AMBIENT-IT

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

Updated on 26/09/2024

Sign up

Apache Beam training

2 days (14 hours)

Presentation

Are you looking to master a powerful, flexible solution for processing large quantities of data in streams or batches? Our Apache Beam training course gives you a comprehensive introduction to this open source framework, designed to unify data processing in a variety of environments.

Apache Beam lets you create portable data processing pipelines that can run on multiple engines such as Google Cloud Dataflow, Apache Spark and Flink.

Thanks to its SDKs compatible with different languages (Python, Java), it is the ideal tool for Data Engineers, Data Scientists and developers looking to process large volumes of data in a scalable way.

During this course, you'll learn how to design efficient pipelines, manage data sources and sinks, and optimize your real-time processing with advanced techniques such as windowing and triggers.

You'll also acquire the skills needed to run and deploy your pipelines in cloud environments, such as Google Cloud Platform.

As with all our courses, Apache Beam will be presented with its latest features (at the time of writing).

Objectives

- Understand the fundamentals of Apache Beam and its place in the data processing ecosystem
- How Apache Beam compares to other technologies
- Master the architecture of Apache Beam and its various components
- Design, structure and execute data processing pipelines in flow and batch mode

- Optimize data processing with windowing and late element management

Target audience

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

Prerequisites

- Knowledge of fundamental concepts in data processing and data engineering
- Experience with a programming language (ideally Python or Java)
- Understanding the principles of databases and data lakes

APACHE BEAM TRAINING PROGRAM

INTRODUCTION TO APACHE BEAM

- Overview of data processing technologies and positioning of Apache Beam
- Comparison with other technologies such as Spark, Flink and Google Cloud Dataflow
- Installing and configuring the Apache Beam environment

APACHE BEAM ARCHITECTURE AND FEATURES

- Understanding Beam's architecture and key components
- Details of available SDKs and their compatibility with programming languages
- Presentation of the different runners (Direct, Dataflow, Flink, Spark)
- Supported distributed processing back-ends

APACHE BEAM PROGRAMMING MODEL

- Data flow in a Beam pipeline and basic terminologies
- Creation of a simple WordCount pipeline to illustrate the concept
- Analysis of essential transformations :
  - ParDo
  - GroupByKey
  - Windowing

PIPELINE DEVELOPMENT AND IMPLEMENTATION

- Writing and structuring a Beam pipeline

- Using complex transformations and creating composite transformations
- Managing data sources and sinks: Reading and writing data

PIPELINE EXECUTION AND DEPLOYMENT

- Pipeline execution methods: local, cloud and clustered
- Pipeline configuration for different runtime environments
- Deployment on Google Cloud Platform with Dataflow

WINDOWING, TRIGGERS AND LATE ELEMENT MANAGEMENT

- Window strategies :
  - Tumbling
  - Sliding
  - Session
  - Global Windows
- Using triggers to manage late elements
- Understanding Watermarks and their impact on real-time data processing

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 is used to check correct acquisition.

skills.

Sanction

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