

# Hadoop - Développement

3 j (21 heures)

Ref : BDDH

## Public

Chefs de projets, développeurs, data scientists et toute personne souhaitant comprendre les techniques de développement avec MapReduce dans l'environnement Hadoop

## Pré-requis

Avoir la connaissance d'un langage de programmation objet comme Java

## Moyens pédagogiques

Formation réalisée en présentiel ou à distance selon la formule retenue  
Exposés, cas pratiques, synthèse, assistance post-formation pendant trois mois  
Un poste par stagiaire, vidéoprojecteur, support de cours fourni à chaque stagiaire

## Modalités de suivi et d'évaluation

Feuille de présence émargée par demi-journée par les stagiaires et le formateur  
Exercices de mise en pratique ou quiz de connaissances tout au long de la formation permettant de mesurer la progression des stagiaires  
Questionnaire d'évaluation de la satisfaction en fin de stage  
Auto-évaluation des acquis de la formation par les stagiaires  
Attestation de fin de formation

## Objectifs

- Connaître les principes du framework Hadoop
- Mettre en œuvre les fonctionnalités du framework Hadoop
- Développer des algorithmes parallèles efficaces avec MapReduce
- Utiliser la technologie MapReduce pour paralléliser des calculs

## Programme détaillé

### INTRODUCTION

---

- Les fonctionnalités du framework Hadoop
- Le projet et les modules
- Hadoop Common
- HDFS

YARN

Spark

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

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

OutputFormat

Type personnalisé : création d'un Writable spécifique

Utilisation

Contraintes

## OUTILS COMPLEMENTAIRES

---

Mise en oeuvre du cache distribué

Paramétrage d'un job

ToolRunner

Transmission de propriétés

Accès à des systèmes externes

S3

HDFS

HAR

Répartition du job sur la ferme au travers de YARN

## **STREAMING**

---

Définition du streaming MapReduce

Création d'un job MapReduce dans Python

Répartition sur la ferme

Avantages et inconvénients

Liaisons avec des systèmes externes

Introduction au pont Hadoop

Suivi d'un job en streaming

## **PIG**

---

Pattern et best practices MapReduce

Introduction à Pig

Caractéristiques du langage : latin

Installation / lancement

Ecriture d'un script Pig

Les fonctions de bases

Ajouts de fonctions personnalisées

Les UDF

Mise en oeuvre

## **HIVE**

---

Simplification du requêtage

Syntaxe de base

Création de tables

Ecriture de requêtes

Comparaison Pig / Hive

## **SECURITE EN ENVIRONNEMENT HADOOP**

---

Mécanisme de gestion de l'authentification

Configuration des ACL

---