TRENDS | Academy
Learning | Evolving | Empowering

Serverless Data Processing with Dataflow

Duration: 3 Days

Prerequisites:

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

Course Description:

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.

Target Audience:

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

Course Outlines:

Module 1: Introduction
- Course Introduction
- Beam and Dataflow Refresher

Module 2: Beam Portability
- Beam Portability
- Runner v2
- Container Environments
- Cross-Language Transforms

Module 3: Separating Compute and Storage with Dataflow
- Dataflow
- Dataflow Shuffle Service
- Dataflow Streaming Engine
- Flexible Resource Scheduling

Module 4: IAM, Quotas, and Permissions
- IAM
- Quota

Module 5: Security
- Data Locality
- Shared VPC
- Private IPs
- CMEK

Module 6: Beam Concepts Review
- Beam Basics
- Utility Transforms
- DoFn Lifecycle

Module 7: Windows, Watermarks, Triggers
- Windows
- Watermarks
- Triggers

Module 8: Sources and Sinks
- Sources and Sinks
- Text IO and File IO
- BigQuery IO
- PubSub IO
- Kafka IO
- Bigable IO
- Avro IO
- Splittable DoFn

Module 9: Schemas
- Beam Schemas
- Code Examples

Module 10: State and Timers
- State API
- Timer API
- Summary

Module 11: Best Practices
- Schemas
- Handling unprocessable Data
- Error Handling
- AutoValue Code Generator
- JSON Data Handling
- Utilize DoFn Lifecycle
- Pipeline Optimizations

Module 12: Dataflow SQL and DataFrames
- Dataflow and Beam SQL
- Windowing in SQL
- Beam DataFrames

Module 13: Beam Notebooks
- Beam Notebooks

Module 14: Monitoring
- Job List
- Job Info
- Job Graph
- Job Metrics
- Metrics Explorer

Module 15: Logging and Error Reporting
- Logging
- Error Reporting

Module 16: Troubleshooting and Debug
- Troubleshooting Workflow
- Types of Troubles

Module 17: Performance
- Pipeline Design
- Data Shape
- Source, Sinks, and External Systems
- Shuffle and Streaming Engine

Module 18: Testing and CI/CD
- Testing and CI/CD Overview
- Unit Testing
- Integration Testing
- Artifact Building
- Deployment

Module19: Reliability
- Introduction to Reliability
- Monitoring
- Geolocation
- Disaster Recovery
- High Availability

Module 20: Flex Templates
- Classic Templates
- Flex Templates
- Using Flex Templates
- Google-provided Templates

Module 20: Summary
- Summary

REGISTER NOW!
training@trends.com.ph
(+632) 8863-2123
www.trendsacademy.com.ph