Cloud Contraptions
TRAINING & CONSULTING

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

Practical Apache Spark for Data Pipelines

Duration

21 hours

Target Audience

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

Executive Summary

This three-day course equips participants with practical skills to develop, manage, and optimize Apache Spark pipelines on GCP Dataproc serverless through targeted lectures, hands-on labs, and a capstone project. By the end, attendees will understand Spark batch and streaming use-cases, master its execution model and core data structures, and build reusable, performance-tuned pipelines for diverse data workloads.

Description

The course equips participants with practical skills to develop, manage, and optimize Apache Spark pipelines on GCP Dataproc serverless. Through targeted lectures, hands-on labs, and a capstone project, attendees will master Spark's architecture, data structures, pipeline development, and tuning to maintain and expand DPP data pipelines, create reusable code, and address batch and streaming contexts.

Objectives

- Understand use-cases and benefits of Spark Batch and Structured Streaming.
- Gain working knowledge of Spark's execution model to support pipelines.

Copyright 2025 Cloud Contraptions LLC – All Rights Reserved

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

- Develop reusable code for batch and streaming contexts.
- Build and optimize Spark pipelines on GCP Dataproc serverless.
- Master core data structures, operations, and performance tuning.

Duration

21 hours of intensive training with live instruction delivered over three to five days to accommodate varied scheduling needs.

Training Materials

Students receive comprehensive courseware, including slides, code samples, and lab guides with pre-configured datasets.

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
- Spark Fundamentals Overview
- Pipeline Development Overview
- Advanced Spark and Optimization Overview

Spark Architecture and Use-Cases

- Spark topology: master, driver, worker nodes, executors
- Use-cases for Batch and Structured Streaming
- Spark’s role in data engineering

Core Data Structures

- DataFrames and Spark SQL basics
- Overview of Datasets and RDDs
- Core operations: filtering, aggregations, joins

Copyright 2025 Cloud Contraptions LLC – All Rights Reserved

To discuss this course and customizations:
Call: 434-509-5680 or Email: sales@cloudcontraptions.com

Hands-On: DataFrame Processing

- Load a CSV dataset into a DataFrame
- Apply transformations
- Query with Spark SQL

Spark Execution Model

- Partitioning
- Lazy Execution
- Fault Tolerance
- Checkpointing
- Serialization

Batch and Streaming Pipelines

- Designing Batch Pipelines
- Structured Streaming Fundamentals
- Building Reusable Code Components

Hands-On: Batch & Streaming Pipelines

- Create a batch pipeline for a log dataset, including a reusable data cleaning function
- Build a streaming pipeline for a simulated real-time dataset (e.g., sensor data)

Advanced Features

- Broadcast Variables
- Accumulators
- Serialization Challenges

Performance Tuning

- Resource management: memory, CPU, partitioning
- Optimization: caching, shuffle reduction

Pipeline Optimization Capstone

- Optimize a batch or streaming pipeline
- Utilize reusable code components

Case Study and Wrap-up

- Discuss real-world Spark applications
- Review takeaways

Copyright 2025 Cloud Contraptions LLC – All Rights Reserved