Course : Spark Advanced

Machine Learning and industrialization of analytical workflows

Practical course - 3d - 21h00 - Ref. SPN

Price : 2010 € E.T.

Spark is a distributed computing framework for complex Big Data processing and analysis. If you've already used Spark, we'd like to take your analyses a step further with machine learning, and introduce you to MLOps for deploying and industrializing analytical models.

Teaching objectives

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

- Learn advanced data analysis with Spark
- Performing machine learning (ML) processing with Spark
- Understanding Docker and its usefulness in industrializing analytical workflows
- Detailing and implementing the analytical cycle with Spark
- Learn how to industrialize your analysis workflow
- Discover MLOps

Intended audience

Professionals who want to use Spark for batch and real-time analytics.

Prerequisites

Connaissances des API Spark, notamment RDD et DataFrame. Connaissances des algorithmes d'apprentissage supervisés et non supervisés. Maîtrise d'un des langages suivants : Scala, Python.

Practical details

Hands-on work

Alternating theory and practical work. 60% exercises for greater depth. Practical feedback.

Course schedule

PARTICIPANTS

Professionals who want to use Spark for batch and real-time analytics.

PREREQUISITES

Connaissances des API Spark, notamment RDD et DataFrame. Connaissances des algorithmes d'apprentissage supervisés et non supervisés. Maîtrise d'un des langages suivants : Scala, Python.

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 Introduction

- A reminder of the Spark API.
- Docker concepts and their use in data analysis.
- Docker containers.

Hands-on work
Get to grips with the working environment, create Docker containers.

2 The analytical cycle with Spark

- Ingestion of data.
- Exploration.
- Data preparation.
- Learning.
- Industrialization.

Storyboarding workshops
Presentation of case studies and discussion of the different stages of the cycle.

3 Ingestion of data.

- Data loading.
- Batch processing.
- Streaming treatments.
- Data formats: images, binary, structured, Graph...

Hands-on work
Load data from various sources.

4 Data mining

- Descriptive statistics.
- Identify outliers and empty data.
- Identify invalid values and other anomalies.

Hands-on work
Identify anomalies in a dataset.

5 Preparation and feature engineering (data transformation process)

- Data cleansing.
- Pipelines.
- Transformer les valeurs numériques, catégoriques, binaires et texte.
- Création de nouvelles features.
- Réduction de dimensions.
- Vectorisation.

Hands-on work
Prepare data for analysis.

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 ML lifecycle with MLflow

- Life cycle of a machine learning project.
- Introducing the MLflow open-source platform.
- MLflow's main components: Tracking, Models and Projects.
- Parameters, metrics, tags and artifacts.

Hands-on work
Creating and using a machine learning project.

7 Machine learning

- MLlib, Spark's machine learning library and available algorithms.
- Divide a dataset.
- Configure a template and run it.
- Interpretation and validation of learning outcomes.
- Introduction to Spark Streaming.

Hands-on work
Implementing machine learning.

8 Case studies

- Make recommendations.
- Make sales forecasts.
- Semantic analysis.
- Computer vision with Spark and PyTorch.
- Analyse temps réel avec Spark et Kafka.

Case study
Carry out the various case studies proposed.

Dates and locations

REMOTE CLASS
2026 : 23 Mar., 18 May, 28 Sep.

PARIS LA DÉFENSE
2026 : 23 Mar., 18 May, 28 Sep.