Data Integration with Cloud Data Fusion

Duration: 2 Days

Prerequisites:

To get the most out of this course, participants are encouraged to have:
• Completed "Big Data and Machine Learning Fundamentals"

Course Description:

This 2-day course introduces learners to Google Cloud's data integration capability
using Cloud Data Fusion. In this course, we discuss challenges with data integration
and the need for a data integration platform (middleware). We then discuss how
Cloud Data Fusion can help to effectively integrate data from a variety of sources
and formats and generate insights. We take a look at Cloud Data Fusion's main
components and how they work, how to process batch data and real time streaming
data with visual pipeline design, rich tracking of metadata and data lineage, and how to deploy data pipelines on various execution engines.

Target Audience:

This course is primarily intended for the following participants:
• Data Engineer
• Data Analysts

Course Outlines:

Module 1: Introduction to data integration and Cloud Data Fusion
- Data integration: what, why, challenges
- Data integration tools used in industry
- User personas
- Introduction to Cloud Data Fusion
- Data integration critical capabilities
- Cloud Data Fusion UI components

Module 2: Building Pipelines
- Cloud Data Fusion architecture
- Core concepts
- Data pipelines and directed acyclic graphs (DAG)
- Pipeline Lifecycle
- Designing pipelines in Pipeline Studio

Module 3: Designing Complex Pipelines
- Branching, Merging and Joining
- Actions and Notifications
- Error handling and Macros
- Pipeline Configurations, Scheduling, Import and Export

Modul 4: Pipeline Execution Environment
- Schedules and triggers
- Execution environment: Compute profile and provisioners
- Monitoring pipelines

Module 5: Building Transformations and Preparing Data with Wrangler
- Wrangler
- Directives
- User-defined directives

Module 6: Connectors and streaming pipelines
- Understand the data integration architecture.
- List various connectors.
- Use the Cloud Data Loss Prevention (DLP) API.
- Understand the reference architecture of streaming pipelines.
- Build and execute a streaming pipeline

Module 7: Metadata and data lineage
- Metadata
- Data lineage

Module 8: Summary
- Course Summary