

Delta Lake – Implémenter des lakehouses et pipelines de données

4 jours (28 heures)

Perfectionnement

Présentiel ou classe à distance

à partir de

3 140 € HT par personne (TVA en vigueur) **Repas inclus en présentiel**

Contenu mis à jour le 25/09/2026. Document téléchargé le 07/10/2026.

Objectifs

A l'issue de cette formation, vous serez capable de :

- Mettre en oeuvre les techniques modernes de gestion et d'ingénierie des données avec Delta Lake
- Déterminer comment les transactions ACID (Atomicité, Cohérence, Isolation et Durabilité) fiabilisent les Data Lakes à l'échelle
- Exécuter des travaux de streaming ou en batch de façon concurrente sur votre Delta Lake
- Utiliser le "Time Travel" pour des rollbacks ou pour explorer des anciennes versions
- Construire des pipelines de données robustes selon l'architecture en médaille.

Les compétences visées par cette formation font l'objet d'une évaluation.

À qui s'adresse cette formation ?

Ingénieurs Data, architectes Data, Data Scientists, Data Analysts, chefs de projets, chefs de produits Data, ingénieurs DevOps, MLOps et/ou AIOps.

Prérequis

Avoir une expérience en BDD et/ou analytiques Big Data. Avoir connaissance du langage Python.

Programme

1 Pendant votre formation

Delta Lake – Implémenter des lakehouses et pipelines de données

Jour 1

Evolution des architectures de données

- Des entrepôts de données aux Data Lakes et aux lakehouses
- Bénéfices des lakehouses
- L'architecture en médaillon
- L'écosystème Delta : stockage, partage, connecteurs

Exemple de travaux pratiques (à titre indicatif)

- Prendre en main Delta Lake avec PySpark

Bien démarrer avec Delta Lake

- Delta Lake On-Premise ou sur le Cloud
- Choisir son interface
- Librairies natives de Delta Lake
- Apache Spark avec Delta Lake
- Le format de fichier Parquet
- Les logs des transactions : implémentation de l'atomicité
- Opérations sur des tables Delta : create, read, write, overwrite, delete, update, merge
- Optimisation de la performance : partitionnement, Data Skipping, compaction, ZORDER BY, liquid clustering
- Utilisation du "Time Travel" et limites
- Rétention des données
- VACUUM
- Change Data Feed (CDF)
- Manipulation des schémas (validation, évolution, updates)
- Opérations sur des données de streaming
- Delta Sharing

Exemple de travaux pratiques (à titre indicatif)

- Effectuer la mise en oeuvre à travers des opérations sur des tables et des métadonnées

Jour 2

Conception architecturale d'un lakehouse Delta Lake

- Architecture à deux tiers
- La couche de stockage
- Gestion des données

- Implémentation de SQL Analytics avec Spark SQL
- Intégration de fonctionnalités supportant le Machine Learning
- Architecture en médaillon
- Bronze Layer
- Silver Layer
- Gold Layer

Exemple de travaux pratiques (à titre indicatif)

- *Utiliser Spark SQL et les fonctionnalités avancées de Machine Learning avec Delta Lake*

L'écosystème Delta Lake

- Connecteurs
- Apache Flink
- Flink DataStream Connector : DeltaSource API, DeltaSink API, Kafka Delta Ingest
- Trino Connector : Show Catalogs, créer un schéma, travailler avec des tables
- Intégrer Apache Kafka dans Delta Lake

Exemples de travaux pratiques (à titre indicatif)

- *Prendre en main Flink et Trino sur Delta Lake, créer des tables et un catalogue de données*

Jour 3

Le streaming de et vers un Delta Lake

- Streaming vs batch
- Delta comme source
- Delta comme puits
- Options de streaming
- Ecritures idempotentes de flux
- Métriques de performance
- Auto Loader et Live Tables

Exemples de travaux pratiques (à titre indicatif)

- *Créer, optimiser et effectuer la maintenance de tables en utilisant leurs propriétés*

Patrons de conception (design patterns) pour Delta Lake

- Optimisation des coûts : solutions à haut débit, Smart Device Integration
- Ingestion efficace de flux
- Delta Rust
- Combinaison et coordination de Data Stores opérationnels

Exemple de travaux pratiques (à titre indicatif)

- *Mettre en oeuvre des patrons de conception pour répondre à un cas d'usage (optimisation des coûts, débit d'ingestion de flux élevé, intégration de Data Stores)*

Jour 4

Data Engineering avec Delta Lake et Spark et stratégies de déploiement

- Stratégies de déploiement
- Développement, test et scheduling du master pipeline
- Monitoring des pipelines
- Gérer l'évolution des schémas
- Gouvernance : créer le catalogue de données
- Exemple de provisioning de l'infrastructure avec l'IaC (Infrastructure as Code)
- Implémenter le CI/CD pour les pipelines de données

Exemples de travaux pratiques (à titre indicatif)

- Effectuer un brainstorming sur la stratégie de déploiement à partir d'un cas d'usage
- Concevoir et tester des pipelines de données
- Mettre en oeuvre un déploiement avec l'IaC
- Automatiser des pipelines avec le CI/CD

Migration vers Delta Lake

- Introduction à la migration de données
- Importance de la préparation des données
- Evaluer l'infrastructure existante et des capacités de données
- Planifier la migration
- Créer un plan de migration détaillé
- Approche incrémentielle vs migration Big Bang
- Gérer les métadonnées
- Aspects sécuritaires et de conformité
- Tests post-migration pour garantir l'intégrité et la performance
- Optimisation post-migration et ajustements

Exemples de travaux pratiques (à titre indicatif)

- Planifier une migration d'un entrepôt de données vers Delta Lake
- Mettre en oeuvre une migration dans un environnement Open Source ou Cloud

Financement

Plusieurs dispositifs de financement sont accessibles via les OPCO (Opérateurs de Compétences), organismes agréés par le ministère du Travail dont le rôle est d'accompagner, collecter et gérer les contributions des entreprises au titre du financement de la formation professionnelle.

Pour plus d'information, une équipe de gestionnaires SKOLAE Formation spécialisée vous accompagne dans le choix de vos formations et la gestion administrative.

[Nous contacter pour en savoir plus →](#)

Modalités pédagogiques

Formation délivrée en présentiel ou distanciel* (blended-learning, e-learning, classe virtuelle, présentiel à distance).

Le formateur alterne entre méthode** démonstrative, interrogative et active (via des travaux pratiques et/ou des mises en situation).

Variables suivant les formations, les moyens pédagogiques mis en oeuvre sont :

- Ordinateurs Mac ou PC (sauf pour certains cours de l'offre Management), connexion internet fibre, tableau blanc ou paperboard, vidéoprojecteur ou écran tactile interactif (pour le distanciel)
- Environnements de formation installés sur les postes de travail ou en ligne
- Supports de cours et exercices

En cas de formation intra sur site externe à M2I, le client s'assure et s'engage également à avoir toutes les ressources matérielles pédagogiques nécessaires (équipements informatiques...) au bon déroulement de l'action de formation visée conformément aux prérequis indiqués dans le programme de formation communiqué.

* nous consulter pour la faisabilité en distanciel

** ratio variable selon le cours suivi

[Accessibilité aux personnes en situation de handicap](#)

Sessions

À distance

Du 2 février 2027 au 5 février 2027

Paris La Défense

Du 23 février 2027 au 26 février 2027