Formation Hadoop pour le traitement distribué des Big Data

APERÇU / PLAN DU COURS

DESCRIPTION DU COURS

Brève description :

Les entreprises produisent d'énormes quantités de données chaque jour. Ces données sont stockées puis traitées et analysées pour en tirer de la valeur. Grâce aux plateformes de stockage et de traitements distribués de type Hadoop, il est devenu plus facile pour les ingénieurs de répondre aux problématiques du Big Data avec une grande efficacité et à un coût réduit.

Ce séminaire vous offre l'occasion de vous essayer au traitement distribué de données massives via la plateforme Hadoop et ses outils comme Hive et Spark.

Objectifs	Les participants seront capables de : 1. Stocker des données sur Hadoop 2. Décrire les différentes étapes du modèle de programmation MapReduce 3. Charger des données sur Hive et les traiter à l'aide du langage HiveQL 4. Charger des données dans Spark et les traiter à l'aide du langage Spark SQL
Dates	**14 au 17 mars 2023**
Durée	**4 jours**

PUBLIC CIBLE

- Développeurs
- Experts en business intelligence
- Analystes de données

PRE-REQUIS

- Notions en administration système sous Linux.
- Notions en programmation orientée objet
- Connaissances de base du langage SQL

AGENDA ET CONTENU DU COURS

Contenu Pédagogique	**L'écosystème du Big Data** ▪ Les origines et caractéristiques du Big Data ▪ Les avancées technologiques ▪ L'informatique parallèle ▪ Aperçu des plateformes Big Data ▪ Aperçu des services cloud de AWS ▪ **Démonstration** : Requêtes SQL sur des données massives avec Amazon Athena **Le paradigme MapReduce** ▪ Historique ▪ Le pattern MapReduce ▪ Cas d'utilisation de MapReduce ▪ **Travaux pratiques :** programmation MapReduce **Hadoop** ▪ Historique ▪ Socle technique ▪ Hadoop – YARN ▪ Ecosystème Hadoop ▪ **Travaux pratiques** : programmation Hadoop MapReduce **Apache Hive** ▪ Présentation ▪ Langage de définition de données ▪ Interrogation des données ▪ Gestion des données complexes ▪ **Travaux pratiques** : analyse de données avec HiveQL **Spark SQL** ▪ Apache Spark ▪ Les RDD ▪ Installation de Apache Spark en local ▪ **Travaux pratiques** : calcul distribué avec PySpark ▪ Spark SQL et DataFrames ▪ **Travaux pratiques** : analyse de données avec Spark SQL

MODES D'ANIMATION PEDAGOGIQUE

- Exposés théoriques ;
- Questions-réponses ;
- Exercices d'application

- Démonstrations
- Travaux pratiques

SUPPORTS PEDAGOGIQUES

- Support de cours ;
- Les Travaux Pratiques : sur le poste de travail et dans le Cloud de AWS via le service Amazon Elastic Map Reduce (EMR) ;
- Le langage de programmation : Python.