Programme de formation Databricks

Notre contact :

Jean-Noël SAUNOIS
contact@stere-informatique.fr
46, rue de Lagny
93100 Montreuil

PRESENTATION : Cette formation vous apprend à concevoir un processus ETL complet et sécurisé sur la plateforme Spark. Vous y découvrirez l'extraction, le chargement, la transformation des données et la création de Dashboards, le tout avec des pipelines collaboratifs, multi-langages et optimisés grâce à l'architecture Databricks.

OBJECTIFS

A l'issu de cette formation, les participants seront capables de :

- 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.

DURÉE DE LA FORMATION

- 2 jours, soit 14 heures de formation.

PRÉREQUIS À LA FORMATION

- Connaissance de Scala, SQL et idéalement Python.

PUBLIC VISÉ

- Développeurs, Data Engineer, Architectes, Administrateurs système, Data miners, Data scientists, Data Analyst, Business intelligence Analyst, Market intelligence Analyst.

CONTENU DE LA FORMATION

1. **Introduction et prise en main (2h)**
   - Présentation de Databricks : concepts clés et avantages
   - Databricks vs Apache Spark : différences et complémentarité
   - Découverte de l'interface et des notebooks Databricks
   - Création et gestion des clusters, tables, pools et jobs
2. **Extraction et ingestion des données (3h)**
   - Importer et structurer ses données
   - Ajout de schémas et gestion des tables via SQL
   - Utilisation de Python pour l'ingestion et le contrôle des données
3. **Transformation des données (3h)**
   - Transformer ses données avec Python et Scala
   - Manipuler et modifier les données avec Spark SQL
   - Utilisation de l'API DataFrame pour optimiser les traitements
4. **Chargement et stockage (1h)**
   - Charger des données depuis des fichiers nested XML et JSON
   - Travailler avec les tables Delta pour assurer la fiabilité des données

5. **Visualisation et déploiement (2h)**
    - Création et présentation de Dashboards interactifs
    - Développement de jobs pour automatiser les mises à jour
    - Créer un projet avec IntelliJ IDE : configuration, dépendances, externalisation des propriétés et déploiement des jobs
    - Bonnes pratiques pour mettre à jour et maintenir vos contenus Databricks
6. **Optimisation et performances (2h)**
    - Nouveautés de la plateforme et bonnes pratiques d'optimisation
    - Concepts clés de Spark Streaming : fenêtres temporelles (window functions), watermarking et agrégations temps réel
    - Automatisation avec Delta Live Tables : création et gestion de flux de données fiables
    - Orchestration de workflows et pipelines de bout en bout
7. **Gouvernance et sécurité (1h)**
    - Découverte du Unity Catalog : gestion centralisée des métadonnées
    - Mise en place de la gouvernance et de la sécurité dans un environnement Lakehouse

MOYENS PÉDAGOGIQUES ET TECHNIQUES

- Accueil des stagiaires dans une salle dédiée à la formation.
- Supports de formation projetés sur écran.
- Etude de cas concrets.
- Mise à disposition en ligne de documents supports à la suite de la formation.

Formation à distance : utilisation de la plateforme de visioconférence client et partage d'écran pour le support de formation.

Attention : Nous ne fournissons pas le matériel informatique pour la formation. Les stagiaires doivent être équipés d'ordinateur et d'une connexion internet.

MODALITÉS D'ÉVALUATION

- Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
- Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MODALITÉS ET DÉLAIS D'ACCES

- L'inscription doit être finalisée 24 heures avant le début de la formation.

TARIFICATION

Inter entreprise :

- 1200€ HT par personne pour 2 jours de formation.

Intra entreprise :

- Contactez-nous pour un devis personnalisé.