Programmer avec Apache Spark de Databricks

Maîtriser les Fondamentaux de Spark sur la Plateforme Databricks

DESCRIPTION

Apache Spark est le moteur d'analyses unifiées ultra-rapide incontournable pour le big data.

Adopté massivement par des acteurs majeurs comme **Netflix**, **Yahoo** et **eBay**, il traite quotidiennement des pétaoctets de données. Cette formation de deux jours vous apprend à exploiter la puissance de Spark spécifiquement au sein de l'écosystème **Databricks**.

Ce cours propose un panorama complet, conçu pour être immédiatement applicable. Notre approche alterne entre :

- Présentations théoriques pour bâtir des fondations solides.
- Exercices pratiques pour vous rendre opérationnel sur la plateforme.

Ce module est construit pour couvrir les piliers de la data engineering moderne sur Databricks

- **Les fondements d'Apache Spark** : l'architecture Spark, les DataFrames, Spark SQL et PySpark.
- **La manipulation de données avec les APIs Spark** : DataFrame et Spark SQL pour ingérer, transformer et manipuler efficacement de grands volumes de données.
- **La plateforme Databricks** : les composants clés de la plateforme Databricks et son architecture (Clusters, Notebooks, Workspace) pour gérer vos projets.
- **Le stockage optimisé avec Delta Lake** : les caractéristiques et les avantages déterminants du format Delta Lake (transactions ACID, Time Travel) qui est au cœur de la plateforme.

À l'issue de la session, vous conservez tous les notebooks de cours. Vous pouvez réexécuter vos travaux pratiques sur le service cloud Databricks Free Edition (gratuit) ou les exporter en code source pour n'importe quel environnement Spark.

Ce cours est aligné sur la certification "Databricks Data Engineer Associate".

Il couvre les trois piliers de l'examen : les fondamentaux de Spark, l'utilisation de la plateforme Databricks et la compréhension de Delta Lake. Le passage de l'examen se fait après la formation et reste optionnel.

OBJECTIFS PEDAGOGIQUES

Stage pratique
Data Engineering

Code :
ASPWD

Durée :
2 jour(s) (14,00 heures)

Exposés : **40 %**
Cas pratiques : **50 %**
Echanges d'expérience : **10 %**

Inter-entreprises :
Prochaines sessions disponibles sur notre site web.
Tarif : 1 800,00 € HT / participant

Intra-entreprise :
Tarifs et dates sur demande.

- Décrire les concepts fondamentaux d'Apache Spark (Architecture, RDD, DataFrames, Spark SQL, PySpark).
- Utiliser les APIs Spark (DataFrame et Spark SQL) pour ingérer, transformer et manipuler des données.
- Identifier les composants clés de la plateforme Databricks et son architecture.
- Utiliser l'espace de travail Databricks pour organiser le code (Notebooks, Repos/Git) et gérer les ressources de calcul (Clusters).
- Expliquer les caractéristiques et les avantages du format de stockage Delta Lake.
- Mettre en œuvre les opérations clés de Delta Lake (création de tables, transactions ACID : MERGE, UPDATE) et utiliser le "Time Travel"

PUBLIC CIBLE

Data engineers et data analysts ayant l'expérience des traitements Big Data, qui souhaitent apprendre à utiliser Apache Spark pour effectuer leurs traitements Big Data, construire des jobs Spark à destination de la production et comprendre mais aussi déboguer des applications Spark.

PRE-REQUIS

- Une première expérience avec Apache Spark est conseillée

METHODE PEDAGOGIQUE

Formation avec apports théoriques, échanges sur les contextes des participants et retours d'expérience du formateur, complétés de travaux pratiques et de mises en situation. La formation mélange les langages Python et Scala.

Chaque sujet abordé comprend une partie d'exposé couplée à une mise en pratique de Spark au travers d'un environnement type notebook web. Inspiré d'outils tels IPython/Jupyter, les notebooks permettent aux participants de développer des jobs, des requêtes d'analyse et des représentations visuelles s'appuyant sur leur propre cluster Spark, le tout depuis leur navigateur web.

