Ambient IT
140 Avenue Jean Lolive
93500 Pantin

Mail : formation@ambient-it.net
Tél : +33 (0)1 48 10 09 46

AMBIENT-IT

Formation Spark Tuning Avancé

Durée 4 jours (28 heures)

S'inscrire

Présentation

Conçu en 2009 aux États Unis, Apache Spark est un moteur d'analyse unifié pour le traitement de grande quantité de données à grande échelle. Cet outil se démarque par sa simplicité d'utilisation malgré sa capacité à délivrer des analyses sophisitiquées.

Cette formation Spark Tuning est destinée aux administrateurs voulant optimiser les performances de leur système de gestion de données. L'ajustement et l'optimisation des ressources (CPU cores et mémoires) joue un rôle important pour le maintien d'un système informatique de bonne qualité.

Après une introduction au langage Scala, et une explication de Spark, nous étudierons l'api RDD, les dataframes, le Spark Streaming. Nous verrons ensuite Spark en production et finirons sur une introduction au Machine Learning.

À chaque fois, des exercices pratiques sur des clusters de machines avec des datasets significatifs permettront d'assimiler par la pratique les concepts présentés.

Comme toutes nos formations, celle-ci vous présentera la dernière version stable en date (Spark 3.3 à la date de l'article).

Objectifs

- Être capable d'installer et d'utiliser Spark 3 et ses nouveautés de manière autonome
- Être capable d'utiliser Scala comme langage principal dans Spark
- Comprendre et optimiser les dataframes
- Appréhender le tuning sur Spark en production en utilisant les bonnes pratiques

Public visé

- Développeurs
- Architectes
- Administrateurs systèmes
- DevOps

Pré-requis

- Connaissances de base d'un système Unix
- Connaissance de Python

Programme de la formation Spark et Tuning avancé

Jour 1 – Introduction Scala et Spark

- Pourquoi Scala est le langage du Bigdata ?
- Introduction au paradigme fonctionnel
- Installation des environnements
- Hands-on Scala
- Syntaxe
- Pattern matching
- API collection
- Les types fonctionnels
- Pourquoi Spark ?
- Architecture de Spark

Jour 2 – Comprendre et utiliser Spark

L'API RDD

- Présentation des RDD
- PairedRDD
- Manipulation de l'api RDD (transformations , actions ....)
- L'import et l'export depuis et vers : cSv, Avro et Elasticsearch

Dataframe

- Présentation des Dataframes
- L'api Dataframe et UDF
- SqlContext
- Utilisation de SQL avec des Dataframes
- Les Datasets

Jour 3 – Dataframe et optimisation

Optimisation

- L’analyse du DAG via Spark-UI
- Pattern d’optimisation
- Cache et persistance
- Impact de la localité des données sur les performances

Spark streaming

- StreamingContext
- DStream
- Continuous Aggregations
- Analyse temps-réel depuis un Apache Kafka
- Les problématiques des garanties de livraison
- Spark vs Flink

Jour 4 – Spark en prod et conclusion

Spark en production

- Spark en cluster : Yarn, Mesos, Standalone
- Yarn client vs Yarn cluster
- Stockage (HDFS, S3, Cassandra ....)

Architecture

- Architecture Lambda
- Architecture Kappa

Introduction au Machine Learning (Optionnel)

- Les classes d'algorithmes pour le ML : supervisé et non supervisé
- Les algorithmes de ML
- Comment fonctionne l'algorithme de la régression linéaire et / ou de la régressionlogistique
- Mise en pratique d’un algorithme de régression linéaire ou la régression logistique

2 modules spécifiques sont disponibles en Intra-entreprise uniquement

Module for Data Engineer - Spark Scala

Jour 1 – RDD & Dataframes

L’API RDD

- Présentation des RDD
- PairedRDD
- Manipulation de l’api RDD (transformations , actions, etc.)
- L’import et l’export depuis et vers : cSv, Parquet

Dataframe

- Présentation des Dataframes
- L’api Dataframe et UDF
- Utilisation de SQL avec des Dataframes
- Les Datasets

Jour 2 – Mise en prod & Optimisation

Optimisation

- L’analyse du DAG via Spark-UI
- Pattern d’optimisation
- Cache et persistance
- Impact de la localité des données sur les performances

Spark en production

- Spark en cluster : Yarn, Mesos, Standalone
- Yarn client vs Yarn cluster
- Stockage (HDFS, S3, Cassandra, etc.)

Module for Data Scientist - Spark Python

Jour 1 – RDD & Dataframes

L’API RDD

- Présentation des RDD
- PairedRDD
- Manipulation de l’api RDD (transformations, actions, etc.)
- L’import et l’export depuis et vers : CSV, Parquet

Dataframe

- Présentation des Dataframes
- L’api Dataframe et UDF
- Utilisation de SQL avec des Dataframes
- Les Datasets

Jour 2 – Spark ML/ MLlib

Algorithmes

- Les classes d’algorithmes pour le ML : supervisé et non supervisé
- Les algorithmes de ML
- Comment fonctionne l’algorithme de la régression linéaire, la régression
- Logistique, Random Forest...
- Clustering : KNN, K-mean

MLlib

- Introduction à MLlib 2.0
- Pipelines : Transformer, Estimator, Model
- Cross-Validation
- Hyperparameters tuning
- ML persistence: sauvegarde et chargement des pipelines

Sociétés concernées

Cette formation s’adresse aux entreprises, petites ou grandes, souhaitant former ses équipes à une

nouvelle technologie informatique avancée.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiple permet de vérifier l'acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.