

## ECOSYSTÈME HADOOP & DISTRIBUTIONS

# Data Analyst - Analyse de données en environnement Hadoop

**REF : SIHA001**

**DUREE : 21h**

**Présentiel Classe virtuelle**

### **PUBLIC**

Cette formation Data Analyst - Analyse de données en environnement Hadoop est destinée aux personnes qui devront manipuler les données dans un cluster Apache Hadoop.

Modalités et délais d'accès : les inscriptions sont fermées 24h avant la 1ère journée de formation.

Accessibilité : Si vous avez des contraintes particulières liées à une situation de handicap, veuillez nous contacter au préalable afin que nous puissions, dans la mesure du possible, adapter l'action de formation.

### **PREREQUIS**

Cette formation Data Analyst - Analyse de données en environnement Hadoop nécessite d'avoir une expérience dans la manipulation de données. Une connaissance préliminaire d'Hadoop n'est pas exigée.

### **MODALITES PEDAGOGIQUES**

1 poste et 1 support par stagiaire

8 à 10 stagiaires par salle

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

### **MODALITES D'EVALUATION**

Auto-évaluation des acquis par le stagiaire via un questionnaire en ligne

Attestation de fin de stage remise au

## **OBJECTIFS PEDAGOGIQUES**

Cette formation Data Analyst – Analyse de données en environnement Hadoop vous permettra de :

- Identifier le fonctionnement d'Hadoop Distributed File System (HDFS) et YARN/MapReduce
- Explorer HDFS
- Suivre l'exécution d'une application YARN
- Définir le fonctionnement et utiliser les différents outils de manipulation de la donnée :
- Hue : Utilisation de l'interface unifiée
- Hive, Pig : Les générateurs de MapReduce
- Tez : L'optimisation des générateurs de MapReduce
- Sqoop : Comment importer les données de l'entreprise dans un cluster Hadoop?
- Oozie : Comment organiser les exécutions des différentes applications ?

## **PROGRAMME**

### **Introduction**

- Présentation générale d'**Hadoop**
- Exemples d'utilisation dans différents secteurs
- Historique et chiffres clés : Quand parle-t-on de **Big Data** ?

### **L'écosystème d'Hadoop**

- Le système de fichier HDFS
- Le paradigme **MapReduce** et l'utilisation à travers YARN

### **Manipulation des données dans un cluster Hadoop**

- Hue : Comment fonctionne cette interface web ?
- Hive : Pourquoi Hive n'est pas une base de données ?
- Requête sur Hive
- Utilisation de HCatalog
- Utilisation avancée sur Hive
- Utilisation de fonctions utilisateurs
- Paramétrage de requête
- Pig : Fonctionnement de Pig
- Programmation avec Pig Latin
- Utilisation du mode Local
- Utilisation de fonctions utilisateurs
- Tez : Qu'est-ce que Tez ?
- Comment et quand l'utiliser ?
- Oozie : Fonctionnement de Oozie
- Création de Workflows avec Oozie
- Manipulation des Workflows
- Ajout d'éléments d'exploitation dans les Workflows
- Ajout de conditions d'exécution
- Paramétrage des Workflows
- Sqoop : A quoi sert Sqoop ?
- Chargement des données depuis une **base de données**



stagiaire

## relationnelle

- Chargement des données depuis **Hadoop**
- Utilisation et paramétrage avancée
- Les particularités des distributions : Impala, Hawq
- Quelles sont les bonnes pratiques d'utilisation des différents outils ?

Version du : 29/11/2021