multisoft
Systems

Apache Hudi Training

COURSE CONTENT

GET IN TOUCH

Multisoft Systems
B - 125, Sector - 2, Noida

(+91) 9810-306-956

info@multisoftsystems.com

www.multisoftsystems.com

About Multisoft

Train yourself with the best and develop valuable in-demand skills with Multisoft Systems. A leading certification training provider, Multisoft collaborates with top technologies to bring world-class one-on-one and certification trainings. With the goal to empower professionals and business across the globe, we offer more than 1500 training courses, which are delivered by Multisoft's global subject matter experts. We offer tailored corporate training; project Based Training, comprehensive learning solution with lifetime e-learning access, after training support and globally recognized training certificates.

About Course

Apache Hudi is a powerful open-source data lake framework that enables near real-time data ingestion, incremental processing, and efficient storage management. Multisoft Systems' Apache Hudi Training is designed to help data engineers, analysts, and big data professionals gain expertise in managing large-scale data lakes with Hudi.

Module 1: Introduction to Apache Hudi

- ✓ Overview of Apache Hudi
- ✓ Need for Hudi in Big Data Ecosystems
- ✓ Key Features and Advantages
- ✓ Comparison with Delta Lake & Apache Iceberg
- ✓ Use Cases and Industry Applications

Module 2: Hudi Architecture and Components

- ✓ Understanding Hudi's Architecture
- ✓ Hudi Table Types: Copy-on-Write (COW) & Merge-on-Read (MOR)
- ✓ Data Ingestion & Storage Mechanism
- ✓ Indexing in Hudi
- ✓ Role of Timeline Server & Commit Protocol

Module 3: Setting Up Apache Hudi

- ✓ System Requirements and Installation
- ✓ Hudi Configuration & Prerequisites
- ✓ Deploying Hudi on Apache Spark
- ✓ Working with Hudi on AWS, Azure, GCP

Module 4: Hudi Data Ingestion and Writing

- ✓ Writing Data to Hudi Tables
- ✓ Bulk Insert, Upsert, and Delete Operations
- ✓ Schema Evolution in Hudi
- ✓ Partitioning and Clustering
- ✓ Optimizing Write Performance

Module 5: Querying and Reading Data in Hudi

- ✓ Querying Hudi Tables using Apache Spark

- ✓ Integration with Presto, Hive, and Trino
- ✓ Snapshot and Incremental Queries
- ✓ Querying Data Lake with Hudi

Module 6: Hudi Data Management and Optimizations

- ✓ Compaction and Cleaning Policies
- ✓ Clustering for Performance Enhancement
- ✓ Metadata Management in Hudi
- ✓ Performance Tuning Strategies

Module 7: Apache Hudi Integration with Big Data Ecosystem

- ✓ Hudi with Apache Spark
- ✓ Integration with Apache Flink
- ✓ Using Hudi with AWS Glue, EMR, Databricks
- ✓ Combining Hudi with Kafka for Streaming Data

Module 8: Data Governance and Security in Hudi

- ✓ Managing Metadata & Schema Evolution
- ✓ Role-based Access Control (RBAC)
- ✓ Data Lineage and Auditing
- ✓ Implementing Security Best Practices

Module 9: Advanced Use Cases and Best Practices

- ✓ Real-time Data Processing with Hudi
- ✓ Implementing Change Data Capture (CDC)
- ✓ Scaling Hudi for Large-Scale Workloads
- ✓ Troubleshooting Common Issues

Module 10: Hands-on Project

- ✓ End-to-End Data Pipeline with Hudi
- ✓ Implementing Incremental Processing
- ✓ Performance Benchmarking