

Machine Learning distribué avec Spark ML

Référence : IA048

Durée : 3 jours (21 heures)

Certification : Aucune

CONNAISSANCES PRÉALABLE

- Connaissances en Python/Scala, Spark et ML

PROFIL DES STAGIAIRES

- Data engineers, data scientists

OBJECTIFS

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

CERTIFICATION PRÉPARÉE

- Aucune

MÉTHODES PÉDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique numérique pendant le stage
- La formation est constituée d'apports théoriques interactifs, de démonstrations en direct, de travaux dirigés, d'ateliers collaboratifs, de mises en situation professionnelles, d'un projet fil rouge et d'études de cas réels
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

FORMATEUR

- Consultant-formateur expert en Machine Learning et plus particulièrement sur Spark ML

MÉTHODES D'ÉVALUATION DES ACQUIS

- Validation des prérequis envoyée en amont de la formation. En l'absence de réponse d'un ou plusieurs participants, un temps sera consacré en ouverture de session pour prendre connaissance du positionnement de chaque stagiaire sur les objectifs pédagogiques évalués
- Validation des compétences tout au long de la formation via exercices pratiques, corrections collectives, quiz techniques et restitution des travaux réalisés durant le projet fil rouge
- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation des compétences acquises envoyée au stagiaire
- Attestation de fin de stage adressée avec la facture

ACCESSIBILITÉ DE LA FORMATION

- Si vous avez besoin d'un aménagement pour suivre votre formation dans de bonnes conditions, nous vous invitons à contacter notre référent handicap au 01 71 19 70 30 ou à l'adresse referent.handicap@edugroupe.com. Un échange en amont nous permettra d'étudier avec vous les adaptations possibles.

CONTENU DU COURS

1. Jour 1 – Fondamentaux du Machine Learning distribué avec Spark (7h)

2. Architecture Spark et principes du ML distribué (2h)

- Rappels sur l'écosystème Apache Spark
- Architecture Driver / Executors / Cluster Manager
- Fonctionnement des DataFrames et RDD
- Introduction à Spark MLlib
- Concepts de parallélisation et distribution des calculs
- Pipelines ML et workflow distribué
- Comparaison Spark ML vs Scikit-learn
- 💡 *Mises en pratique : Exploration d'un cluster Spark ; Création d'un pipeline Spark ML simple ; Analyse des performances d'un traitement distribué*

3. Préparation et transformation des données pour le ML (3h)

- Chargement et ingestion de données massives
- Nettoyage et traitement des valeurs manquantes
- Encodage des variables catégorielles
- Standardisation et normalisation
- Feature engineering distribué
- VectorAssembler et gestion des features
- Gestion des jeux d'entraînement et de test
- Création de pipelines de transformation
- 💡 *Mises en pratique : Préparation d'un dataset Big Data réel ; Construction d'un pipeline complet de preprocessing ; Optimisation des transformations sur gros volumes*

4. Premiers modèles supervisés avec Spark ML (2h)

- Régression linéaire distribuée
- Classification avec Logistic Regression
- Arbres de décision et Random Forest
- Paramètres d'entraînement distribués
- Gestion des performances et scalabilité
- Lecture et interprétation des métriques
- Mises en pratique : Entraînement de modèles de classification ; Comparaison des performances de plusieurs algorithmes ; Analyse des résultats et des métriques

5. Jour 2 – Optimisation et évaluation des modèles ML distribués (7h)

6. Optimisation des performances Spark ML (3h)

- Optimisation mémoire et partitionnement
- Caching et persistance des données
- Broadcast variables
- Optimisation des jointures
- Gestion des ressources cluster
- Monitoring des jobs Spark
- Spark UI et analyse des performances
- Bonnes pratiques d'optimisation ML
- 💡 *Mises en pratique : Analyse de jobs lents avec Spark UI ; Optimisation d'un pipeline ML ; Comparaison avant/après optimisation*

7. Évaluation et tuning des modèles (2h30)

- Métriques de classification et régression
- Cross-validation distribuée
- TrainValidationSplit
- Grid Search distribué
- Gestion des hyperparamètres
- Prévention du surapprentissage
- Sélection et comparaison des modèles
- 💡 *Mises en pratique : Mise en œuvre d'une validation croisée ; Optimisation d'un Random Forest ; Construction d'un benchmark de modèles*

8. Machine Learning non supervisé avec Spark (1h30)

- Clustering K-Means distribué
- Réduction de dimension
- Cas d'usage Big Data
- Analyse des clusters
- Évaluation des performances de clustering
- 💡 *Mises en pratique : Segmentation d'un dataset volumineux ; Visualisation et interprétation des clusters*

9. Jour 3 – Industrialisation et déploiement des modèles ML distribués (7h)

10. Industrialisation des pipelines ML (2h30)

- Structuration des pipelines ML
- Sauvegarde et chargement de modèles
- Gestion des versions
- Gestion des dépendances
- Organisation du code Python/Scala
- Logs et traçabilité
- Intégration dans des workflows Data Engineering
- 💡 *Mises en pratique : Sauvegarde et réutilisation d'un pipeline ; Structuration d'un projet ML distribué ; Mise en place d'un workflow reproductible*

11. Déploiement et intégration des modèles (2h)

- Déploiement batch et streaming
- Introduction à Spark Structured Streaming
- Inférence temps réel
- Intégration avec Hadoop/Kafka
- Déploiement sur cluster
- Concepts MLOps appliqués à Spark
- Monitoring des modèles en production
- 💡 *Mises en pratique : Déploiement d'un modèle sur cluster Spark ; Exécution d'inférences distribuées ; Simulation d'un pipeline temps réel*

12. Atelier fil rouge : projet ML distribué complet (2h30)

- Cadrage du cas d'usage
- Préparation des données
- Construction et optimisation du modèle
- Évaluation et tuning
- Déploiement simplifié du pipeline
- Analyse des performances techniques et métier
- 💡 *Mises en pratique : Réalisation d'un projet complet en groupe ; Présentation des résultats ; Revue collective et recommandations d'amélioration*