AMBIENT-IT

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

Updated 05/02/2025

Sign up

Spark Training Streaming

3 days (21 hours)

Presentation

Spark is a framework for performing distributed calculations on a cluster computers. This course introduces the latest version 3.5.4, which brings a host of new features and impressive performance improvements!

Apache Spark can rapidly process large quantities of data on a massive scale. It has recently become one of the world's leading distributed computing frameworks. It has the advantage of integrating different programming languages such as Java, Scala, Python and R.

Our training course covers the advanced concepts of Spark Streaming and its integration with Kafka, as well as all the best practices for successful deployment in production. Practical work is carried out in Scala (or Python as an option).

Objectives

- Handling large volumes of data using Spark Streaming best practices
- Understanding the advanced concepts of the new Spark Streaming API v4
- Integrating and cohabiting Kafka with Spark Streaming
- Be able to use Spark Streaming in production

Target audience

- Developers
- Data Engineer
- Architects
- System administrators
- DevOps

Prerequisites

- Ideally, you should have taken our Spark ML or Spark Tuning Advanced training courses.
- Basic knowledge of a Unix system
- Knowledge of Scala, Git & Kafka

Spark Streaming training program

Day 1

Introduction to Spark (in a streaming context)

- Spark architecture
- Internal operation (Stage, Task, Scheduler ...)
- Batch vs Stream
- The microbatch model
- DStreams API with Scala

Structured Streaming

- Introduction to the Structured Streaming API.
- API source
- API Sink
- Functional API
- SQL streaming
- Streaming Json, Csv, Packet sources
- Calculating streaming aggregates

Day 2

Introduction to apache Kafka

- Internal operation (Topic, partition, Offset ...)
- Producer
- Consumers
- Partitioning
- Commit des offsets

Spark streaming integration with Kafka

- Streaming in Source and Sink
- Calculate aggregates in real time
- Stream-static and Stream-Stream joins
- Watermarks
- Windowing (tumbling, sliding, reduce...)

Day 3

Stateful Streaming

- State store
- GroupState operators
- Timeouts

Spark streaming in production

- State checkpointing and fault-tolerance.
- Monitoring via Spark-UI
- Tuning

Schematics management with Avro (Optional + 1 day on request)

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, 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.