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

Updated on 12/04/2025

Register

Lakehouse Architecture Training

3 days (21 hours)

Overview

The Lakehouse Architecture Training offers a modern, unified approach to data management, combining the flexibility of a data lake with the reliability of a data warehouse.

This architecture is based on open table formats such as Apache Iceberg, enabling centralized governance, high performance, and true scalability.

This training course will enable you to understand and master the foundations of a Lakehouse platform: data organization, batch and streaming ingestion, history via snapshots, governance, security, optimization, and monitoring.

It will show you how to structure an environment capable of supporting BI, analytics, and machine learning uses, while maintaining a high level of quality and consistency.

Throughout the modules, you will learn how to build a robust Lakehouse architecture, deploy Apache Iceberg tables, manage their evolution, optimize their performance, and supervise their operation. You will see how Lakehouse integrates with DataOps and MLOps practices, and how it becomes a common foundation for technical and business teams.

By the end of this training, you will be able to design a complete and operational Lakehouse, industrialize your pipelines, structure your data into governed domains, and improve the quality and availability of your company's data assets.

Like all our training courses, this one is based on the latest stable version of Apache Iceberg and focuses on a decidedly practical and operational approach.

Objectives

- Understand the principles of a modern and scalable Lakehouse architecture.
- Deploy and govern Apache Iceberg tables.
- Implement reliable batch and streaming ingestion.
- Leverage time travel, history, and schema evolution.
- Optimize performance and operating costs.
- Industrialize usage via DataOps and MLOps.

Target audience

- CIO
- Data managers and data architects
- DevOps
- DataOps
- MLOps
- Project Managers

Prerequisites

- Good knowledge of data architectures
- Experience with a distributed processing engine
- Basic knowledge of object storage
- Proficiency in SQL and BI/Analytics applications

Lakehouse Architecture Training

[Day 1 - Morning]

Lakehouse fundamentals and open formats

- Understanding the benefits of Lakehouse for simplifying data access.
- Identify the limitations of traditional data lake architectures.
- Adopt Iceberg for reliability, openness, and scalability.
- Easily manage history with snapshots.
- Structure data using Bronze/Silver/Gold zones.
- Hands-on workshop: Reading metadata from an Iceberg table.

[Day 1 - Afternoon] Catalogs

and integration into the IS

- Understanding the role of catalogs in centralizing governance.
- Choose between Hive, Glue, Nessie, or REST Catalog depending on the IS.
- Structure namespaces and environments in a consistent manner.

- Integrate Spark, Trino, or Flink without proprietary dependencies.
- Secure and standardize table discovery.
- Hands-on workshop: Creating and registering tables in a catalog.

Schema, partitioning, and evolution

- Define schemas that are readable and suitable for BI/ML uses.
- Choose effective partitioning to optimize reads.
- Evolve the schema without breaking existing pipelines.
- Simplify the integration of data producers and consumers.
- Ensure consistency and compatibility between versions.
- Hands-on workshop: Modify an Iceberg schema in real conditions.

[Day 2 - Morning]

Batch, micro-batch, and streaming ingestion

- Choosing the right ingestion mode according to business criticality.
- Feed Iceberg reliably with Spark Streaming or Flink.
- Avoiding small files to stabilize performance.
- Manage late data and incremental updates.
- Simplify data merges and corrections.
- Hands-on workshop: End-to-end ingestion of a stream into Iceberg.

[Day 2 - Afternoon]

Snapshots, time travel, and history

- Use time travel to analyze data evolution.
- Easily restore or compare a past state.
- Understand the business impact of full versioning.
- Track changes and errors via the integrated history.
- Effectively manage retention and cleanup.
- Hands-on workshop: Diagnosing an incident using time travel.

Governance, security, and Data Mesh

- Define a Lakehouse-compatible governance framework.
- Apply masking, RBAC, and consistent access rules.
- Ensure lineage, auditability, and traceability for all domains.
- Structure data into products in a Data Mesh.
- Enhance quality and reliability through testing and validation.
- Hands-on workshop: Creating a secure Data Mesh domain.

[Day 3 - Morning]

Performance and optimization

- Identify factors influencing Iceberg performance.
- Speed up queries through pruning and file organization.
- Optimize storage to reduce costs.
- Monitor tables to detect drifts and anomalies.
- Stabilize BI and Data Science workloads.
- Hands-on workshop: Optimizing a heavy table.

[Day 3 - Afternoon]

Lakehouse for MLOps and DataOps

- Using Lakehouse as a foundation for ML features.
- Versioning data to improve model reliability.
- Automating DataOps pipelines: testing, promotion, CI/CD.
- Align Data, DevOps, and business teams around a common framework.
- Accelerate ML experimentation and industrialization.
- Hands-on workshop: MLOps pipeline using versioned Iceberg data.

Monitoring, incidents, and operations

- Implement observability and metrics for the Lakehouse.
- Monitor table health: manifests, snapshots, compaction.
- Anticipate incidents and define recovery scenarios.
- Optimize costs through targeted FinOps policies.
- Structure a clear runbook for DataOps teams.
- Hands-on workshop: Creating a comprehensive operational runbook.

Target companies

This training is intended for both individuals and companies, large or small, wishing to train their teams in new advanced IT technology or to acquire specific business knowledge or modern methods.

Positioning at the start of training

The positioning at the start of the training complies with Qualiopi quality criteria. Upon final registration, the learner receives a self-assessment questionnaire that allows us to assess their estimated level of knowledge of different types of technologies, their expectations and personal objectives for the upcoming training, within the limits imposed by the selected format. This questionnaire also allows us to anticipate certain connection or internal security issues within the company (intra-company or virtual classroom) that could be problematic for the monitoring and smooth running of the training session.

Teaching methods

Practical training: 60% practical, 40% theory. Training materials distributed in digital format to all participants.

Organization

The course alternates between theoretical input from the trainer, supported by examples and discussion sessions, and group work.

Assessment

At the end of the session, a multiple-choice questionnaire is used to verify that the skills have been correctly acquired.

Certification

A certificate will be issued to each trainee who has completed the entire training course.