AMBIENT-IT

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

Updated on 16/05/2023

Sign up

Dask training: Scaling Python tools

2 days (14 hours)

Presentation

The Dask training course will enable you to process a large data set and learn parallel programming in Python. Dask transforms Python code from local multicore machines to large distributed clusters in the cloud.

During this course, you will acquire the knowledge and skills you need to interact with Dask. The software offers a familiar user interface, mirroring the APIs of various ecosystem libraries such as PyData and Pandas.

Dask has a task scheduler that runs task graphs in parallel. As a result, custom workloads and high-level collections will be fed by you through low-level collections.

Like all our training courses, this one will be presented with the latest stable version and all its new features (2023.1.0 released on January 13, 2023 at the time of writing).

Objectives

- How to use DataFrames
- Easier scaling of PyData libraries
- Scaling different types of code
- Understand how to deploy tasks locally using multiple Python processes

Target audience

- Data Analyst
- Data Miners

- Developers

Prerequisites

- Programming in Python
- Knowledge of data processing and manipulation

Dask training program

Introduction

- Introducing Dask
- Installation and configuration
- Best practices

Fundamentals

- Task cluster deployment
- Real-time task management
- Deploying Dask
  - Creating Dask displayboards
  - Overlapping calculations
  - In-house design
  - Sparse tables
- Creating and storing Dask DataFrames
  - DataFrame and Parquet
  - DataFrame and SQL
  - Indexing in DataFrames
  - In-house design

Internal

- Dask Internal
- Understanding the costs associated with code
- Planning
- Working with task graphs
- Debugging and performance

Environmental management

- Maintaining consistent environments
- Temporary installations
- Send files directly to compute nodes
- Redefining Dask objects

- Single-machine schedulers

Selecting the collection backend

- Change the default main library
- Define a new collection backend
- Define a separate method for each distributable creation routine

Reference

- API
- Command line interface
- Development guidelines
- Manager's instructions
- Change log
- Configuration

Xarray with dask arrays

- Launching the client to provide a dashboard
- Opening a dataset
- Perform standard Xarray operations
- Automatic parallelization and customized workflows
- Parallel and continuous computing
- Data retention in memory

Companies concerned

This course is aimed at companies, large or small, wishing to train their teams in a new, advanced computer technology.

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.