual no sólo acelera el proceso de análisis sino también mantiene a variedad y originalidad de la información.

Por otro lado, estas arquitecturas facilitan la colaboración interdisciplinaria entre los distintos equipos que participan en los proyectos de ciencia de datos, dando trazabilidad y permitiendo la gestión adecuada de gobernanza de datos como lo son el control de versiones, los metadatos, y la asignación de roles y políticas de acceso para promover la calidad de los datos.

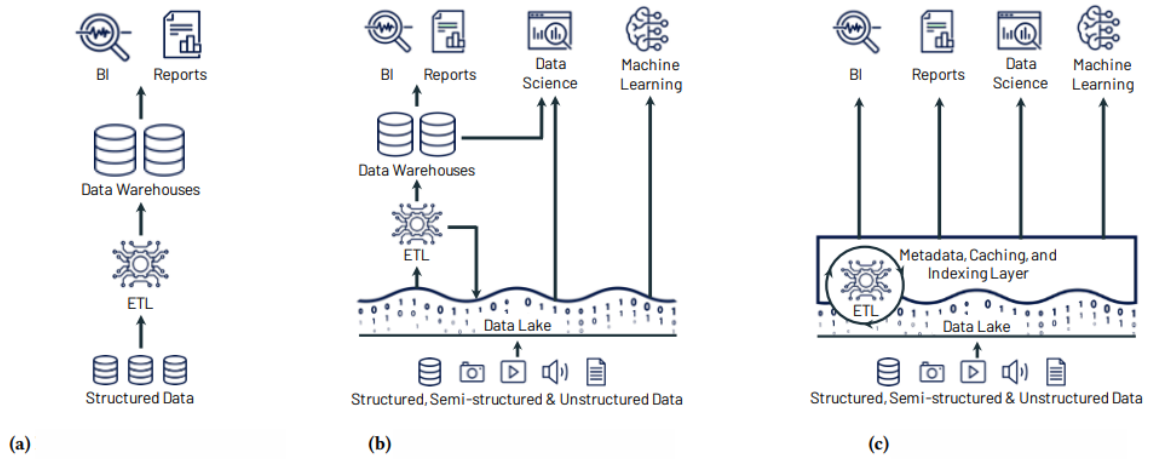

Figura 1. *Esquema de Estructuras de Almacenamiento. a) Esquema de Data Warehouse. b) Esquema de Data Lake. c) Esquema de Data LakeHouse. Tomado de (Armbrust et al., 2021)*

3.2. Modelo de datos, normalización y no normalización y gobernanza de datos

Como se ha venido abordando a lo largo de este informe, la cuarta revolución industrial ha traído consigo una generación masiva de datos que deben ser adecuadamente gestionados garantizando para que estos tengan un impacto real en los diferentes contextos de uso. En este sentido, para que los análisis de los datos sean efectivos y confiables, estos grandes conjuntos de datos deben contar con modelos de datos adecuados que estructuren y organicen la información coherentemente. Un modelo de datos adecuado (según el contexto de uso) facilita la manipulación, compresión y la eficiencia de los análisis mejorando así la toma de decisiones a partir de ellos (Koçyiğit, Kayabay , Eren, Gökalp, & Gökalp, 2021)

Cuando se habla de un modelo de datos, se hace referencia a una representación de cómo se relacionan los datos entre sí. Este concepto aporta claridad sobre qué tipo de información se almacena, como se accede y como se relacionan con otros datos (tablas). Uno de los diseños de modelos de datos abordado en el curso, es el modelo en estrella. Este se basa en una tabla central, denominada tabla de hechos que contiene los datos transaccionales (ventas, ingresos, cantidades, registros, etc), y varias tablas al rededor, conocidas como tablas de hecho, que contienen la información de atributos como productos, clientes, tiempo, ubicación, entre otras. Este enfoque facilita el acceso a los datos y su simplicidad mejora el rendimiento de consultas al evitar relaciones complejas entres las tablas que componen la base de datos (Kimball & Ross, 2013)

Por otro lado, el curso también enfatiza en dos conceptos clave en la estructuración de los modelos de datos: Por un lado, está la normalización la cual dividir los datos en tablas más pequeñas y relacionadas entre sí, buscando disminuir la redundancia y asegurar la integridad de los datos, y es utilizada sobre todo en contextos de bases de datos operacionales. Por su parte, la no normalización consiste en combinar datos en tablas más grandes replicando algunas columnas, reduciendo las uniones entre tablas, lo cual es importante cuando se trabaja con grandes volúmenes de datos (Inmon, 2005).

