FORMATION · BIG DATA

Databricks – Lakehouse et traitements Spark

Durée de la formation : 2 jours
Tarif : 1 590 € HT par participant
Réf : BI310
Délai d'accès : 2 à 4 semaines
Niveau : Intermédiaire

Prérequis

Pratique de Python ou du SQL. Des notions de traitement de données distribuées sont un plus.

Public concerné

Data engineers, data scientists, développeurs

Objectifs pédagogiques

- Comprendre l'architecture lakehouse et la place de Databricks
- Traiter les données avec Spark et Delta Lake
- Organiser les traitements en couches et les automatiser
- Gouverner les données avec Unity Catalog

Objectifs opérationnels

- Utiliser les notebooks et les clusters
- Lire, transformer et écrire des tables Delta
- Mettre en œuvre une architecture médaillon
- Planifier des traitements avec les Jobs
- Gérer les droits et le lignage dans Unity Catalog

Méthodes mobilisées et modalités d'évaluation

Apports théoriques, démonstrations et nombreux travaux pratiques sur des cas concrets. À l'issue de chaque module, le formateur évaluera chacun des participants en fonction des cas pratiques et exercices effectués.

Databricks – Lakehouse et traitements Spark

Databricks est une plateforme lakehouse construite autour d'Apache Spark et de Delta Lake : ingénierie des données, analytique et machine learning y partagent les mêmes données.

Cette formation vous rend autonome pour construire des traitements de données fiables sur Databricks, de l'ingestion aux tables prêtes pour l'analyse, avec la gouvernance d'Unity Catalog.

Les travaux pratiques sont réalisés dans un espace de travail Databricks d'entraînement.

Inscription

Contactez-nous au 01 86 95 71 60 ou à l'adresse formation@adbi.fr pour connaître les dates de notre prochaine session. Inscription au plus tard 10 jours avant le début de la session.

Accessibilité

Nos salles de formation sont accessibles aux personnes à mobilité réduite. N'hésitez pas à nous informer si vous avez des besoins particuliers et s'il y a des besoins d'aménagements spécifiques lors de votre inscription.

Modes de participation

Nos formations sont proposées sous 3 formats :

- Chez nous, près de la place d'Italie (Paris 13e)
- Chez vous, dans vos locaux
- À distance, en classe virtuelle

Attestation

Une attestation de fin de formation sera remise à tous les participants à l'issue de leur parcours.

Programme :

Jour 1

- **Découvrir Databricks**
  - Le lakehouse : principes et bénéfices
  - Espace de travail, notebooks, clusters et entrepôts SQL
  - Databricks sur Azure, AWS et Google Cloud
- **Spark et Delta Lake**
  - DataFrames PySpark et Spark SQL
  - Tables Delta : transactions, historique, voyage dans le temps
  - MERGE, mises à jour et optimisation des fichiers
  - Travaux pratiques : premières tables Delta

Jour 2

- **Construire des pipelines**
  - Architecture médaillon : bronze, silver, gold
  - Ingestion incrémentale (Auto Loader)
  - Pipelines déclaratifs : principes
  - Jobs, dépendances et planification
- **Gouvernance et exploitation**
  - Unity Catalog : catalogues, schémas, droits, lignage
  - Coûts : dimensionnement et arrêt automatique des clusters
- **Atelier de synthèse**
  - Pipeline lakehouse de bout en bout