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

Updated on 05/05/2026

Sign up

Distributed Machine Learning with Spark ML Training

3 days (21 hours)

Overview

Spark MLlib is Apache Spark's distributed machine learning library. It enables the processing of massive data volumes where traditional tools fall short, leveraging distributed computing for training and inference.

Our Distributed Machine Learning with Spark ML training will enable you to master the Spark MLlib ecosystem, design robust ML pipelines, and optimize your processing for scaling.

You will learn to transform raw data into actionable features (Feature Engineering), train complex algorithms (Random Forests, Gradient Boosting, ALS), and manage the lifecycle of your models using tools like MLflow.

By the end of the course, you will be able to develop, evaluate, and deploy high-performance predictive models on production clusters, while mastering performance issues related to data partitioning and shuffling.

Like all our courses, this one will introduce you to **the latest stable version** of the technology and its new features.

Objectives

- Develop distributed ML models
- Optimize computational performance
- Prepare and transform data for ML
- Evaluate and deploy models

Target audience

- Data engineers
- Data scientists

Prerequisites

- Knowledge of Python/Scala, Spark, and ML

Software requirements

- At least 8 GB of RAM, 16 GB if possible
- Linux (Ubuntu, Fedora, etc.), macOS, or Windows (preferably with WSL2)
- A local Spark cluster or an environment such as Databricks Community
- A code editor or Jupyter Notebooks

Course outline for our Distributed Machine Learning with Spark ML training

[Day 1 - Morning]

Spark MLlib Architecture and Pipelines

- Understanding the distributed architecture of MLlib
- Mastering DataFrames for ML
- Key Concepts: Transformers, Estimators, and Pipelines
- Data Type Management (Vector, Dense, Sparse)
- Serialization and persistence of workflows
- Hands-on workshop: Setting up a complete classification pipeline.

[Day 1 - Afternoon]

Data Preparation and Feature Engineering at Scale

- Cleaning and imputation of distributed data
- Encoding: StringIndexer, OneHotEncoder
- Feature assembly with VectorAssembler
- Dimensionality Reduction (PCA) and Variable Selection
- Standardization and scaling (MinMaxScaler, StandardScaler)
- Hands-on workshop: Preparing a large dataset for training.

Regression and Classification Algorithms

- Linear and Logistic Regression
- Decision trees and ensembles (Random Forest, GBT)
- Multi-class evaluation and handling of class imbalance
- Model interpretability in a distributed environment
- Analysis of Residuals and Prediction Errors
- Hands-on Workshop: Training and Comparing Classification Models.

[Day 2 - Morning]

Clustering and Recommendation Systems

- Unsupervised learning with K-means
- Bisecting K-means and Gaussian Mixture Models
- Collaborative filtering with ALS (Alternative Least Squares)
- Large-scale similarity measures
- Cold Start Recommendation Optimization
- Hands-on workshop: Building a distributed recommendation engine.

[Day 2 - Afternoon]

ML performance optimization

- Impact of Shuffling on Training Performance
- Caching and Checkpointing Strategies
- Data partitioning and task parallelism
- Monitoring via the Spark UI and Bottleneck Detection
- Memory Resource Management for Large Models
- Hands-on Workshop: Auditing and Optimizing a Slow ML Job

Model Tuning and Selection

- Distributed cross-validation
- Grid search with ParamGridBuilder
- Hyperparameter optimization and success metrics
- BinaryClassificationEvaluator vs MulticlassClassificationEvaluator
- Saving and Exporting the Best Models
- Hands-on workshop: Fine-tuning a model to maximize accuracy.

[Day 3 - Morning]

Large-scale NLP and Text Mining

- Text preprocessing: Tokenizer, StopWordsRemover
- Vector representation: TF-IDF and Word2Vec
- Sentiment analysis and text classification
- Using N-grams for Context
- NLP Architectures with Spark
- Hands-on Workshop: Semantic Analysis of a Text Data Stream.

[Day 3 - Afternoon] Industrialization

and MLOps

- Model lifecycle with MLflow (Tracking, Registry)
- Persistence in ML format and portability
- Batch Inference vs. Stream Inference
- Introduction to Spark Serving and real-time architectures
- Monitoring model drift in production
- Hands-on workshop: Experiment tracking and model deployment.

Case studies and capstone project

- Designing an end-to-end ML architecture
- Choosing algorithms based on volume and latency
- Automation of the training pipeline
- Production best practices (CI/CD for ML)
- Deployment checklist
- Hands-on workshop: Final project—Industrialization of a complex business problem.

Target Audience

This training is intended for both individuals and companies, large or small, seeking to train their teams in new advanced IT technologies or to acquire specific business knowledge or modern methodologies.

Assessment upon enrollment

The pre-training assessment complies with Qualiopi quality standards. Upon final registration, the learner receives a self-assessment questionnaire that allows us to evaluate their estimated proficiency in various types of technologies, as well as their expectations and personal goals for the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could pose challenges for monitoring and ensuring the smooth running of the training session.

Teaching Methods

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

Organization

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

Assessment

At the end of the session, a multiple-choice questionnaire is used to verify that the skills have been properly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training program.