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

PySpark training: process your mass data quickly

3 days (21 hours)

Presentation

Our PySpark training course will teach you how to manipulate large volumes of data quickly, and take advantage of the power of the Python API to master Apache Spark.

Our training course is divided into the modules you need to understand the Apache Spark ecosystem and how to use PySpark. We'll start with a presentation of Hadoop (its architecture and components). We'll then guide you through the installation of this Big Data framework and the configuration of PySpark.

You'll discover how to use the Python API on Spark to manipulate your data, so you can master your entire ETL process (data extraction, loading and transformation). In addition, there is a module dedicated to the use of Pandas, to give you a more in-depth understanding of the tool. You'll also learn how to use Spark for machine learning, Spark Streaming and, of course, Spark SQL.

Our training course will introduce you to the latest version of Apache Spark, Spark 3.5.

Objectives

- Understand the role of Hadoop and Spark in Big Data.
- Master the architecture and operation of Hadoop
- Installing and interacting with Spark
- Using Spark SQL to manipulate DataFrames
- Apply PySpark and Pandas for data manipulation

Target audience

- Data analysts

- Data scientists
- Data engineers
- Developers

Prerequisites

- Knowledge of SQL
- Basic knowledge of mathematics and statistics
- Basic knowledge of Python

PySpark training program

Introducing Hadoop

- What is Hadoop?
- Its role in Big Data
- Architectural overview
- How does Hadoop work?
- Main modules
  - HDFS
  - YARN
  - MapReduce
  - Hadoop Common

Introducing Spark

- Spark vs Hadoop
- The differences with MapReduce
- Why use Spark?
- Features
  - MLlib
  - Streaming
  - SQL
  - GraphX
- How does Spark work?
- Data sets
  - RDD
  - DataFrames
  - Data Sets

How do I install Spark?

- Local
- On a distributed infrastructure
- In the Cloud
- First interaction with Spark

Spark SQL

- Introduction to Spark SQL
- Creating DataFrames
- Handling DataFrames
- Data loading
- Data storage
- Differences between SQL API and dataframe API
- Explanation of how catalyst works, and diagnostic and debugging tools.

Using PySpark

- Introducing PySpark
- Using SparkSQL to manipulate data
- Load data in different formats
- Transforming your data
- Practical work: Loading and transforming data with PySpark

Pandas API

- Install Pandas
- Transform and apply
- How do data types change?
- The hints
- Good development practices

Spark.ml

- Supervised learning
- Random trees
- Create personalized recommendations
- Text data processing
- Automate analysis with pipelines

Spark Streaming

- DStream
- Data sources
- Using the API
- Modifying data

Troubleshooting

- Exceptions linked to the absence of memory

- Repeated failure of the Spark task
- Spark Shell command fails
- FileAlreadyExistsException
- Too Large Frame" error
- Spark jobs fail due to compilation failures

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.