

Data Processing with PySpark

Learn to process big data with PySpark, from RDDs and DataFrames to SQL functions, external sources and streaming.

11

MODULES

111

TOPICS

COURSE OVERVIEW

This course covers the fundamentals of Apache Spark, its architecture, and installation on various platforms before introducing PySpark and its core concepts such as SparkSession, SparkContext and RDDs. It then examines RDD creation, persistence, caching, partitioning and transformations, followed by Spark SQL and DataFrame operations including case classes, UDFs and conversions between RDDs, DataFrames and DataSets. It proceeds through SQL and built-in functions for aggregation, windowing, dates and string manipulation, then working with external sources such as Hive, MySQL, CSV and JSON, including joins and handling nulls. The course concludes with Spark Streaming concepts and an overview of local and cluster deployment modes.

WHO THIS COURSE IS FOR

Data Engineers

Data Analysts

Big Data Developers

Software Developers

Database Administrators

LEARNING OBJECTIVES

- Explain Apache Spark architecture and compare it with Hadoop MapReduce
- Describe PySpark features and its advantages over Scala
- Apply SparkSession, SparkContext and RDD operations in PySpark
- Implement RDD persistence, caching, partitioning and lineage
- Build DataFrames from RDDs and apply SQL functions for data analysis
- Apply built-in functions for string, date and timestamp manipulation
- Integrate Spark with external sources such as Hive, MySQL, CSV and JSON
- Configure Spark Streaming and deployment modes for data processing

11 Modules, 111 Topics

01

Introduction to Apache Spark

TOPICS

Introduction to Big Data · What is Apache Spark? · Evaluation of Apache Spark · Features · Spark Architecture · Spark Vs Hadoop Map Reduce · Spark SQL Vs HIVE

02

Installation

TOPICS

Installation on MAC · Installation on Windows · With Scala and IntelliJ · Creating DataBricks Account · DataBricks Compute, Notebook, tables

03

PySpark Introduction

TOPICS

Why Pyspark? · Need for Pyspark · Spark Python Vs Scala · Pyspark features · Real-life usage of PySpark · Web/Application

04

PySpark Basics

TOPICS

SparkSession · SparkContext · Stage · Executor · RDD · Parallelize

05

RDDs (Resilient Distributed Datasets)

TOPICS

- Parallelize
- Read Text File
- Read CSV
- Create RDD
- RDD Persistence and Caching Mechanism
- RDD Features
- RDD Limitations
- RDD Lineage
- Action
- Pair Functions- Paired RDD
- Repartition and Coalesce
- Shuffle Partitions
- Cache vs Persist

5.1 PERSISTENCE OPTIONS

- MEMORY_ONLY
- MEMORY_SER_ONLY
- DISK_ONLY
- DISK_SER_ONLY
- MEMORY_AND_DISK_ONLY

5.2 CORE COMPUTING

- Fault Tolerance model in spark
- Different ways of creating a RDD
- Word Count Example
- Creating spark objects(RDDs) from Scala Objects(lists)
- Increasing the no of partitions
- Aggregations Over Structured Data:
 - reduceByKey()

RDDs (Resilient Distributed Datasets) (continued)

5.3 GROUPINGS AND AGGREGATIONS

- Single Grouping and Single Aggregation
- Single Grouping and multiple Aggregation
- Multi Grouping and Single Aggregation
- Multi Grouping and Multi Aggregation

5.4 VARIOUS ACTIONS AND TRANSFORMATIONS

- countByKey()
- countByValue()
- sortByKey()
- zip()
- Union()
- Distinct()

06

SQL - DataFrame

TOPICS

- Introduction
- Making data Structured
- Case Classes
- ways to extract case class objects
- using function
- using map with multiple expressions
- using map with single expression
- Sql Context
- Data Frames API
- DataSet API
- RDD vs DataFrame vs DataSet
- Create a DataFrame
- Create an empty DataFrame
- Convert RDD to DataFrame
- Convert DataFrame to Pandas
- union() & unionAll()
- unionByName()
- UDF (User Defined Function)
- map()

07

SQL Functions

TOPICS

Aggregate Functions · Window Functions · Date and Timestamp Functions · JSON Functions · Read & Write JSON file

08

Built-In Functions

TOPICS

- when()
- expr()
- lit()
- split()
- concat_ws()
- substring()
- translate()
- regexp_replace()
- overlay()
- to_timestamp()
- to_date()

09

External Sources

TOPICS

- Working with sql statements
- Spark and Hive Integration
- Spark and mysql Integration
- Working with CSV
- Working with JSON
- Transformations and actions on dataframes
- Narrow, wide transformations
- Addition of new columns, dropping of columns ,renaming columns
- Addition of new rows, dropping rows
- Handling nulls
- Joins

10

Spark Streaming

TOPICS

What is spark streaming? · Dataframe in context of spark streaming · Reading data from files · Running the stream with the data

11

Deployment Modes

TOPICS

Local Mode · Cluster Modes(Standalone , YARN

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