Curso Serverless Data Processing with Dataflow

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

WE ARE CAS

Duración
21 horas

Modalidad
Aula Virtual

Learning by doing

Curso Oficial

Acerca de:

El **curso Serverless Data Processing with Dataflow** está destinado a profesionales de **Big Data** que deseen ampliar su comprensión de Dataflow para avanzar en sus aplicaciones de procesamiento de datos.

Comenzando con los fundamentos, este curso explica cómo Apache Beam y Dataflow funcionan juntos para satisfacer sus necesidades de procesamiento de datos sin el riesgo de depender de un proveedor. La sección sobre el desarrollo de canalizaciones (pipelines) cubre cómo convertir la lógica comercial en aplicaciones de procesamiento de datos que puedan ejecutarse en Dataflow. El curso culmina con un enfoque en las operaciones, que revisa las lecciones más importantes para operar una aplicación de datos en Dataflow, incluido el monitoreo, la resolución de problemas, las pruebas y la confiabilidad.

Dirigido a:

- Ingeniero de datos.
- Analistas de datos y científicos de datos que aspiren a desarrollar habilidades de ingeniería de datos.

Objetivos:

- Demostrar cómo Apache Beam y Dataflow funcionan juntos para satisfacer las necesidades de procesamiento de datos de tu organización.
- Resumir los beneficios de Beam Portability Framework y activarlo para sus canalizaciones de Dataflow.
- Habilitar Shuffle y Streaming Engine, para canalizaciones por lotes y de transmisión respectivamente, para obtener el máximo rendimiento.
- Habilitar Flexible Resource Scheduling para obtener un rendimiento más rentable.
- Seleccionar la combinación correcta de permisos de IAM para el trabajo de Dataflow.
- Implementar las mejores prácticas para un entorno de procesamiento de datos seguro.
- Seleccionar y ajustar I/O de tu elección para la canalización de Dataflow.
- Utilizar esquemas para simplificar tu código Beam y mejorar el rendimiento de su canalización.
- Desarrollar una canalización Beam utilizando SQL y DataFrames.
- Realizar monitoreo, resolución de problemas, pruebas y CI/CD en canalizaciones de Dataflow.

Requisitos:

- Haber completado **Building Batch Data Pipelines**.
- Haber completado **Building Resilient Streaming Analytics Systems**.

Material del curso:

Documentación oficial para el **curso Serverless Data Processing with Dataflow**.

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 Temas:

- Course Introduction
- Beam and Dataflow Refresher

Objetivos:

- Introduce the course objectives.
- Demonstrate how Apache Beam and Dataflow work together to fulfill your organization's data processing needs.

Módulo 2: Beam Portability Temas:

- Beam Portability
- Runner v2
- Container Environments
- Cross-Language Transforms

Objetivos:

- Summarize the benefits of the Beam Portability Framework.
- Customize the data processing environment of your pipeline using custom containers.
- Review use cases for cross-language transformations.
- Enable the Portability framework for your Dataflow pipelines.

Módulo 3: Separating Compute and Storage with Dataflow Temas:

- Dataflow
- Dataflow Shuffle Service
- Dataflow Streaming Engine
- Flexible Resource Scheduling

Objetivos:

- Enable Shuffle and Streaming Engine, for batch and streaming pipelines respectively, for maximum performance.
- Enable Flexible Resource Scheduling for more cost-efficient performance.

Módulo 4: IAM, Quotas, and Permissions Temas:

- IAM
- Quota

Objetivos:

- Select the right combination of IAM permissions for your Dataflow job.
- Determine your capacity needs by inspecting the relevant quotas for your Dataflow jobs.

Módulo 5: Security Temas:

- Data Locality
- Shared VPC
- Private IPs
- CMEK

Objetivos:

- Select your zonal data processing strategy using Dataflow, depending on your data locality needs.
- Implement best practices for a secure data processing environment.

