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

Mis à jour le 05/05/2026

S'inscrire

Formation Machine Learning distribué avec Spark ML

3 jours (21 heures)

Présentation

Spark MLlib est la bibliothèque de Machine Learning distribué d'Apache Spark. Elle permet de traiter des volumes de données massifs là où les outils traditionnels saturent, en s'appuyant sur le calcul distribué pour l'entraînement et l'inférence.

Notre formation Machine Learning distribué avec Spark ML vous permettra de maîtriser l'écosystème Spark MLlib, de concevoir des pipelines ML robustes et d'optimiser vos traitements pour le passage à l'échelle.

Vous apprendrez à transformer des données brutes en caractéristiques exploitables (Feature Engineering), à entraîner des algorithmes complexes (Random Forests, Gradient Boosting, ALS) et à gérer le cycle de vie de vos modèles avec des outils comme MLflow.

À l'issue de la formation, vous serez en mesure de développer, évaluer et déployer des modèles prédictifs performants sur des clusters de production, tout en maîtrisant les problématiques de performance liées au partitionnement et au shuffling des données.

Comme toutes nos formations, celle-ci vous présentera **la dernière version stable** de la technologie et ses nouveautés.

Objectifs pédagogiques

- Développer des modèles ML distribués
- Optimiser les performances de calcul
- Préparer et transformer des données pour ML
- Évaluer et déployer les modèles

Public visé

- Data engineers
- Data scientists

Pré-requis

- Connaissances en Python/Scala, Spark et ML

Pré-requis logiciels

- 8 Go de RAM au minimum, 16 Go si possible
- Linux (Ubuntu, Fedora, etc.), macOS ou Windows (avec WSL2 de préférence)
- Un cluster Spark local ou un environnement type Databricks Community
- Un éditeur de code ou Jupyter Notebooks

Programme de notre formation Machine Learning distribué avec Spark ML

[Jour 1 - Matin]

Architecture de Spark MLlib et Pipelines

- Comprendre l'architecture distribuée de MLlib
- Maîtriser les DataFrames pour le ML
- Concepts clés : Transformers, Estimators et Pipelines
- Gestion des types de données (Vector, Dense, Sparse)
- Sérialisation et persistance des workflows
- Atelier pratique : Mise en place d'un pipeline complet de classification.

[Jour 1 - Après-midi]

Préparation et Feature Engineering à l'échelle

- Nettoyage et imputation de données distribuées
- Encodage : StringIndexer, OneHotEncoder
- Assemblage de caractéristiques avec VectorAssembler
- Réduction de dimensionnalité (PCA) et sélection de variables
- Standardisation et mise à l'échelle (MinMaxScaler, StandardScaler)
- Atelier pratique : Préparation d'un dataset massif pour l'entraînement.

Algorithmes de Régression et Classification

- Régression Linéaire et Logistique distribuée
- Arbres de décision et ensembles (Random Forest, GBT)
- Évaluation multi-classe et gestion du déséquilibre
- Interprétabilité des modèles en environnement distribué
- Analyse des résidus et des erreurs de prédiction
- Atelier pratique : Entraînement et comparaison de modèles de classification.

[Jour 2 - Matin]

Clustering et Systèmes de Recommandation

- Apprentissage non-supervisé avec K-means
- Bisecting K-means et Gaussian Mixture Models
- Filtrage collaboratif avec ALS (Alternative Least Squares)
- Mesures de similarité à grande échelle
- Optimisation des recommandations à froid (Cold Start)
- Atelier pratique : Création d'un moteur de recommandation distribué.

[Jour 2 - Après-midi]

Optimisation des performances ML

- Impact du Shuffling sur les performances d'entraînement
- Stratégies de Caching et de Checkpointing
- Partitionnement des données et parallélisme des tâches
- Monitoring via la Spark UI et détection des bottlenecks
- Gestion des ressources mémoire pour les gros modèles
- Atelier pratique : Audit et optimisation d'un job ML lent.

Tuning et Sélection de Modèles

- Validation croisée distribuée (Cross-Validation)
- Recherche par grille avec ParamGridBuilder
- Optimisation des hyperparamètres et métriques de succès
- BinaryClassificationEvaluator vs MulticlassClassificationEvaluator
- Sauvegarde et export des meilleurs modèles
- Atelier pratique : Tuning fin d'un modèle pour maximiser la précision.

[Jour 3 - Matin]

NLP et Text Mining à grande échelle

- Prétraitement de texte : Tokenizer, StopWordsRemover
- Représentation vectorielle : TF-IDF et Word2Vec
- Analyse de sentiments et classification de texte
- Utilisation des N-grams pour le contexte
- Architectures NLP avec Spark
- Atelier pratique : Analyse sémantique sur un flux de données textuelles.

[Jour 3 - Après-midi]

Industrialisation et MLOps

- Cycle de vie du modèle avec MLflow (Tracking, Registry)
- Persistance au format ML et portabilité
- Inférence en mode Batch vs Inférence en mode Stream
- Introduction à Spark Serving et architectures temps réel
- Suivi de la dérive (drift) des modèles en production
- Atelier pratique : Tracking d'expériences et déploiement d'un modèle.

Cas pratiques et Projet de synthèse

- Conception d'une architecture ML de bout en bout
- Choix des algorithmes selon la volumétrie et la latence
- Automatisation du pipeline d'entraînement
- Bonnes pratiques de production (CI/CD pour le ML)
- Checklist de mise en service
- Atelier pratique : Projet final - Industrialisation d'une problématique métier complexe.

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.