DataTips

Formation

Spark Performance Tuning

DataTipsLearning.com

Tarif : 1 500 € HT

Durée : 2 Jours (14h)

Lieu : PARIS/ Distancielle/ Votre Local

Contact :

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

Plus de détails :

Objectifs
Public / Prérequis
Programme détaillé
Méthode pédagogique

Spark Performance Tuning fait référence au processus d'ajustement des paramètres pour ajuster la mémoire, les cores et les instances utilisés par le système. Ce processus garantit que Spark a des performances sans faille et empêche également le goulot d'étranglement des ressources dans Spark.

Plusieurs moyens d'optimisation possible, il y a certains à la charge de Spark et son configuré par défaut, et il y a d'autre qui ont à la charge du développeur.

Objectifs

- Comprendre le concept d'optimisation.
- Comprendre les composants de base de Spark.
- Utiliser les différent API de Spark, Dataframe et Dataset.
- Différencier les actions, transformation et lazy évaluation.
- Utiliser les différents types de jointure.
- Comprendre comment calculer les ressources.
- Apprendre à bien partitionner les données, et les cacher.

Public

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

Prérequis

- Connaissance d'au moins un langage de programmation.
- Connaissance de base d'Apache Spark.

Programme détaillé

1. Concepts et architecture de base de Spark
2. L'anatomie d'un job Spark Job : DAG, Jobs, Stages & Tasks
3. Transformations, Actions, et Lazy Evaluation
4. L'API DataFrame
5. Higher-Order Functions dans DataFrames et Spark SQL
6. L'API Dataset
7. Traitement en parallèle
8. La configuration du shuffle
9. Le Partitioning et Bucketing
10. Le Catalyst Optimizer
11. Scheduling d'un job Spark
12. Les modes d'exécution : cluster, client et local
13. User-Defined Functions et Aggregate Functions (UDFs, UDAFs)
14. Choisir un type de jointure

15. Shared Variables
16. Spark Partitioner Object : Hash, Range et Custom Partitioning
17. Iterator-to-Iterator Transformations avec mapPartitions
18. Cluster Sizing
19. Ajuster Spark Settings
20. Calculer les ressources à allouer à une application Spark
21. Dynamic Allocation
22. Utiliser l'interface Spark UI
23. Memory Management pour les Datasets et les DataFrames
24. Serialization et Deserialization
25. Caching et Persistence
26. Adaptive Query Execution
27. Dynamic Partition Pruning (DPP)
28. Catalog Plugin API et DataSourceV2
29. Accelerator-Aware Scheduler
30. Dynamically Coalescing Shuffle Partitions
31. Dynamically Switching Join Strategies
32. Dynamically Optimizing Skew Joins
33. Les concepts I/O
34. Data locality
35. Types des fichiers et des compressions
36. Memory Pressure et Garbage Collection
37. Tester une Application Spar
38. Quelques techniques de Debugging

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.