+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 PySpark : traitez rapidement vos données en masse

3 jours (21 heures)

Présentation

Notre formation PySpark vous apprendra à manipuler rapidement un grand volume de données, bénéficiez de la puissance de l'API de Python pour maîtriser Apache Spark.

Notre formation se compose en différents modules nécessaires pour comprendre l'écosystème d'Apache Spark et l'utilisation de PySpark. En effet, nous débuterons par une présentation d'Hadoop (son architecture et ses composants). Ensuite, nous vous guiderons sur l'installation de ce framework Big Data ainsi que la configuration de PySpark.

Vous découvrirez comment utiliser l'API de Python sur Spark pour manipuler vos données, ainsi, vous maitriserez l'intégralité de votre processus ETL (extraction, chargement et transformation des données). De plus, un module est dédié à l'utilisation de Pandas pour approfondir l'utilisation de l'outil. Vous saurez également utiliser Spark pour le machine learning, Spark Streaming et évidemment Spark SQL.

Notre formation vous présentera la dernière version d'Apache Spark, Spark 3.5.

Objectifs

- Comprendre le rôle d'Hadoop et de Spark dans le Big Data.
- Maîtriser l'architecture et le fonctionnement d'Hadoop
- Installer et interagir avec Spark
- Utiliser Spark SQL pour manipuler les DataFrames
- Appliquer PySpark et Pandas pour la manipulation de données

Public visé

- Data analysts

- Data scientists
- Data engineers
- Développeurs

Pré-requis

- Connaissances en SQL
- Connaissances de base en mathématiques et statistiques
- Connaissances de base de Python

Programme de notre formation PySpark

Présentation d'Hadoop

- Qu'est-ce qu'Hadoop ?
- Son rôle dans le Big Data
- Présentation de son architecture
- Comment Hadoop fonctionne ?
- Les modules principaux
  - HDFS
  - YARN
  - MapReduce
  - Hadoop Common

Présentation de Spark

- Spark vs Hadoop
- Les différences avec MapReduce
- Pourquoi utiliser Spark ?
- Les fonctionnalités
  - MLlib
  - Streaming
  - SQL
  - GraphX
- Comment fonctionne Spark ?
- Les ensembles de données
  - RDD
  - DataFrames
  - Data Sets

Comment installer Spark ?

- En local
- Sur une infrastructure distribuée
- Sur le Cloud
- Première interaction avec Spark

Spark SQL

- Introduction à Spark SQL
- Création de DataFrames
- Manipulation des DataFrames
- Chargement des données
- Stockage des données
- Différences entre l'API SQL et l'API dataframe
- Explication du fonctionnement de catalyst, et outils de diagnostique et debugging.

Utiliser PySpark

- Présentation de PySpark
- Utilisation de SparkSQL pour manipuler des données
- Charger des données de différents formats
- Transformer ses données
- TP : Chargement et transformation de données avec PySpark

L'API Pandas

- Installer Pandas
- Transform et apply
- Comment les types de données changent ?
- Les hints
- Les bonnes pratiques de développement

Spark.ml

- Apprentissage supervisé
- Random trees
- Créer des recommandations personnalisées
- Traitement de données textuelles
- Automatiser ses analyses avec des pipelines

Spark Streaming

- DStream
- Les sources de données
- Utiliser l'API
- Modifier des données

Troubleshooting

- Exceptions liées à l'absence de mémoire

- Échec répété de la tâche Spark
- Échec de la commande Spark Shell
- FileAlreadyExistsException
- Erreur "Too Large Frame"
- Les jobs Spark échouent à cause d'échecs de compilation

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.