

## Hadoop, installation et administration

**SII-315**    **4 Jours (28 Heures)**

### Description

La plateforme Apache Hadoop permet de faciliter la création d'applications distribuées. Ce stage vous permettra de comprendre son architecture et vous donnera les connaissances nécessaires pour installer, configurer et administrer un cluster Hadoop. Vous apprendrez également à l'optimiser et le maintenir dans le temps.

### À qui s'adresse cette formation ?

#### Pour qui

Administrateurs de cluster Hadoop, développeurs.

#### Prérequis

Aucune

### Les objectifs de la formation

- Découvrir les concepts et les enjeux liés à Hadoop
- Optimiser la plateforme
- Comprendre le fonctionnement de la plateforme et de ses composants
- Installer la plateforme et la gérer

### Programme de la formation

#### Présentation du Framework Apache Hadoop

- Enjeux du Big Data et apports du framework Hadoop.
- Présentation de l'architecture Hadoop.
- Description des principaux composants de la plateforme Hadoop.
- Présentation des distributions principales du marché et des outils complémentaires (Cloudera, MapR, Dataiku).
- ).
- Avantages/inconvénients de la plateforme.

## Préparations et configuration du cluster Hadoop

- Principes de fonctionnement de Hadoop Distributed File System (HDFS).
- Principes de fonctionnement de MapReduce.
- Design "type" du cluster.
- Critères de choix du matériel.
- Travaux pratiques Configuration du cluster Hadoop.

## Installation d'une plateforme Hadoop

- Type de déploiement.
- Installation d'Hadoop.
- Installation d'autres composants (Hive, Pig, HBase, Flume).
- ).
- Travaux pratiques Installation d'une plateforme Hadoop et des composants principaux.

## Gestion d'un cluster Hadoop

- Gestion des noeuds du cluster Hadoop.
- Les TaskTracker, JobTracker pour MapReduce.
- Gestion des tâches via les schedulers.
- Gestion des logs.
- Utiliser un manager.
- Travaux pratiques Lister les jobs, statut des queues, statut des jobs, gestion des tâches, accès à la web UI.

## Gestion des données dans HDFS

- Import de données externes (fichiers, bases de données relationnelles) vers HDFS.
- Manipulation des fichiers HDFS.
- Travaux pratiques Importer des données externes avec Flume, consulter des bases de données relationnelles avec Sqoop.

## Configuration avancée

- Gestion des autorisations et de la sécurité.
- Reprise sur échec d'un name node (MRV1).
- NameNode high availability (MRV2/YARN).
- Travaux pratiques Configuration d'un service-level authentication (SLA) et d'un Access Control List (ACL).

## Monitoring et optimisation Tuning

- Monitoring (Ambari, Ganglia.
- ).
- Benchmarking/profiling d'un cluster.
- Les outils Apache GridMix, Vaaidya.
- Choisir la taille des blocs.
- Autres options de tuning (utilisation de la compression, configuration mémoire.
- ).
- Travaux pratiques Appréhender les commandes au fil de l'eau de monitoring et d'optimisation de cluster.