Publication date : 09/30/2025

Course : Spark Java, developing applications for Big Data

Practical course - 3d - 21h00 - Ref. SPK
Price : 2010 € E.T.

4,7 / 5

Often presented as the successor to Hadoop, Spark simplifies the programming of big data processing, enabling the use of Scala, Python or Java . This training course will teach programmers how to process data streams in real time, and how to carry out batch processing (from SQL to machine learning).

Teaching objectives

At the end of the training, the participant will be able to:

- Master the fundamental concepts of Spark
- Developing applications with Spark Streaming
- Setting up a Spark cluster
- Exploiting data with Spark SQL
- A first approach to machine learning

Intended audience

Project managers, data scientists, developers, architects.

Prerequisites

Good knowledge of Java. Knowledge of big data.

Practical details

Hands-on work

Practical application of the concepts covered in the course using the Java language.

Course schedule

PARTICIPANTS

Project managers, data scientists, developers, architects.

PREREQUISITES

Good knowledge of Java. Knowledge of big data.

TRAINER QUALIFICATIONS

The experts leading the training are specialists in the covered subjects. They have been approved by our instructional teams for both their professional knowledge and their teaching ability, for each course they teach. They have at least five to ten years of experience in their field and hold (or have held) decision-making positions in companies.

ASSESSMENT TERMS

The trainer evaluates each participant's academic progress throughout the training using multiple choice, scenarios, hands-on work and more.
Participants also complete a placement test before and after the course to measure the skills they've developed.

1 Introducing Apache Spark

- History of the framework.
- The different versions of Spark (Scala, Python and Java).
- Comparison with the Apache Hadoop environment.
- The different Spark modules.

Hands-on work
Install and configure Spark. Run a first example with word counting.

2 Programming with Resilient Distributed Datasets (RDD)

- RDD presentation.
- Create, manipulate and reuse RDDs.
- Accumulators and broadcast variables.
- Use partitions.

Hands-on work
Handling different datasets with RDDs and using the API provided by Spark.

3 Handling structured data with Spark SQL

- SQL, DataFrames and datasets.
- The different types of data sources.
- Interoperability with RDDs.
- Spark SQL performance.
- JDBC/ODBC server and Spark SQL CLI.

Hands-on work
Dataset manipulation via SQL queries. Connection with an external database via Java DataBase Connectivity (JDBC) Open Database Connectivity (OBDC).

4 Spark on a cluster

- The different types of architecture: standalone, Apache Mesos or Hadoop YARN.
- Set up a cluster in standalone mode.
- Pack an application with its dependencies.
- Deploy applications with Spark-submit.
- Size a cluster.

Hands-on work
Setting up a Spark cluster.

5 Real-time analysis with Spark Streaming

- Operating principle.
- Introducing Discretized Streams (DStreams).
- The different types of source.
- API handling.
- Comparison with Apache Storm.

Hands-on work
Log consumption with Spark Streaming.

TEACHING AIDS AND TECHNICAL RESOURCES

• The main teaching aids and instructional methods used in the training are audiovisual aids, documentation and course material, hands-on application exercises and corrected exercises for practical training courses, case studies and coverage of real cases for training seminars.
• At the end of each course or seminar, ORSYS provides participants with a course evaluation questionnaire that is analysed by our instructional teams.
• A check-in sheet for each half-day of attendance is provided at the end of the training, along with a course completion certificate if the trainee attended the entire session.

TERMS AND DEADLINES

Registration must be completed 24 hours before the start of the training.

ACCESSIBILITY FOR PEOPLE WITH DISABILITIES

Do you need special accessibility accommodations? Contact Mrs. Fosse, Disability Manager, at psh-accueil@orsys.fr to review your request and its feasibility.

6 Graph manipulation with GraphX

- Introducing GraphX.
- The different operations.
- Create graphs.
- Vertex and Edge RDD.
- Presentation of different algorithms.

Hands-on work
Handling the GraphX API through various examples.

7 Machine learning with Spark

- Introduction to machine learning.
- The different classes of algorithms.
- Introducing SparkML and MLlib.
- Implementing different algorithms in MLlib.

Hands-on work
Using SparkML and MLlib.

Dates and locations

REMOTE CLASS
2026 : 15 June, 14 Sep., 23 Nov.

PARIS LA DÉFENSE
2026 : 15 June, 14 Sep., 23 Nov.