+33 (0)1 48 10 09 46
formation@ambient-it.net
140 Avenue Jean Lolive 93500 Pantin - France

Mis à jour le 01/07/2025

S'inscrire

Formation Apache Hudi

3 jours (21 heures)

Présentation

Notre formation Apache Hudi vous permettra de transformer un simple data lake en véritable lakehouse transactionnel. Ce framework open source, initialement développé par Uber, apporte des capacités ACID, la mise à jour incrémentale et le time-travel — autant d'atouts majeurs pour vos pipelines temps réel, vos démarches RGPD et vos analyses de données à haute fréquence.

Apprenez à gérer des flux CDC massifs sans sacrifier les performances ni la gouvernance, ainsi qu'à remonter l'historique complet de vos données, exécuter des requêtes snapshot ou incrémentales, et prouver la conformité de vos jeux de données. Maîtrisez les modes Copy-on-Write et Merge-on-Read, la compaction intelligente et l'indexation fine pour des temps de réponse minimalistes.

À l'issue de cette formation, vos ingénieurs sauront installer, sécuriser et exploiter Apache Hudi, optimiser les performances de leurs pipelines et répondre instantanément aux exigences métier et réglementaires.

Comme toutes nos formations, celle-ci s'appuie sur la dernière version stable disponible — Apache Hudi 1.0 LTS — et couvre en détail ses nouveautés et bonnes pratiques.

Objectifs

- Installer et configurer Apache Hudi
- Ingestion batch & streaming
- Exploiter le time-travel et les lectures incrémentales
- Optimiser les performances
- Mettre en place une gouvernance et une sécurité complètes

Public visé

- Data Engineers
- Data Architects
- DevOps / SRE

Pré-requis

- Lecture/écriture dans l’un des langages : Python, Scala ou Java
- Bonne pratique d’Apache Spark
- Connaissance du stockage objet (S3, GCS ou HDFS) et du format Parquet
- À l’aise avec Docker, CLI Linux, Git

Programme de notre Formation Apache Hudi

Pourquoi Apache Hudi ?

- Évolution des formats de tables
- Besoins d’upserts, deletes et time-travel dans un data lake
- Principes ACID et timeline transactionnelle
- Cas d’usage typiques : CDC temps réel, conformité RGPD, IoT
- Positionnement dans l’écosystème Spark / Flink / Trino

Anatomie d’Hudi

- Modes de tables : Copy-on-Write (CoW) vs Merge-on-Read (MoR)
- Timeline, commits, instant states
- Metadata Table & Record-Level Index
- Services internes : compaction, clustering, cleaning
- Connecteurs query-engine (Hive, Presto/Trino, Spark SQL)

Mise en route sandbox

- Pré-requis Docker & Spark stand-alone
- Déploiement rapide via docker-compose
- Création d’une table CoW partitionnée
- Premier chargement batch (API DataSource)
- Atelier pratique : Quick Start Hudi Docker – créer et interroger sa première table

Upserts, Deletes & Time-Travel avec Spark

- Clés essentielles : recordkey, precombine, partitionpath
- Écriture incrémentale : gestion des conflits, ordering field
- Lecture snapshot vs incrémentale
- Requêtes “à l’instant T” (time-travel)
- Atelier pratique : Insérer 1 M de lignes, rejouer un CDC, interroger T-1

Ingestion streaming avec Spark Structured Streaming & Flink

- Architecture exactly-once sur stream
- Gestion de l’ordre d’arrivée et late events
- Configurer un sink Flink SQL vers table MoR
- Stratégies de compaction asynchrone
- Atelier pratique : Pipeline Kafka ? Flink ? Hudi en live

Performance et optimisation

- Indexation : Bloom, HFile, Global Secondary, Record-Level
- Compaction : triggers automatiques vs manuels
- Clustering pour réécriture de partitions chaudes
- Tuning write-amp & small files
- Benchmarks lecture vs écriture (CoW / MoR)

Sécurité, gouvernance & conformité

- Soft-delete, purge physique et rollback
- Gestion des schémas Avro & compatibilité évolutive
- Chiffrement côté stockage (SSE-S3, KMS)
- Intégration Lake Formation / tag-based ACL
- Atelier pratique : Supprimer des enregistrements RGPD et vérifier via snapshot

Déploiement cloud & opérations

- Bonnes pratiques S3 / GCS (consistency, retries, caching)
- Monitoring Prometheus & Hudi CLI
- Stratégies de backup / restore et migration 0.x ? 1.x
- Coûts opératoires : write-amp, compaction, vacuum policies
- Modèles IaC (Terraform, Helm) pour production

Intégration analytique & projet final

- Exposer les tables dans Hive Metastore / Glue Catalog
- Query federation : Trino, Athena, Spark SQL
- Data Lakehouse : Hudi + Presto dashboard temps réel
- CI/CD : tests d’intégrité, validation schema

Sociétés concernées

Cette formation s’adresse à la fois aux particuliers ainsi qu’aux entreprises, petites ou grandes, souhaitant former ses équipes à une nouvelle technologie informatique avancée ou bien à acquérir des connaissances métiers spécifiques ou des méthodes modernes.

Positionnement à l'entrée en formation

Le positionnement à l'entrée en formation respecte les critères qualité Qualiopi. Dès son inscription définitive, l'apprenant reçoit un questionnaire d’auto-évaluation nous permettant d’apprécier son niveau estimé sur différents types de technologies, ses attentes et objectifs personnels quant à la formation à venir, dans les limites imposées par le format sélectionné. Ce questionnaire nous permet également d’anticiper certaines difficultés de connexion ou de sécurité interne en entreprise (intraentreprise ou classe virtuelle) qui pourraient être problématiques pour le suivi et le bon déroulement de la session de formation.

Méthodes pédagogiques

Stage Pratique : 60% Pratique, 40% Théorie. Support de la formation distribué au format numérique à tous les participants.

Organisation

Le cours alterne les apports théoriques du formateur soutenus par des exemples et des séances de réflexions, et de travail en groupe.

Validation

À la fin de la session, un questionnaire à choix multiples permet de vérifier l’acquisition correcte des compétences.

Sanction

Une attestation sera remise à chaque stagiaire qui aura suivi la totalité de la formation.