EduGroupe
Accompagner pour réussir

Spark, Développer des applications pour le BigData

Référence : **PYCB037B** Durée : **3 jours (21 heures)** Certification : **Aucune**

Connaissances préalables

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

Profil des stagiaires

- Chefs de projet, Data Scientists, Développeurs, Architectes...

Objectifs

- 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

Certification préparée

- Aucune

Méthodes pédagogiques

- 6 à 12 personnes maximum par cours, 1 poste de travail par stagiaire
- Remise d'une documentation pédagogique papier ou numérique pendant le stage
- La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions

Formateur·rice

- Consultant-Formateur expert Bigdata

Méthodes d'évaluation des acquis

- Auto-évaluation des acquis par le stagiaire via un questionnaire
- Attestation des compétences acquises envoyée au stagiaire
- Attestation de fin de stage adressée avec la facture

Contenu du cours

1. 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

2. 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
- Exemple d'atelier : Mise en oeuvre avec Spark sur Hadoop HDFS et Yarn. Soumission de jobs, supervision depuis l'interface web

3. 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
- Exemple d'atelier : traitement de flux DStreams en Scala. Watermarking. Gestion des micro-batches
- Intégration de Spark Streaming avec Kafka
- Exemple d'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

4. 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
- Exemple d'atelier : Mise en pratique en Java, Scala et Python. Notion de contexte Spark. Extension aux sessions Spark

5. 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
- Exemple d'ateliers : écriture d'un ETL entre HDFS et HBase. Extraction, modification de données dans une base distribuée. Collections de données distribuées. Exemples

6. Support Cassandra

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

7. Spark GraphX

- Fourniture d'algorithmes, d'opérateurs simples pour des calculs statistiques sur les graphes
- Exemple d'atelier : exemples d'opérations sur les graphes

8. 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
- Exemple d'atelier : mise en oeuvre d'une régression logistique sur Spark

Notre référent handicap se tient à votre disposition au 01.71.19.70.30 ou par mail à referent.handicap@edugroupe.com pour recueillir vos éventuels besoins d'aménagements, afin de vous offrir la meilleure expérience possible.