AMBIENT-IT

+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Updated 05/02/2025

Sign up

Spark 3 & Machine Learning training

3 days (21 hours)

Presentation

Spark is a framework for performing distributed calculations on a cluster computers. This course introduces the latest version 3.5.4, which brings a host of new features and impressive performance improvements!

Created in 2009 at Berkeley, it is becoming the preferred "Big Data" platform, gradually replacing the Hadoop ecosystem, thanks to unified APIs in Java, Scala, Python and R that make it very easy to use.

The course reviews the main Spark components, as well as the new :

- Spark Core
- Spark SQL
- Spark Streaming
- Spark ML
- GraphFrame
- SparkR
- Deep Learning pipeline

Our Spark and Machine Learning course also Spark's integration with HDFS. It presents the Spark API. Practical work is carried out in Scala by default (or Python as an option).

Objectives

- Be able to use Spark 3 and its new features independently
- Understand the concept of Machine Learning and the fundamental concepts of Spark, and be able to use them.
- Handling large volumes of data using best practices in Spark 4
- Understanding documentation, APIs and the Big Data ecosystem
- Integrating Spark into a Hadoop ecosystem

- Create real-time analysis applications with Spark Streaming
- Parallel programming on a cluster
- Mastering Spark SQL

Target audience

- Developers
- Architects
- System administrators
- DevOps

Prerequisites

- Basic knowledge of a Unix system
- Knowledge of Scala or Python & Git
- Stats-oriented culture
- Test My Knowledge

Technical requirements

- Have Visual Studio Code installed

Spark and Machine Learning training program

Day 1 - Understanding and using Spark 3

Big Data context and issues - Distributed computing

- Why Spark? What's new in version 2 & 3
- Standalone installation, test with jupyter
- Spark Core (MapReduce replacement)
- RDD Resilient Distributed Datasets
- PairedRDD
- Spark Context VS Spark Session
- DAG Directed Acyclic Graph
- RDD Objects, DAG Scheduler, Task Scheduler, Worker
- Hadoop and HDFS
- NameNode & DataNode
  - core-site, hdfs-site
  - Spark on a cluster
- Spark Standalone: Cluster Manager, Worker, Executor, Spark Context
- Mesos (Private Cluster), Marathon, YARN
- Structured API

Spark SQL (HIVE replacement)

- SQLContext
- HiveContext
  - DataFrames
  - Spark Structure, Schema and Partitioning

Day 2 - Understanding Machine Learning and its integration into Spark 3

Introduction to Machine Learning (ML)

- Supervised learning
- Unsupervised learning
- Clustering: KNN, K-mean
- Regression: Regression tree
- Classification: Random Forest, SVM, AUC, ROC curve

Spark ML - Introduction

- Pipelines: Transformer, Estimator, Model
- ML persistence
- MLlib in R & PySpark

DataVisualisation

- Matplotlib
- Seaborn
- Plotly
- Bokeh

GraphFrame

- Package presentation

Day 3 - Spark 3 in advanced mode: Handling large-scale data

Spark Streaming

- Structured Streaming API
- StreamingContext
- Static and Dynamic Datasets
  - Continuous Aggregations
  - Encoders
- Real-Time Analytics of a log file
  - Catalyst Optimizer and Tungsten Engine for greater efficiency
- Create agents, sources, channels and sinks

- Serialization with Avro RPC

SparkR

- Deep Learning pipeline

package presentation

- Package presentation
- The concept of transfer

learning Conclusion

- Lambda VS Kappa architecture

Companies concerned

This course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer technology, or to acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level of proficiency in different types of technology, as well as his or her expectations and personal objectives for the training to come, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical input from the trainer, supported by examples, brainstorming sessions and group work.

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Sanction

A certificate will be issued to each trainee who completes the course.