DataTips

Formation

Écosystème Hadoop

DataTipsLearning.com

Tarif : 2 400 € HT

Durée : 4 Jours (28h)

Lieu : PARIS/ Distancielle/ Votre Local

Contact :

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

Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

L'écosystème Hadoop n'est ni un langage de programmation ni un service, c'est une plate-forme ou un Framework qui résout les problèmes de Big Data. Vous pouvez le considérer comme une suite qui englobe un certain nombre de services (ingestion, stockage, analyse et maintenance) à l'intérieur.

L'écosystème Hadoop est composé souvent des éléments suivants : HDFS, YARN, MapReduce, Spark, Pig, Hive, HBase et Oozie.

Objectifs

- Comprendre le Big Data.
- Quels sont les 5V ?
- HDFS présentation & architecture.
- YARN présentation & architecture.
- Stratégie de réplication, balancer et modes.
- Analyse de données avec MapReduce.
- Les types de fichier et les algorithmes de compression.
- Utiliser l'entrepôt de données Apache Hive.
- Développer une base de données NoSQL HBase.
- Comparer MapReduce et Spark.

Public

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

Prérequis

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

Programme détaillé

I. L'écosystème Hadoop
   01. Comprendre le Big Data
   02. Définitions du Big Data
   03. Les 5V
   04. Présentation de Hadoop
   05. Comparaison avec les autres systèmes
   06. Historique d'Apache Hadoop
   07. Les composants de Hadoop
   08. Scalabilité horizontale et verticale

09. Exécuter un job distribué MapReduce
10. Data Flow
11. Hadoop Streaming

II. HDFS
01. Qu'est-ce que HDFS
02. Pourquoi HDFS
03. Les concepts de HDFS
04. Architecture HDFS
05. Mode d'écriture
06. Mode de lecture
07. Les types d'accès
08. Stratégie de réplication, balancer et modes
09. Problématique des petits fichiers
10. Les commandes CLI de HDFS
11. Les formats de stockage
12. Les algorithmes de compression
13. Fédération HDFS
14. HDFS High Availability
15. Modèle de cohérence
16. Copie parallèle avec distcp
17. Un cluster HDFS équilibré
18. Intégrité des données
19. Sérialisation
20. Sécurité

III. MapReduce
01. Présentation de MapReduce
02. Architecture de MapReduce
03. Utilisation de MapReduce en Java
04. MapReduce dans Hadoop 1.0 & 2.0
05. MapReduce vs Spark
06. Les tests unitaires avec MR
07. Exécution sur un cluster et en local
08. Soumission et monitoring d'un job
09. Remote Debugging
10. Optimisation d'un job MR
11. Cycle de vie d'un job
12. Shuffle & Sort
13. Les types de MapReduce
14. Les Formats d'entrée et de sortie
15. Partitions et map des données

IV. YARN
01. Limites de MapReduce1 MR1
02. Concept général de YARN
03. Applications YARN
04. Anatomie d'une application YARN

05. Architecture de MR1
06. Architecture de YARN
07. Options du Scheduler
08. Lancer une Application Spark avec YARN
09. YARN vs MESOS
10. YARN optimisation : Executor, cores & mémoire
11. Demandes de ressources
12. Durée de vie des applications
13. Création d'applications YARN
14. Planification des retards
15. Sécurité d'une application YARN
16. Fédération YARN
17. Cache partagé YARN

V. Hive
01. Qu'est-ce que Hive
02. Architecture de Hive
03. Hive vs les bases de données traditionnelles
04. Syntaxe HiveQL
05. Les types de données
06. Les formats de données Hive
07. Partitionner
08. Hive vs Hive 2
09. Hive vs Impala
10. Utiliser Hive avec Spark
11. Installation de Hive
12. Exécution Hive
13. Requêtage de données
14. Les fonctions UDF
15. HiveServer2 & Beeline
16. Hive CLI
17. Hive HCatalog
18. Hive WebHCat
19. Hive Logging
20. Les opérations DDL
21. Les opérations DML
22. Les opérations SQL

VI. Sqoop
01. Introduction
02. Fonctionnement de Sqoop
03. Connecteurs Sqoop
04. Exemple d'import
05. Exemple d'export
06. Les fonctions principales
07. Les arguments
08. Code généré
09. Consistance des imports

23. HBase et HDFS
24. Monitoring HBase
25. Test unitaire d'une Applications HBase

IX. Spark
01. Les bases de Spark
02. MapReduce vs Spark
03. Installation de Spark
04. Les RDD
05. Création d'une Applications Spark
06. Configuration d'une Applications Spark
07. DataFrames & Apache Spark SQL
08. Les actions et transformations
09. Spark Streaming, ML & graphe
10. Les variables broadcast et accumulators
11. Anatomie d'un job Spark
12. Les exécuteurs et les gestionnaires de cluster

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.