KOENIG
step forward

KOENIG SOLUTIONS · COURSE OUTLINE

# Big Data Basics: Hadoop, PySpark & Kafka

Learn to store, process and analyse big data using Hadoop, PySpark, Spark SQL, Spark ML and Kafka.

| 6 | 27 | 1 |
|---|---|---|
| MODULES | TOPICS | HANDS-ON ACTIVITIES |

## COURSE OVERVIEW

This course begins with an introduction to big data concepts, the five Vs and the Hadoop ecosystem, then covers HDFS storage and distributed processing with Map Reduce. It moves into PySpark foundations, including Spark configuration and resilient distributed datasets, before addressing Spark SQL alongside Hadoop Hive. The course then introduces machine learning with Spark ML, covering linear regression, logistic regression and random forest, and closes with Kafka architecture, workflow and cluster configuration alongside Spark.

## WHO THIS COURSE IS FOR

- **Data Engineers**
- **Big Data Analysts**
- **Software Developers**
- **Data Scientists**
- **System Administrators**

## LEARNING OBJECTIVES

- Explain big data concepts and the five Vs of big data
- Describe the components of the Hadoop ecosystem
- Apply HDFS for distributed big data storage
- Implement distributed processing using Map Reduce
- Build data pipelines with PySpark and resilient distributed datasets
- Compare Spark SQL with Hadoop Hive
- Develop machine learning models with Spark ML
- Configure a Kafka cluster and describe its workflow

COURSE SYLLABUS

# 6 Modules, 27 Topics

01

## Big Data Introduction

TOPICS

Big Data Overview · Five Vs of Big Data · What is Big Data and Hadoop · Introduction to Hadoop · Components of Hadoop Ecosystem · Big Data Analytics Introduction

02

## HDFS and Map Reduce

TOPICS

- HDFS - Big Data Storage
- Distributed Processing with Map Reduce
- Mapping and reducing stages concepts
- **Key Terms:** Output Format, Partitioners, Combiners, Shuffle, and Sort

HANDS-ON

- Map Reduce task

03

## PySpark Foundation

TOPICS

PySpark Introduction · Spark Configuration · Resilient distributed datasets (RDD) · Working with RDDs in PySpark · Aggregating Data with Pair RDDs

04

## Spark SQL and Hadoop Hive

TOPICS

Introducing Spark SQL · Spark SQL vs Hadoop Hive · Working with Spark SQL Query Language

05

## Machine Learning with Spark ML

TOPICS

- Introduction to MLlib Various ML algorithms supported by Mlib
- ML model with Spark ML
- Linear regression
- Logistic regression
- Random forest

06

### Kafka and Spark

TOPICS

Kafka architecture · Kafka workflow · Configuring Kafka cluster · Operations

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