Además de los conceptos anteriormente mencionados, se abordó la gobernanza de datos la cual se basa en un conjunto de políticas para garantizar los estándares de calidad y asignación de roles que aseguren que los datos de una entidad o empresa sean gestionados de forma segura, consistente y responsablemente a lo largo del ciclo de vida tanto de los datos como de los proyectos. Esto permite la correcta calidad de los datos, su disponibilidad para diferentes usuarios y el uso adecuado según normativas tanto internas como externas.

Uno de los procesos usuales en la ejecución de proyectos del área de datos es la limpieza de información, en este sentido se suelen encontrar información duplicada, registros faltantes, valores no coherentes con las variables de análisis, por mencionar algunas. Garantizar un proceso de gobernanza de datos adecuado, genera una disminución del preprocesamiento de la información, ya que existen procesos definidos para validación, limpieza y generación de la información antes de ser guardada. Esto mejora considerablemente tanto el tiempo de ejecución de los proyectos como los resultados obtenidos a partir de los análisis (Khatri & Brown, 2010)

Si bien estos conceptos no son de naturaleza del rol específico de la ingeniería de datos, tanto la definición adecuada del modelo de datos a utilizar, así como las técnicas de normalización, desnormalización y un correcto marco de gobernanza de datos impactan directamente en el normal desempeño de los científicos de datos.

3.3. Procesos óptimos para mejorar los proyectos de ciencia de datos

Los procesos ETL (Por sus siglas en inglés, Extract, Transform, Load) son parte fundamental de la práctica diaria del científico de datos, estos representan los procesos para recopilar información de diferentes fuentes, estandarizarlos según el contexto del proyecto y cargarlos en un sistema donde puedan ser consultados de forma directa y eficiente. El conocimiento profundo de estos procesos por parte del científico de datos le permite participar activamente en las diferentes etapas del ciclo de vida de los proyectos, desde la recolección de información hasta la puesta en producción de modelos analíticos y tomas de decisiones. Este ciclo de vida comprende fases como obtención, limpieza, aplicación de modelos, validación e interpretación. La extracción se refiere a obtener datos de diferentes orígenes. Por su parte, la transformación consiste en limpiar, estandarizar, identificar valores faltantes y demás transformaciones necesarias según lo establecido para cada variable. Finalmente, la carga lleva estos datos a un repositorio de final (Kimball & Caserta, 2011)

Estructurar de manera correcta un proceso de ETL en el proyecto de datos, garantiza la calidad de estos, conduciendo así a obtener modelos más precisos, generar reportes confiables y visualizaciones mucho más útiles y dicientes. Por ejemplo, analizar valores atípicos, corregir inconsistencias entre fuentes y tomar decisiones para abordar los valores nulos, entre otras transformaciones disminuye los errores y se garantiza la confianza en los resultados analíticos (Batini, Cappiello, Francalanci, & Maurino, 2009). Así pues, la ETL no es un proceso propio de la ingeniería de datos, sino más bien un puente que conecta

a estos con los científicos de datos y se puede considerar el punto de partida para iniciar de forma adecuada y óptima los proyectos analíticos con gran confianza en los resultados.

4. Conclusiones

El almacenamiento eficiente de datos es clave en la era del Big Data, donde la variedad, velocidad y volumen de la información requieren arquitecturas modernas como el *Data Lakehouse*. Esta solución híbrida permite trabajar con datos estructurados y no estructurados de forma flexible y escalable, optimizando tiempos y recursos (Janssen, Ilayperuma, Jayasinghe, Bukhsh, & Daneva, 2024). Además, facilita la gobernanza y colaboración entre equipos, lo que convierte al almacenamiento no solo en una necesidad técnica, sino en una herramienta estratégica para la toma de decisiones y la innovación organizacional (Koçyiğit, Kayabay , Eren, Gökalp, & Gökalp, 2021).

