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

Mis à jour le 04/12/2025

S'inscrire

Formation Architecture Lakehouse

3 jours (21 heures)

Présentation

La Formation Architecture Lakehouse vous propose une approche moderne et unifiée de la gestion des données, en combinant la flexibilité d'un Data Lake et la fiabilité d'un Data Warehouse.

Cette architecture s'appuie sur les formats de tables ouverts tels qu'Apache Iceberg, permettant une gouvernance centralisée, des performances élevées et une véritable évolutivité.

Cette formation vous permettra de comprendre et de maîtriser les fondations d'une plateforme Lakehouse : organisation des données, ingestion batch et streaming, historique via snapshots, gouvernance, sécurité, optimisation et supervision.

Elle vous montrera comment structurer un environnement capable d'alimenter à la fois les usages BI, Analytics et Machine Learning, tout en conservant un haut niveau de qualité et de cohérence.

Au fil des modules, vous apprendrez à construire une architecture Lakehouse robuste, à déployer des tables Apache Iceberg, à gérer leur évolution, à optimiser leurs performances et à superviser leur exploitation. Vous verrez comment le Lakehouse s'intègre dans les pratiques DataOps et MLOps, et comment il devient un socle commun pour les équipes techniques et métiers.

À l'issue de cette formation, vous serez en mesure de concevoir un Lakehouse complet et opérationnel, d'industrialiser vos pipelines, de structurer vos données en domaines gouvernés et d'améliorer la qualité comme la disponibilité du patrimoine Data de votre entreprise.

Comme toutes nos formations, celle-ci s'appuie sur la dernière version stable d'Apache Iceberg et privilégie une approche résolument pratique et opérationnelle.

Objectifs

- Comprendre les principes d’une architecture Lakehouse moderne et évolutive.
- Déployer et gouverner des tables Apache Iceberg.
- Mettre en œuvre une ingestion fiable en batch et streaming.
- Exploiter le time travel, l’historique et l’évolution de schéma.
- Optimiser les performances et les coûts d’exploitation.
- Industrialiser les usages via DataOps et MLOps.

Public visé

- DSI
- Responsables Data et Architectes Data
- DevOps
- DataOps
- MLOps
- Chefs de projets

Pré-requis

- Bonnes connaissances des architectures Data
- Pratique d’un moteur de traitement distribué
- Notions de stockage objet
- Maîtrise du SQL et des usages BI / Analytics

Formation Architecture Lakehouse

[Jour 1 - Matin]

Fondamentaux du Lakehouse et formats ouverts

- Comprendre l’intérêt du Lakehouse pour simplifier l’accès à la donnée.
- Identifier les limites des architectures Data Lake classiques.
- Adopter Iceberg pour fiabilité, ouverture et évolutivité.
- Gérer facilement l’historique grâce aux snapshots.
- Structurer la donnée via les zones Bronze / Silver / Gold.
- Atelier pratique : Lecture des métadonnées d’une table Iceberg.

[Jour 1 - Après-midi]

Catalogues et intégration au SI

- Comprendre le rôle des catalogues pour centraliser la gouvernance.
- Choisir entre Hive, Glue, Nessie ou REST Catalog selon le SI.
- Structurer namespaces et environnements de manière cohérente.

- Intégrer Spark, Trino ou Flink sans dépendance propriétaire.
- Sécuriser et normaliser la découverte des tables.
- Atelier pratique : Création et enregistrement de tables dans un catalogue.

Schéma, partitionnement et évolution

- Définir des schémas lisibles et adaptés aux usages BI/ML.
- Choisir un partitionnement efficace pour optimiser les lectures.
- Faire évoluer le schéma sans casser les pipelines existants.
- Simplifier l'intégration des producteurs et consommateurs Data.
- Garantir cohérence et compatibilité entre versions.
- Atelier pratique : Modifier un schéma Iceberg en conditions réelles.

[Jour 2 - Matin]

Ingestion Batch, Micro-batch et Streaming

- Choisir le bon mode d'ingestion selon la criticité métier.
- Alimenter Iceberg avec Spark Streaming ou Flink de manière fiable.
- Éviter les small files pour stabiliser les performances.
- Gérer les données en retard et les mises à jour incrémentales.
- Simplifier merges et corrections de données.
- Atelier pratique : Ingestion end-to-end d'un flux dans Iceberg.

[Jour 2 - Après-midi]

Snapshots, time travel et historique

- Exploiter le time travel pour analyser l'évolution des données.
- Restaurer ou comparer un état passé simplement.
- Comprendre l'impact métier du versionnement complet.
- Suivre changements et erreurs via l'historique intégré.
- Gérer efficacement rétention et nettoyage.
- Atelier pratique : Diagnostic d'un incident via time travel.

Gouvernance, sécurité et Data Mesh

- Définir un cadre de gouvernance compatible Lakehouse.
- Appliquer masquage, RBAC et règles d'accès cohérentes.
- Assurer lignée, audit et traçabilité pour tous les domaines.
- Structurer la donnée en produits dans un Data Mesh.
- Renforcer qualité et fiabilité via tests et validations.
- Atelier pratique : Création d'un domaine Data Mesh sécurisé.

[Jour 3 - Matin]

Performance et optimisation

- Identifier les facteurs influençant les performances Iceberg.
- Accélérer les requêtes via pruning et organisation de fichiers.
- Optimiser le stockage pour réduire les coûts.
- Surveiller les tables pour détecter dérives et anomalies.
- Stabiliser workloads BI et Data Science.
- Atelier pratique : Optimisation d’une table lourde.

[Jour 3 - Après-midi]

Lakehouse pour MLOps et DataOps

- Utiliser le Lakehouse comme socle pour les features ML.
- Versionner les données pour fiabiliser les modèles.
- Automatiser pipelines DataOps : tests, promotion, CI/CD.
- Aligner équipes Data, DevOps et métiers sur un référentiel commun.
- Accélérer expérimentation et industrialisation ML.
- Atelier pratique : Pipeline MLOps utilisant des données Iceberg versionnées.

Supervision, incidents et exploitation

- Mettre en place observabilité et métriques pour le Lakehouse.
- Suivre santé des tables : manifestes, snapshots, compaction.
- Anticiper incidents et définir des scénarios de reprise.
- Optimiser les coûts via politiques FinOps ciblées.
- Structurer un runbook clair pour les équipes DataOps.
- Atelier pratique : Création d’un runbook opérationnel complet.

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.