BEST

Développer des applications pour Spark avec Hadoop Cloudera

Formation officielle "Cloudera Developer Training for Spark and Hadoop"

DESCRIPTION

Apache Spark s'est imposé ces dernières années comme le framework big data de référence, et comme un outil central de l'écosystème hadoop. Cette formation intensive emmène le participant de la découverte de Spark jusqu'à l'utilisation de ses fonctionnalités avancées.

La démarche pédagogique équilibre apports théoriques sur les structures fondamentales Spark (RDD, DataFrame, DataSets) et de nombreux travaux pratiques. Les participants manipulent la console interactive pour prototyper. Ensuite, ils codent, déploient et monitorent des applications sur un cluster. Le programme intègre les évolutions majeures de la nouvelle version Spark 2, et des cas d'usages complexes de traitement en flux (streaming).

Au cours de la formation, un panorama de l'écosystème hadoop est 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 Sqoop et Kafka sont proposés, afin que les participants maitrisent 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 sur un cluster hadoop.

OBJECTIFS PÉDAGOGIQUES

Identifier et utiliser les outils appropriés à chaque situation dans un écosystème hadoop
Utiliser Apache Spark et l'intégrer dans l'écosystème hadoop
Utiliser Sqoop, Kafka, Flume, Hive et Impala

PUBLIC CIBLE

Développeur
Analyste

PRÉ-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

MÉTHODE PÉDAGOGIQUE

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

Stage pratique en présentiel
CLOUDERA

Code :
CLSPH

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

14 - 17 déc. 2020
Formation à distance / 2 695 eur

18 - 21 jan. 2021
Paris / 2 695 eur

15 - 18 mar. 2021
Paris / 2 695 eur

25 - 28 mai 2021
Paris / 2 695 eur

24 - 27 août 2021
Paris / 2 695 eur

Tarif & dates intra :
Sur demande

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 « CCA Spark and Hadoop Developer » de Cloudera.

PROFILS 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.

POUR ALLER PLUS LOIN :

- Site internet Cloudera
- Livre blanc "Hadoop, feuille de route"
- Toutes nos formations Hadoop Cloudera
- Toutes nos formations sur la Data Science

Programme pédagogique détaillé par journée

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 d'HDFS
- Utiliser HDFS

LE TRAITEMENT DISTRIBUÉ SUR UN CLUSTER HADDOP

- L'architecture de YARN
- Travailler avec YARN

LES BASES DE SPARK

- Introduction à Spark
- Démarrer et utiliser la console Spark
- Introduction aux Datasets et DataFrames Spark
- Les opérations sur les DataFrames

MANIPULATION DES DATAFRAMES ET DES SCHEMAS

- 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 sur les colonnes nommées
- Les requêtes de groupement et d'aggrégation
- 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 LES 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 ALGORITHMES ITÉRATIFS AVEC SPARK

- D'autres 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 Spark Streaming
- Créer des streaming DataFrames
- Transformer des DataFrames
- Exécuter des requêtes de streaming

STRUCTURED STREAMING AVEC KAFKA

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

AGGREGATION ET JOINTURES SUR DES STREAMING DATAFRAMES

- Aggregation sur des streaming DataFrames
- Jointure sur des streaming DataFrames

Suppléments

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