Módulo 6: Beam Concepts Review Temas:

- Beam Basics

- Utility Transforms
- DoFn Lifecycle

Objetivos: Review main Apache Beam concepts (Pipeline, PCollections, PTransforms, Runner, reading/writing, Utility PTransforms, side inputs), bundles and DoFn Lifecycle. **Módulo 7: Windows, Watermarks, Triggers**Temas:

- Windows
- Watermarks
- Triggers

Objetivos:

- Implement logic to handle your late data.
- Review different types of triggers.
- Review core streaming concepts (unbounded PCollections, windows).

Módulo 8: Sources and Sinks Temas:

- Sources and Sinks
- Text IO and File IO
- BigQuery IO
- PubSub IO
- Kafka IO
- Bigable IO
- Avro IO
- Splittable DoFn

Objetivos:

- Write the I/O of your choice for your Dataflow pipeline.
- Tune your source/sink transformation for maximum performance.
- Create custom sources and sinks using SDF.

Módulo 9: Schemas Temas:

- Beam Schemas
- Code Examples

Objetivos:

- Introduce schemas, which give developers a way to express structured data in their Beam pipelines.
- Use schemas to simplify your Beam code and improve the performance of your pipeline.

Módulo 10: State and Timers Temas:

- State API
- Timer API
- Summary

Objetivos:

- Identify use cases for state and timer API implementations.
- Select the right type of state and timers for your pipeline.

Módulo 11: Best Practices Temas:

- Schemas
- Handling unprocessable Data
- Error Handling
- AutoValue Code Generator
- JSON Data Handling
- Utilize DoFn Lifecycle
- Pipeline Optimizations

Objetivos: Implement best practices for Dataflow pipelines. **Módulo 12: Dataflow SQL and DataFrames**Temas:

- Dataflow and Beam SQL
- Windowing in SQL
- Beam DataFrames

Objetivos: Develop a Beam pipeline using SQL and DataFrames. **Módulo 13: Beam Notebooks**Temas: Beam Notebooks Objetivos:

- Prototype your pipeline in Python using Beam notebooks.
- Launch a job to Dataflow from a notebook.

Módulo 14: Monitoring Temas:

- Job List
- Job Info
- Job Graph
- Job Metrics
- Metrics Explorer

Objetivos:

- Navigate the Dataflow Job Details UI.
- Interpret Job Metrics charts to diagnose pipeline regressions.
- Set alerts on Dataflow jobs using Cloud Monitoring.

Módulo 15: Logging and Error Reporting Temas:

- Logging
- Error Reporting

Objetivos: Use the Dataflow logs and diagnostics widgets to troubleshoot pipeline issues. **Módulo 16: Troubleshooting and Debug**Temas:

- Troubleshooting Workflow
- Types of Troubles

Objetivos:

- Use a structured approach to debug your Dataflow pipelines.
- Examine common causes for pipeline failures.

Módulo 17: Performance Temas:

- Pipeline Design
- Data Shape
- Source, Sinks, and External Systems
- Shuffle and Streaming Engine

Objetivos:

- Understand performance considerations for pipelines.
- Consider how the shape of your data can affect pipeline performance.

Módulo 18: Testing and CI/CD Temas:

- Testing and CI/CD Overview
- Unit Testing
- Integration Testing
- Artifact Building
- Deployment

Objetivos:

- Testing approaches for your Dataflow pipeline.
- Review frameworks and features available to streamline your CI/CD workflow for Dataflow pipelines.

Módulo 19: Reliability Temas:

- Introduction to Reliability
- Monitoring
- Geolocation
- Disaster Recovery
- High Availability

Objetivos: Implement reliability best practices for your Dataflow pipelines. **Módulo 20: Flex Templates**Temas:

- Classic Templates
- Flex Templates
- Using Flex Templates
- Google-provided Templates

Objetivos: Using flex templates to standardize and reuse Dataflow pipeline code. **Módulo 21: Summary**Temas:

Summary Objetivos: Quick recap of training topics

CAS TRAINING

UN ESPACIO PARA CRECER

cas-training.com