Formation Spark, développer des applications pour le Big Data

N° ACTIVITÉ : 11 92 18558 92 | TÉLÉPHONE : 01 85 77 07 07 | E-MAIL : inscription@hubformation.com

Référence	CB037
Durée	3 jours (21h)
Tarif	2 510 €HT
Repas	69 €HT(en option)

Objectifs

| Identifier le fonctionnement de Spark et son utilisation dans un environnement Hadoop.
| Savoir intégrer Spark dans un environnement Hadoop, traiter des données Cassandra, HBase, Kafka, Flume, Sqoop, S3.
| Prépare l'examen "Certification Hadoop avec Spark pour développeur de Cloudera"

Public

| Chefs de projet, data scientists, développeurs.

Prérequis

| Connaissances de Java ou Python, notions de calculs statistiques
| Avoir des bases de Hadoop

Programme de la formation

Introduction
| Présentation Spark, origine du projet,apports, principe de fonctionnement.Langages supportés.

Premiers pas
| Utilisation du shell Spark avec Scala ou Python.Modes de fonctionnement. Interprété, compilé.
| Utilisation des outils de construction. Gestion des versions de bibliothèques.

Règles de développement
| Mise en pratique en Java, Scala et Python.Notion de contexte Spark
| Différentes méthodes de création des RDD : depuis un fichier texte, un stockage externe.
| Manipulations sur les RDD (Resilient Distributed Dataset).Fonctions, gestion de la persistence.

Cluster
| Différents cluster managers : Spark en autonome, avec Mesos, avec Yarn, avec Amazon EC2
| Architecture : SparkContext, Cluster Manager, Executor sur chaque noeud.Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job
| Mise en oeuvre avec Spark et Amazon EC2.Soumission de jobs, supervision depuis l'interface web

Traitements
| Lecture/écriture de données : Texte, JSon, Parquet, HDFS, fichiers séquentiels.
| Jointures. Filtrage de données, enrichissement.Calculs distribués de base. Introduction aux traitements de données avec map/reduce.
| Travail sur les RDDs. Transformations et actions. Lazy execution. Impact du shuffle sur les performances.
| RDD de base, key-pair RDDs.Variables partagées : accumulateurs et variables

SESSIONS PROGRAMMÉES

A DISTANCE (FRA)

du 1er au 3 septembre 2025
du 1er au 3 décembre 2025

PARIS

du 1er au 3 septembre 2025
du 1er au 3 décembre 2025

VOIR TOUTES LES DATES

broadcast.

Intégration hadoop
| Présentation de l'écosystème Hadoop de base : HDFS/Yarn.Travaux pratiques avec YARN
| Création et exploitation d'un cluster Spark/YARN.Intégration de données sqoop, kafka, flume vers une architecture Hadoop.
| Intégration de données AWS S3.

Support Cassandra
| Description rapide de l'architecture Cassandra. Mise en oeuvre depuis Spark.Exécution de travaux Spark s'appuyant sur une grappe Cassandra.

DataFrames
| Spark et SQL
| Objectifs : traitement de données structurées,.L'API Dataset et DataFrames
| Optimisation des requêtes.Mise en oeuvre des Dataframes et DataSet.Comptabilité Hive
| Travaux pratiques: extraction, modification de données dans une base distribuée.Collections de données distribuées.Exemples.

Streaming
| Objectifs , principe de fonctionnement : stream processing.Source de données : HDFS, Flume, Kafka, ...
| Notion de StreamingContexte, DStreams, démonstrations.Travaux pratiques : traitement de flux DStreams en Scala.

Machine Learning
| Fonctionnalités : Machine Learning avec Spark,algorithmes standards, gestion de la persistence, statistiques.
| Support de RDD.Mise en oeuvre avec les DataFrames.

Spark GraphX
| Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes
| Travaux pratiques : exemples d'opérations sur les graphes.

Méthode pédagogique

Chaque participant travaille sur un poste informatique qui lui est dédié. Un support de cours lui est remis soit en début soit en fin de cours. La théorie est complétée par des cas pratiques ou exercices corrigés et discutés avec le formateur. Le formateur projette une présentation pour animer la formation et reste disponible pour répondre à toutes les questions.

Méthode d'évaluation

Tout au long de la formation, les exercices et mises en situation permettent de valider et contrôler les acquis du stagiaire. En fin de formation, le stagiaire complète un QCM d'auto-évaluation.

Suivre cette formation à distance

Voici les prérequis techniques pour pouvoir suivre le cours à distance :

| Un ordinateur avec webcam, micro, haut-parleur et un navigateur (de préférence Chrome ou Firefox). Un casque n'est pas nécessaire suivant l'environnement.
| Une connexion Internet de type ADSL ou supérieure. Attention, une connexion Internet ne permettant pas, par exemple, de recevoir la télévision par Internet, ne sera pas suffisante, cela engendrera des déconnexions intempestives du stagiaire et dérangera toute la classe.
| Privilégier une connexion filaire plutôt que le Wifi.
| Avoir accès au poste depuis lequel vous suivrez le cours à distance au moins 2 jours avant la formation pour effectuer les tests de connexion préalables.
| Votre numéro de téléphone portable (pour l'envoi du mot de passe d'accès aux supports de cours et pour une messagerie instantanée autre que celle intégrée à la classe virtuelle).
| Selon la formation, une configuration spécifique de votre machine peut être attendue, merci de nous contacter.
| Pour les formations incluant le passage d'une certification la dernière journée, un voucher vous est fourni pour passer l'examen en ligne.
| Pour les formations logiciel (Adobe, Microsoft Office...), il est nécessaire d'avoir le logiciel installé sur votre machine, nous ne fournissons pas de licence ou de version test.
| Horaires identiques au présentiel.

Accessibilité

Les sessions de formation se déroulent sur des sites différents selon les villes ou les dates, merci de nous contacter pour vérifier l'accessibilité aux personnes à mobilité réduite.
Pour tout besoin spécifique (vue, audition...), veuillez nous contacter au 01 85 77 07 07.