Spark, traitement des données

Description :

Apache Spark est un moteur d'analyse unifié (Unified Analytics Engine), créé pour le traitement rapide des données. Ce framework open source de calcul distribué permet l'analyse et le traitement de données à grandes échelles. Le framework permet le développement d'applications de traitement hautement performantes.

Profil de l'intervenant :

Consultant formateur expert Big Data

Objectifs :

A l'issue de la formation, les stagiaires sauront :

- Maîtriser les concepts fondamentaux de Spark
- 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
- Ce qu'est le Machine Learning (première approche)

Publics :

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

Durée :

3 jours

Prérequis :

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

Méthode pédagogique de cette formation :

La formation est constituée d'apports théoriques, d'exercices pratiques et de réflexions. Remise d'une documentation pédagogique papier ou numérique pendant le stage 6 à 8 personnes maximum par cours.

Méthode d'évaluation des acquis de la formation :

MAJ DEC 2022

Auto évaluation des acquis par le stagiaire via une questionnaire. Attestation de fin de stage signée remise au stagiaire en fin de formation

Information handicap :

Cette formation est accessible aux personnes en situation de handicap. Chaque situation étant unique, nous vous demandons de préciser à l'inscription votre handicap. Nous pourrons ainsi confirmer l'ensemble des possibilités d'accueil et vous permettre de suivre la formation dans les meilleures conditions en accord avec votre employeur. Pour toutes informations complémentaires, nous vous conseillons la structure suivante https://www.agefiph.fr/

Programme de cette formation :

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
- 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