



## **Informe académico del curso Data Management and Storage in the Cloud**

José Leonardo Herrera Gutiérrez

Informe final del curso de extensión para optar al título de Especialista en Analítica y  
Ciencia de Datos

Asesora

Maria Bernarda Salazar Sánchez, Doctora (PhD) en Ingeniería Electrónica

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

| Cita                       | (Herrera Gutiérrez, 2025)                                                                                                                                                                            |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Referencia</b>          | Herrera Gutiérrez, J. L. (2025). <i>Informe final académico del curso Data Management and Storage in the Cloud</i> [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia. |
| <b>Estilo APA 7 (2020)</b> |                                                                                                                                                                                                      |



Grupo de Investigación Intelligent Information Systems Lab In2Lab

Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).



Centro de Documentación Ingeniería (CENDOI)

**Repositorio Institucional:** <http://bibliotecadigital.udea.edu.co>

Universidad de Antioquia - [www.udea.edu.co](http://www.udea.edu.co)

**Rector:** John Jairo Arboleda Céspedes.

**Decano:** Julio Cesar Saldarriaga Molina

**Jefe departamento:** Danny Alejandro Múnera Ramírez

**Coordinadora del Programa:** Maria Bernarda Salazar Sánchez

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

## **Dedicatoria**

A mi familia, por la vida y el apoyo. A mi entorno, por las enseñanzas. A mi pareja, por el amor.

## **Agradecimientos**

Mi más profundo agradecimiento a cada persona que, con empeño, ha compartido su conocimiento conmigo, directa o indirectamente. A las instituciones donde me he educado. Agradezco por cada oportunidad tomada y perdida; pasada, presente y futura.

## **Tabla de contenido**

|                                 |    |
|---------------------------------|----|
| Resumen .....                   | 7  |
| 1. Introducción.....            | 8  |
| 2. Justificación Académica..... | 9  |
| 3. Descripción técnica.....     | 10 |
| 4. Conclusiones.....            | 12 |
| Referencias .....               | 13 |

## **Lista de Tablas**

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
| Tabla 1. Comparación de herramientas en Google Cloud Platform.....                   | 9  |
| Tabla 2. Contenidos abordados en cada módulo del curso .....                         | 10 |
| Tabla 3. Descripción de los contenidos abordados en las sesiones de laboratorio..... | 11 |

## **Siglas, acrónimos y abreviaturas**

|                |                                                   |
|----------------|---------------------------------------------------|
| <b>APA</b>     | American Psychological Association                |
| <b>Cms.</b>    | Centímetros                                       |
| <b>ERIC</b>    | Education Resources Information Center            |
| <b>Esp.</b>    | Especialista                                      |
| <b>GCP</b>     | Google Cloud Platform                             |
| <b>IA</b>      | Inteligencia Artificial                           |
| <b>IEEE</b>    | Institute of Electrical and Electronics Engineers |
| <b>MP</b>      | Magistrado Ponente                                |
| <b>MSc</b>     | Magister Scientiae                                |
| <b>Párr.</b>   | Párrafo                                           |
| <b>PhD</b>     | Philosophiae Doctor                               |
| <b>PBQ-SF</b>  | Personality Belief Questionnaire Short Form       |
| <b>PostDoc</b> | PostDoctor                                        |
| <b>UdeA</b>    | Universidad de Antioquia                          |

## **Resumen**

En este informe se evalúa el curso de extensión Data Management and Storage in the Cloud emitido por Google, y su eficacia en resolver retos de manejo de datos como la escalabilidad e integración con herramientas analíticas. En este reporte se tiene como objetivo revisar la importancia del curso con respecto a las necesidades en la industria de soluciones en la nube. Este documento, presentado en formato impreso, se enfoca en analizar cómo el curso cubre habilidades demandadas en la industria.

En este reporte se examina la eficiencia de usar proyectos prácticos para que los estudiantes tengan un acercamiento a tareas específicas; también se evalúan la integración de las herramientas de Google Cloud para ocuparse de retos reales en la industria. Además, se explora el alineamiento del curso con las tendencias actuales en big data e inteligencia artificial. Mediante este análisis, en este informe se enfatiza la importancia del curso en la capacitación de profesionales y estudiantes en el uso de arquitecturas híbridas y ecosistemas en la nube.

