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 de 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
HADOOP CLOUDERA

Code :
CLSPH

Durée :
4 jours (28 heures)

Certification :
300 € HT

Exposés :
40%

Cas pratiques :
50%

Échanges d'expérience :
10%

Sessions à venir :

4 - 7 fév. 2019
Paris / 2 695 eur

3 - 6 juin 2019
Paris / 2 695 eur

12 - 15 nov. 2019
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 HORTONWORKS
- 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
- Ingestion et stockage de données
- Traitement de données
- Exploration et analyse de données
- Autres outils de l'écosystème
- TP « Introduction. Exécuter des requêtes avec Impala »

HDFS : LE SYSTÈME DE FICHIERS HADOOP

- Les composants d'un cluster Hadoop
- L'architecture d'HDFS
- Utiliser HDFS
- TP « Utiliser HDFS avec la ligne de commande et l'interface graphique HUE »

LE TRAITEMENT DISTRIBUÉ SUR UN CLUSTER HADDOP

- L'architecture de YARN
- Travailler avec YARN
- TP « Lancer des jobs YARN et les monitorer »

LES BASES DE SPARK

- Introduction à Spark
- Démarrer et utiliser la console Spark
- Introduction aux Datasets et DataFrames Spark
- Les opérations sur les DataFrames
- TP « Explorer des DataFrames avec la console Spark »

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
- TP « Travailler avec des DataFrames et des schémas »

Jour 2

ANALYSER DES DONNÉES AVEC DES REQUÊTES SUR DATAFRAMES

- Requêter des DataFrames avec des expression sur les colonnes nommées
- Les requêtes de groupement et d'aggrégation
- Les jointures
- TP « Analyser des données avec des requêtes sur les DataFrame »

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
- TP « Travailler avec des RDD »

TRANSFORMER LES DONNÉES AVEC DES RDD

- Écrire et passer des fonctions de transformation
- Fonctionnement des transformations de Spark
- Conversion entre RDD et DataFrames
- TP « Transformer des données avec des RDD »

AGGREGATION 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
- TP « Joindre des données en utilisant des RDD de paires »

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
- Comparaison de Spark SQL, Impala et Hive-on-Spark
- TP « Requêter des tables et des vues avec Spark SQL »

Jour 3

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
- TP « Travailler avec des Datasets en Scala »

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 Sparks
- TP « Écrire, configurer et lancer une application Spark »

LE TRAITEMENT DISTRIBUÉ AVEC SPARK

- Rappels sur le fonctionnement de Spark avec YARN
- Le partitionnement des données dans les RDD
- Jobs, Étapes et Tâches
- Le plan d'exécution, et l'optimisation avec Catalyst
- TP « Suivre et explorer l'exécution de requêtes »

PERSISTANCE DE LA DONNÉE DISTRIBUÉE

- La persistance des DataFrames et des DataSets
- Les niveaux de persistances
- Voir les RDD persistés
- TP « Persistance des données distribuées »

LES ALGORITHMES ITÉRATIFS AVEC SPARK

- D'autres cas d'usages courants de Spark
- Les algorithmes itératifs en Spark
- Machine Learning avec Spark
- TP « Implémenter un algorithme itératif avec Spark »

Jour 4

SPARK STREAMING : LES FONDAMENTAUX

- Introduction à Spark Streaming
- Les Dstreams
- Développer des applications streaming
- TP « Écrire une application streaming avec Spark »

SPARK STREAMING : TRAITEMENT MULTI-BATCH

- Les opérations multi-batch
- Le découpage temporel
- Les opérations à état
- Les fenêtres glissantes
- Introduction au streaming structuré
- TP « Traiter des batchs multiples en streaming »

SPARK STREAMING : LES SOURCES DE DONNÉES

- Panorama des sources de données streaming
- Flume et Kafka comme sources streaming
- TP « Traiter en streming des messages Kafka »

Suppléments

IMPORTER DES DONNÉES RELATIONNELLES AVEC SQOOP

- Généralités sur Sqoop
- Imports et Exports
- TP « Importer de la données depuis MySQL avec Sqoop »

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
- TP « Produire et consommer des messages Kafka »

CAPTURER DES DONÉES AVEC FLUME

- Introduction à Flume
- Architecture de Flume : Sources, Sinks et Channels
- Configuration de Flume
- TP « Collecter des logs de serveur web avec Flume »

INTEGRATION DE FLUME ET DE KAFKA (FLAFKA)

- Cas d'usage de l'intégration Flume/Kafka
- Configuration
- TP « Envoyer des messages de Flume à Kafka »