Curso Data Engineering on Google Cloud

Calle de la Basílica, 19
28020 Madrid
(34) 915 53 61 62
www.cas-training.com

WE ARE CAS

Duración
28 horas

Modalidad
Aula Virtual

Learning by doing

Curso Oficial

Acerca de:

Con el **curso Data Engineering on Google Cloud** obtendrás experiencia práctica con el diseño y la creación de sistemas de **procesamiento de datos en Google Cloud**. Este curso utiliza conferencias, demostraciones y laboratorios prácticos para mostrarte cómo diseñar sistemas de procesamiento de datos, crear canalizaciones de datos de un extremo a extremo, analizar datos e implementar el (machine learning). Este curso cubre datos estructurados, no estructurados y de transmisión.

Dirigido a:

Este curso está destinado a desarrolladores que sean responsables de:

- Extracción, carga, transformación, limpieza y validación de datos.
- Diseño de pipelines y arquitecturas para el procesamiento de datos.
- Integración de capacidades de análisis y machine learning en canalizaciones (pipelines) de datos.
- Consulta de conjuntos de datos, visualización de resultados de consultas y creación de informes.

Objetivos:

- Diseñar y crear sistemas de procesamiento de datos en Google Cloud.
- Procesar datos por lotes y de transmisión mediante la implementación de canalizaciones (pipelines) de datos de escalado automático en Dataflow.
- Obtener información empresarial a partir de conjuntos de datos extremadamente grandes con BigQuery.
- Aprovechar los datos no estructurados con las APIs de Spark y ML en Dataproc.
- Habilitar conocimientos instantáneos a partir de la transmisión de datos.
- Comprender las APIs de ML y BigQuery ML, y aprender a usar AutoML para crear modelos potentes sin codificación.

Requisitos:

- Haber completado el curso Google Cloud Big Data and Machine Learning Fundamentals o tener una experiencia equivalente.
- Tener competencia básica con un lenguaje de consulta común como SQL.
- Tener experiencia con actividades de modelado de datos y ETL (extracción, transformación, carga).
- Tener experiencia en el desarrollo de aplicaciones utilizando un lenguaje de programación común como Python.
- Estar familiarizado con el machine learning y/o estadísticas.

Material del curso:

Documentación oficial para el **curso Google Cloud Big Data and Machine Learning Fundamentals**.

Perfil del docente:

- Formador certificado por Google Cloud.
- Más de 5 años de experiencia profesional.
- Más de 4 años de experiencia docente.
- Profesional activo en empresas del sector IT.

Metodología:

- “Learning by doing” se centra en un contexto real y concreto, buscando un aprendizaje en equipo para la resolución de problemas en el sector empresarial.
- Aulas con grupos reducidos para que el profesional adquiera la mejor atención por parte de nuestros instructores profesionales.
- El programa de estudios como partners oficiales es confeccionado por nuestro equipo de formación y revisado por las marcas de referencia en el sector.
- La impartición de las clases podrá ser realizada tanto en modalidad Presencial como Virtual.

Contenidos:

Módulo 1: Introduction to Data Engineering Temas:

- Explore the role of a data engineer
- Analyze data engineering challenges
- Introduction to BigQuery
- Data lakes and data warehouses
- Transactional databases versus data warehouses
- Partner effectively with other data teams
- Manage data access and governance
- Build production-ready pipelines
- Review Google Cloud customer case study

Objetivos:

- Understand the role of a data engineer
- Discuss benefits of doing data engineering in the cloud
- Discuss challenges of data engineering practice and how building data pipelines in the cloud helps to address these
- Review and understand the purpose of a data lake versus a data warehouse, and when to use which

Módulo 2: Building a Data Lake Temas:

- Introduction to data lakes
- Data storage and ETL options on Google Cloud
- Building a data lake using Cloud Storage
- Securing Cloud Storage
- Storing all sorts of data types
- Cloud SQL as a relational data lake

Objetivos:

- Understand why Cloud Storage is a great option for building a data lake on Google Cloud
- Learn how to use Cloud SQL for a relational data lake

Módulo 3: Building a Data Warehouse Temas:

- The modern data warehouse
- Introduction to BigQuery
- Getting started with BigQuery
- Loading data
- Exploring schemas
- Schema design
- Nested and repeated fields
- Optimizing with partitioning and clustering

Objetivos:

- Discuss requirements of a modern warehouse
- Understand why BigQuery is the scalable data warehousing solution on Google Cloud
- Understand core concepts of BigQuery and review options of loading data into BigQuery

Módulo 4: Introduction to Building Batch Data Pipelines Temas:

- EL, ELT, ETL
- Quality considerations

- How to carry out operations in BigQuery
- Shortcomings
- ETL to solve data quality issues

Objetivos:

- Review different methods of loading data into your data lakes and warehouses: EL, ELT, and ETL
- Discuss data quality considerations and when to use ETL instead of EL and ELT

