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

Updated on 01/07/2025

Sign up

Apache Hudi Training

3 days (21 hours)

Presentation

Our Apache Hudi training course will enable you to transform a simple data lake into a true transactional lakehouse. This open source framework, originally developed by Uber, provides ACID capabilities, incremental updating and time-travel - all major assets for your real-time pipelines, RGPD initiatives and high-frequency data analysis.

Learn how to manage massive CDC flows without sacrificing performance or governance, as well as how to trace the full history of your data, run snapshot or incremental queries, and prove the compliance of your datasets. Master Copy-on-Write
Copy-on-Write and Merge-on-Read modes, intelligent compaction and fine-grained indexing for minimal response times.

By the end of this course, your engineers will know how to install, secure and operate Apache Hudi, optimize the performance of their pipelines and respond instantly to business and regulatory requirements.

Like all our training courses, this one is based on the latest stable release - Apache Hudi 1.0 LTS - and covers in detail its new features and best practices.

Objectives

- Install and configure Apache Hudi
- Batch & streaming ingestion
- Exploit time-travel and incremental reads
- Optimize performance
- Implement complete governance and security

Target audience

- Data Engineers
- Data Architects
- DevOps / SRE

Prerequisites

- Reading/writing in one of the following languages: Python, Scala or Java
- Good working knowledge of Apache Spark
- Knowledge of object storage (S3, GCS or HDFS) and Parquet format
- Comfortable with Docker, Linux CLI, Git

Our Apache Hudi training program

Why Apache Hudi?

- Evolution of table formats
- Upserts, deletes and time-travel requirements in a data lake
- ACID principles and transactional timelines
- Typical use cases: real-time CDC, RGPD compliance, IoT
- Positioning in the Spark / Flink / Trino ecosystem

Anatomy of Hudi

- Table modes: Copy-on-Write (CoW) vs. Merge-on-Read (MoR)
- Timeline, commits, instant states
- Metadata Table & Record-Level Index
- Internal services: compaction, clustering, cleaning
- Query-engine connectors (Hive, Presto/Trino, Spark SQL)

Getting started in sandbox

- Docker & Spark stand-alone prerequisites
- Rapid deployment via docker-compose
- Creating a partitioned CoW table
- First batch load (DataSource API)
- Practical workshop: Quick Start Hudi Docker - creating and querying your first table

Upserts, Deletes & Time-Travel with Spark

- Essential keys: recordkey, precombine, partitionpath
- Incremental writing: conflict management, ordering field
- Snapshot vs. incremental reading
- Time-travel queries
- Practical workshop: Inserting 1 M rows, replaying a CDC, querying T-1

Streaming ingestion with Spark Structured Streaming & Flink

- Exactly-once stream architecture
- Managing arrival order and late events
- Setting up a SQL sink Flink to MoR table
- Asynchronous compaction strategies
- Practical workshop: Pipeline Kafka ? Flink ? Hudi live

Performance and optimization

- Indexing: Bloom, HFile, Global Secondary, Record-Level
- Compaction: automatic vs. manual triggers
- Clustering for rewriting hot partitions
- Tuning write-amp & small files
- Read vs. write benchmarks (CoW / MoR)

Security, governance & compliance

- Soft-delete, physical purge and rollback
- Avro schema management & scalable compatibility
- Storage-side encryption (SSE-S3, KMS)
- Lake Formation / tag-based ACL integration
- Practical workshop: Delete RGPD records and verify via snapshot

Cloud deployment & operations

- S3 / GCS best practices (consistency, retries, caching)
- Prometheus & Hudi CLI monitoring
- Backup / restore strategies and migration 0.x ? 1.x
- Operating costs: write-amp, compaction, vacuum policies
- IaC models (Terraform, Helm) for production

Analytical integration & final project

- Exposing tables in Hive Metastore / Glue Catalog
- Query federation: Trino, Athena, Spark SQL
- Data Lakehouse: Hudi + Presto real-time dashboard
- CI/CD: integrity testing, schema validation

Companies concerned

This course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced IT technology, or to acquire specific business knowledge or modern methods.

Positioning on entry to training

Positioning at the start of training complies with Qualiopi quality criteria. As soon as registration is finalized, the learner receives a self-assessment questionnaire which enables us to assess his or her estimated level on different types of technology, as well as his or her expectations and personal objectives with regard to the training to come, within the limits imposed by the selected format. This questionnaire also enables us to anticipate any connection or security difficulties within the company (intra-company or virtual classroom) which could be problematic for the follow-up and smooth running of the training session.

Teaching methods

Practical training: 60% hands-on, 40% theory. Training material distributed in digital format to all participants.

Organization

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

Validation

At the end of the session, a multiple-choice questionnaire verifies the correct acquisition of skills.

Certification

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