Formation Spark avec PySpark en Python

Description

L'environnement Apache Spark est aujourd'hui central dans l'approche big data de la donnée. Cette formation spark avec PySpark en python vous permet de maîtriser les principes de l'environnement Apache Spark et l'utilisation du package PySpark pour gérer des données, appliquer des algorithmes de machine learning ou accélérer vos processus. Cette formation Spark s'adresse à tous ceux qui veulent manipuler Apache Spark en utilisant le langage python.

Deux jours intensifs basés sur des applications réelles pour la préparation, le traitement et l'analyse des données dans l'environnement Apache Spark.
Formation Spark en petits groupes avec maximum 6 participants pour plus d'échanges avec nos formateurs !

PLUS D'INFORMATIONS :
Formation disponible en intra ou en inter-entreprises
Inscription : https://www.stat4decision.com/fr/formations/formation-spark-avec-python/

Durée

3.00 jours (21.00 heures)

Profils des stagiaires et prérequis

Profils :

- Data scientist désirant monter en compétence sur l'utilisation d'Apache Spark
- Développeur étant amené à automatiser des traitements de données massives

Prérequis :

- Connaissances de base en traitement de données (statistique et tables de données)
- Une connaissance de Python est fortement conseillée.

Objectifs pédagogiques

- Comprendre l'environnement Apache Spark
- Savoir utiliser le package PySpark pour communiquer avec Spark
- Maîtriser l'utilisation de Spark SQL
- Maîtriser l'utilisation de Spark.ml

Programme détaillé

- Rappels sur Python et la manipulation des données
- Introduction à l'environnement Big Data et à Spark
    - Pour qui ? Pour quoi faire ? Comment ?
    - Comment installer Apache Spark
    - Pyspark un package Python pour gérer votre environnement Apache Spark
    - Quelle infrastructure pour utiliser Spark en entreprise ?

    - Les principes de l'environnement : RDD, DataFrame, DataSet…
- Installation de Spark :
    - Sur une infrastructure distribuée
    - En local
    - En cloud (exemples avec Amazon AWS et Microsoft Azure)
- Spark pour la manipulation des données
    - Utilisation de SparkSQL et des DataFrames pour manipuler des données
    - Charger des données depuis Hadoop, depuis des fihiers csv…
    - Transformer des données (création de DataFrames, ajout de colonnes, filtres…)
    - Cas pratiques de chargement et de modifications de données avec Spark et PySpark
- L'utilisation de spark.ml pour le machine learning
    - Apprentissage supervisé : Forêts aléatoires avec Spark
    - Mise en place d'un outil de recommandation
    - Traitement de données textuelles
    - Automatiser vos analyses avec des pipelines
- Introduction et utilisation de Spark Streaming avec PySpark

Organisation de la formation

Moyens pédagogiques et techniques

- Accueil des stagiaires dans une salle dédiée à la formation.
- Documents supports de formation projetés.
- Exposés théoriques
- Etude de cas concrets
- Mise à disposition en ligne de documents supports à la suite de la formation.
- Outils utilisés : Apache Spark en local et en cloud.
  Nous utilisons un cluster basé sur Google Cloud Platform (GCP).

Dispositif de suivi de l'exécution et d'évaluation des objectifs de la formation

- Mises en situation.
- Cas pratiques validés par le formateur.
- Évaluation des connaissances (quizz / tests).