Módulo 5: Executing Spark on Dataproc Temas:

- The Hadoop ecosystem
- Run Hadoop on Dataproc
- Cloud Storage instead of HDFS
- Optimize Dataproc

Objetivos:

- Review the parts of the Hadoop ecosystem
- Learn how to lift and shift your existing Hadoop workloads to the cloud using Dataproc
- Understand considerations around using Cloud Storage instead of HDFS for storage
- Learn how to optimize Dataproc jobs

Módulo 6: Serverless Data Processing with Dataflow Temas:

- Introduction to Dataflow
- Why customers value Dataflow
- Dataflow pipelines
- Aggregating with GroupByKey and Combine
- Side inputs and windows
- Dataflow templates
- Dataflow SQL

Objetivos:

- Understand how to decide between Dataflow and Dataproc for processing data pipelines
- Understand the features that customers value in Dataflow
- Discuss core concepts in Dataflow
- Review the use of Dataflow templates and SQL

Módulo 7: Manage Data Pipelines with Cloud Data Fusion and Cloud Composer Temas:

- Building batch data pipelines visually with Cloud Data Fusion
- Components
- UI overview
- Building a pipeline
- Exploring data using Wrangler
- Orchestrating work between Google Cloud services with Cloud Composer
- Apache Airflow environment
- DAGs and operators
- Workflow scheduling
- Monitoring and logging

Objetivos:

- Discuss how to manage your data pipelines with Data Fusion and Cloud Composer
- Understand Data Fusion's visual design capabilities
- Learn how Cloud Composer can help to orchestrate the work across multiple Google Cloud services

Módulo 8: Introduction to Processing Streaming Data Temas: Process Streaming Data Objetivos:

- Explain streaming data processing
- Describe the challenges with streaming data
- Identify the Google Cloud products and tools that can help address streaming data challenges

Módulo 9: Serverless Messaging with Pub/Sub Temas:

- Introduction to Pub/Sub
- Pub/Sub push versus pull
- Publishing with Pub/Sub code

Objetivos:

- Describe the Pub/Sub service
- Understand how Pub/Sub works
- Gain hands-on Pub/Sub experience with a lab that simulates real-time streaming sensor data

Módulo 10: Dataflow Streaming Features Temas:

- Steaming data challenges
- Dataflow windowing

Objetivos:

- Understand the Dataflow service
- Build a stream processing pipeline for live traffic data
- Demonstrate how to handle late data using watermarks, triggers, and accumulation

Módulo 11: High-Throughput BigQuery and Bigtable Streaming Features Temas:

- Streaming into BigQuery and visualizing results
- High-throughput streaming with Cloud Bigtable
- Optimizing Cloud Bigtable performance

Objetivos:

- Learn how to perform ad hoc analysis on streaming data using BigQuery and dashboards
- Understand how Cloud Bigtable is a low-latency solution
- Describe how to architect for Bigtable and how to ingest data into Bigtable
- Highlight performance considerations for the relevant services

Módulo 12: Advanced BigQuery Functionality and Performance Temas:

- Analytic window functions
- Use With clauses
- GIS functions
- Performance considerations

Objetivos:

- Review some of BigQuery's advanced analysis capabilities
- Discuss ways to improve query performance

Módulo 13: Introduction to Analytics and AI Temas:

- What is AI?
- From ad-hoc data analysis to data-driven decisions
- Options for ML models on Google Cloud

Objetivos:

- Understand the proposition that ML adds value to your data
- Understand the relationship between ML, AI, and Deep Learning
- Identify ML options on Google Cloud

Módulo 14: Prebuilt ML Model APIs for Unstructured Data Temas:

- Unstructured data is hard

- ML APIs for enriching data

Objetivos:

- Discuss challenges when working with unstructured data
- Learn the applications of ready-to-use ML APIs on unstructured data

Módulo 15: Big Data Analytics with Notebooks Temas:

- What's a notebook?
- BigQuery magic and ties to Pandas

Objetivos:

- Introduce Notebooks as a tool for prototyping ML solutions Learn to execute BigQuery commands from Notebooks

Módulo 16: Production ML Pipelines with Kubeflow Temas:

- Ways to do ML on Google Cloud
- Vertex AI Pipelines
- AI Hub

Objetivos:

- Describe options available for building custom ML models
- Understand the use of tools like Vertex AI Pipelines

Módulo 17: Custom Model Building with SQL in BigQuery ML Temas:

- BigQuery ML for quick model building
- Supported models

Objetivos:

- Learn how to create ML models by using SQL syntax in BigQuery
- Demonstrate building different kinds of ML models using BigQuery ML

Módulo 18: Custom Model Building with AutoML Temas:

- Why AutoML?
- AutoML Vision
- AutoML NLP
- AutoML tables

Objetivos:

- Explore various AutoML products used in machine learning
- Learn to use AutoML to create powerful models without coding

CAS TRAINING

UN ESPACIO PARA CRECER

cas-training.com