Spark, mise en œuvre des outils de Machine Learning

Description :

Apache Spark est un moteur d'analyse unifié (Unified Analytics Engine), créé pour le traitement rapide des données. Ce framework open source de calcul distribué permet l'analyse et le traitement de données à grandes échelles. Le framework permet le développement d'applications de traitement hautement performantes.

Profil de l'intervenant :

Consultant formateur expert Big Data

Objectifs :

A l'issue de la formation, les stagiaires sauront :

- Mettre en oeuvre les outils de Machine Learning sur Spark
- Créer des modèles et les exploiter.

Publics :

Développeurs, Chefs de projet, Data Scientists, Développeurs, Architectes...

Durée :

2 jours

Prérequis :

Avoir suivi le formation Spark Traitement des données ou équivalent et bonnes connaissances du langage Python, Java ou Scala

Méthode pédagogique de cette formation :

La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions. Remise d'une documentation pédagogique papier ou numérique pendant le stage 6 à 8 personnes maximum par cours.

Méthode d'évaluation des acquis de la formation :

Auto évaluation des acquis par le stagiaire via une questionnaire. Attestation de fin de stage signée remise au stagiaire en fin de formation

Information handicap :

Cette formation est accessible aux personnes en situation de handicap. Chaque situation étant unique, nous vous demandons de préciser à l'inscription votre handicap. Nous pourrons ainsi confirmer l'ensemble des possibilités d'accueil et vous permettre de suivre la formation dans les meilleures conditions en accord avec votre employeur. Pour toutes informations complémentaires, nous vous conseillons la structure suivante https://www.agefiph.fr/

Programme de cette formation :

Introduction

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

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.

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)

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.

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.
- Atelier : régression logistiques, forêts aléatoires, k-moyennes.
- Recommandations, recommendForAllUsers(), recommendForAllItems();

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.

Spark/GraphX

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

IA

- Introduction aux réseaux de neurones.
- Les types de couches : convolution, pooling et pertes.