FORMATION · BIG DATA

Spark et Machine Learning

Durée de la formation : 3 jours
Tarif : 1 990 € HT par participant
Réf : BI190
Délai d'accès : 2 à 4 semaines
Niveau : Avancé

Prérequis

Bonne pratique de Python et du SQL. Des notions de statistiques et de machine learning (régression, classification) sont recommandées.

Public concerné

Data engineers, data scientists, développeurs

Objectifs pédagogiques

- Comprendre le calcul distribué et l'architecture de Spark
- Manipuler de gros volumes avec les DataFrames et Spark SQL
- Construire des pipelines de machine learning avec Spark ML
- Optimiser et industrialiser les traitements

Objectifs opérationnels

- Charger, nettoyer et transformer des données avec PySpark
- Préparer des variables (features) à grande échelle
- Entraîner, évaluer et régler un modèle
- Sauvegarder un pipeline et l'appliquer à de nouvelles données
- Diagnostiquer un traitement lent dans l'interface Spark

Méthodes mobilisées et modalités d'évaluation

Apports théoriques, démonstrations et nombreux travaux pratiques sur des cas concrets. À l'issue de chaque module, le formateur évaluera chacun des participants en fonction des cas pratiques et exercices effectués.

Spark et Machine Learning

Apache Spark permet de traiter de très gros volumes de données et d'y entraîner des modèles de machine learning en répartissant les calculs sur plusieurs machines.

Cette formation vous apprend à manipuler les données avec PySpark et Spark SQL, puis à construire, évaluer et déployer des pipelines de machine learning avec Spark ML.

Les exercices sont réalisés en Python, sur des jeux de données de taille réaliste.

Inscription

Contactez-nous au 01 86 95 71 60 ou à l'adresse formation@adbi.fr pour connaître les dates de notre prochaine session. Inscription au plus tard 10 jours avant le début de la session.

Accessibilité

Nos salles de formation sont accessibles aux personnes à mobilité réduite. N'hésitez pas à nous informer si vous avez des besoins particuliers et s'il y a des besoins d'aménagements spécifiques lors de votre inscription.

Modes de participation

Nos formations sont proposées sous 3 formats :

- Chez nous, près de la place d'Italie (Paris 13e)
- Chez vous, dans vos locaux
- À distance, en classe virtuelle

Attestation

Une attestation de fin de formation sera remise à tous les participants à l'issue de leur parcours.

Programme :

Jour 1

- **Spark et le calcul distribué**
  - Pourquoi distribuer les calculs : limites d'une seule machine
  - Architecture : driver, exécuteurs, partitions, tâches
  - Modes d'exécution : local, cluster, plateformes managées
- **Manipuler les données**
  - DataFrames : lecture de CSV, JSON, Parquet et Delta
  - Transformations, actions et évaluation paresseuse
  - Spark SQL, fonctions de fenêtrage et jointures
  - Travaux pratiques : préparation d'un jeu de données volumineux

Jour 2

- **Machine learning avec Spark ML**
  - Rappels : apprentissage supervisé et non supervisé
  - Transformers, estimators et pipelines
  - Préparation des variables : encodage, normalisation, assemblage
- **Modèles et évaluation**
  - Régression, classification, clustering
  - Validation croisée et recherche d'hyperparamètres
  - Métriques d'évaluation et interprétation
  - Travaux pratiques : un modèle de prédiction de bout en bout

Jour 3

- **Performances**
  - Partitionnement, cache et persistance
  - Lire l'interface Spark : étapes, mélanges (shuffles), déséquilibres
  - Bonnes pratiques d'écriture des traitements
- **Industrialiser**
  - Sauvegarder et recharger un pipeline
  - Suivi des expériences et des modèles (MLflow)
  - Traitements en flux avec Structured Streaming : principes
- **Atelier de synthèse**
  - De la donnée brute au modèle déployé en lot