

Machine Learning avec Spark (Spark ML)

Utiliser le Framework de la fondation Apache

 Présentiel ou en classe à distance



3 jours (21 h)

Prix inter : 2.190,00 € HT
Forfait intra : 8.190,00 € HT

Réf.: BI052

Apparu dans la version 1.2 de Spark, la bibliothèque d'apprentissage automatique (Machine Learning) contient tous les algorithmes nécessaires à l'exploration de données (Data Mining) par apprentissage statistique. L'ensemble des API proposés par Spark ML permet de créer et régler des flux de travaux complets (Pipelines) combinant plusieurs algorithmes d'apprentissage automatique. Durant cette formation de 3 jours, les participants découvriront l'étendue des possibilités offertes par Spark ML et s'approprieront les différentes méthodes de mises en oeuvre pour un déploiement adapté aux besoins de leur entreprise ou organisations.

A qui s'adresse cette formation ?



Pour qui

- Chefs de projet et spécialistes big data souhaitant mettre en oeuvre Spark ML



Prérequis

- Avoir suivi la formation "Les bases de l'apprentissage Machine (Machine Learning)" (BI105)
- Disposez-vous des connaissances nécessaires pour suivre cette formation ? Testez-vous !

Programme

1 - L'écosystème SPARK et l'apprentissage Big Data

- Enjeux machine learning et Big Data
- L'écosystème Apache Spark
- Les différentes briques de base
- Focus SPARK SQL
- Dataframes et Datasets
- Lab : Mise en oeuvre de l'écosystème SPARK pour l'apprentissage machine Big Data

2 - Le chargement de données d'entraînement massives

- Chargement générique de données
- Chargement de fichiers de formats spécifiques
- Interrogation de bases HIVE
- Interrogation de bases externes
- Lab : Chargement de données de sources diverses sur un cluster SPARK

3 - L'exploration de données d'entraînement massives

- Réalisation de statistiques de base avec SPARK
- Exploitation des librairies graphiques statistiques dans un cadre Big Data
- Lab : Exploration de données d'entraînement sur un cas concret

4 - Le "Pipelining"

- Le concept de Pipeline Spark
- Les composants d'un Pipeline
- Le fonctionnement d'un Pipeline
- La gestion des paramètres
- Persistance et chargement de Pipelines
- Lab : Création d'un premier pipeline d'apprentissage machine avec SPARK

5 - Le prétraitement et l'ingénierie des variables prédictives

- Extraction de variables prédictives
- Transformation de variables
- Sélection de variables prédictives
- Hachage de variables
- Lab : Prétraitement et ingénierie des variables prédictives sur un cas concret

6 - La création de modèles d'apprentissage Big Data

- Classification de données massives
- Régression de données massives
- Clustering de données massives
- Systèmes de recommandation Big Data
- Règles d'association de données massives
- Lab : Réalisation de modèles d'apprentissage sur des cas concrets Big Data

7 - L'optimisation du réglage des modèles d'apprentissage

- Réglage des hyper-paramètres des modèles
- Validation croisée
- Séparation des données (entraînement, validation)
- Lab : Optimisation du réglage de modèles d'apprentissage sur cas concrets

8 - Déploiement de modèles d'apprentissage Big Data

- Création d'application prédictive en batch
- Création d'application prédictive en streaming
- Mise en oeuvre concrète sur un cluster Big Data
- Bonnes pratiques de déploiement
- Lab : Création d'applications en batch et en streaming sur cas concrets



Les objectifs de la formation

- Être capable de mettre en oeuvre Spark ML pour de l'apprentissage machine et l'analyse de données
- Comprendre comment récupérer les données et les préparer pour un traitement efficient
- Savoir modéliser les données et optimiser les modèles
- Être en mesure d'industrialiser et de déployer une application de prédiction



Evaluation

- Pendant la formation, le formateur évalue la progression pédagogique des participants via des QCM, des mises en situation et des travaux pratiques. Les participants passent un test de positionnement avant et après la formation pour valider leurs compétences acquises.



Les points forts de la formation

- Un tour d'horizon de l'écosystème Spark ML.
- Un focus sur la mise en oeuvre d'algorithmes à l'aide de la brique Machine Learning de Spark qui permet de réaliser de l'apprentissage sur un socle Big Data.
- Des ateliers pour expérimenter le framework sur des cas concrets.



Dates et villes 2026 - Référence BI052



Dernières places disponibles



Session garantie

A distance

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

du 30 mars au 1 avr.

du 24 août au 26 août

Toulouse

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Aix-en-Provence

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Strasbourg

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Marseille

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Sophia Antipolis

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Lille

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Rouen

du 19 janv. au 21 janv.

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Nantes

du 30 mars au 1 avr.

du 1 juin au 3 juin

du 24 août au 26 août

Paris

du 30 mars au 1 avr.

du 24 août au 26 août

du 1 juin au 3 juin

du 2 nov. au 4 nov.

Rennes

du 30 mars au 1 avr.

du 1 juin au 3 juin

du 24 août au 26 août

Lyon

du 30 mars au 1 avr.

du 24 août au 26 août

du 2 nov. au 4 nov.

Bordeaux

du 30 mars au 1 avr.

du 24 août au 26 août

du 2 nov. au 4 nov.