

> Especialización en

# GOOGLE DATA ENGINEERING



## ¿Por qué llevar esta especialización?

¿Sabías que, gracias a la visión centrada en la simplicidad, velocidad y analítica avanzada, Google Cloud Platform redefine la ingeniería de datos con herramientas como BigQuery, Dataflow y Pub/Sub, pensadas para que los datos fluyan de forma eficiente y escalable desde su origen hasta el análisis?

Por ello, DMC INSTITUTE ha diseñado la Especialización Google Data Engineering que te permitirá dominar el diseño de pipelines de datos (ETLs) modernos para el procesamiento por lotes así como en tiempo real, aprovechando la escalabilidad automática y la arquitectura sin servidores de Google.



## Sobre esta Especialización

**14**

sesiones

**56**

horas  
académicas

**37**

talleres  
prácticos

## ¿Cómo impulsamos tu carrera?

- Sesiones 80% enfocadas en la práctica.
- Enfoque en **Casos Aplicados a Negocio**, enfrentando los retos del mercado.
- Énfasis en **habilidades técnicas**.
- **Mentoría especializada** con docentes expertos.
- Acompañamiento **constante**.



# Objetivos de la Especialización

## OBJETIVO GENERAL:

- Utiliza los servicios de la plataforma de Google Cloud Platform (GCP) para el diseño e implementación de Lakehouses para entornos de analítica o ciencia de datos, con soporte para manejo de datos masivos (Big Data), fuentes batch y streaming y diferentes formatos de tabla como Delta, Iceberg y Hudi.

## ¿Cuáles son los requisitos?



### Conocimientos / Habilidades

- Comprensión de bases de datos relacionales y sus principales conceptos: tabla, relación, cardinalidad.
- Conocimiento de un lenguaje de programación.



### Experiencia Laboral

Mínimo un año de experiencia trabajando en áreas de datos, o elaborando reportes e informes.



### Tecnológicos

- Contar con una laptop o computadora de escritorio con sistema operativo Windows y disponibilidad de micrófono y cámara web.
- Contar con una cuenta de Google Cloud Platform (GCP)

## Dirigido a

- **Ingenieros de datos Jr.**, que actualmente busquen:
  - Especializarse en el manejo de la plataforma GCP para la implementación de sus soluciones de datos.
- **Ingenieros de datos expertos en sistemas on-premise o en otras plataformas cloud**, que estén buscando:
  - Migrar sus soluciones a la nube GCP, o crear entornos híbridos con ella.
- **Analistas y científicos de datos**, que estén buscando:
  - Incorporar a sus proyectos de visualización y/o de machine learning un motor de datos robusto basado en la nube de Google.
  - Complementar o enrumbar su carrera hacia la ingeniería de datos en la nube.
- **Profesionales de las áreas de sistemas, informática, desarrollo de software y afines**, que estén buscando:
  - Iniciar y especializarse en la carrera de ingeniería de datos on cloud, de rápido crecimiento y alta demanda laboral.



# Malla Curricular

## I. GCP Fundamentals

### 1. Cloud Fundamentals

- Cloud Computing. Definición y ventajas. Diferencias respecto a sistemas on-premise.
- Clasificación clásica de servicios: IaaS, PaaS, SaaS.
- Evolución de los servicios cloud. Everything as a Service (XaaS).
- Servicios dedicados vs Serverless. Diferencias principales, casos de uso.
- Modelos de costo generales. Por almacenamiento, por capacidad de cómputo, por flujo de datos, entre otros.

### 2. GCP Fundamentals

- **Taller:** Exploración de los componentes principales de la consola de GCP.
- **Taller:** Introducción a los servicios de cómputo, creación y testeo de una máquina virtual.
- **Taller:** Introducción a los servicios de almacenamiento, creación y testeo de un Cloud Storage.
- **Taller:** Introducción a los servicios de bases de datos, creación y testeo de bases de datos relacionales.
- **Taller:** Introducción a los servicios de seguridad. Configuración y uso de IAM y KMS.
- **Taller:** Introducción a los servicios de monitoreo. Recomendaciones para minimizar el billing.

## II. GCP Lakehouse Fundamentals (Low-code)

### 3. Introducción a Data Warehousing

- Data Warehouse vs. BD relacional. Repaso de motores DW disponibles.
- Arquitecturas modernas para Warehouses.
- Modelos tabulares y multidimensionales. Aplicaciones en analytics y data science.

### 4. Configuración y uso de un Datalake básico

- Base de datos vs. Datalake vs. Lakehouse. Diferencias conceptuales y aplicaciones.
- **Taller:** Configuración de una arquitectura de dos capas en Cloud Storage (Raw, Curated).
- **Taller:** Control de acceso mediante GCS y ACLs.
- **Taller:** Exposición de un Cloud Storage para lectura y escritura desde Google Cloud Shell.

### 5. Ingesta desde archivos con Data Fusion (GUI)

- Ingesta Batch vs. Streaming. Diferencias conceptuales y casos de uso.
- **Taller:** Implementación de un ETL para ingesta batch de archivos (CSV) desde otro Cloud Storage.
- **Taller:** Automatización del ETL basada en detección de archivos nuevos en el bucket (Eventarc, Cloud Functions).
- **Taller:** Automatización del ETL basada en calendarización (Cloud Scheduler).

### 6. Ingesta desde base de datos SQL con Data Fusion (GUI)

- **Taller:** Implementación de un ETL para ingesta incremental desde una base SQL de GCP (Cloud SQL).
- **Taller:** Automatización del ETL basada en calendarización (Cloud Scheduler).

