DataTips

Formation

Apache Spark

DataTipsLearning.com

Tarif :
1 500 € HT

Durée :
3 Jours (21h)

Lieu :
PARIS/ Distancielle/ Votre Local

Contact :

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

Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

Les données sont considérées actuellement le nouvel or noir, pour chercher facilement l'or noire, c'est Apache Spark qui se présente comme le candidat le plus puissant dans le domaine.

La puissance de ce moteur de traitement vient de deux faits, le premier c'est qu'on peut utiliser Spark avec différents langages de programmation sans que la performance se dégrade, notamment Scala, SQL & Python.

Le deuxième, Spark traite différentes structures de données comme les DataFrames et les graphes. Aussi différents types de traitements le batch et le streaming.

Objectifs

- Comprendre le principe des traitements distribués.
- Comparer Spark à MapReduce.
- Maîtriser les concepts fondamentaux de Spark.
- Utiliser les dataframes et les datasets.
- Comprendre l'optimisation de Spark et Spark UI.
- Comprendre les différents transformations et actions.
- Résoudre un problème data science avec Spark ML.
- Gérer des données en temps réel avec Spark Structured Streaming.
- Concevoir une application avec Spark et lancer avec spark-submit.

Public

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

Prérequis

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

Programme détaillé

1. Introduction Apache Spark
2. Spark vs MapReduce
3. Spark Architecture
4. Spark RDD
5. Dataframes et Datasets
6. Accumulators & Broadcasts
7. Spark SQL
8. Read data
9. Write Data
10. Transformations & Actions

11. Spark Partition
12. Spark Cache
13. Catalyst Optimizer
14. Spark Broadcasting
15. Spark Adaptive Query Execution (AQE)
16. Création de cluster Spark
17. Jobs, Stages et Tasks
18. Optimisation des jobs spark
19. Lancer job via spark-submit
20. Différents types de déploiement
21. Création d'un modèle avec SparkML
22. Spark Structured Streaming
23. Spark GraphFrames

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 avec des corrections et surtout 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.