

KOENIG SOLUTIONS • COURSE OUTLINE

Hadoop Developer with Spark

Learn to store, process and analyse data across Hadoop and Spark, from HDFS and Hive to Spark SQL, MLlib and Streaming.

40 Hours

DURATION

21

MODULES

76

TOPICS

18

HANDS-ON EXERCISES

COURSE OVERVIEW

This course covers the Apache Hadoop ecosystem, including HDFS file storage and distributed processing with YARN, before moving into Apache Spark fundamentals such as the Spark shell, DataFrames and Datasets. It addresses querying and transforming data with DataFrame queries, Apache Hive, Hive QL and Spark SQL, alongside working with RDDs, pair RDDs and distributed processing concepts such as partitioning, shuffle, skew and persistence. It also covers writing, configuring and deploying Spark applications, building machine learning models with Spark ML and MLlib, and processing live data streams with Spark Streaming using sources such as Flume and Kafka.

WHO THIS COURSE IS FOR

Hadoop Developers

Spark Developers

Data Engineers

ETL Developers

Data Analysts

Big Data Architects

PREREQUISITES

What to Expect: This course is designed for developers and engineers who have programming experience, but prior knowledge of Hadoop and/or Spark is not required.

- Apache Spark examples and hands-on exercises are presented in Scala and Python. The ability to program in one of those languages is required.
- Basic familiarity with the Linux command line is assumed
- Basic knowledge of SQL is helpful

LEARNING OBJECTIVES

- Distribute, store, and process data in a Hadoop cluster
- Write, configure, and deploy Spark applications on a cluster
- Use the Spark shell for interactive data analysis
- Process and query structured data using Spark SQL and Hive Query Language
- Understand a wide variety of learning algorithms and Build an end-to-end Machine Learning Model with MLlib in pySpark
- Use Spark Streaming to process a live data stream

21 Modules, 76 Topics

01

Introduction to Apache Hadoop and the Hadoop Ecosystem

TOPICS

Apache Hadoop Overview · Data Ingestion and Storage · Data Processing · Data Analysis and Exploration · Other Ecosystem Tools · Introduction to the Hands-On Exercises

02

Apache Hadoop File Storage

TOPICS

Apache Hadoop Cluster Components · HDFS Architecture · Using HDFS

03

Distributed Processing on an Apache Hadoop Cluster

TOPICS

YARN Architecture · Working With YARN

04

Apache Spark Basics

TOPICS

What is Apache Spark? · Starting the Spark Shell · Using the Spark Shell · Getting Started with Datasets and DataFrames · DataFrame Operations

05

8 Exercises

Working with DataFrames and Schemas

TOPICS

Introduction to DataFrames · Eager and Lazy Execution

06

Analyzing Data with DataFrame Queries

TOPICS

Querying DataFrames Using Column Exp · Grouping and Aggregation Queries · Joining DataFrames

07

Introduction to Apache Hive

TOPICS

About Hive · Transforming data with Hive QL

Working with Apache Hive

5 Exercises

09**Exercise**

Hive and Spark Integration

TOPICS

Hive and Spark Integration

10

RDD Overview

TOPICS

RDD Overview · RDD Data Sources · Creating and Saving RDDs · RDD Operations

11

Transforming Data with RDDs

TOPICS

Writing and Passing Transformation Functions · Transformation Execution · Converting Between RDDs and DataFrames

12

Aggregating Data with Pair RDDs

TOPICS

Key-Value Pair RDDs · Map-Reduce · Other Pair RDD Operations

13

Querying Tables and Views with Apache Spark SQL

TOPICS

- Querying Tables in Spark Using SQL
- Querying Files and Views
- The Catalog API
- Comparing Spark SQL, Apache Impala, and Apache Hive-on-Spark

14

Working with Datasets in Scala

TOPICS

Datasets and DataFrames · Creating Datasets · Loading and Saving Datasets · Dataset Operations

15

Writing, Configuring, and Running Apache Spark Applications

TOPICS

Writing a Spark Application · Building and Running an Application · Application Deployment Mode · The Spark Application Web UI · Configuring Application Properties

16

Distributed Processing

TOPICS

Review: Apache Spark on a Cluster · RDD Partitions ·
Example: Partitioning in Queries · Stages and Tasks ·
Job Execution Planning · **Example:** Catalyst Execution
Plan · **Example:** RDD Execution Plan

17

Distributed Processing Challenges

TOPICS

Shuffle · Skew · Order

18

Distributed Data Persistence

TOPICS

DataFrame and Dataset Persistence · Persistence Storage Levels · Viewing Persisted RDDs

19

4 Exercises

Machine Learning with Spark ML

TOPICS

- Common Apache Spark Use Cases
- **Iterative Algorithms in Apache Spark:** Machine Learning, Graph Processing
- Introduction to MLlib- Various ML algorithms supported by Mlib
- ML model with Spark ML

20

Apache Spark Streaming: Introduction to DStreams

TOPICS

Apache Spark Streaming Overview · **Example:**
Streaming Request Count · DStreams · Developing
Streaming Applications

21

Apache Spark Streaming: Processing Multiple Batches

TOPICS

Multi-Batch Operations · Time Slicing · State
Operations · Sliding Window Operations · **Preview:**
Structured Streaming

22

Apache Spark Streaming: Data Sources

TOPICS

Streaming Data Source Overview · Apache Flume and Apache Kafka Data Sources · **Example:** Using a Kafka Direct
Data Source

Practise What You Learn

#	MODULE	ACTIVITY
05	Working with DataFrames and Schemas	■ Introducing DataFrames ■ Reading and Writing DataFrames ■ Working with Columns ■ Working with Complex Types ■ Combining and Splitting DataFrames ■ Summarizing and Grouping DF ■ Working with UDFs ■ Working with Windows
08	Working with Apache Hive	■ Working with Partitions ■ Working with Buckets ■ Working with Skew ■ Using Serdes to Ingest Text Data ■ Using Complex Types to Denormalize Data
09	Hive and Spark Integration	Spark Integration with Hive
19	Machine Learning with Spark ML	■ Implement Linear regression ■ Implement logistic regression ■ Implement Random Forest ■ Implement k-means

WHAT IS INCLUDED

Everything You Need to Succeed

- Koenig course material for every module
- Delivery by certified instructors and subject matter experts
- Hands-on exercises throughout the course
- Course completion certificate from Koenig

ABOUT KOENIG

Who We Are

Established in 1993, Koenig Solutions is a globally recognised training provider with 35+ technology partnerships, 5,000+ training programmes and a network of certified instructors and subject matter experts, supporting individuals and enterprise teams worldwide.

FLEXIBLE DELIVERY

Learn the Way That Suits You

Classroom Training

Live Online Classes (ILO)

Fly-Me-A-Trainer (FMAT)

Flexi (Self-Paced)

1-on-1 Training

Ready to Get Started?

Talk to a learning advisor about schedules, group training and customised delivery.

Visit koenig-solutions.com

✉ info@koenig-solutions.com