

# Hadoop Coudera : administration

Référence : PYCB032B

Durée : 3 jours

Certification : **Aucune**

## CONNAISSANCES PREALABLES

- Connaissance des commandes des systèmes unix/linux.

## PROFIL DES STAGIAIRES

- Chefs de projet, administrateurs et toute personne souhaitant mettre en œuvre un système distribué avec Hadoop.
- Les travaux pratiques sont réalisés selon le choix des participants sur une distribution Hadoop Cloudera.

## OBJECTIFS

- Connaître les principes du framework Hadoop et savoir l'installer et le configurer.
- Maîtriser la configuration et la gestion des services avec Cloudera Manager.

## CERTIFICATION PREPAREE

Aucune

## METHODES PEDAGOGIQUES

- Mise à disposition d'un poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques, de réflexions et de retours d'expérience
- Le suivi de cette formation donne lieu à la signature d'une feuille d'émargement

## FORMATEUR

Consultant-Formateur expert Bigdata

## METHODE D'EVALUATION DES ACQUIS

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation de fin de stage adressée avec la facture

## CONTENU DU COURS

### Introduction

- Les fonctionnalités du framework Hadoop
- Les différentes versions
- Distributions : Apache, Cloudera, Hortonworks, EMR, MapR, DSE
- Spécificités de chaque distribution
- Les apports de la distribution Cloudera
- Architecture et principe de fonctionnement
- Terminologie : NameNode, DataNode, ResourceManager, NodeManager
- Rôle des différents composants
- Le projet et les modules : Hadoop Common, HDFS, YARN, Spark, MapReduce, Oozie, Pig, Hive, HBase, ...

### Les outils Hadoop

- Infrastructure/Mise en œuvre : Avro, Ambari, Zookeeper, Pig, Tez, Oozie, Falcon, Pentaho
- Vue d'ensemble

- Gestion des données
- Exemple de sqoop
- Restitution : webhdfs, hive, Hawq, Mahout, ElasticSearch ...
- Outils complémentaires : Spark, SparkQL, SparkMLib, Storm, BigTop, Zebra
- Outils complémentaires de développement : Cascading, Scalding, Flink/Pachyderm
- Outils complémentaires d'analyse : RHadoop, Hama, Chukwa, kafka

### Installation et configuration

- Présentation de Cloudera Manager
- Trois modes d'installation : local, pseudo-distribué, distribué
- Première installation
- Mise en œuvre avec un seul noeud Hadoop
- Présentation de Cloudera Manage

- Configuration de l'environnement, étude des fichiers de configuration : core-site.xml, hdfs-site.xml, mapred-site.xml, yarn-site.xml et capacity-scheduler.xml
- Création des users pour les daemons hdfs et yarn, droits d'accès sur les exécutable et répertoires
- Lancement des services
- Démarrage des composants : hdfs, hadoop-daemon, yarn-daemon, etc.
- Gestion de la grappe, différentes méthodes : ligne de commandes, API Rest, serveur http intégré, APIS natives
- Exemples en ligne de commandes avec hdfs, yarn, mapred
- Présentation des fonctions offertes par le serveur http
- Travaux pratiques : Organisation et configuration d'une grappe hadoop avec Cloudera Manager - Traitement de données - Requêtage SQL avec Impala

### **Administration Hadoop**

- Outils complémentaires à yarn et hdfs : jConsole, jconsole yarn
- Exemples sur le suivi de charges, l'analyse des journaux
- Principe de gestion des noeuds, accès JMX
- Travaux pratiques : mise en œuvre d'un client JMX

- Administration HDFS : présentation des outils de stockage des fichiers, fsck, dfsadmin
- Mise en œuvre sur des exemples simples de récupération de fichiers
- Gestion centralisée de caches avec Cacheadmin

### **Sécurité**

- Mécanismes de sécurité et mise en œuvre pratique : Activation de la sécurité avec Kerberos dans core-site.xml, et dans hdfs-site.xml pour les NameNode et DataNode
- Sécurisation de yarn avec la mise en œuvre d'un proxy et d'un Linux Container Executor

### **Exploitation**

- Installation d'une grappe Hadoop. Lancement des services
- Lancement des services
- Principe de la supervision des éléments par le NodeManager
- Présentation de Ganglia, Kibana
- Travaux pratiques : Visualisation des alertes en cas d'indisponibilité d'un noeud - Configuration des logs avec log4j.