AMBIENT-IT

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

Updated on 16/05/2024

Sign up

Spark Tuning Advanced training

4 days (28 hours)

Presentation

Designed in the USA in 2009, Apache Spark is a unified analysis engine for processing large quantities of data on a massive scale. This tool stands out for its ease of use, despite its ability to deliver sophisticated analyses.

This Spark Tuning training course is designed for administrators who want to optimize the performance of their data management system. Tuning and optimizing resources (CPU cores and memory) plays an important role in maintaining a high quality IT system.

After an introduction to the Scala language, and an explanation of Spark, we'll look at the RDD api, dataframes and Spark Streaming. We'll then take a look at Spark in production, and finish with an introduction to Machine Learning.

Each time, practical exercises on clusters of machines with significant datasets will enable you to assimilate the concepts presented.

Like all our training courses, this one will introduce you to the latest stable release, Spark 3.4.3.

Objectives

- Be able to install and use Spark 3 and its new features independently
- Be able to use Scala as Spark's main language
- Understanding and optimizing dataframes
- Understanding Spark tuning in production using best practices

Target audience

- Developers
- Architects
- System administrators
- DevOps

Prerequisites

- Initial experience in developing and putting Spark processing into production
- Basic knowledge of a Unix system
- Knowledge of Scala or Python & Git

Spark and Advanced Tuning training program

Day 1 - Introduction to Scala and Spark

- Why is Scala the language of Bigdata?
- Introduction to the functional paradigm
- Setting up environments
- Hands-on Scala
- Syntax
- Pattern matching
- API collection
- Functional types
- Why Spark?
- Spark architecture

Day 2 - Understanding and using Spark

RDD API

- RDD presentation
- PairedRDD
- Handling the RDD api (transformations , actions.....)
- Import and export to and from: cSv, Avro and Elasticsearch

Dataframe

- Introducing Dataframes
- Dataframe and UDF api
- SqlContext
- Using SQL with Dataframes
- Datasets

Day 3 - Dataframe and optimization

Optimization

- DAG analysis via Spark-UI
- Optimization pattern
- Cache and persistence
- Impact of data locality on Spark streaming

performance

- StreamingContext
- DStream
- Continuous Aggregations
- Real-time analysis from an Apache Kafka
- Delivery guarantee issues
- Spark vs Flink

Day 4 - Spark in prod and conclusion

Spark in production

- Spark in cluster: Yarn, Mesos, Standalone
- Yarn customer vs. Yarn cluster
- Storage (HDFS, S3, Cassandra .....)

Architecture

- Lambda Architecture
- Kappa Architecture

Introduction to Machine Learning (Optional)

- Classes of ML algorithms: supervised and unsupervised
- ML algorithms
- How the linear regression and/or logistic regression algorithm works
- Practical application of a linear regression algorithm or logistic regression

2 specific modules are available on an intra-company basis only Module for

Data Engineer - Spark Scala

Day 1 - RDD & Dataframes

RDD API

- RDD presentation
- PairedRDD
- Handling the RDD api (transformations, actions, etc.)
- Import and export to and from: cSv, Parquet

Dataframe

- Introducing Dataframes
- Dataframe and UDF api
- Using SQL with Dataframes
- Datasets

Day 2 - Production & Optimization

Optimization

- DAG analysis via Spark-UI
- Optimization pattern
- Cache and persistence
- Impact of data locality on Spark performance in

production

- Spark in cluster: Yarn, Mesos, Standalone
- Yarn customer vs. Yarn cluster
- Storage (HDFS, S3, Cassandra, etc.)

Module for Data Scientist - Spark Python

Day 1 - RDD & Dataframes

RDD API

- RDD presentation
- PairedRDD
- Handling the RDD api (transformations, actions, etc.)
- Import and export to and from: CSV, Parquet

Dataframe

- Introducing Dataframes
- Dataframe and UDF api
- Using SQL with Dataframes
- Datasets

Day 2 - Spark ML/ MLlib

Algorithms

- Classes of ML algorithms: supervised and unsupervised
- ML algorithms
- How does the linear regression algorithm work?
- Logistics, Random Forest...
- Clustering: KNN, K-mean

MLlib

- Introduction to MLlib 2.0
- Pipelines: Transformer, Estimator, Model
- Cross-Validation
- Hyperparameters tuning
- ML persistence: saving and loading pipelines

Companies concerned

This training 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.