+44 7989 401397
info@olsensoft.com

# Big Data (4 days)

## Course overview

This course takes a detailed look at how to implement Big Data solutions using Apache Spark. The course describes the problems that Big Data is designed to solve, and explains how Hadoop addresses these issues via HDFS, Yarn, and the Spark API.

We show plenty examples to help you understand how to create and use RDDs from various data sources, such as flat files, NoSQL databases, and relational databases. We also explore the key Spark APIs layered on top of RDDs, including Spark Streaming via DataFrames, Spark SQL, and Spark Machine Learning and Spark Graph Processing.

## What you'll learn

- Big Data principles
- Creating and using RDDs
- Spark Streaming
- Spark SQL
- Spark Machine Learning
- Spark Graph Processing

## Prerequisites

- Solid experience in Scala (or Python/Java)

## Course details

- Introduction to Big Data: Introduction to Hadoop; Data serialization; Column-based storage; Messaging systems; NoSQL; Distributed SQL query engine
- Introduction to Apache Spark: Key features of Spark; Spark architecture; Application execution; Resilient Distributed Datasets; Spark API; Caching; Spark jobs
- Interactive Data Analysis with Spark Shell: Key concepts; REPL commands; Using Scala; Number analysis; Log analysis
- Writing Spark Applications: Writing a Hello world application; Compiling and running an application; Monitoring and debugging an application
- Spark Streaming: Overview of Spark streaming; Spark streaming API; Creating a discretized stream; Processing a discretized stream; Output operations
- Spark SQL: Overview of Spark SQL; Performance considerations; Usage scenarios; Spark SQL API; Built-in functions
- Machine Learning with Spark: Overview of Machine Learning; Spark Machine Learning Libraries (MLlib API); Spark ML

- Graph Processing with Spark: Overview of graphs; Overview of GraphX API; Using GraphX API
- Cluster Managers: Standalone cluster manager; Apache Mesos; YARN