### 7. Procesamiento y flujo de datos en el Datalake con Data Fusion (GUI)

- **Taller:** Implementación de un ETL para procesamiento de datos de Raw a Curated.

### 8. Conversión a Lakehouse con BigQuery y buenas prácticas

- Revisión conceptual de arquitecturas con soporte Push-down para optimización de respuesta.
- Revisión conceptual sobre escalamiento de Integration Runtime (IR) y paralelización de pipelines.
- **Taller:** De Datalake a Lakehouse con BigQuery.
- **Taller:** Test de conexión desde Looker y desde Power BI Desktop.

## III. GCP Scalable Lakehouses

### 9. Introducción a la arquitectura Medallion

- Arquitectura dos capas (Raw, Curated) vs. Arquitectura Medallion (Bronze, Silver, Gold). Diferencias conceptuales, aplicaciones, relación con Big Data.
- Almacenamiento para Medallion. Delta vs. Hudi vs. Iceberg. Comparativa con formatos Parquet, AVRO.
- Enfoques ETL vs. ELT para datalakes modernos.

### 10. BigQuery para Lakehousing

- BigQuery como Datalake. Casos de uso, ventajas y limitaciones.
- **Taller:** Implementación de una Datalake con arquitectura Medallion sobre BigQuery.

### 11. Ingesta Batch en BigQuery

- **Taller:** Uso de lenguaje SQL para leer, transformar y cargar en BigQuery (bronze) data desde Cloud Storage.
- **Taller:** Uso de Datastream para cargar datos en BigQuery (bronze) desde un motor SQL (incremental).

### 12. Procesamiento y flujo de datos en BigQuery (+Dataform)

- **Taller:** Uso de lenguaje SQL para la implementación de ETLs de Bronze a Silver.
- **Taller:** Uso de lenguaje SQL para la implementación de ETLs de Silver a Gold.
- **Taller:** Automatización de ETLs según calendario con Schedule Queries.
- **Taller:** Automatización y orquestación de ETLs con Dataform.
- **Taller:** Test de conexión desde Looker Studio.

# Malla Curricular

## 13. Cloud Storage para Lakehousing

• Cloud Storage como Datalake. Casos de uso, ventajas y limitaciones.

• **Taller:** Implementación de un Datalake con arquitectura Medallion sobre un Cloud Storage.

## 14. Ingesta Batch con Dataflow

• Data Fusion vs. Dataflow vs. Dataflow Batch. Casos de uso.

• **Taller:** Pipeline Dataflow para poblar la capa Bronze desde otro Cloud Storage.

• **Taller:** Pipeline Dataflow para poblar la capa Bronze desde una tabla SQL (incremental).

• **Taller:** Automatización de los pipelines basada en calendarización.

## 15. Ingesta Real-time (streaming) en Medallion con PubSub

• Fuentes streaming en ingeniería de datos.

• PubSub para manejo de streaming en GCP.

• Características del enfoque Productor-Receptor.

• ¿Databricks para ingesta streaming? ¿cuándo sí y cuando no?

• **Taller:** Ingesta streaming en el Datalake Medallion mediante PubSub.

## 16. Procesamiento y flujo de datos en Medallion con PySpark sobre Dataproc

• Introducción al procesamiento distribuido y al servicio Dataproc. Diferencias respecto a Databricks.

• **Taller:** PySpark sobre Dataproc para la implementación de ETLs para el Datalake (flujo entre capas).

• **Taller:** Orquestación y automatización de los ETL con Composer (Airflow).

## 17. Procesamiento y flujo de datos en Medallion con PySpark sobre Databricks

• **Taller:** Montaje del Datalake en el entorno de Google Databricks.

• **Taller:** PySpark en Databricks para transformaciones complejas a través de Medallion.

• **Taller:** Uso de Jobs Databricks para automatización de flujos.

## 18. Conversión a Lakehouse con BigQuery y buenas prácticas

• Formatos de tabla recomendados para cada capa de Medallion.

• Datalake (BigQuery) vs. Databricks como motores

independientes de datos. Consideraciones y casos de uso.

• Técnicas de optimización en Databricks. Particionado, Z-Ordering, auto-optimize, Vacuum en Delta Lake.

• **Taller:** Exposición del Datalake mediante BigQuery.

• **Taller:** Test de conexión desde Looker Studio y Power BI Desktop.

## Nuestra Propuesta de Capacitación

### Las metodologías que aplicamos



#### Desarrollo de competencias clave en el mundo de los datos

Analiza • Innova • Transforma



#### Aprendizaje Secuencial

- Descubre conocimiento de vanguardia
- Explora con la guía del experto
- Aplica lo aprendido



#### Aprendizaje basado en práctica (Learning by Doing)

- Resuelve retos
- Aprende en base a proyectos
- Analiza casos



## ¿Qué certificado obtendrás?

- Certificado por aprobación de la Especialización en GCP Data Engineering, por un total de 56 horas académicas.



## ¿Por qué elegirnos?

-  Somos los primeros en Perú en apostar por el desarrollo de profesionales y empresas en data & analytics con más de 16 años de experiencia.
-  Las empresas worldclass de Latam confían en nosotros para acompañarlas en su transformación hacia el enfoque data driven.
-  Nuestros docentes son destacados expertos en data & analytics que lideran equipos de alto rendimiento en las empresas más grandes de Latam.
-  Nuestra metodología "Aprende haciendo" ha logrado que nuestra comunidad de +25K profesionales en todo Latam mejoren su situación laboral.
-  Tenemos el portafolio más completo con +150 capacitaciones sincrónicas y asincrónicas que se ajustan a diferentes perfiles y niveles de conocimiento.





Visita nuestra web

[www.dmc.pe](http://www.dmc.pe)

Síguenos en:

