DataTips

Formation

Cloudera

DataTipsLearning.com

Tarif :
1500 € HT

Durée :
3 Jours
(21h)

Lieu :
PARIS/
Distancielle/
Votre Local

Contact :

contact@datatipslearning.com
+33 0751386021
Réf. DTPSL-CLD-2022
S'inscrire

Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

Cloudera a été cofondée en 2008. En 2018 Cloudera fusionne avec Hortonworks et s'oriente vers un monde du Cloud. Cloudera met en avant sa plateforme de gestion de données : la Cloudera Data Plateform, un mix plus efficace entre l'ancienne plateforme Cloudera CDH et celle d'Hortonworks HDP.

La CDP permet une gestion unifiée, centralisée et hybride des infrastructures Big Data. Orientée aussi bien IT que métiers, elle élargie pour ses clients les capacités d'analyse de la donnée sur tous les cas d'usage Data existant.

Objectifs

- Les bases de l'environnement Hadoop, MapReduce, Spark et HDFS.
- Comment les données sont distribuées, stockées et traitées dans un cluster Hadoop ?
- Gérer un cluster via Cloudera Manager.
- Créer, configurer et déployer des applications Apache Spark sur un cluster Hadoop.
- Traiter et requêter des données structurées à l'aide de Spark SQL.
- Charger des données dans le cluster l'aide de Flume, ou Sqoop.
- Utiliser Spark Streaming pour traiter un flux de données.
- Utiliser Apache Flume et Apache Kafka pour ingérer des données pour Spark Streaming.
- Dépanner, diagnostiquer, mettre au point et résoudre les problèmes sur Hadoop.

Public

- Data Engineer
- Data Scientist
- Data Analyste
- Data Architectes

Prérequis

- Connaissance d'au moins un langage de programmation.
- Quelques notions de base de données : Table, Ligne, Colonne.

Programme détaillé

01. Introduction au Big Data
02. Présentation d'Apache Hadoop
03. HDFS : le système de fichiers Hadoop
04. Les composants d’un cluster Hadoop
05. Architecture et utilisation de HDFS
06. Architecture et utilisation de YARN
07. Traitement des données sur un cluster Apache Hadoop

08. Importation de données relationnelles avec Apache Sqoop
09. Requêter des données Hadoop avec Apache Impala
10. Stockage de données : HDFS et Apache HBase
11. Stockage de données : Apache Kudu
12. Scripting Apache Pig
13. Principes de base d'Apache Spark
14. Les RDD de Spark
15. DataFrames et Apache Spark SQL
16. Les opérations DataFrames
17. Persistance des données distribuées
18. Développer et exécuter des applications Apache Spark
19. Traitement parallèle dans Apache Spark
20. Configurer une application Apache Spark
21. Spark Web UI
22. Comparer Spark SQL, Apache Impala, et Apache Hive sur Spark
23. Traitement des messages avec Apache Kafka
24. Capturer des données avec Apache Flume
25. Intégration d'Apache Flume et d'Apache Kafka
26. Spark Structured Streaming

Méthode pédagogique

La formation se compose d'une partie théorique, et également une partie pratique représentant 60% de de la formation.
La partie pratique contient plusieurs exercices sous forme de notebook Databricks avec les corrections, avec aussi un projet à la fin de la formation comme simulation d'une prod.

Chaque jour, une évaluation rapide des connaissances est effectuée avant de commencer les nouvelles parties de la formation.
A la fin, une synthèse globale est délivré aux stagiaires, renforcé par un projet prod.
Finalement, une évaluation QCM est proposée.

Un support de cours sera remis à chaque stagiaire comprenant les slides, les exercices et les corrigés et un git du projet prod.
Une feuille de présence est fournie en fin de formation avec une certificat de complétion de formation pour chaque stagiaire.

Le formateur est un Data Engineer expert, qui intervient sur le sujet depuis plusieurs années en formation mais aussi en conseil.