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

Updated on 01/10/2024

Sign up

Apache Avro training

2 days (14 hours)

Presentation

Are you looking for a reliable, high-performance solution for serializing, storing and exchanging large-scale data? Our Apache Avro training course offers a comprehensive introduction to this open source serialization format, optimized for processing massive data in distributed environments.

Apache Avro is designed to enable efficient serialization of binary data, while facilitating interchangeability between systems and languages, making it an ideal tool for Data Engineers and Big Data Developers.

Its flexible layout means it can be adapted to changing data structures without compromising compatibility with existing systems.

During this course, you'll learn how to create and manage Avro schemas, optimally serialize and deserialize data, and integrate Avro into your distributed data processing pipelines.

You'll also discover how to use Apache Avro with other tools from the Hadoop and Kafka ecosystems to process massive data in real time or in batch mode.

This course will enable you to develop essential skills for managing large-scale data, while familiarizing you with best practices for optimizing and scaling schemas in Big Data environments.

As with all our training courses, it will be presented with the latest resources available.

Objectives

- Understanding the basic principles and architecture of Apache Avro
- Structuring and validating Avro schemas for data serialization
- Mastering serialization and deserialization processes with Avro
- Integrating Apache Avro into distributed systems such as Hadoop, Kafka, and Spark
- Manage the evolution of schematics while ensuring compatibility with existing systems
- Optimize the performance of data processing pipelines with Avro

Target audience

- Data Engineers
- Data Scientists
- Big Data developers
- Data architects

Prerequisites

- Knowledge of the basic concepts of massive data processing
- Experience with a programming language such as Java or Python
- Understanding of database and distributed systems concepts
- Familiarity with Big Data technologies such as Hadoop, Spark, or Kafka is a plus

APACHE AVRO TRAINING PROGRAM

Introduction to Apache Avro

- Data serialization: definition and importance
- History and conception of Apache Avro in the Hadoop ecosystem
- Comparison with other serialization formats (JSON, Protobuf, Thrift)
- Avro file structure: schema and data
- Self-description principles: schema stored with data
- Benefits of Apache Avro for large-scale data processing

The Avro schematic model

- Understanding Avro schema syntax (JSON)
- Primitive (int, long, float, etc.) and complex (records, arrays, maps) data types
- Schema validation: ensuring data integrity
- Schema evolution management (adding/removing fields)
- Schema compatibility (backward, forward, full compatibility)
- Practical examples of Avro schematics in real-life applications

Serialization and deserialization with Apache Avro

- Serialization process: convert objects into Avro binary or JSON format

- Deserialization: transforming Avro data into readable objects
- Advantages of binary format for performance and file size reduction
- Use of libraries in different languages (Java, Python, etc.)
- Serialization with and without schematics: differences and use cases
- Optimizing performance when serializing massive data

Apache Avro in distributed systems

- Using Avro with Apache Hadoop: storing and processing Avro files in HDFS
- Integration with Apache Kafka for event logs and streaming pipelines
- Avro with Apache Spark: reading and writing Avro data in Spark applications
- Use cases in Big Data architectures (streaming and batch)
- Avro file management in multi-cluster environments (replication, backup)
- Performance optimization in distributed systems with Avro

Advanced tools and practices with Apache Avro

- Avro tools: avro-tools for file manipulation and conversion
- Avro data compression: compatible formats (Snappy, Deflate, etc.)
- Best practices for schema evolution in production environments
- Avro data security (encryption, access management)
- Schema debugging and validation in distributed systems
- Monitor and manage Apache Avro performance in data pipelines

Companies concerned

This course is aimed at both individuals and companies, large or small, wishing to train their teams in a new advanced computer 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 of proficiency in different types of technology, as well as his or her expectations and personal objectives for 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 course: 60% Practical, 40% Theory. Training material distributed in digital format to all participants.

Organization

The course alternates theoretical inputs from the trainer supported by examples and

brainstorming sessions and group work.

Validation

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

Sanction

A certificate will be issued to each trainee who completes the course.