Ambient IT
128 Boulevard Macdonald
75019 Paris

Mail : formation@ambient-it.net
Tel : +33 (0)1 40 34 11 53

Formation Spark & 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 2.3.2 sortie en septembre 2018, 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 4 des 5 principaux composants de Spark : 1. Spark Core 2. Spark SQL 3. Spark Streaming 4. Spark ML Le module non présenté est celui sur les graphes (GraphX) [En option pour cette formation] La formation présente aussi l'intégration de Spark avec HDFS. Elle présente l'API Python de Spark (Les travaux pratiques sont réalisés en Python ou Scala au choix du participant).

Objectifs

- Être capable d'installer et d'utiliser Spark 2 et ses nouveautés de manière autonome
- Comprendre le concept de Machine Learning et être capable de l'utiliser dans Spark 2
- Manipuler des volumes importants de données en utilisant les bonnes pratiques dans Spark 2

Public visé

Développeurs, Architectes, Administrateurs systèmes, DevOps

Pré-requis

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

Programme de la formation Spark et Machine Learning

Jour 1 - Comprendre et utiliser Spark 2

Pourquoi Spark ?

Installation

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

Spark SQL (Remplaçant de HIVE)

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

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

Spark ML (Remplaçant de Mahout)

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

Jour 3 - Spark 2 en mode avancé : Manipuler les données à grande échelle

Spark Streaming

- StreamingContext
- Static et Dynamic Datasets
- Continuous Aggregations
- Avantages de DStreams
- Encoders
- Analyse temps-réel d'un fichier de log (Real-Time Analytics)

- Gagner en efficacité grâce à Catalyst Optimizer et Tungsten Engine
- Intégration avec Flume
    - Création d'agents, de sources, channel et sink
    - Sérialisation avec Avro RPC

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

Conclusion

- Lambda VS Kappa architecture

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.