

# Développer des applications pour Spark avec Hadoop Cloudera avec Certification

## INFORMATIONS GÉNÉRALES

**Type de formation :** Formation continue

**Éligible au CPF :** Non

**Domaine :** IA, Big Data et Bases de données

**Action collective :** Non

**Filière :** Big Data

**Rubrique :** NoSQL et Hadoop

**Code de formation :** BD019

## € Tarifs

**Prix public :** 4000 €

### Tarif & financement :

Nous vous accompagnons pour trouver la meilleure solution de financement parmi les suivantes :

**Le plan de développement des compétences de votre entreprise :** rapprochez-vous de votre service RH.

**Le dispositif FNE-Formation.**

**L'OPCO** (opérateurs de compétences) de votre entreprise.

**France Travail:** sous réserve de l'acceptation de votre dossier par votre conseiller Pôle Emploi.

**CPF -MonCompteFormation**

Contactez nous pour plus d'information : [contact@aston-institut.com](mailto:contact@aston-institut.com)

## PRÉSENTATION

### Objectifs & compétences

- Identifier et utiliser les outils appropriés à chaque situation dans un écosystème hadoop
- Utiliser Apache Spark et l'intégrer dans l'écosystème hadoop
- Utiliser Sqoop, Kafka, Flume, Hive et Impala

### Public visé

Développeur Analyste

### Pré-requis

- Être à l'aise pour programmer dans l'un de ces langages : Scala et/ou Python
- Connaissance de base des lignes de commande Linux requise
- La connaissance de base du SQL est un plus

## 📍 Lieux & Horaires

**Durée :** 28 heures

### Délai d'accès :

Jusqu'à 8 jours avant le début de la formation

## PROGRAMME

### 1.Introduction à Hadoop et à son écosystème

- 1.1. Introduction générale à hadoop
- 1.2. Traitement de données
- 1.3. Introduction aux exercices pratiques

### 2.HDFS : le système de fichiers Hadoop

- 2.1. Les composants d'un cluster hadoop
- 2.2. L'architecture d'HDFS
- 2.3. Utiliser HDFS

### 3.Le traitement distribué sur un cluster Hadoop

- 3.1. L'architecture de YARN
- 3.2. Travailler avec YARN
- 4.Les bases de Spark
- 4.1. Introduction à Spark
- 4.2. Démarrer et utiliser la console Spark
- 4.3. Introduction aux Datasets et DataFrames Spark
- 4.4. Les opérations sur les DataFrames

### 5.Manipulation des dataframes et des schemas

- 5.1. Créer des DataFrames depuis diverses sources de données
- 5.2. Sauvegarder des DataFrames
- 5.3. Les schémas des DataFrames
- 5.4. Exécution gloutonne et paresseuse de Spark

### 6.Analyser des données avec des requêtes sur dataframes

- 6.1. Requête des DataFrames avec des expressions sur les colonnes nommées
- 6.2. Les requêtes de groupement et d'agrégation

## 📅 Prochaines sessions

Consultez-nous pour les prochaines sessions.

### 6.3. Les jointures

#### **7.Les RDD - Structure fondamentale de Spark**

- 7.1. Introduction aux RDD
- 7.2. Les sources de données de RDD
- 7.3. Créer et sauvegarder des RDD
- 7.4. Les opérations sur les RDD

#### **8.Transformers les données avec des RDD**

- 8.1. Écrire et passer des fonctions de transformation
- 8.2. Fonctionnement des transformations de Spark
- 8.3. Conversion entre RDD et DataFrames

#### **9.Agrégation de données avec les RDD de paires**

- 9.1. Les RDD clé-valeur
- 9.2. Map-Reduce : principe et usage dans Spark
- 9.3. Autres opérations sur les RDD de paires

#### **10.Requêtage de tables et de vues avec Spark SQL**

- 10.1. Requêter des tables en Spark en utilisant SQL
- 10.2. Requêter des fichiers et des vues
- 10.3. L'API catalogue de Spark

#### **11.Travailler avec des Datasets Spark en Scala**

- 11.1. Les différences entre Datasets et DataFrames
- 11.2. Créer des Datasets
- 11.3. Charger et sauvegarder des Datasets
- 11.4. Les opérations sur les Datasets

#### **12.Écrire, configurer et lancer des applications Spark**

- 12.1. Écrire une application Spark
- 12.2. Compiler et lancer une application
- 12.3. Le mode de déploiement d'une application
- 12.4. L'interface utilisateur web des applications Spark
- 12.5. Configurer les propriétés d'une application

#### **13.Le traitement distribué avec Spark**

- 13.1. Rappels sur le fonctionnement de Spark avec YARN
- 13.2. Le partitionnement des données dans les RDD
- 13.3. Exemple : le partitionnement dans les requêtes
- 13.4. Jobs, étapes et tâches
- 13.5. Exemple : le plan d'exécution de Catalyst

## **MODALITÉS**

### **Modalités**

Jusqu'à 8 jours avant le début de la formation, sous condition d'un dossier d'inscription complet

### **Méthode**

**Fin de formation :** entretien individuel.

**Satisfaction des participants :** questionnaire de satisfaction réalisé en fin de formation.

**Assiduité :** certificat de réalisation.

**Validations des acquis :** grille d'évaluation des acquis établie par le formateur en fin de formation.