

# Formation Analyse de données en environnement Hadoop

**Objectifs :** Connaître les fonctionnements d'Hadoop Distributed File System (HDFS) et YARN / MapReduce  
Savoir explorer HDFS et suivre l'exécution d'une application YARN  
Déterminer les fonctionnements et l'utilisation des différents outils de manipulation des données

**Compétences visées :** - Comprendre ce que sont Hadoop et YARN

- Connaître les différents outils et les Framework dans un environnement Hadoop
- Appréhender MapReduce
- Comprendre comment exécuter une tâche de MapReduce sur YARN
- Exécuter des modifications en masse avec PIG
- Savoir écrire des requêtes pour HIVE afin d'analyser ses données
- Savoir utiliser Sqoop pour transférer les données entre Hadoop et une base de données relationnelle
- Découverte d'autres briques : automatiser vos process avec Oozie
- Utiliser une base de données No-SQL (HBase)

**Durée :** 3 jour(s) (21 heures)

**Public :** Data Scientists, Développeurs décisionnels, ...

**Méthode pédagogique :** Pédagogie active mêlant exposés, exercices et applications pratiques

**Modalités d'évaluation :** Un formulaire d'auto-évaluation proposé en amont de la formation nous permettra d'évaluer votre niveau et de recueillir vos attentes. Ce même formulaire soumis en aval de la formation fournira une appréciation de votre progression.

Des exercices pratiques seront proposés à la fin de chaque séquence pédagogique pour l'évaluation des acquis.

En fin de formation, vous serez amené(e) à renseigner un questionnaire d'évaluation à chaud.

Une attestation de formation vous sera adressée à l'issue de la session.

Trois mois après votre formation, vous recevrez par email un formulaire d'évaluation à froid sur l'utilisation des acquis de la formation.

**Accessibilité :** Vous souhaitez suivre notre formation Formation par ville et êtes en situation de handicap ? Merci de nous contacter afin que nous puissions envisager les adaptations nécessaires et vous garantir de bonnes conditions d'apprentissage

**Tarifs :**

- Présentiel : 1950 € HT
  - Distanciel : 1800 € HT
- (-10% pour 2 inscrits, -20% dès 3 inscrits)

**Option(s) :**

- Forfait déjeuners : 60 € HT

## Nos prochaines sessions

### Distance

du 9 au 11 décembre 2024

du 26 au 28 février 2025

du 8 au 10 décembre 2025

### Lyon

du 12 au 14 mai 2025

du 20 au 22 octobre 2025

### Paris

du 2 au 4 avril 2025

du 22 au 24 septembre 2025

### Toulouse

du 26 au 28 mai 2025

du 22 au 24 octobre 2025

## Programme :

### - Hadoop

- Comprendre Hadoop et son écosystème
- Quels impacts de l'arrivée d'Hadoop dans un SI traditionnel ?
- Le Hadoop Distributed File System (HDFS)
- Introduction aux données dans HDFS
- MapReduce Framework et YARN

### - Pig

- Introduction à Pig
- Programmation Pig avancée
- Troubleshooting et optimisation avec Pig
- Résolution des problèmes avec Pig
- Utiliser l'UI Web d'Hadoop
- Démo optionnelle : résolution d'un « Failed Job » avec l'UI Web
- Echantillonnage de données et débogage
- Vue d'ensemble des performances
- Comprendre le plan d'exécution
- Astuces pour améliorer la performance de vos « Pig Jobs »

## - Hive

- Programmation Hive
- Utilisation de HCatalog
- Programmation Hive avancée
- Etendre Hive
- Transformation de données avec des Scripts personnalisés
- Fonctions définies par l'utilisateur
- Paramétrer les requêtes
- Exercices « Hands-On » : transformation de données avec Hive
- Programmation Hive avancée (suite)
- Analyse de données et statistiques

## - Sqoop

- Import/Export avec Sqoop (SGBDR <-> HDFS)
- Sqoop, fonctions avancées
- Définition de workflow avec Oozie

## - Optionnel : (sous réserve de temps)

- Introduction à H-Base
- Exemple d'ingestion de données avec l'ETL Talend
- Créer son propre cluster Hadoop (plateforme de test)

*Date de dernière modification : 5 novembre 2024*