***Keywords— computación en la nube, datos distribuidos, inteligencia artificial, GCP, arquitectura en la nube.***

## **1. Introducción**

El crecimiento exponencial de los datos disponibles y su complejidad asociada ha requerido metodologías avanzadas para el almacenamiento y manejo de los datos. Los sistemas tradicionales como las bases de datos relacionales tienen problemas de escalabilidad y para manejar datos no estructurados, por lo cual han surgido arquitecturas nuevas como los data Lakehouses que se caracterizan por su flexibilidad y rendimiento. El curso de Data Management and Storage in the Cloud [1] aborda estos retos ofreciendo herramientas y habilidades para utilizar los ecosistemas en la nube.

En el curso se resuelven problemas como el gobierno de datos ineficiente, los sistemas de almacenamiento fragmentados, y la falta de integración con herramientas de Inteligencia Artificial. Su audiencia se compone de profesionales en datos que buscan conocimiento de soluciones en la nube, especialmente en sus aplicaciones prácticas para analítica y técnicas de aprendizaje automático (ML, por sus siglas en inglés machine learning). El curso prepara a los estudiantes en BigQuery, administración de datos con Dataplex y análisis predictivo usando Vertex AI. El curso tiene una duración de 26 horas y está dividido en cuatro módulos: (1) Introducción a la administración y el almacenamiento de datos en la nube; (2) Componentes claves de la organización de datos; (3) Pasos para encontrar datos, y (4) Técnicas para acceder a los datos. Incorpora vídeos, textos y laboratorios para poner en práctica los conocimientos teóricos en todos los módulos.

Este informe académico explora cómo el curso cubre las necesidades de diferentes industrias desde el uso de plataformas en la nube como una herramienta para ser competitivos con el uso de técnicas de aprendizaje automático, automatización y manejo de grandes volúmenes de datos. El análisis realizado aborda la revisión de material escrito, audiovisuales y actividades prácticos en modalidad de laboratorios, donde se evalúa las habilidades del estudiante o profesional en términos de liderar el manejo de datos en un entorno tecnológico con el uso de herramientas de inteligencia artificial.

## 2. Justificación Académica

En la actualidad, el manejo de grandes volúmenes de información requiere el uso de datos estructurados y no estructurados a través de sistemas distribuidos. Durante la mayor parte del tiempo, las bases de datos relacionales han sido la principal forma de organizar los datos, debido a la gran cantidad de transacciones que se realizan en todas las industrias. Sin embargo, con el crecimiento de los datos no estructurados, los sistemas NoSQL han cobrado una relevancia dada su capacidad para escalar y manejar este tipo de datos. Los avances recientes como los data lakehouse, combinan los beneficios de los data warehouse en gobierno de datos y los data lakes en flexibilidad [2].

Este curso implementa herramientas en sus laboratorios como Gemini for Workspace para consultas asistidas con inteligencia artificial, y Dataplex para el gobierno de datos [3]. Además, se enfoca en el linaje de datos, un tema importante en el cumplimiento de normativas en diferentes industrias. El curso aborda temas importantes que frecuentemente son omitidos al combinar conceptos teóricos como normalización y desnormalización utilizando herramientas prácticas en la nube. Su enfoque en procesamiento de datos en tiempo real (batch, streaming) y la optimización en casos mediante tablas particionadas, ayuda a los estudiantes a prepararse para retos de la industria.

| Herramienta | Usos principales                    | Características                                    | Integración con IA                          |
|-------------|-------------------------------------|----------------------------------------------------|---------------------------------------------|
| BigQuery    | Analíticas con SQL                  | Detección automática de esquemas, particionamiento | Gemini para la optimización de consultas    |
| Dataproc    | Procesamiento de datos distribuidos | Integración con Apache Spark y Hadoop              | Pipelines de ML con Vertex AI               |
| Dataplex    | Gobierno de metadatos               | Linaje de datos                                    | Etiquetamiento automático con modelos de IA |

*Tabla 1. Comparación de herramientas en Google Cloud Platform*

Este curso se especializa en los ecosistemas de Google Cloud, y ofrece laboratorios de situaciones reales. Se combinan frameworks teóricos con herramientas en la nube, mediante sus laboratorios se abordan temas como el procesamiento en tiempo real, la optimización de costos, dotando a los profesionales la habilidad para aplicarlos en la industria.

