Tél : (+33)1 55 60 76 72
Email : formation@oxiane.com

Spark : traitement de données

Traitement de données

Détails

- Code : DB-SPK
- Durée : 3 jours (21 heures)

Public
- Chefs de projets
- Développeurs
- Data Scientist

Pré-requis
- Connaissances de Java ou Python
- Notions de calculs statistiques et des bases Hadoop ou avoir suivi le stage *Hadoop, l&#039écosystème*

Objectifs
- Comprendre le fonctionnement de Spark et son utilisation dans un environnement Hadoop
- Savoir intégrer Spark dans un environnement Hadoop
- Savoir traiter des données Cassandra, HBase, Kafka, Flume, Sqoop, S3

Programme

Introduction
- Présentation Spark
- Origine du projet
- Apports
- Principe de fonctionnement
- Langages supportés
- Mise en oeuvre sur une architecture distribuée
- Architecture : clusterManager, driver, worker, ...

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
- Mise en pratique en Java, Scala et Python
- Notion de contexte Spark
- Extension aux sessions Spark

Cluster
- Différents cluster managers : Spark interne, avec Mesos, avec Yarn, avec Amazon EC2
- Architecture : SparkContext, SparkSession, Cluster Manager, Executor sur chaque nœud
- Définitions : Driver program, Cluster manager, deploy mode, Executor, Task, Job
- Mise en œuvre 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

Support Cassandra
- Description rapide de l'architecture Cassandra
- Mise en œuvre 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 œuvre 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
- Watermarking
- Gestion des microbatches
- Travaux pratiques : 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

Intégration Hadoop
- Rappels sur l'écosystème Hadoop de base : HDFS/Yarn
- 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

Machine Learning

- Fonctionnalités : Machine Learning avec Spark, algorithmes standards, gestion de la persistance, statistiques
- 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

Modalité

- Stage pratique en présentiel
- Stage pratique en distanciel
- Nombre de stagiaires minimum : 4
- Nombre de stagiaires maximum : 10

Méthodes pédagogiques

- Exposés
- Cas pratiques

Profils des intervenants

- Toutes nos formations sont animées par des consultants-formateurs expérimentés et reconnus par leurs pairs.

Modalités d'évaluation

- Evaluation des acquis de la formation par le biais de cas pratiques et/ou mises en situation.
- Attestation de formation remise à chaque participant.

Démarche qualité

- Questionnaire d'évaluation de satisfaction à chaud complété par chaque participant à l'issue de la formation.

Moyens pédagogiques

- Salle équipée de PC (1 poste par stagiaire), vidéo-projecteur.
- Espace de pause.

Dernière mise à jour le 14/01/2022