AMBIENT-IT

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

Updated 04/11/2024

Sign up

Apache Arrow training

2 days (14 hours)

Presentation

Efficiently process your in-memory data with our Apache Arrow training course to enable smooth exchanges between your different frameworks.

At the end of this course, you'll learn how to use Apache Arrow to manipulate massive data efficiently, integrate multiple data processing tools and perform complex analytical operations.

What's more, this tool will teach you to master several advanced concepts of memory management, performance optimization and large-scale data parallelism.

What's more, this technology guarantees data portability and simplifies software development.

As with all our training courses, this one will introduce you to the latest version of Apache Arrow (at the time of writing: Apache Arrow 15.

Objectives

- Explore the internal architecture of Apache Arrow
- Developing and testing features with Apache Arrow
- Integrate continuous integration and packaging practices
- Analytical operations

Target audience

- Developers
- Engineers
- Data Analyst

Prerequisites

- Basic programming skills (Python, Java)
- Understanding of data processing and file manipulation concepts

APACHE ARROW TRAINING PROGRAM

INTRODUCTION TO APACHE ARROW

- Introducing Apache Arrow and its ecosystem
- Importance of Apache Arrow in Big Data and data analysis
- Key concepts: columns, tables, schemas, and data types
- Advantages of using Apache Arrow for in-memory data processing
- Installing Apache Arrow and configuring the development environment

ARCHITECTURE AND DESIGN

- Overview of Apache Arrow's internal architecture
- Understanding memory representation and columnar data format
- The Arrow library and its various programming languages (C++, Python, Java, etc.)
- Exploring IPC (Inter-Process Communication) and Flight (gRPC) interfaces
- Using datasets and RecordBatches

CONTRIBUTE TO APACHE ARROW

- Guide for new contributors: How to get started
- Process for reporting bugs and suggesting features
- Create and submit your first Pull Request (PR)
- Best practices for collaborative work with Git and GitHub
- Pull request life cycle and code review

DEVELOPMENT AND TESTING

- Configuring the development environment for different languages
- Compiling Arrow libraries: steps and troubleshooting
- Write, run and debug unit and integration tests
- Applying Apache Arrow coding style conventions
- Use of everyday development tools such as Archery

CONTINUOUS INTEGRATION AND PACKAGING

- Introducing continuous integration in the Arrow project

- Run Docker builds to validate changes
- Using Crossbow for packaging and testing
- Understanding automated build tools and processes
- Solving common problems in continuous integration

TUTORIALS AND ADDITIONAL RESOURCES

- Practical tutorials for using Arrow in Python and R
- Access additional resources to deepen your knowledge
- Help with documentation: how to contribute and improve documents
- Examples of advanced use of Apache Arrow in real-life situations
- Discuss the latest updates and upcoming features

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 enabling 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 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 course: 60% Practical, 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.

Sanction

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