Data Analyse Pig, Hive et Spark

Durée: 3 Jours **Réf de cours: GKDAPHS**

Résumé:

Cette formation permet aux participants d'acquérir les compétences et connaissances nécessaires pour utiliser les outils permettant de traiter et d'analyser des données sur Hadoop.
Elle leur permettra de développer des compétences en analyse de données en se focalisant sur Pig, Hive et Spark.
Mise à jour : 26.01.2023

Public visé:

Cette formation s'adresse à toute personne souhaitant manipuler et analyser des données dans un système Hadoop.

Objectifs pédagogiques:

- Comprendre ce que sont Hadoop et YARN
- Pouvoir manipuler des données sous Hadoop
- Savoir manipuler les données PIG
- Savoir analyser les données avec HIVE

Pré-requis:

Avoir des connaissances sur les systèmes d'information, les bases de données et les concepts de programmation.

Contenu:

Introduction

- Introduction au Big Data - Comprendre les concepts clés et les enjeux du Big Data
- Introduction à Hadoop – Principales distributions de Hadoop
- La plateforme Hadoop

Architecture et composants de la plateforme Hadoop

- HDFS
- NameNode / DataNode / RessourceManager
- Paradigme MapReduce et YARN
- Les technologies émergentes

Traitement des données avec Pig

- Description et caractéristiques de Pig : Présentation Pig, Différence entre Pig et MapReduce, Cas d'utilisation de Pig
- Traitement des données : Modélisation des données, Programmation avec Pig Latin, Transformations dans la syntaxe Pig Latin, Fonctions de chargement et de stockage
- Travaux pratiques

Requêtage des données avec Hive

- Description et caractéristiques de Hive
- Utilisation de Hcatalog
- Analyse des données avec Hive
- Management des données Hive : Formats de données Hive, Création des bases de données et des tableaux de management, Tableaux auto-managés, Simplification des requêtes avec Views, Stockage des résultats de requêtes, Contrôle 'accès aux données
- Traitement de texte avec Hive : Fonctions String, Utilisation des expressions habituelles dans Hive

Apache Spark SQL

- Présentation générale
- Caractéristiques – Architecture
- Les bases de Spark
- DataFrame et DataSets
- Les RDD
- Le SQL Contexte
- Opérations sur le DataFrames et les DataSets
- Comparaison entre Spark SQL et Hive

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.