Développer des applications avec Apache Spark

Formation officielle Cloudera Data Engineering: Developing Applications with Apache Spark

DESCRIPTION

Apache Spark s'est imposé ces dernières années comme le framework de référence pour traiter les données massives distribuées, et comme un outil central de l'écosystème Hadoop. Cette formation intensive embarque le participant de la découverte de Spark jusqu'à l'utilisation de ses fonctionnalités avancées.

La démarche pédagogique alterne apports théoriques sur les structures fondamentales Spark (RDD, DataFrame, DataSets) et de nombreux travaux pratiques pour favoriser la prise en main.

Les apprenants sont ainsi amenés à manipuler la console interactive pour prototyper, avant de coder, déployer et suivre des applications sur un cluster.

Un temps est également consacré aux évolutions majeures de la nouvelle version de Spark avec, entre autres, des cas d'usages complexes de traitement en flux (streaming).

Au cours de la formation, un panorama de l'écosystème Hadoop est ainsi dressé, en insistant sur les concepts essentiels des environnements distribués : stockage sur HDFS, calcul avec Map-Reduce et gestion des ressources via YARN.

Des compléments sur l'ingestion de données avec Hive et Kafka sont proposés, afin que les apprenants maîtrisent l'ensemble des outils nécessaires pour développer des applications Spark. Ils disposent ainsi d'une expertise complète pour préparer des données massives et les analyser avec Spark et/ou sur un cluster Hadoop.

OBJECTIFS PEDAGOGIQUES

- Appréhender l'écosystème Apache Hadoop et son intégration dans le cycle de vie du traitement des données.
- S'approprier les schémas selon lesquels les données sont distribuées, stockées et traitées dans un cluster Hadoop
- Implémenter, configurer et déployer des applications Apache Spark sur un cluster Hadoop

Stage pratique
Data Engineering

Code :
CLSPH

Durée :
4 jour(s) (28,00 heures)

Exposés : **40.00 %**
Cas pratiques : **50.00 %**
Echanges d'expérience : **10.00 %**

Inter-entreprises :
Prochaines sessions disponibles sur notre site web.
Tarif : 2 700,00 € HT / participant

Intra-entreprise :
Tarifs et dates sur demande.

- Manipuler Spark shell et les applications Spark pour explorer, traiter et analyser des données distribuées
- Requêter des données à l'aide de Spark SQL, les DataFrames et les Datasets
- Créer et exécuter des dataframes avec Spark Streaming pour traiter un flux de données en continu (streaming)

PUBLIC CIBLE

- Développeur
- Analyste

PRE-REQUIS

- Être à l'aise pour programmer dans l'un de ces langages : Scala et/ou Python
- Connaissance de base des lignes de commande Linux requise
- La connaissance de base du SQL est un plus
- Aucune expérience préalable avec hadoop n'est nécessaire

METHODE PEDAGOGIQUE
Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience pratique du formateur, complétés de travaux pratiques et de mises en situation. Les exemples Apache Spark et les exercices de "hands-on" sont présentés avec Scala et Python. A la suite de la formation, les stagiaires auront la possibilité de passer l'examen Certification "CDP Data Developer " de Cloudera.

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

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION
L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique. Une évaluation à chaud sur la satisfaction des stagiaires est réalisée systématiquement en fin de session et une attestation de formation est délivrée aux participants mentionnant les objectifs de la formation, la nature, le programme et la durée de l'action de formation ainsi que la formalisation des acquis.

PROGRAMME PEDAGOGIQUE DETAILLE

Jour 1

INTRODUCTION À HADOOP ET À SON ÉCOSYSTÈME

- Introduction générale à Hadoop
- Traitement de données
- Introduction aux exercices pratiques

HDFS : LE SYSTÈME DE FICHIERS HADOOP

- Les composants d'un cluster Hadoop
- L'architecture HDFS
- Utiliser HDFS

LE TRAITEMENT DISTRIBUÉ SUR UN CLUSTER HADOOP

- L'architecture de YARN
- Travailler avec YARN

LES BASES DE APACHE SPARK

- Introduction à Apache Spark
- Démarrer et utiliser Spark Shell
- Introduction aux Datasets et DataFrames Spark
- Les opérations sur les DataFrames

MANIPULATION DES DATAFRAMES ET DES SCHÉMAS

- Créer des DataFrames depuis diverses sources de données
- Sauvegarder des DataFrames
- Les schémas des DataFrames
- Exécution gloutonne et paresseuse de Spark

Jour 2

ANALYSER DES DONNÉES AVEC DES REQUÊTES SUR DATAFRAMES

- Requêter des DataFrames avec des expressions de colonnes
- Les requêtes de groupement et d'agrégation
- Les jointures

LES RDD - STRUCTURE FONDAMENTALE DE SPARK

- Introduction aux RDD

- Les sources de données de RDD
- Créer et sauvegarder des RDD
- Les opérations sur les RDD

TRANSFORMER LES DONNÉES AVEC DES RDD

- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark
- Conversion entre RDD et DataFrames

AGRÉGATION DE DONNÉES AVEC LES RDD DE PAIRES

- Les RDD clé-valeur
- Map-Reduce : principe et usage dans Spark
- Autres opérations sur les RDD de paires

Jour 3

REQUÊTAGE DE TABLES ET DE VUES AVEC SPARK SQL

- Requêter des tables en Spark en utilisant SQL
- Requêter des fichiers et des vues
- L'API catalogue de Spark

TRAVAILLER AVEC DES DATASETS SPARK EN SCALA

- Les différences entre Datasets et DataFrames
- Créer des Datasets
- Charger et sauvegarder des Datasets
- Les opérations sur les Datasets

ÉCRIRE, CONFIGURER ET LANCER DES APPLICATIONS SPARK

- Écrire une application Spark
- Compiler et lancer une application
- Le mode de déploiement d'une application
- L'interface utilisateur web des applications Spark
- Configurer les propriétés d'une application

LE TRAITEMENT DISTRIBUÉ AVEC SPARK

- Rappels sur le fonctionnement de Spark avec YARN
- Le partitionnement des données dans les RDD
- Exemple : le partitionnement dans les requêtes

- Jobs, étapes et tâches
- Exemple : le plan d'exécution de Catalyst
- Exemple : le plan d'exécution de RDD

PERSISTANCE DE LA DONNÉE DISTRIBUÉE

- La persistance des DataFrames et des Datasets
- Les niveaux de persistances
- Voir les RDD persistés

LES CAS D'USAGE DE SPARK

- Les cas d'usages courants de Spark
- Les algorithmes itératifs en Spark
- Machine Learning avec Spark
- Exemple : K-means

Jour 4

INTRODUCTION À SPARK STRUCTURED STREAMING

- Introduction à Apache Spark Streaming
- Créer des streaming DataFrames
- Transformer des streaming DataFrames
- Exécuter des requêtes de streaming

STRUCTURED STREAMING AVEC KAFKA

- Introduction
- Recevoir des messages Kafka
- Envoyer des messages Kafka

AGRÉGATION ET JOINTURES SUR DES STREAMING DATAFRAMES

- Aggrégation sur des streaming DataFrames
- Jointure sur des streaming DataFrames

LE TRAITEMENT DE MESSAGES AVEC KAFKA

- Introduction à Kafka
- Passer à l'échelle avec Kafka
- L'architecture d'un cluster Kafka
- La ligne de commande Kafka

Accessibilité

L'inclusion est sujet important pour OCTO Academy.
Nos référent·es sont à votre disposition pour faciliter l'adaptation de votre formation à vos besoins spécifiques.
Pour les contacter : academy.accessibilite@octo.com