Modern Lakehouse & Data Platform Architecture

Duration: 3 Days

Description

Modern Lakehouse & Data Platform Architecture is an advanced technical training program that addresses transition processes from the traditional Hadoop ecosystem to modern data platforms, next-generation data architectures, and open table formats. Within the training, Data Lake, Data Warehouse, and Lakehouse architectures are compared, and the architectural components, use cases, and advantages provided by Apache Iceberg and Delta Lake technologies to enterprise data platforms are examined. Features such as ACID Transactions, Schema Evolution, Time Travel, and data versioning are covered with practical examples of how they are used in large-scale data management. The program also focuses on transforming existing Hadoop, HDFS, and Hive-based infrastructures into modern Lakehouse architectures, data and metadata migration strategies, performance optimization, and enterprise architecture design. It aims to enable participants to technically evaluate modern Lakehouse architectures and open table formats, use Apache Iceberg and Delta Lake technologies, and develop appropriate modernization strategies for transitioning from existing Hadoop-based data platforms to modern, scalable, and sustainable data architectures. **Delegates will learn how to**

- Explain the architectural differences between the traditional Hadoop ecosystem and modern data platforms
- Compare Data Lake, Data Warehouse, and Lakehouse approaches
- Evaluate the key features and use cases of open table formats
- Understand Apache Iceberg and Delta Lake architectures
- Apply ACID Transactions, Schema Evolution, and Time Travel features
- Manage data consistency, versioning, and performance optimization processes
- Develop migration strategies from Hadoop-based data platforms to Lakehouse architectures
- Design modern data platform architectures suitable for enterprise requirements

Audience

- Data Engineer / Senior Data Engineer
- Data Architect / Enterprise Data Architect
- Data Platform Engineer
- Big Data Engineer

- Solution Architect
- Technical teams responsible for the development, management, and modernization of data platforms

Outline

1. Modern Data Platform Architecture

- Transition from traditional data platforms to modern data architectures
- Core components of the Hadoop ecosystem
- HDFS, Hive, and Apache Spark architectures
- Architectural and operational limitations of Hadoop-based data platforms
- Modern Data Platform concept
- Separation of Storage and Compute layers
- Scalability, flexibility, and sustainability
- On-Premise, Cloud, and Hybrid data platforms

2. Data Lake, Data Warehouse & Lakehouse Architecture

- Data Warehouse architecture and key features
- Data Lake architecture and use cases
- Comparison of Data Lake and Data Warehouse
- Limitations of traditional Data Lake architectures
- Emergence of the Lakehouse architecture
- Core components of the Lakehouse architecture
- Batch and Streaming data processing approaches
- Data management in the Lakehouse architecture
- Architecture selection criteria for enterprise data platforms

3. Open Table Formats

- Open Table Format concept
- The role of Parquet and similar file formats
- Limitations of traditional file-based data storage
- Metadata and Transaction Log structures
- Overview of Apache Iceberg, Delta Lake, and Apache Hudi
- Comparison of open table formats
- Storage Engine and Query Engine relationship
- Interoperability in open table formats
- Vendor independence and technology selection criteria

4. Apache Iceberg Architecture

- Introduction to Apache Iceberg architecture

- Iceberg Table structure
- Metadata Files and Manifest Files
- Snapshot architecture
- Hidden Partitioning
- Partition Evolution
- Schema Evolution
- Snapshot Management
- Time Travel and data versioning
- Data reading and writing operations with Iceberg
- Using Iceberg with Apache Spark and Spark SQL
- Application: Creating and managing Iceberg tables

5. Delta Lake Architecture

- Introduction to Delta Lake architecture
- Delta Table structure
- Delta Transaction Log
- ACID Transactions
- Schema Enforcement
- Schema Evolution
- Time Travel and Versioning
- MERGE, UPDATE, and DELETE operations
- OPTIMIZE and Compaction approaches
- Delta Lake performance features
- Using Delta Lake with Apache Spark
- Application: Data management operations on Delta tables

6. ACID Transactions & Data Consistency

- ACID principles
- Transaction management in distributed data platforms
- Atomicity, Consistency, Isolation, and Durability
- Snapshot Isolation
- Optimistic Concurrency Control
- Concurrent data reading and writing operations
- Data consistency and reliability
- Comparison of Apache Iceberg and Delta Lake transaction mechanisms
- Application: Transaction and data consistency scenarios

7. Schema Evolution & Time Travel

- Schema Evolution concept

- Column addition, deletion, and modification operations
- Schema Compatibility
- Data versioning approaches
- Snapshot-based data management
- Time Travel queries
- Historical Data Analysis
- Rollback and data recovery scenarios
- Data retention and version cleanup policies
- Application: Schema Evolution and Time Travel operations

8. Hadoop to Modern Lakehouse Migration

- Assessment of the existing Hadoop ecosystem
- Analysis of HDFS and Hive-based data architectures
- Hadoop modernization strategies
- Identification of existing data sources and dependencies
- Migration from Hive tables to Apache Iceberg or Delta Lake
- Data migration and transformation approaches
- Metadata and catalog migration
- Modernization of existing ETL/ELT processes
- Data validation and consistency checks
- Phased migration and parallel operation strategies
- Migration risks and rollback planning
- Creating a roadmap for migration from Hadoop to Lakehouse

9. Lakehouse Performance & Optimization

- Partitioning strategies
- File Size Optimization
- Small Files problem
- Compaction
- Query Performance Optimization
- Metadata Optimization
- Management of Storage and Compute resources
- Performance and cost optimization
- Comparison of Apache Iceberg and Delta Lake performance approaches

10. Enterprise Lakehouse Architecture Workshop

- Assessment of the sample Hadoop architecture
- Identification of technical limitations in the existing architecture
- Designing the target Lakehouse architecture

- Apache Iceberg and Delta Lake selection criteria
- Identification of data storage and processing layers
- Creating the Metadata and catalog architecture
- Evaluation of ACID Transactions, Schema Evolution, and Time Travel scenarios
- Determination of the data migration and modernization strategy
- Preparing the technical roadmap for the target architecture

Prerequisites

- Knowledge of SQL and data processing processes
- Familiarity with the Hadoop ecosystem or distributed data processing technologies
- Basic experience with ETL/ELT processes and data storage architectures
- Knowledge of Apache Spark or similar big data processing technologies is recommended