



— Phirio —

Spark ML

DS033

Durée: 2 jours

1 670 €

20 au 21 janvier

28 au 29 avril

7 au 8 juin

6 au 7 octobre

11 au 12 décembre

Public :

Chefs de projet, architectes

Objectifs :

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

Connaissances préalables nécessaires :

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

Programme :

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 formes 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 persistance, statistiques.



— Phirio —

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.

L'approche du Deep Learning avec Spark. `Deeplearning4j` sur Spark.