

Serverless Data Processing with Dataflow

This training is intended for big data practitioners who want to further their understanding of Dataflow in order to advance their data processing applications. Beginning with foundations, this training explains how Apache Beam and Dataflow work together to meet your data processing needs without the risk of vendor lock-in. The section on developing pipelines covers how you convert your business logic into data processing applications that can run on Dataflow. This training culminates with a focus on operations, which reviews the most important lessons for operating a data application on Dataflow, including monitoring, troubleshooting, testing, and reliability.

**DURATION**

3 days

**LEVEL**

300 - Advanced

**FORMAT**

Instructor-led, On-demand

What you'll learn

- 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.

Overview	21 Modules · 21 Labs · 81 Videos · 18 Quizzes
Who this course is for	• Data Engineer, Data Analysts and Data Scientists aspiring to develop Data Engineering skills
Products	• Dataflow Prime • Google Cloud Observability • Dataflow • Cloud Operations
Prerequisites	• Completed “Building Batch Data Pipelines” • Completed “Building Resilient Streaming Analytics Systems”

Module 01

Introduction

Topics	• Course Introduction • Beam and Dataflow Refresher
Objectives	• Introduce the course objectives. • Demonstrate how Apache Beam and Dataflow work together to fulfill your organization’s data processing needs.

Module 02

Separating Compute and Storage with Dataflow

Topics	• Dataflow
--------	--

Module 03

Beam Portability

Topics	• Beam Portability • Runner v2 • Container Environments • Cross-Language TransformS
Objectives	• 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.
Activities	• Quiz

Module 04

Dataflow Shuffle Service & Streaming Engine

Topics	• Dataflow Shuffle Service • Dataflow Streaming Engine • Flexible Resource Scheduling
Objectives	• Enable Shuffle and Streaming Engine, for batch and streaming pipelines respectively, for maximum performance. • Enable Flexible Resource Scheduling for more cost-efficient performance.
Activities	• Quiz

Module 05

IAM, Quotas, and Permissions

Topics	• IAM • Quota
Objectives	• Select the right combination of IAM permissions for your Dataflow job. • Determine your capacity needs by inspecting the relevant quotas for your Dataflow jobs.
Activities	• Quiz

Module 06

Security

Topics	• Data Locality • Shared VPC • Private IPs • CMEK
Objectives	• Select your zonal data processing strategy using Dataflow, depending on your data locality needs. • Implement best practices for a secure data processing environment.
Activities	• Hands-on lab and quiz

Module 07

Beam Concepts Review

Topics	• Beam Basics • Utility Transforms • DoFn Lifecycle
Objectives	• Review main Apache Beam concepts (Pipeline, PCollections, PTransforms, Runner, reading/writing, Utility PTransforms, side inputs), bundles and DoFn Lifecycle.
Activities	• Hands-on lab and quiz

Module 08

Windows, Watermarks, Triggers

Topics	• Windows • Watermarks • Triggers
Objectives	• Implement logic to handle your late data. • Review different types of triggers. • Review core streaming concepts (unbounded PCollections, windows).
Activities	• Hands-on lab and quiz

Module 09

Sources and Sinks

Topics

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

Objectives

- 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.

Activities

- Quiz
-

Module 10

Schemas

Topics

- Beam Schemas
- Code Examples

Objectives

- 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.

Activities

- Hands-on lab and quiz
-

Module 11 State and Timers

Topics	• State API • Timer API • Summary
Objectives	• Identify use cases for state and timer API implementations. • Select the right type of state and timers for your pipeline.
Activities	• Quiz

Module 12 Best Practices

Topics	• Schemas • Handling unprocessable Data • Error Handling • AutoValue Code Generator • JSON Data Handling • Utilize DoFn Lifecycle • Pipeline Optimizations
Objectives	• Implement best practices for Dataflow pipelines.
Activities	• Hands-on lab and quiz

Module 13 Dataflow SQL and DataFrames

Topics	• Dataflow and Beam SQL • Windowing in SQL • Beam DataFrames
Objectives	• Develop a Beam pipeline using SQL and DataFrames.
Activities	• Hands-on lab and quiz

Module 14

Beam Notebooks

Topics	• Beam Notebooks
Objectives	• Prototype your pipeline in Python using Beam notebooks. • Launch a job to Dataflow from a notebook.
Activities	• Quiz

Module 15

Monitoring

Topics	• Job List • Job Info • Job Graph • Job Metrics • Metrics Explorer
Objectives	• Navigate the Dataflow Job Details UI. • Interpret Job Metrics charts to diagnose pipeline regressions. • Set alerts on Dataflow jobs using Cloud Monitoring.
Activities	• Quiz

Module 16

Logging and Error Reporting

Topics	• Logging • Error Reporting
Objectives	• Use the Dataflow logs and diagnostics widgets to troubleshoot pipeline issues.
Activities	• Quiz

Module 17

Troubleshooting and Debug

Topics	• Troubleshooting Workflow • Types of Troubles
Objectives	• Use a structured approach to debug your Dataflow pipelines. • Examine common causes for pipeline failures.
Activities	• Hands-on lab and quiz

Module 18

Performance

Topics	• Pipeline Design • Data Shape • Source, Sinks, and External Systems • Shuffle and Streaming Engine
Objectives	• Understand performance considerations for pipelines. • Consider how the shape of your data can affect pipeline performance.
Activities	• Quiz

Module 19

Testing and CI/CD

Topics	• Testing and CI/CD Overview • Unit Testing • Integration Testing • Artifact Building • Deployment
Objectives	• Testing approaches for your Dataflow pipeline. • Review frameworks and features available to streamline your CI/CD workflow for Dataflow pipelines.
Activities	• Hands-on labs and quiz

Module 20

Reliability

Topics

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

Objectives

- Implement reliability best practices for your Dataflow pipelines.

Activities

- Quiz
-

Module 21

Flex Templates

Topics

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

Objectives

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

Activities

- Hands-on labs and quiz
-

Module 22

Summary

Topics

- Quick recap of training topics

Objectives

- Synthesize core Apache Beam programming concepts across batch and streaming data pipelines
- Formulate production data processing architectures leveraging Cloud Dataflow managed capabilities