Serverless Data Processing with Dataflow (SDPF)

ID GO-SDPF **Duración** 3 días

Quién debería asistir

- Data engineer.
- Data analysts and data scientists aspiring to develop data engineering skills

Prerrequisitos

To get the most out of this course, participants should have completed the following courses:

- Building Batch Data Pipelines
- Building Resilient Streaming Analytics Systems

Objetivos del curso

- Demonstrate how Apache Beam and Dataflow work together to fulfill your organization's data processing needs.
- Summarize the benefits of the Beam Portability Framework and enable it for your Dataflow pipelines.
- Enable Shuffle and Streaming Engine, for batch and streaming pipelines respectively, for maximum performance.
- Enable Flexible Resource Scheduling for more cost-efficient performance.
- Select the right combination of IAM permissions for your Dataflow job.
- Implement best practices for a secure data processing environment.
- Select and tune the I/O of your choice for your Dataflow pipeline.
- Use schemas to simplify your Beam code and improve the performance of your pipeline.
- Develop a Beam pipeline using SQL and DataFrames.
- Perform monitoring, troubleshooting, testing and CI/CD on Dataflow pipelines.

Esquema Detallado del Curso

Module 1: Introduction

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

Module 2: Beam Portability

- 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.

Module 3: Separating Compute and Storage with Dataflow

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

Module 4: IAM, Quotas, and Permissions

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

Module 5: Security

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

Module 6: Beam Concepts Review

- Review main Apache Beam concepts (Pipeline, PCollections, PTransforms, Runner, reading/writing, Utility PTransforms, side inputs), bundles and DoFn Lifecycle.

Module 7: Windows, Watermarks, Triggers

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

Module 8: Sources and Sinks

- 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.

Module 9: Schemas

- 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.

Module 10: State and Timers

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

Module 11: Best Practices

- Implement best practices for Dataflow pipelines.

Module 12: Dataflow SQL and DataFrames

- Develop a Beam pipeline using SQL and DataFrames.

Module 13: Beam Notebooks

- Prototype your pipeline in Python using Beam notebooks.
- Use Beam magics to control the behavior of source recording in your notebook.
- Launch a job to Dataflow from a notebook.

Module 14: Monitoring

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

Module 15: Logging and Error Reporting

- Use the Dataflow logs and diagnostics widgets to troubleshoot pipeline issues.

Module 16: Troubleshooting and Debug

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

Module 17: Performance

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

Module 18: Testing and CI/CD

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

Module 19: Reliability

- Implement reliability best practices for your Dataflow pipelines.

Module 20: Flex Templates

- Using flex templates to standardize and reuse Dataflow pipeline code.

Module 21: Summary

- Summary.

Centros de Entrenamiento Mundial