En este contexto, la eficiencia en el manejo de datos no solo depende de dónde y cómo se almacenan, sino también de cómo se estructuran y gestionan. Modelos de datos bien diseñados, como el modelo en estrella, junto con decisiones acertadas sobre normalización o desnormalización, permiten optimizar consultas, reducir redundancias y mantener la integridad de la información (Kimball & Ross, 2013). A su vez, la implementación de políticas claras de gobernanza de datos asegura la calidad, trazabilidad y disponibilidad de la información, reduciendo el esfuerzo en tareas de limpieza y preparación (Inmon, 2005).

Complementariamente, los procesos ETL desempeñan un papel esencial como puente entre la ingeniería y la ciencia de datos, permitiendo trasladar información desde múltiples fuentes hacia entornos preparados para el análisis (Azzabi, Alfughi, & Ouda, 2024). La correcta implementación de estos procesos garantiza que los datos sean estandarizados, consistentes y útiles desde el inicio del ciclo de vida analítico.

En suma, el éxito de los proyectos de ciencia de datos en la era del Big Data depende de la integración de una arquitectura de almacenamiento moderna, un modelado de datos adecuado, una sólida gobernanza de la información y procesos ETL bien estructurados. Estos elementos, al trabajar en conjunto, permiten garantizar la calidad, accesibilidad y utilidad de los datos, optimizando el análisis y facilitando una toma de decisiones precisa, ágil y confiable.
En este contexto, el curso “Data Management and Storage in the Cloud de Google Cloud” representa un complemento valioso en la formación de especialistas en analítica y ciencia de datos, ya que no solo proporciona conocimientos técnicos clave, sino que también desarrolla habilidades para adaptarse a un entorno tecnológico en constante evolución como lo es el cloud. La gestión de datos es, sin duda, un campo dinámico que continúa transformándose para dar respuesta a las crecientes demandas de almacenamiento, procesamiento y análisis de información en contextos cada vez más complejos.

Referencias

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021, January). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In Proceedings of CIDR. *8*, 28. https://15721.courses.cs.cmu.edu/spring2023/papers/02-modern/armbrust-cidr21.pdf

Azzabi, S., Alfughi, Z., & Ouda, A. (2024). Data Lakes: A Survey of Concepts and Architectures. *MDPI*. https://www.mdpi.com/2073-431X/13/7/183

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR). 41(3), 1–52. https://doi.org/10.1145/1541880.1541883

Camargo Vega, J. J., Camargo Ortega,, J. F., & Joyanes Aguilar, L. (2015). Conociendo big data. Revista Facultad de Ingeniería. 24(38), 63-77. http://www.scielo.org.co/scielo.php?pid=S0121-11292015000100006&script=sci_arttext

Freitas, N., Rocha, A. D., & Barata, J. (s.f.). Data management in industry:concepts, systematic review and future directions. *Hournal of Intelligent Manufacturing*.

Google LLC. (2024). *Data Management and Storage in the Cloud - Español*. Google Cloud Skills Boost.

Harby, A. A., & Farhana , Z. (2025). Data Lakehouse: A survey and experimental study. *ScienceDirect*.

Inmon, W. (2005). Building the Data Warehouse. *John Wiley & Sons*.

Janssen, N., Ilayperuma, T., Jayasinghe, J., Bukhsh, F., & Daneva, M. (2024). The evolution of data storage architectures: examining the secure value of the Data Lakehouse. *Journal of Data, Information and Management*.

Khatri, V., & Brown, C. V. (2010). *Designing data governance. Communications of the ACM.* https://doi.org/10.1145/1629175.1629210

Kimball, R., & Ross, M. (2013). *he Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling (3rd ed.).* Wiley.

Kimball, R., & Caserta, J. (2011). *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Wiley.*

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling. John Wiley & Sons.* https://books.google.es/books?hl=es&lr=&id=4rFXzk8wAB8C&oi=fnd&pg=PR27&dq=The+Data+Warehouse+Toolkit:+The+Definitive+Guide+to+Dimensional+Modeling.+John+Wiley+%26+Sons.&ots=3q6RMfZ3NM&sig=SP-bJkqnOH3Yf94liZihXYHZpdk

Koçyiğit, A., Kayabay , K., Eren, P., Gökalp, M., & Gökalp, E. (2021). Data-driven manufacturing: An assessment model for data science maturity. *ELSEVIER*. https://doi.org/10.1016/j.jmsy.2021.07.011

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges.

Yébenes Serrano, J. R. (2022). *Marco para la construcción de sistemas de gobernanza de datos en entornos de industria 4.0.*