Spark, développer des applications pour le Big Data

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 : Hive - Spark

PRÉSENTATION

Objectifs & compétences

A l'issue de la formation, le stagiaire sera capable de maîtriser le framework Spark pour traiter des données hétérogènes et optimiser les calculs.

Maîtriser les concepts fondamentaux de Spark
Savoir intégrer Spark dans un environnement Hadoop
Développer des applications d'analyse en temps réel avec Spark Streaming
Faire de la programmation parallèle avec Spark sur un cluster
Manipuler des données avec Spark SQL
Avoir une première approche du Machine Learning

Public visé

Chefs de projet, Data Scientists, Développeurs, Architectes…

Pré-requis

Avoir des connaissances de Java ou Python et des notions de calculs statistiques

PROGRAMME

Maîtriser les concepts fondamentaux de Spark
Présentation Spark, origine du projet,apports, principe de fonctionnement. Langages supportés.
Modes de fonctionnement : batch/Streaming.
Bibliothèques : Machine Learning, IA
Mise en oeuvre sur une architecture distribuée. Architecture : clusterManager, driver, worker, …
Architecture : SparkContext, SparkSession, Cluster Manager, Executor sur chaque noeud.
Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job

Savoir intégrer Spark dans un environnement Hadoop
Intégration de Spark avec HDFS, HBase,
Création et exploitation d'un cluster Spark/YARN. Intégration de données sqoop, kafka, flume vers une architecture Hadoop et traitements par Spark.
Intégration de données AWS S3.
Différents cluster managers : Spark interne, avec Mesos, avec Yarn, avec Amazon EC2

Atelier : Mise en oeuvre avec Spark sur Hadoop HDFS et Yarn. Soumission de jobs, supervision depuis l'interface web

Développer des applications d'analyse en temps réel avec Spark Streaming
Objectifs , principe de fonctionnement: stream processing. Source de données : HDFS, Flume, Kafka, …
Notion de StreamingContext, DStreams, démonstrations.

Atelier : traitement de flux DStreams en Scala. Watermarking. Gestion des micro-batches.

Intégration de Spark Streaming avec Kafka

Code de formation : BD021

€ Tarifs

Prix public : 1990 €

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

Lieux & Horaires

Durée : 21 heures

Délai d'accès : Jusqu'a 8 jours avant le début de la formation, sous condition d'un dossier d'insciption complet

Prochaines sessions

Consultez-nous pour les prochaines sessions.

Atelier : mise en oeuvre d'une chaîne de gestion de données en flux tendu : IoT, Kafka, SparkStreaming, Spark. Analyse des données au fil de l'eau.

Faire de la programmation parallèle avec Spark sur un cluster
Utilisation du shell Spark avec Scala ou Python. Modes de fonctionnement. Interprété, compilé.
Utilisation des outils de construction. Gestion des versions de bibliothèques.

Atelier : Mise en pratique en Java, Scala et Python. Notion de contexte Spark. Extension aux sessions Spark.

Manipuler des données avec Spark SQL
Spark et SQL
Traitement de données structurées. L'API Dataset et DataFrames
Jointures. Filtrage de données, enrichissement. Calculs distribués de base. Introduction aux traitements de données avec map/reduce.
Lecture/écriture de données : Texte, JSon, Parquet, HDFS, fichiers séquentiels.
Optimisation des requêtes. Mise en oeuvre des Dataframes et DataSet. Compatibilité Hive

Atelier : écriture d'un ETL entre HDFS et HBase

Atelier : extraction, modification de données dans une base distribuée. Collections de données distribuées. Exemples.

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

Spark GraphX
Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes

Atelier : exemples d'opérations sur les graphes.

Avoir une première approche du Machine Learning
Machine Learning avec Spark, algorithmes standards supervisés et non-supervisés (RandomForest, LogisticRegression, KMeans, ...)
Gestion de la persistance, statistiques.
Mise en oeuvre avec les DataFrames.

Atelier : mise en oeuvre d'une régression logistique sur Spark

MODALITÉS

Modalités

Modalités : en présentiel, distanciel ou mixte . Toutes les formations sont en présentiel par défaut mais les salles sont équipées pour faire de l'hybride. – Horaires de 9H à 12H30 et de 14H à 17H30 soit 7H – Intra et Inter entreprise.
Pédagogie : essentiellement participative et ludique, centrée sur l'expérience, l'immersion et la mise en pratique. Alternance d'apports théoriques et d'outils pratiques.
Ressources techniques et pédagogiques : Support de formation au format PDF ou PPT Ordinateur, vidéoprojecteur, Tableau blanc, Visioconférence : Cisco Webex / Teams / Zoom.
Pendant la formation : mises en situation, autodiagnostics, travail individuel ou en sous-groupe sur des cas réels.

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'evalution des acquis établie par le formateur en fin de formation.