Introduction à Spark

Formation Informatique / Bureautique / Sécurité

Apache Spark est un framework open source de calcul distribué en mémoire permettant le traitement de grands volumes. Le but de cette formation est de présenter le framework Spark et d'apprendre à l'utiliser avec le langage Python pour traiter des problèmes de Big Data.

OBJECTIFS

- Comprendre le principe de fonctionnement de Spark
- Apprendre à utiliser l'API PySpark pour interagir avec Spark en Python
- Apprendre à utiliser les méthodes de Machine Learning avec la librairie MLlib de Spark
- Apprendre à traiter les flux de données avec Spark Streaming
- Apprendre à manipuler les données avec Spark SQL

PUBLIC

Développeur, Data Analyst, Data Scientists, Architectes Big Data et toute autre personne souhaitant acquérir des connaissances dans le domaine de la Data Science et sur Spark

PRE-REQUIS

- Une première expérience en programmation Python, avoir des connaissances en SQL, avoir des connaissances en mathématiques et statistiques.

PROGRAMME

Introduction à Hadoop

- L'ère du Big Data
- Architecture et composants de la plateforme Hadoop
- HDFS
- NameNode / DataNode / RessourceManager
- Paradigme MapReduce et YARN

Introduction à Spark

- Qu'est-ce que Spark ?
- Spark vs MapReduce
- Fonctionnement : RDD, DataFrames, Data Sets
- Comment interagir avec Spark
- PySpark : programmer avec Spark en Python

Manipulation des données

- Formats basiques (fichiers textes, JSON, CSV, SequencesFiles, fichiers compressés)
- Interagir avec des sources de données externes : connecteurs Hive, JDC, Hbase, ElasticSearch, …

Spark Streaming

- Introduction à Spark Streaming
- La notion de « DStream »
- Principales sources de données
- Utilisation de l'API
- Manipulation des données

Spark SQL

- Initiation à Spark SQL
- Création de DataFrames
- Manipulation des DataFrames (opérations basiques, agrégations & Groupby, Missing Data)
- Chargement et stockage de données (avec Hive, JSON, etc…)

Spark ML avec MLlib

- Modélisation Statistique & Apprentissage
- Types de données (Vector / LabeledPoint / Model)
- Préparation des données
- Utilisation d'algorithme de MLlib (k-means / Régression logistique / arbre de discrimination / forêt aléatoire)
- Exemple de création d'un modèle et de son évaluation avec Spark MLlib sur un jeu de données

GraphX et GraphFrames

- Présentation de GraphX
- Principe de création des graphes
- API GraphX
- Présentation GraphFrames
- GraphX vs GraphFrames

Travaux pratiques

A retenir

Durée : **3 jours** soit 21h.
Réf. **GKDSP**

01 42 93 52 72

Dates des sessions

Cette formation est également proposée en formule INTRA-ENTREPRISE.

Inclus dans cette formation

Coaching Après-COURS

Pendant 30 jours, votre formateur sera disponible pour vous aider. CERTyou s'engage dans la réalisation de vos objectifs.

100% SATISFACTION GARANTIE

Votre garantie 100% SATISFACTION

Notre engagement 100% satisfaction vous garantit la plus grande qualité de formation.

- **Alternance d'apports théoriques, d'exercices pratiques et de mise en situation sous forme de travaux pratiques permettant de tester les différentes notions abordées avec le langage Python**

Horaires, Planning et Déroulement de cette formation

Horaires

- Formation de 9h00 (9h30 le premier jour) à 17h30.
- Deux pauses de 15 minutes le matin et l'après-midi.
- 1 heure de pause déjeuner

DEROULEMENT

- Les horaires de fin de journée sont adaptés en fonction des horaires des trains ou des avions des différents participants.
- Une attestation de suivi de formation vous sera remise en fin de formation.
- Cette formation est organisée pour un maximum de 14 participants.

PROCHAINES FORMATIONS

Réussir la Certification Gestion de Projet PMP du PMI
Réussir la Certification PRINCE2 Foundation
Réussir les Certifications PRINCE2 Foundation et PRINCE2 Practitioner
Réussir la Certification ITIL Foundation
Réussir la Certification Agile certifié SCRUM Master
Réussir les Certifications TOGAF Certified et TOGAF Foundation

Retrouvez cette formation sur notre site :
Introduction à Spark