Big Data – Spark pour les développeurs

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

Découvrir les concepts clés du Big Data
Comprendre l'écosystème technologique d'un projet Big Data
Evaluer les techniques de gestion des flux de données massives
Implémenter des modèles d'analyses statistiques pour répondre aux besoins métiers
Découvrir les outils de Data Visualisation

Public visé

Architectes
Développeurs
Analystes Informatique

Pré-requis

Connaissances de l'informatique et des principes d'architectures classiques
Connaissances de base des modèles relationnels
Connaissances de base du développement informatique

PROGRAMME

Découverte de Scala et Introduction à Spark
Introduction à Scala

Travailler avec :
Les variables
Les types de données
Les contrôles de flux
L'interpréteur Scala
Les collections et les méthodes standards (map(), etc.)

Travailler avec :
Les fonctions
Les méthodes
Les fonctions littérales
Définition et description des notions de :
Classe
Objet
Case Class
Introduction et origine de Spark
L'écosystème de Spark
Spark vs Hadoop
Télécharger et installer Spark
Le Shell Spark et le SparkContext.

Travaux pratiques :
Mise en place de l'environnement
Démarrage de l'interpréteur Scala
Découverte de Spark
Découverte de Spark Shell

RDDs et l'architecture de Spark, Spark SQL, les DataFrames et DataSets

Code de formation : NE080

€ Tarifs

Prix public : 2100 €

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 : 28 heures

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

informations :
NULL

Prochaines sessions

Consultez-nous pour les prochaines sessions.

Le concept de RDD, leur cycle de vie, l'évaluation « Lazy »
Le partitionnement et les transformations des RDD
Travailler avec les RDD
Créer et transformer (map, filter, etc.)
Vue d'ensemble des RDDs
SparkSession, charger et sauvegarder des données, les formats de données (JSON, CSV, Parquet, text, ...)
Introduction aux DataFrames et DataSets
Travailler avec les DataFrames, la DSL de requêtage :
Column
Filtrering
Grouping
Aggregation
Lancement de requêtes SQL sur les RDDs
Travailler avec l'API des DataSets
Transformer et partitionner (flatMap(), explode() et split())
Synthèse Datasets vs DataFrames vs RDDsLa virtualisation du poste de travail

Shuffling, Transformations et performances, Amélioration des performances

Travailler avec :
Grouping
Reducing
Joining
Shuffling, Les dépendances Narrow vs Wide, et les impacts de performances
Découverte de l'optimiseur de requêtes Catalyst (explain(), Query Plans, Problèmes des lambda)
L'optimiseur Tungsten (Format Binaire, la Cache Awareness, ...)
Présentation du Caching :
Concept
Type de stockage
Préconisations
Minimisation du Shuffling pour l'amélioration des performances
Utiliser les « Accumulators » et les « Broadcast Variables »
Recommandations générales de performance :
Utilisation du Spark UI
Transformations efficaces
Stockage de données
Monitoring

Travaux Pratiques
Le Groupe Shuffling
Découverte de Catalyst
Découverte de Tungsten
Caching, Joins, Shuffles, Broadcasts, Accumulators
Recommandations générales sur les Broadcasts

Créer des applications standalone et Spark Streaming
La Core API, Le builder SparkSession
Configurer et créer une SparkSession
Construire et Lancer des application (sbt.build.sbt et spark-submit)
Le cycle de vie des applications (Driver, Executors, et Tasks)
Les gestionnaires de cluster (Standalone, YARN, Mesos)
Gestions des Logs et Debug
Introduction aux Bases de Spark Streaming
Spark Streaming (1.0+)
DStreams, Receivers, Batching
Transformations Stateless
Transformations Windowed
Transformations Stateful
Structured Streaming (2+)
Applications continues
Le concept de Table, Result Table
Étapes de Structured Streaming
Les Sources et Sinks
Consommation de données en provenance de Kafka
Présentation de Kafka
Le format « Kafka » de Structured Streaming
Travailler sur le Stream

Travaux Pratiques
La soumission de travaux Spark
Fonctionnalités additionnelles de Spark
Spark Streaming
Spark Structured Streaming

Spark Structured Streaming avec Kafka
Optionnel : La « Sessionization » des Structured Streaming
Optionnel : L'analyse de séries temporelles avec PySpark

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.