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

Updated 10/29/2024

Sign up

Apache Parquet training

2 days (14 hours)

Presentation

Apache Parquet is the technology for you! It's an open source file format, optimized for storing and processing large quantities of analytical data. It uses columnar storage, allowing you to quickly read specific data sets without loading the entire file.

During this course, you'll explore the features of Apache Parquet, including its internal structure and metadata organization, which optimize data processing. You'll also learn how to configure your files for optimum performance, define schemas and use Bloom filters to narrow results and speed up queries.

The program will also cover advanced features, such as available compression options (e.g. Snappy, Gzip) to reduce data weight while maintaining data integrity, and page indexes, which facilitate efficient data retrieval. You'll also discover how to deal with errors and data corruption thanks to checksums, essential for guaranteeing the reliability of analyses.

At the end of this course, you'll be able to configure, optimize and maintain Parquet files adapted to various use cases. You'll master data encoding and the use of indexes, essential skills for processing large databases quickly and accurately. You'll also have the foundations to contribute to the Apache Parquet project if you wish, thereby enriching the open source ecosystem.

As with all courses, Apache Parquet will be presented with its latest version.
: Apache Parquet 1.14.

Objectives

- Master the advantages of Apache Parquet for processing big data
- Efficiently configure Parquet files for optimum performance

- Use encodings and filters to optimize queries
- Select and apply the appropriate compression methods
- Contribute to the development and evolution of the Apache Parquet project

Target audience

- Data analysts
- Data engineers
- Architects
- Developers

Prerequisites

- Good understanding of databases and structured file manipulation
- Experience with a programming language such as Python or Java
- Basic knowledge of massive data processing (Hadoop, Spark, etc.).
- Familiarity with columnar storage and compression concepts

APACHE PARQUET TRAINING PROGRAM

INTRODUCTION TO APACHE PARQUET

- Introducing Apache Parquet
- Advantages of using Parquet to process large amounts of data
- Comparison with other file formats such as CSV, JSON and ORC
- Typical use cases

CONFIGURATION AND METADATA

- Understanding the internal structure of a Parquet file
- Optimum configuration for enhanced performance
- Managing and using metadata
- Parquet extensibility for specific needs
- Practical examples of Parquet file configuration

DATA TYPES AND ENCODING

- Supported data types
- Nested encoding for structured data
- Using Bloom filters to optimize queries
- Handling null values in datasets
- Practical workshops on schema definition and data encoding

DATA PAGE COMPRESSION AND MANAGEMENT

- Available compression methods and how to choose them
- Importance of checksums for data integrity
- Segmentation of data into column chunks for efficient retrieval
- Using page indexes to improve query performance
- Error and data corruption recovery scenarios

DEVELOPER'S GUIDE AND PROJECT CONTRIBUTION

- Exploring Parquet sub-projects
- Process for compiling and building Parquet from source code
- How to contribute to Parquet-Java and to Parquet's overall development
- Parquet release and update process
- Creating a development environment for contributions

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 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.