Apache Hudi : Ingestion Incrémentale et Gestion des Données en Temps Quasi-Réel sur Data Lake

Lien : https://innov-systems.com/formation/apache-hudi-ingestion-incrementale-et-gestion-des-donnees-en-temps-quasi-reel-sur-data-lake

DURÉE
5 jours (35h)

RÉFÉRENCE
BDI147

CATÉGORIE
Big Data : fondamentaux et architecture

Innov Systems

OBJECTIFS DE LA FORMATION

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

- Comprendre l’architecture et les concepts d’Apache Hudi
- Mettre en œuvre l’ingestion incrémentale des données
- Gérer les mises à jour et suppressions dans un Data Lake
- Exploiter les mécanismes CDC pour les données en continu
- Concevoir des architectures de données quasi temps réel
- Intégrer Hudi avec Spark, Kafka et les moteurs SQL
- Optimiser les performances et le stockage
- Garantir la qualité et la cohérence des données
- Mettre en place une gouvernance efficace des Data Lakes
- Réduire les coûts opérationnels liés au traitement des données
- Développer des pipelines Big Data modernes et évolutifs
- Étudier des cas concrets d’implémentation Apache Hudi

POUR QUI ?

- Data engineers
- Architectes Big Data
- Data architects
- Ingénieurs plateformes de données
- Développeurs Spark
- Administrateurs Hadoop
- DevOps et DataOps engineers
- Consultants Big Data
- Responsables data et analytics
- Machine learning engineers
- Toute personne impliquée dans la gestion de Data Lakes et plateformes de données

Innov Systems

Programme détaillé

1 / Introduction à Apache Hudi et aux Data Lakes

- Évolution des architectures Big Data
- Limites des Data Lakes traditionnels
- Présentation d'Apache Hudi et de ses cas d'usage

2 / Fondamentaux d'Apache Hudi

- Architecture et composants clés
- Concepts de tables Hudi
- Gestion des métadonnées et du stockage

3 / Ingestion incrémentale des données

- Principes de l'ingestion continue
- Capture des changements (CDC)
- Chargement incrémental des données

4 / Gestion des mises à jour et suppressions

- Opérations Upsert et Delete
- Gestion des enregistrements évolutifs
- Maintien de la cohérence des données

5 / Types de tables Apache Hudi

- Copy-On-Write (COW)
- Merge-On-Read (MOR)
- Choix du modèle selon les besoins métier

Innov Systems

6 / Traitement des données en temps quasi-réel

- Streaming et micro-batch
- Intégration avec Apache Kafka
- Réduction de la latence des traitements

7 / Intégration avec l’écosystème Big Data

- Apache Spark et Hudi
- Hive, Presto, Trino et moteurs SQL
- Interopérabilité avec les plateformes cloud

8 / Optimisation des performances et maintenance

- Compaction et clustering
- Gestion du stockage et des coûts
- Tuning des performances Hudi

9 / Gouvernance et qualité des données

- Gestion des versions de données
- Auditabilité et traçabilité
- Bonnes pratiques de gouvernance Data Lake

10 / Études de cas et implémentation pratique

- Mise en place d’un Data Lake moderne avec Hudi
- Cas d’usage temps réel en entreprise
- Bonnes pratiques d’industrialisation

Innov Systems

Approche pédagogique

- Support Ecrit et Projection
- Exposés Interactifs, Podcasts et Vidéos
- Brainstorming et Jeux de Rôle
- Cas Pratiques et Labs inclus pour leur impact opérationnel
- Test de Validation des Acquis des Connaissances

Prochaines dates programmées

30 Nov. au 04 Déc. 2026	Distanciel
25 au 29 Jan. 2027	Présentiel - Casablanca
22 au 26 Mars 2027	Présentiel - Casablanca
17 au 21 Mai 2027	Présentiel - Casablanca
12 au 16 Juil. 2027	Présentiel - Casablanca

Autres dates possibles sur demande. Contactez-nous pour organiser une session intra-entreprise.

Réservation & Renseignements

Téléphone : +212 522 247 210
Email : contact@innov-systems.com
Web : https//www.innov-systems.com

Innov Systems — Tous droits réservés