### 3. Descripción técnica

El curso de Data Management and Storage in the Cloud está estructurado para unificar fundamentos teóricos con habilidades prácticas, iniciando con conceptos básicos de almacenamiento de datos, prosiguiendo con arquitecturas más complejas y finalizando con la integración con técnicas de aprendizaje automático. El primer módulo, extracción y almacenamiento de datos, introduce conceptos principales de bases de datos relacionales, sistemas NoSQL, y métodos de ingestión de datos como APIs y conectores. Este conocimiento es importante para entender cómo se obtienen los datos estructurados y no estructurados, así como su transformación y transferencia a través de sistemas en la nube.

El segundo módulo, componentes clave de la organización de datos, se centra en los frameworks de gobierno de datos, manejo de metadatos, y arquitecturas modernas. En este módulo se exploran los esquemas normalizados y desnormalizados, se comparan los modelos de estrella y copo de nieve, al tiempo que se profundiza en arquitecturas híbridas como los data lakehouses. Un tema importante es la integración de capas de almacenamiento (data lakes) y capas de analítica (data warehouses), y como los catálogos de metadatos en herramientas como Dataplex ayudan a unificar el gobierno de datos en sistemas distribuidos. En los laboratorios prácticos, se simulan escenarios reales, como combinar archivos CSV provenientes de un data lake con tablas de un data warehouse usando BigQuery para abordar preguntas de negocios.

En el tercer módulo llamado pasos para encontrar datos, se usa el panel explorador de BigQuery para buscar datasets públicos. En los laboratorios, los usuarios identifican fuentes de datos en batch y streaming, desarrollando habilidades de interpretación de metadatos. En este módulo también se introduce el Analytics Hub, donde se muestra cómo las organizaciones comparten los datasets de forma segura.

En el módulo final, técnicas para acceder a los datos, se une la ingeniería de datos con analíticas avanzadas. En esta parte, se optimiza el rendimiento de las consultas usando técnicas de particionamiento y clustering en BigQuery, reduciendo costos. Además, se integran flujos de trabajo de aprendizaje automático usando Vertex AI y Dataproc, usando notebooks de Jupyter en Python para automatizar el entrenamiento de modelos predictivos [4]. Por ejemplo, en el laboratorio se guía a los usuarios a generar data sintética con inteligencia artificial generativa para simular escenarios de pruebas, abordando retos éticos y logísticos.

| Módulo | Temas principales                             | Herramientas usadas     | Habilidades                                         |
|--------|-----------------------------------------------|-------------------------|-----------------------------------------------------|
| 1      | BD relacionales, APIs, datos anidados         | BigQuery                | Ingesta de datos, diseño de esquemas                |
| 2      | Gobierno de datos, arquitectura de Lakehouses | Dataplex                | Administración de metadatos                         |
| 3      | Descubrimiento de datos, linaje de datos      | Analytics Hub, Dataplex | Cumplimiento de datos                               |
| 4      | Particionado de datos, integración de ML      | Dataproc, Vertex AI     | Optimización de consultas, entrenamiento de modelos |

*Tabla 2. Contenidos abordados en cada módulo del curso*

En el curso se emplea la metodología de aprendizaje basado en proyectos a través del caso de estudio ficticio de TheLook eCommerce a lo largo de todos los módulos. Se realizan tareas considerando distintos roles, como por ejemplo mejorar la eficiencia de consultas siendo un analista de datos en la nube. La plataforma de la nube utilizada para los laboratorios se llama Qwiklabs, y se cuentan con información complementaria como glosarios, documentos, y quizzes en formato evaluación. Las pruebas prácticas ofrecen retroalimentación para afianzar conceptos; mientras que el proyecto final agrupa todas las habilidades requeridas para diseñar una tabla particionada, optimizar consultas y generar un reporte de ventas mediante BigQuery y Dataproc [5]. Este contenido crea un ambiente de aprendizaje que permite a los profesionales aplicar las herramientas necesarias en los ecosistemas en la nube.

