

PySpark Development

Learn to process distributed data with PySpark, from RDDs and DataFrames to streaming and ETL pipelines.

5 Days

DURATION

7

MODULES

56

TOPICS

8

HANDS-ON ACTIVITIES

COURSE OVERVIEW

This comprehensive PySpark course offers a hands-on, project-driven introduction to distributed data processing with Apache Spark. Beginning with core concepts, the curriculum guides learners through the Spark ecosystem, resilient distributed datasets (RDDs), and their transformations and actions. Participants gain practical skills through real-world projects such as weather analytics, university datasets, and customer profiling. The course then transitions to advanced techniques using DataFrames, SQL-like operations, collaborative filtering with ALS models, and hyperparameter tuning. Learners also explore Spark Streaming for real-time analytics and apply Spark ETL processes for efficient data extraction, transformation, and loading. By the end of the course, participants will be equipped to design scalable, end-to-end data processing pipelines using PySpark's modern capabilities.

WHO THIS COURSE IS FOR

Data Engineers

Data Analysts

Data Scientists

Big Data Developers

ETL Developers

LEARNING OBJECTIVES

- Describe the Apache Spark stack and execution process
- Create RDDs and apply transformations such as map and filter
- Apply RDD actions including reduce, aggregate and sortBy
- Build DataFrames and perform SQL-like transformations
- Implement collaborative filtering using ALS models
- Apply hyperparameter tuning and cross validation to models
- Configure Spark Streaming for real-time data processing
- Develop Spark ETL pipelines for extraction and loading

7 Modules, 56 Topics

01

Fundamentals of PySpark

TOPICS

A Brief Primer on PySpark · Brief Introduction to Spark · Apache Spark Stack · Spark Execution Process · Newest Capabilities of PySpark · Cloning GitHub Repository

02

2 Projects

Resilient Distributed Datasets

TOPICS

- Resilient Distributed Datasets
- Creating RDDs
- Schema of an RDD
- Understanding Lazy Execution
- Introducing Transformations - .map(...)
- Introducing Transformations - .filter(...)
- Introducing Transformations - .flatMap(...)
- Introducing Transformations - .distinct(...)
- Introducing Transformations - .sample(...)
- Introducing Transformations - .join(...)
- Introducing Transformations - .repartition(...)

03

3 Projects

Resilient Distributed Datasets and Actions

TOPICS

- Introducing Actions - .collect(...)
- Introducing Actions - .reduce(...) and .reduceByKey(...)
- Introducing Actions - .count()
- Introducing Actions - .foreach(...)
- Introducing Actions - .aggregate(...) and .aggregateByKey(...)
- Introducing Actions - .coalesce(...)
- Introducing Actions - .combineByKey(...)
- Introducing Actions - .histogram(...)
- Introducing Actions - .sortBy(...)
- Introducing Actions - Saving Data
- Introducing Actions - Descriptive Statistics

04**2 Projects**

DataFrames and Transformations

TOPICS

- Creating DataFrames
- Specifying Schema of a DataFrame
- Interacting with DataFrames
- The .agg(...) Transformation
- The .sql(...) Transformation
- Creating Temporary Tables
- Joining Two DataFrames
- Performing Statistical Transformations
- The .distinct(...) Transformation

05**Projects**

Collaborative Filtering and Techniques

TOPICS

- Collaborative filtering
- Utility Matrix
- Explicit and Implicit Rating
- Expected Results
- Dataset
- Joining Dataframe
- Train and Test Data
- ALS model
- Optimization Hyperparameter tuning and cross validation
- Best model and evaluate prediction

06

Spark Streaming

TOPICS

- Basics of Spark Streaming
- Creating Stream Processing Context
- Reading Real-Time Data
- Simple RDD and DataFrame Operations on Streams
- Viewing and Aggregating Stream Data

07

Spark ETL

TOPICS

Introduction to ETL · Dataset · Preprocessing, extraction, transformation · Loading Data and cleaning

Practise What You Learn

#	MODULE	ACTIVITY
02	Resilient Distributed Datasets	■ Project 1: Count Data Project (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition) ■ Project 2: Weather Temperature Crunch (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying map, filter, faltmap, distinct, join and repartition on instream data)
03	Resilient Distributed Datasets and Actions	■ Project 3: Students/Professor University Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions.) ■ Project 4: Customer Data Datasets (ingestion of dataset, doing a preprocessing and exploratory dataset though the data set, applying RDD actions through specified applicability) ■ Project 5: Movie ratings
04	DataFrames and Transformations	■ Project 6: CompanyMegaData (doing all the transformation logics, columunal logic and aggregation and exploratory data analysis) ■ Project 7: University Data (end to end pyspark execution of insight delivery on University Data)
05	Collaborative Filtering and Techniques	Project 8: IMDB Rating project (Optimization logics focused on the project with extensive pyspark logic and clever techniques of manipulation)

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