

Spark for Developers

Duration: 3 Days · **Training format:** Virtual Classroom / Online

Description

This hands-on training course delivers the key concepts and expertise developers need to use Apache Spark to develop high-performance parallel applications. The course covers how to work with “big data” stored in a distributed file system, and execute Spark applications on a Hadoop cluster. After taking this course, participants will be prepared to face real-world challenges and build applications to execute faster decisions, better decisions, and interactive analysis, applied to a wide variety of use cases, architectures, and industries.

Delegates will learn

- Spark fits into the Big Data ecosystem, and how to use Spark for data analysis. The course covers Spark shell for interactive data analysis, Spark internals, Spark APIs, Spark SQL, Spark streaming, and machine learning.

Outline

Spark Basics

- Spark and Hadoop
- Spark Concepts and Architecture
- Spark eco System (core, spark sql, ml, streaming)
- Spark SQL
- RDD

Spark API programming

- Introduction to Spark API / RDD API
- Submitting the First Program to Spark
- Debugging / Logging
- Configuration Properties

Machine Learning Fundamentals

- ML and DL fundamentals
- Regression
- Classification
- Clustering

Machine Learning on bigdata with SparkML

- Feature operations
- Preparing data for ML
- Onehotencoding, scaling..
- Training Models
- classification, regression
- Hyper parameter tuning
- Cross validation, Train Validation Split
- Basic Sentiment analysis on text data

Introduction to Structured Streaming

- Apache Spark Streaming Overview
- Creating Streaming DataFrames
- Transforming DataFrames
- Executing Streaming Queries

Structured Streaming with Apache Kafka

- Receiving Kafka Messages
- Sending Kafka Messages

Aggregating and Joining Streaming DataFrames

- Streaming Aggregation
- Joining Streaming DataFrames

Spark Distributed Processing

- Apache Spark on a Cluster
- RDD Partitions
- Stages and Tasks
- Job Execution Planning

Prerequisites

This course is designed for developers and engineers who have programming experience, but prior knowledge of Spark and Hadoop is not required.