Training: Google Cloud
Serverless Data Processing with Dataflow

Google Cloud

TRAINING GOALS:

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.

Objectives:

- 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.

Audience:

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

CONSPECT:

- Introduction
  - Introduce the course objectives.

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

    - Demonstrate how Apache Beam and Dataflow work together to fulfill your organization's data processing needs.
- 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.
- 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.
- 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.
- Security
    - Select your zonal data processing strategy using Dataflow, depending on your data locality needs.
    - Implement best practices for a secure data processing environment.
- Beam Concepts Review
    - Review main Apache Beam concepts (Pipeline, PCollections, PTransforms, Runner, reading/writing, Utility PTransforms, side inputs), bundles and DoFn
- Windows, Watermarks, Triggers
    - Implement logic to handle your late data.
    - Review different types of triggers.
    - Review core streaming concepts (unbounded PCollections, windows).
- 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.
- 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.
- State and Timers
    - Identify use cases for state and timer API implementations.
    - Select the right type of state and timers for your pipeline.
- Best Practices

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

- Implement best practices for Dataflow pipelines.
- Dataflow SQL and DataFrames
  - Develop a Beam pipeline using SQL and DataFrames.
- 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.
- Monitoring
  - Navigate the Dataflow Job Details UI.
  - Interpret Job Metrics charts to diagnose pipeline regressions.
  - Set alerts on Dataflow jobs using Cloud Monitoring.
- Logging and Error Reporting
  - Use the Dataflow logs and diagnostics widgets to troubleshoot pipeline issues.
- Troubleshooting and Debug
  - Use a structured approach to debug your Dataflow pipelines.
  - Examine common causes for pipeline failures.
- Performance
  - Understand performance considerations for pipelines.
  - Consider how the shape of your data can affect pipeline performance.
- Testing and CI/CD
  - Testing approaches for your Dataflow pipeline.
  - Review frameworks and features available to streamline your CI/CD workflow for Dataflow pipelines.
- Reliability
  - Implement reliability best practices for your Dataflow pipelines.
- Flex Templates
  - Using flex templates to standardize and reuse Dataflow pipeline code.
- Summary
  - Summary

REQUIREMENTS:

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

- “Building Batch Data Pipelines”
- “Building Resilient Streaming Analytics Systems”

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70
BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007

Difficulty level



CERTIFICATE:

The participants will obtain certificates signed by Google Cloud Platform

TRAINER:

Authorized Google Cloud Platform Trainer

Compendium Centrum Edukacyjne Spółka z o.o.
ul. Tatarska 5, 30-103 Kraków, tel.: (12) 298 47 77
ul. Postępu 18B, 02-676 Warszawa, tel.: (22) 417 41 70

BNP Paribas Bank Polska S.A.
EUR: PL 79 1600 1462 1853 6451 3000 0006
USD: PL 52 1600 1462 1853 6451 3000 0007