Google Cloud

Data Integration with Cloud Data Fusion

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.

DURATION
2 days

LEVEL
Intermediate

FORMAT
ILT

What you'll learn

- Identify the need of data integration,
- Understand the capabilities Cloud Data Fusion provides as a data integration platform,
- Identify use cases for possible implementation with Cloud Data Fusion,
- List the core components of Cloud Data Fusion,
- Design and execute batch and real time data processing pipelines,
- Work with Wrangler to build data transformations
- Use connectors to integrate data from various sources and formats,
- Configure execution environment; Monitor and Troubleshoot pipeline execution,
- Understand the relationship between metadata and data lineage

Overview 9 Modules · 8 Labs

Who this course is for This course is primarily intended for the following participants:

- Data Engineer
- Data Analysts

Products Cloud Data Fusion

Prerequisite To get the most out of this course, participants are encouraged to have:

- Completed “Introduction to Data Engineering”

Module 00 Introduction

Topics Course Introduction

Objectives Introduce the course objectives

Activities —

Module 01 Introduction to data integration and Cloud Data Fusion

Topics

- 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

Objectives

- Understand the need for data integration,
- List the situations/cases where data integration can help businesses,
- List the available data integration platforms and tools,
- Identify the challenges with data integration
- Understand the use of Cloud Data Fusion as a data integration platform
- Create a Cloud Data Fusion instance,
- Familiarize with core framework and major components in Cloud Data Fusion

Activities Graded lab, quiz, discussion activity

Module 02 Building pipelines

Topics

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

Objectives

- Understand Cloud Data Fusion architecture
- Define what a data pipeline is
- Understand the DAG representation of a data pipeline,
- Learn to use Pipeline Studio and its components
- Design a simple pipeline using Pipeline Studio,
- Deploy and execute a pipeline

Activities Graded lab and quiz

Module 03 Designing complex pipelines

Topics

- Branching, Merging and Joining
- Actions and Notifications
- Error handling and Macros
- Pipeline Configurations, Scheduling, Import and Export

Objectives

- Perform branching, merging, and join operations.
- Execute pipeline with runtime arguments using macros.
- Work with error handlers.
- Execute pre- and post-pipeline executions with help of actions and notifications.
- Schedule pipelines for execution.
- Import and export existing pipelines.

Activities Graded labs and quiz

Module 04 Pipeline execution environment

Topics

- Schedules and triggers
- Execution environment: Compute profile and provisioners
- Monitoring pipelines

Objectives

- Understand the composition of an execution environment.
- Configure your pipeline's execution environment, logging, and metrics. Understand concepts like compute profile and provisioner.

Objectives

- Create a compute profile.
- Create pipeline alerts.
- Monitor the pipeline under execution.

Activities Quiz

Module 05 Building Transformations and Preparing Data with Wrangler

Topics

- Wrangler
- Directives
- User-defined directives

Objectives

- Understand the use of Wrangler and its main components.
- Transform data using Wrangler UI.
- Transform data using directives/CLI methods.
- Create and use user-defined directives.

Activities Graded lab and quiz

Module 06 Connectors and streaming pipelines

Topics

- 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.

Objectives

- Connectors
- DLP
- Reference architecture for streaming applications
- Building streaming pipelines

Activities Graded lab, quiz, discussion activity

Module 07 Metadata and data lineage

Topics

- Metadata
- Data lineage

Objectives

- List types of metadata.
- Differentiate between business, technical, and operational metadata.
- Understand what data lineage is.

Objectives	• Understand the importance of maintaining data lineage. • Differentiate between metadata and data lineage.
Activities	Graded lab and quiz

Module 08 Summary

Topics	Course Summary
Objectives	Review the course objectives & concepts
Activities	—