+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 05/02/2025

S'inscrire

Formation Spark 3 & Machine Learning

3 jours (21 heures)

Présentation

Spark est un framework pour effectuer des calculs distribués sur un cluster d'ordinateurs. Cette formation présente la toute nouvelle version 3.5.4, qui apporte un lot considérable de nouveautés ainsi qu'une amélioration impressionnante des performances !

Créé en 2009 à Berkeley, il est en train de devenir la plateforme « Big Data » privilégiée, qui remplace peu à peu l'écosystème Hadoop, grâce à des API unifiées en Java, Scala, Python, R qui le rendent très facile d'usage.

La formation passe en revue les principaux composants de Spark, ainsi que les nouveaux packages :

- Spark Core
- Spark SQL
- Spark Streaming
- Spark ML
- GraphFrame
- SparkR
- Deep Learning pipeline

Notre formation Spark et Machine Learning présente aussi l'intégration de Spark avec HDFS. Elle présente l'API de Spark. Les travaux pratiques sont réalisés en Scala par défaut (ou bien Python en option).

Objectifs

- Être capable d'utiliser Spark 3 et ses nouveautés de manière autonome
- Comprendre le concept de Machine Learning et les concepts fondamentaux de Spark, être capable de les utiliser
- Manipuler des volumes importants de données en utilisant les bonnes pratiques dans Spark 4
- Comprendre la documentation, l'API et l'écosystème du Big Data
- Intégrer Spark dans un écosystème Hadoop

- Créer des applications d'analyse en temps réel avec Spark Streaming
- Faire de la programmation parallèle sur un cluster
- Maîtriser Spark SQL

Public visé

- Développeurs
- Architectes
- Administrateurs systèmes
- DevOps

Pré-requis

- Connaissances de base d'un système Unix
- Connaissance de Scala ou Python & Git
- Culture orientée stats
- Tester Mes Connaissances

Pré-requis techniques

- Avoir Visual Studio Code installé

Programme de la formation Spark et Machine Learning

Jour 1 - Comprendre et utiliser Spark 3

Contexte et problématique du Big Data - Calcul distribué

- Pourquoi Spark ? Les nouveautés de la version 2 & 3
- Installation en standalone, test avec jupyter
- Spark Core (Remplaçant de MapReduce)
- RDD Resilient Distributed Datasets
- PairedRDD
- Spark Context VS Spark Session
- DAG Directed Acyclic Graph
- RDD Objects, DAG Scheduler, Task Scheduler, Worker
- Hadoop et HDFS
- NameNode & DataNode
  - core-site, hdfs-site
- Spark sur un Cluster
- Spark Standalone : Cluster Manager, Worker, Executor, Spark Context
- Mesos (Private Cluster), Marathon, YARN
- Structured API

Spark SQL (Remplaçant de HIVE)

- SQLContext
- HiveContext
  - DataFrames
  - Spark Structure, Schéma et partitionnement

Jour 2 - Appréhender le Machine Learning et son intégration dans Spark 3

Introduction au Machine Learning (ML)

- Apprentissage supervisé
- Apprentissage non-supervisé
- Clustering : KNN, K-mean
- Régression : Arbre de régression
- Classification : Random Forest, SVM, AUC, Courbe ROC

Spark ML - Introduction

- Pipelines : Transformer, Estimator, Model
- ML persistence
- MLlib in R & PySpark

DataVisualisation

- Matplotlib
- Seaborn
- Plotly
- Bokeh

GraphFrame

- Présentation du package

Jour 3 - Spark 3 en mode avancé : Manipuler les données à grande échelle

Spark Streaming

- Structured Streaming API
- StreamingContext
- Static et Dynamic Datasets
  - Continuous Aggregations
  - Encoders
- Analyse temps-réel d'un fichier de log (Real-Time Analytics)
  - Gagner en efficacité grâce à Catalyst Optimizer et Tungsten Engine
- Création d'agents, de sources, channel et sink

- Sérialisation avec Avro RPC

SparkR

- Présentation du package

Deep Learning pipeline

- Présentation du package
- Concept de transfert learning

Conclusion

- Lambda VS Kappa architecture

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

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 multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

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