Introduction à Spark

Durée: 3 Jours Réf de cours: GKDSP

Résumé:

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.

Public visé:

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

Objectifs pédagogiques:

- 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

Pré-requis:

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

Contenu:

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

- 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

Méthodes pédagogiques :

Support de cours remis aux participants.

Autres moyens pédagogiques et de suivi:

• Compétence du formateur : Les experts qui animent la formation sont des spécialistes des matières abordées et ont au minimum cinq ans d'expérience d'animation. Nos équipes ont validé à la fois leurs connaissances techniques (certifications le cas échéant) ainsi que leur compétence pédagogique.
• Suivi d'exécution : Une feuille d'émargement par demi-journée de présence est signée par tous les participants et le formateur.
• Modalités d'évaluation : le participant est invité à s'auto-évaluer par rapport aux objectifs énoncés.
• Chaque participant, à l'issue de la formation, répond à un questionnaire de satisfaction qui est ensuite étudié par nos équipes pédagogiques en vue de maintenir et d'améliorer la qualité de nos prestations.

Délais d'inscription :

•Vous pouvez vous inscrire sur l'une de nos sessions planifiées en inter-entreprises jusqu'à 5 jours ouvrés avant le début de la formation sous réserve de disponibilité de places et de labs le cas échéant.
•Votre place sera confirmée à la réception d'un devis ou ""booking form"" signé. Vous recevrez ensuite la convocation et les modalités d'accès en présentiel ou distanciel.
•Attention, si vous utilisez votre Compte Personnel de Formation pour financer votre inscription, vous devrez respecter un délai minimum et non négociable fixé à 11 jours ouvrés.