PROFIL DES INTERVENANTS

Cette formation est dispensée par un·e ou plusieurs consultant·es d'OCTO Technology ou de son réseau de partenaires, expert·es reconnus des sujets traités.

Le processus de sélection de nos formateurs et formatrices est exigeant et repose sur une évaluation rigoureuse leurs capacités techniques, de leur expérience professionnelle et de leurs compétences pédagogiques.

MODALITÉS D'ÉVALUATION ET FORMALISATION À L'ISSUE DE LA FORMATION

L'évaluation des acquis se fait tout au long de la session au travers des ateliers et des mises en pratique.

Afin de valider les compétences acquises lors de la formation, un formulaire d'auto-positionnement est envoyé en amont et en aval de celle-ci.

En l'absence de réponse d'un ou plusieurs participants, un temps sera consacré en ouverture de session pour prendre connaissance du positionnement de chaque stagiaire sur les objectifs pédagogiques évalués.

Une évaluation à chaud est également effectuée en fin de session pour mesurer la satisfaction des stagiaires et un certificat de réalisation leur est adressé individuellement.

PROGRAMME PEDAGOGIQUE DETAILLE

Jour 1

OUVERTURE DE SESSION

- Accueil des participants et tour de table des attentes
- Présentation du déroulé de la formation

INTRODUCTION ET ENVIRONNEMENT DATABRICKS

- Introduction à Spark
  - Historique
  - Place dans le Big Data
- La plateforme Databricks
  - Vue d'ensemble de l'architecture
  - Le Workspace
  - Gestion des clusters
- Le Notebook Databricks
  - Prise en main et bonne pratique

ARCHITECTURE ET CONCEPTS DE SPARK

- Les concepts fondamentaux
  - RDD
  - SparkContext
  - Driver/Executor

- Le modèle d'exécution
  - Stages
  - Tasks
  - Shuffles
  - Transformations vs Actions
- PySpark vs Spark SQL
  - Dans quel cas utiliser quelle solution

LES APIs SPARK DATAFRAMES

- Le DataFrame
  - Création
  - Lecture et écriture de données (formats CSV, JSON, Parquet)

SPARK SQL

- Filtrage et projection
  - Clauses where(), select() et withColumn()
- Jointures (joins)
  - Types de jointures
  - Les meilleures pratiques

Jour 2

LE COEUR DU LAKEHOUSE : DELTA LAKE

- Présentation de Delta Lake
  - Le format open source
  - La gestion des métadonnées
- Les avantages clés
  - ACID
  - Gestion des schémas
  - Time Travel
- Opérations Delta
  - Création de tables
  - Manipulation de données avec MERGE, UPDATE et DELETE

MEILLEURES PRATIQUES DE DATA ENGINEERING SUR DATABRICKS

- Modélisation par couches (Bronze, Silver, Gold)
  - L'architecture LakeHouse et ses principes
- Gestion des versions et collaboration
  - Utilisation des repos Databricks (intégration Git)
- Tests unitaires basiques
  - Écriture de code "Spark testable"

OPTIMISATION ET DIAGNOSTICS

- Analyse de la performance
  - Lecture et interprétation du Spark UI (Stages, Tasks, Exécuteurs)
- Optimisation des requêtes

    - Les techniques de base (coalesce/repartition, taille du fichier)

CONCLUSION

- Bilan et revue des concepts clés évoqués lors de la formation
- Temps d'échange autour des questions et réponses additionnelles
- Ressources complémentaires pour préparer le passage de la certification "Databricks Data Engineer Associate"

Accessibilité

L'inclusion est sujet important pour OCTO Academy.
Nos référent·es sont à votre disposition pour faciliter l'adaptation de votre formation à vos besoins spécifiques.
Pour les contacter : academy.accessibilite@octo.com