

# Introduction to Spark Programming

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

## Description

---

This Introduction to Spark Programming course introduces the Apache Spark distributed computing engine and is suitable for developers, data analysts, architects, technical managers, and anyone who needs to use Spark in a hands-on manner.

The course provides a solid technical introduction to the Spark architecture and how Spark works. It covers the basic building blocks of Spark (e.g., RDDs and the distributed compute engine), as well as higher-level constructs that provide a simpler and more capable interface (e.g., Spark SQL and DataFrames). It also covers more advanced capabilities such as the use of Spark Streaming to process streaming data and provides an overview of Spark ML (machine learning). Finally, the course explores possible performance issues and strategies for optimization.

The course is very hands-on, with many labs. Participants will interact with Spark through the Spark shell (for interactive, ad hoc processing) as well as through programs using the Spark API.

The Apache Spark distributed computing engine is rapidly becoming a primary tool in the processing and analyzing of large-scale data sets. It has many advantages over existing engines, such as Hadoop, including runtime speeds that are 10-100x faster, as well as a much simpler programming model. After taking this course, you will be ready to work with Spark in an informed and productive manner.

## Delegates will learn how to

- Understand the need for Spark in data processing
- Understand the Spark architecture and how it distributes computations to cluster nodes
- Become familiar with basic installation / setup / layout of Spark
- Use the Spark shell for interactive and ad-hoc operations
- Understand RDDs (Resilient Distributed Datasets), and data partitioning, pipelining, and computations
- Understand/use RDD ops such as map(), filter(), reduce(), groupByKey(), join(), etc.
- Understand Spark's data caching and its usage
- Write/run standalone Spark programs with the Spark API
- Use Spark SQL / DataFrames to efficiently process structured data
- Use Spark Streaming to process streaming (real-time) data

- Understand performance implications and optimizations when using Spark
- Become familiar with Spark ML

## Outline

---

What is Apache Spark?

Apache Spark Architecture and Installations

Apache Spark RDD Structure

Apache Spark Project Creation

Data Loading Stages

Transformation Structure Map Method

Transformation Structure FilterMethod

Flat Map Distinct Method

PairRDD and GroupByKey Method

Lazy Evolution Concept

Action Method

What is SparkSQL?

SparkSQL

Data Reading Stages

SparkSQL StructureType

SparkSQL Filter Structure

Spark SQL Group By Structure

Spark SQL API

Spark Tempview Globalview Concepts

Spark and Hadoop(HDFS) Integration

End-to-End Project Development with Spark

CallCenter Data Analysis with Spark

Writing Call Center Results to MongoDB

What is Spark Streaming Real-Time Data Analytics?

Spark Streaming Types

Instant Message Analysis with Spark Streaming

Architecture of the Spark Streaming Example

IoT Analytics with Spark Streaming

Streaming Complete and Update Modes

Streaming Time Window

Message Analysis with Streaming Time Window

Kafka Integration with Streaming

Introduction to Machine Learning with Spark - Spark MLlib?

Spark MLlib Library

What is Estimation (Regression)?

What is Linear Regression?

Implementing Linear Regression with Spark MLlib

Model Evaluation R2 Method

Naive Bayes Algorithm

Naive Bayes Implementation with Spark MLlib

A workshop with Naive Bayes

## Prerequisites

---

Reasonable programming experience.