Tél : +352 27 39 351
Fax : +352 27 39 52 68
Email : formation@oxiane.lu

Hadoop

Développement avec MapReduce

Apache Hadoop est une framework open-source pour le stockage et le traitement distribué de Big Data sur des machines conventionnelles.
Au lieu d'amener les données vers la machine contenant le programme, Hadoop et ses algorithmes basés sur Map/Reduce permettent d'amener la logique de traitement là où les données sont pour les traiter de façon parallèle en diminuant le volume transitant sur le réseau.

À l'issue de cette formation, vous serez capable de mettre en place des clusters Hadoop et de concevoir des algorithmes Map/Reduce .

Détails

- Code : DB-HB
- Durée : 3 jours (21 heures)

Public
- Developpeurs

Pré-requis
- Avoir suivi le stage Java : les bases et avoir mis en pratique les concepts enseignés

Objectifs
- Se familiariser avec l'écosystème Hadoop
- Concevoir, Executer et tester des programmes écrits avec Map/Reduce
- Entrer et sortir des données de formats variés pour les traiter avec Hadoop
- Utiliser Hive pour pouvoir interroger le système de fichiers HDFS avec un langage analogue à SQL
- Utiliser Pig pour produire facilement des programmes Map-Reduce en langage de haut niveau

Programme

Introduction
- Problème des systèmes traditionnels à grande échelle
- Qu'est-ce qu'Hadoop
- Quels problèmes peut-on résoudre avec Hadoop

Concepts fondamentaux et HDFS
- Le projet Hadoop et ses composants
- HDFS, le système de fichiers distribué

MapReduce
- Utilisation de MapReduce
- Analyse de données avec les outils Unix
- Analyse de données avec Hadoop
- Mappers
- Reducers
- Combiners

Clusters Hadoop et écosystème
- Cluster Hadoop: concepts
- Jobs et tasks
- Systèmes de fichiers
- Programmation distribuée: MapReduce, Pig et Spark
- Bases NoSQL: HBase et Cassandra
- Accès SQL à Hadoop: Hive
- Ingestion de données: Flume, Kafka et Sqoop
- Planification des workflows Hadoop: Oozie
- Machine Learning: Mahout et Weka

HDFS
- Motivations et design
- Blocs et noeuds
- Interface en ligne de commande
- Interface Java
- Flux de données
- HBase

Mise en place de clusters Hadoop
- Spécification du cluster
- Configuration et Installation
- Configuration d'Hadoop
- Configuration d'HDFS
- Monitoring et logging
- Maintenance

Entrer et sortir des données d'Hadoop
- ingress et egress: éléments-clés
- Entrer des données de log avec Apache Flume
- Programmation des entrées de données avec Oozie
- Importer/Exporter des données depuis des SGBDR avec Sqoop
- MapReduce et XML
- MapReduce et JSON
- MapReduce et formats personnalisés

L'API Hadoop pour Java

Tests unitaires avec Hadoop

- Pertinence des tests unitaires
- Tester les mappers et reducers: JUnit et MRUnit
- Execution des tests
- LocalJobRunner

Pig

- Faciliter l'écriture de programmes MapReduce avec Pig
- Installation et Exécution
- Le langage de script: Pig Latin
- Fonctions Utlisateurs (UDF)
- Opérateurs de traitement de données

Hive

- Interroger et gérer de larges volumes de données avec Hive
- Installation
- Exécution
- Comparaison avec les bases de données traditionnelles
- HiveQL
- Tables
- Interrogation des données
- Fonction utilsateurs

Réalisation d'une application complète avec Hadoop, Pig et Hive

Modalités

- Type d'action :Acquisition des connaissances
- Moyens de la formation :Formation présentielle – 1 poste par stagiaire – 1 vidéo projecteur – Support de cours fourni à chaque stagiaire
- Modalités pédagogiques :Exposés – Cas pratiques – Synthèse
- Validation :Exercices de validation – Attestation de stages