Serverless Data Processing with Dataflow

Course 4341 – 24 Hours

Overview

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.

On Completion, Delegates will be able to

- 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.

Who Should Attend

- Data Engineer
- Data Analysts and Data Scientists aspiring to develop Data Engineering skills

Prerequisites

- Completed "Building Batch Data Pipelines"
- Completed "Building Resilient Streaming Analytics Systems"

Course Contents

- Introduction
  - Course Introduction
  - Beam and Dataflow Refresher
- Beam Portability
  - Beam Portability
  - Runner v2
  - Container Environments
  - Cross-Language TransformS
- Separating Compute and Storage with Dataflow
  - Dataflow
  - Dataflow Shuffle Service
  - Dataflow Streaming Engine
  - Flexible Resource Scheduling
- IAM, Quotas, and Permissions
  - IAM
  - Quota
- Security
  - Data Locality
  - Shared VPC
  - Private IPs
  - CMEK
- Beam Concepts Review
  - Beam Basics
  - Utility Transforms
  - DoFn Lifecycle
- Windows, Watermarks, Triggers
- Sources and Sinks
  - Sources and Sinks
  - Text IO and File IO
  - BigQuery IO
  - PubSub IO
  - Kafka IO
  - Bigable IO
  - Avro IO
  - Splittable DoFn
- Schemas
  - Beam Schemas
  - Code Examples
- State and Timers
  - State API
  - Timer API
  - Summary
- Best Practices
  - Schemas
  - Handling unprocessable Data
  - Error Handling

    - AutoValue Code Generator
    - JSON Data Handling
    - Utilize DoFn Lifecycle
    - Pipeline Optimizations
- Dataflow SQL and DataFrames
    - Dataflow and Beam SQL
    - Windowing in SQL
    - Beam DataFrames
- Beam Notebooks
- Monitoring
    - Job List
    - Job Info
    - Job Graph
    - Job Metrics
    - Metrics Explorer
- Logging and Error Reporting
- Troubleshooting and Debug
    - Troubleshooting Workflow
    - Types of Troubles
- Performance
    - Pipeline Design
    - Data Shape
    - Source, Sinks, and External Systems
    - Shuffle and Streaming Engine

- Testing and CI/CD
    - Testing and CI/CD Overview
    - Unit Testing
    - Integration Testing
    - Artifact Building
    - Deployment
- Reliability
    - Introduction to Reliability
    - Monitoring
    - Geolocation
    - Disaster Recovery
    - High Availability
- Flex Templates
    - Classic Templates
    - Flex Templates
    - Using Flex Templates
    - Google-provided Templates
- Summary