Publication date : 01/19/2024

# Course : Spark Python, developing applications for big data

*Practical course - 3d - 21h00 - Ref. QNC*
*Price : 1940 CHF E.T.*

★★★★☆ 4,3 / 5

BEST

Spark is a distributed computing framework for manipulating big data. Initially designed to accelerate Hadoop processing, it has become a stand-alone system. It can be programmed in four languages, including Python, which has become predominant. This course introduces you to Spark Python.

## Teaching objectives

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

- ✓ Discover the fundamental concepts of Spark
- ✓ Using Spark's RDD concept
- ✓ Exploiting data with Spark SQL
- ✓ Real-time analysis with Spark Streaming
- ✓ Using Spark with Jupyter notebooks, manipulating data with Pyspark as with Pandas
- ✓ Machine learning with Spark

## Intended audience

Anyone familiar with Python who wants to discover the Apache Foundation's Spark framework.

## Prerequisites

Good Python language skills.

## Practical details

### Exercise

Numerous exercises are used to illustrate the topics.

### Teaching methods

Each topic is illustrated by demonstrations running on a cloud cluster. Participants complete exercises after the concepts have been presented.

## Course schedule

### PARTICIPANTS

Anyone familiar with Python who wants to discover the Apache Foundation's Spark framework.

### PREREQUISITES

Good Python language skills.

### 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 four main components: Spark SQL, Spark Streaming, MLlib and GraphX.
- Python tools and libraries for Spark: PySpark, Jupyter notebooks, Koalas.
- Spark programming concepts.
- Run Spark in a distributed environment.

### Hands-on work

Set up the Python environment for Spark. Implementation of scripts manipulating Spark concepts.

## 2 Using Spark with Python: resilient distributed datasets (RDD)

- Configure your Python environment.
- Connecting to Spark with Python: contexts and sessions.
- RDD overview. Create, manipulate and reuse RDDs.
- The main functions/transformations, implementation of map/reduce algorithms.
- Accumulators and broadcast variables.
- Use partitions.
- Use notebooks and submit Python jobs.

### Hands-on work

Context and session manipulation. RDD creation and reuse. Job submission.

## 3 Handling structured data

- Introduction to Spark SQL and DataFrames and datasets.
- The different types/formats of data sources.returnchariot
- Interoperability with RDDs.
- Use the PySpark Pandas library.

### Tutored hands-on work

Executing queries with Spark SQL. Implementing DataFrames and datasets. DataFrame manipulation.

## 4 Machine learning with Spark

- Introduction to machine learning.
- The different classes of algorithms.
- Introducing MLlib.
- Implementation of the various algorithms in MLlib.

### Hands-on work

Implementation of supervised learning through classification.

### 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.

### 5 Real-time analysis with Spark Streaming

- Understanding the architecture of streaming.
- Introducing Discretized Streams (DStreams).
- The different types of source.
- API manipulation (aggregation, watermarking, etc.).
- Real-time machine learning.

**Hands-on work**

Create real-time statistics from a data source and make predictions using machine learning.

### 6 Graph theory

- Introduction to graph theory (nodes, edges, directed graphs, paths, main algorithms).
- Using the API.
- GraphX and GraphFrame libraries.

**Hands-on work**

Implementation of a page rank algorithm and graph visualization.

## Dates and locations

**REMOTE CLASS**

2026 : 17 June, 17 June, 30 Sep., 30 Sep., 14 Dec.,
14 Dec.