Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: +1 434-509-6890 or Email: sales@cloudcontraptions.com

Practical Apache Spark for Data Pipelines

Class Duration

21 hours of live training delivered over 3-5 days.

Student Prerequisites

- Familiarity with Python (PySpark)
- Basic data processing knowledge
- Access to a GCP account with Dataproc serverless configured (provided if needed)

Target Audience

Data engineers, Python developers, and data professionals who need to develop, manage, and optimize batch and streaming data pipelines with Apache Spark 4 on GCP Dataproc serverless. Teams who also need to schedule and orchestrate these pipelines can continue to Apache Airflow Programming: Developing, Configuring, and Automating Workflows.

Description

The course equips participants with practical skills to develop, manage, and optimize Apache Spark 4 pipelines on GCP Dataproc serverless. By the end, attendees will understand Spark batch and streaming use-cases, master its execution model, Spark Connect architecture, and core data structures - including the VARIANT type for semi-structured data - and build reusable, performance-tuned pipelines for diverse data workloads.

Learning Outcomes

- Understand use-cases and benefits of Spark Batch and Structured Streaming.
- Gain working knowledge of Spark’s execution model and the Spark Connect client-server architecture.
- Develop reusable PySpark code for batch and streaming contexts.
- Build and optimize Spark pipelines on GCP Dataproc serverless.

Copyright 2026 Cloud Contraptions LLC – All Rights Reserved

- Master core data structures - DataFrames, ANSI-mode Spark SQL, and the VARIANT type - plus operations and performance tuning.

Training Materials

Comprehensive courseware is distributed online at the start of class. All students receive a downloadable MP4 recording of the training.

Software Requirements

Students will need access to a GCP account with Dataproc serverless configured. If students are unable to configure access, cloud environment can be provided.

Training Topics

Spark Overview

- Introduction to Apache Spark and its ecosystem
- What's New in Spark 4: ANSI mode, VARIANT, Spark Connect
- Spark Fundamentals Overview
- Pipeline Development Overview
- Advanced Spark and Optimization Overview

Spark Architecture and Use-Cases

- Spark topology: driver, cluster manager, worker nodes, executors
- Spark Connect client-server architecture
- Use-cases for Batch and Structured Streaming
- Spark's role in data engineering

Core Data Structures

- DataFrames and Spark SQL basics (ANSI mode by default)
- VARIANT type for semi-structured data
- Datasets and RDDs as legacy context
- Core operations: filtering, aggregations, joins

Spark Execution Model

- Partitioning
- Lazy Execution
- Fault Tolerance
- Checkpointing

Copyright 2026 Cloud Contraptions LLC – All Rights Reserved

- Serialization

Batch and Streaming Pipelines

- Designing Batch Pipelines
- Structured Streaming Fundamentals
- Stateful Processing with transformWithState
- Python Data Source API for Custom Connectors
- Building Reusable Code Components

Advanced Features

- Broadcast Variables
- Accumulators
- Serialization Challenges

Performance Tuning

- Resource management: memory, CPU, partitioning
- Adaptive Query Execution (AQE)
- Optimization: caching, shuffle reduction

Case Study and Wrap-up

- Discuss real-world Spark applications
- Review takeaways

Copyright 2026 Cloud Contraptions LLC – All Rights Reserved