Ambient IT
140 Avenue Jean Lolive
93500 Pantin

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

AMBIENT-IT

Mis à jour le 26/10/2022

Formation Databricks

S'inscrire

Durée 2 jours (14 heures)

Présentation

Databricks simplifie votre process Big Data. Cette plateforme créée par les fondateurs d'Apache Spark rend plus sûr le déroulement du processus ETL.

En effet, Databricks utilise l'architecture d'Apache Spark en y ajoutant des pipelines fiables et puissants . L'outil fournit une gestion automatisée des clusters et des notebooks de style IPython.

Les pipelines de données peuvent être écrits dans de nombreux langages (Scala, SQL, R, Python) conçus dans des workspaces collaboratifs.

Databricks est également rigoureusement sécurisé grâce à un modèle de protection unifié comportant toutes les fonctionnalités (gestion des identités, cryptage...).

Notre formation Databricks vous présentera comment élaborer un processus ETL complet. Nous commencerons par la présentation du système, puis, l'extraction des données, leurs chargements, leurs transformations et enfin nous évoquerons les dashboards et l'utilisation d'IntelliJ IDE.

Comme toujours, nous vous présenterons la dernière version de l'outil, Databricks 9.

Objectifs

- Bien connaître les spécificités de Databricks
- Extraire les données avec Databricks
- Savoir comment transformer et charger ses données
- Utiliser les dashboards et déployer son processus

Public visé

- Développeurs
- Data Engineer
- Architectes

- Administrateurs système
- Data miners
- Data scientists
- Data analysts
- Business intelligence analysts
- Market inteligence analysts

Pré-requis

- Avoir idéalement suivi nos formations Spark ML ou Spark Tuning Avancé
- Connaissance de Scala, SQL et idéalement Python.

Programme de la formation Databricks

Introduction

- Présentation de l'outil
- Pourquoi utiliser Databricks?
- Databricks vs Apache Spark
- Présentation de l'interface
- Les notebooks
- Créer un cluster et une table
- Créer des jobs
- Créer un pool

Extraire ses données

- Importer ses données
- Ajouter des schemas
- Gérer les tables avec SQL
- Utiliser Python sur Databricks

Transformer ses données

- Transformer ses données avec Scala
- Manipulation de données avec Spark SQL
- Modification de données avec Python
- Utiliser l'API DataFrame

Charger ses données

- Le fichier nested XML
- Le fichier nested json
- Les tables DELTA

Dashboard et déploiement du processus

- Présentation du Dashboard
- Développer un job pour rafraichir le dashboard
- Créer un projet avec IntelliJ IDE
- Créer son application
- Ajouter des dépendances
- Externaliser les propriétés
- Envoyer les jobs

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.