Formation Hadoop – Développement

Objectifs de la formation Développement Hadoop

La gestion des ensembles de données volumineux offre aux entreprises de toutes tailles de nouvelles opportunités et de nouveaux défis à relever. Au cours de cette formation, vous allez acquérir les compétences pratiques de programmation nécessaires pour développer des solutions compatibles avec la plateforme Hadoop d'Apache grâce auxquelles vous pourrez traiter efficacement différents types de Big Data.

Lors de cette formation, vous utiliserez plusieurs produits dédiés au Big Data, Apache Hadoop, MapReduce, le système de fichiers distribué Hadoop (HDFS), HBase, Hive et Pig. Vous étudierez aussi d'autres composants de l'écosystème.

Concrètement après avoir suivi ce cours vous serez en mesure de:

- Développer des algorithmes parallèles efficaces avec MapReduce
- Mettre en œuvre des tâches Hadoop pour extraire des éléments pertinents d'ensembles de données volumineux et variés et apporter ainsi de la valeur à votre entreprise
- Créer, personnaliser et déployer des tâches MapReduce pour synthétiser les données
- Charger des données non structurées des systèmes HDFS et HBase

À qui s'adresse cette formation ?

Public :

Ce cours s'adresse aux Chefs de projets, Développeurs, Data-scientists, et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop.

Prérequis :

Pour suivre cette foramtion dans les meilleures conditions possibles, il vous faut avoir une certaine connaissance d'un langage de programmation objet.

Contenu du cours Développement Hadoop

Introduction

- ✓ Les fonctionnalités du framework Hadoop
- ✓ Le projet et les modules : Hadoop Common, HDFS, YARN, MapReduce
- ✓ Utilisation de yarn pour piloter les jobs mapreduce.

MapReduce

- ✓ Principe et objectifs du modèle de programmation MapReduce.
- ✓ Fonctions map() et reduce().
- ✓ Couples (clés, valeurs).
- ✓ Implémentation par le framework Hadoop.
- ✓ Etude de la collection d'exemples.

Travaux Pratiques:

Rédaction d'un premier programme et exécution avec Hadoop.

Programmation

- ✓ Configuration des jobs, notion de configuration.
- ✓ Les interfaces principales : mapper, reducer,
- ✓ La chaîne de production : entrées, input splits, mapper, combiner, shuffle/sort, reducer, sortie.
- ✓ Partitioner, outputcollector, codecs, compresseurs..
- ✓ Format des entrées et sorties d'un job MapReduce : InputFormat et OutputFormat.

Travaux Pratiques:

Type personnalisés : création d'un writable spécifique. Utilisation. Contraintes.

Outils complémentaires

- ✓ Mise en oeuvre du cache distribué.
- ✓ Paramétrage d'un job : ToolRunner, transmission de propriétés.
- ✓ Accès à des systèmes externes : S3, hdfs, har, ...

Travaux Pratiques:

Répartition du job sur la ferme au travers de yarn.

Streaming

- ✓ Définition du streaming map/reduce.
- ✓ Création d'un job map/reduce en python.
- ✓ Répartition sur la ferme.
- ✓ Avantage et inconvénients.
- ✓ Liaisons avec des systèmes externes.
- ✓ Introduction au pont HadoopR

Travaux Pratiques:

Suivi d'un job en streaming.

Pig

- ✓ Pattern et best practices Map/reduce.
- ✓ Introduction à Pig.
- ✓ Caractéristiques du langage : latin.

Travaux Pratiques:

Installation/lancement de pig.Ecriture de scripts simples pig.

- ✓ Les fonctions de bases.
- ✓ Ajouts de fonctions personnalisées.
- ✓ Les UDF.
- ✓ Mise en oeuvre.

Hive

- ✓ Simplification du requêtage.
- ✓ Syntaxe de base.

Travaux Pratiques:

- ✓ Création de tables. Ecriture de requêtes.
- ✓ Comparaison pig/hive.

Securité en environnement Hadoop

- ✓ Mécanisme de gestion de l'authentification.

Travaux Pratiques:

Configuration des ACLs