| <b>Laboratorio</b>                                                | <b>Objetivo</b>                                                           | <b>Herramientas usadas</b>     | <b>Importancia en la industria</b>                 |
|-------------------------------------------------------------------|---------------------------------------------------------------------------|--------------------------------|----------------------------------------------------|
| Explora los tipos de datos planos y anidados en BigQuery          | Solucionar problemas usando JSON planos y anidados                        | BigQuery                       | Manejo y análisis de datos jerárquicos             |
| Identifica diferentes fuentes de datos por lotes y de transmisión | Clasificar conjuntos de datos                                             | BigQuery, Pub/Sub              | Diseño de pipelines por lotes o streaming          |
| Explora un Lakehouse                                              | Combinar datos en CSV a partir de un Lakehouse                            | BigQuery, Google Cloud Storage | Implementación de arquitecturas híbridas           |
| Explora Dataplex                                                  | Buscar y catalogar tablas que contienen información de ventas             | Dataplex, BigQuery             | Administración de metadatos                        |
| Compara los análisis de datos con BigQuery y Dataproc             | Generar reportes usando SQL y Python                                      | Dataproc, Jupyter              | Workflows analíticos usando diferentes plataformas |
| Administra una tabla particionada en BigQuery                     | Optimizar la eficiencia de consultas usando el particionamiento de tablas | BigQuery                       | Reducción de costos en consultas a gran escala     |

*Tabla 3. Descripción de los contenidos abordados en las sesiones de laboratorio*

#### **4. Conclusiones**

El curso Data Management and Storage in the Cloud sirve de puente entre la educación teórica y las necesidades en la industria. Se enfoca en arquitecturas híbridas como los data lakehouses prepara a los estudiantes para consolidar ecosistemas de datos fragmentados, y a su vez la integración con IA proporciona las herramientas para estar al día con las innovaciones tecnológicas.

A través del aprendizaje por proyectos se asegura que estos conocimientos sean aplicados en casos desde la optimización de consultas hasta el despliegue de modelos de ML. Aunque es un curso que ofrece una estructura sólida y se alinea con las demandas de las empresas, se recomienda a los estudiantes explorar cursos similares que sean aplicados en plataformas como AWS y Azure.

En conclusión, este curso es indispensable para profesionales que trabajen en datos e IA, ya que no solo ayuda a desarrollar las habilidades individuales sino también fortalece a las organizaciones a aprovechar los datos como un recurso valioso.

## Referencias

- [1] Qwiklabs, «Data Management and Storage in the Cloud,» 06 Mayo 2024. [En línea]. Available: [https://www.cloudskillsboost.google/paths/420/course\\_templates/962](https://www.cloudskillsboost.google/paths/420/course_templates/962). [Último acceso: 2025].
- [2] F. Tekiner y S. Pierce, «Open data lakehouse on Google Cloud,» 28 Octubre 2021. [En línea]. Available: <https://cloud.google.com/blog/products/data-analytics/open-data-lakehouse-on-google-cloud>. [Último acceso: 2025].
- [3] G. Verghese, «Understand and trust data with Dataplex data lineage,» 13 Marzo 2023. [En línea]. Available: <https://cloud.google.com/blog/products/data-analytics/announcing-dataplex-data-lineage-in-ga>. [Último acceso: 2025].
- [4] Google, «Introduction to Vertex AI,» [En línea]. Available: <https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform>. [Último acceso: 2025].
- [5] Google, «Introduction to partitioned tables,» [En línea]. Available: <https://cloud.google.com/bigquery/docs/partitioned-tables?hl=es-419>. [Último acceso: 2025].