EduGroupe
Accompagner pour réussir

Spark ML

Référence : **PYDS033** Durée : **2 jours (14 heures)** Certification : **Aucune**

Connaissances préalables

- Connaissance d'un langage de programmation comme Python, Java ou Scala.

Profil des stagiaires

- Architectes
- Chefs de projet

Objectifs

- Savoir mettre en oeuvre les outils de Machine Learning sur Spark, savoir créer des modèles et les exploiter.

Certification préparée

- Aucune

Méthodes pédagogiques

- 6 à 12 personnes maximum par cours, 1 poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions

Formateur·rice

- Consultant-Formateur expert Bigdata

Méthodes d'évaluation des acquis

- 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

Contenu du cours

1. Introduction

- Rappels sur Spark : principe de fonctionnement, langages supportés.

2. DataFrames

- Objectifs : traitement de données structurées
- L'API Dataset et DataFrames
- Optimisation des requêtes
- Mise en oeuvre des Dataframes et DataSet
- Chargement de données, pré-traitement : standardisation, transformations non linéaires, discrétisation
- Génération de données

3. Traitements statistiques de base

- Introduction aux calculs statistiques
- Paramétrisation des fonctions
- Applications aux fermes de calculs distribués
- Problématiques induites
- Approximations
- Précision des estimations
- Exemples sur Spark : calculs distribués de base : moyennes, variances, écart-type, asymétrie et aplatissement (skewness/kurtosis)

4. Machine Learning

- Apprentissage automatique : définition, les attentes par rapport au Machine Learning
- Les valeurs d'observation, et les variables cibles. Ingénierie des variables
- Les méthodes : apprentissage supervisé et non supervisé. Classification, régression
- Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistence, statistiques

5. Mise en oeuvre sur Spark

- Mise en oeuvre avec les DataFrames
- Algorithmes : régression linéaire, k-moyennes, k-voisins, classification naïve bayésienne, arbres de décision, forêts aléatoires, etc.
- Création de jeux d'essai, entraînement et construction de modèles
- Prévisions à partir de données réelles
- Travaux pratiques : régression logistiques, forêts aléatoires, k-moyennes
- Recommandations, recommendForAllUsers(), recommendForAllItems()

6. Modèles

- Chargement et enregistrement de modèles
- Mesure de l'efficacité des algorithmes. Courbes ROC. MulticlassClassificationEvaluator()
- Mesures de performance
- Descente de gradient
- Modification des hyper-paramètres
- Application pratique avec les courbes d'évaluations

7. Spark/GraphX

- Gestion de graphes orientés sur Spark
- Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes
- Travaux pratiques :exemples d'opérations sur les graphes.

8. IA

- Introduction aux réseaux de neurones
- Les types de couches : convolution, pooling et pertes
- L'approche du Deep Learning avec Spark. Deeplearning4j sur Spark

Notre référent handicap se tient à votre disposition au 01.71.19.70.30 ou par mail à referent.handicap@edugroupe.com pour recueillir vos éventuels besoins d'aménagements, afin de vous offrir la meilleure expérience possible.