dask

Parallel Distributed Computing in Python

Ahmed Iqbal

WHAT IS DASK?

01.

Core Library (Collections/API)

The foundational part of Dask, offering efficient parallelized data structures and algorithms.

02.

Distributed Computing

Creation and management of computational clusters for scaling out data processing tasks.

03.

Integrations and Ecosystem

Integrating with various tools and libraries, fostering a rich and diverse ecosystem.

CORE LIBRARY (COLLECTIONS/API)

High-level collections

Dask Dataframes

Dask Array

Dask Bag

Premade clothes, use them right out of the box.

Low-level collections

Dask Delayed & Futures

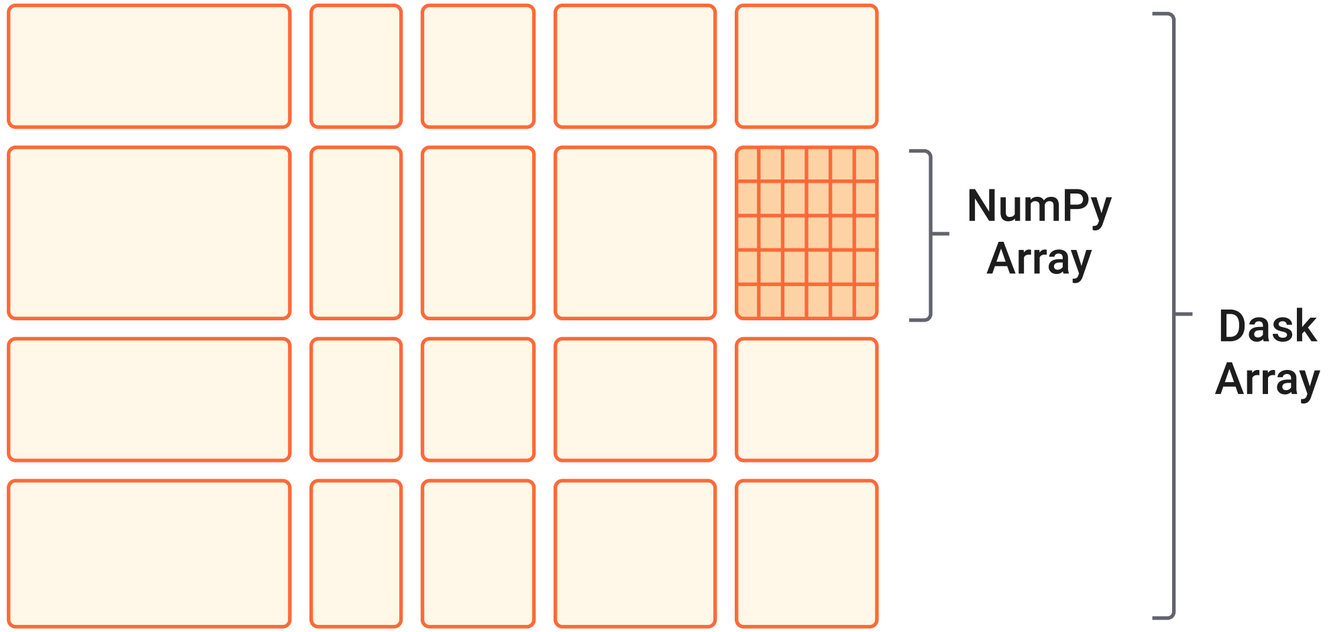
Choose your fabric and stitch your own clothes.

DASK ARRAYS – PARALLELIZED NUMPY

01. **Parallel Computing**

02. **Larger-than-Memory Data Handling**

03. **Blocked Algorithms**



DELAYED – PARALLELIZE ANY CODE

01.

Lazy Evaluation

Delay computation until explicitly triggered, enabling optimization of task execution and resource allocation.

02.

Control Flow

Fine-grained control over task execution order and dependencies, enhancing flexibility in workflow design.

FUTURES - NON-BLOCKING DISTRIBUTED CALCULATIONS

01.

Asynchronous Execution

Enables asynchronous execution of tasks, allowing concurrent computation and efficient resource utilization.

02.

Fault Tolerance

Provides fault tolerance mechanisms to handle failures gracefully, ensuring robustness in distributed computing environments.

DASK FLOW

Collections
(create task graphs)

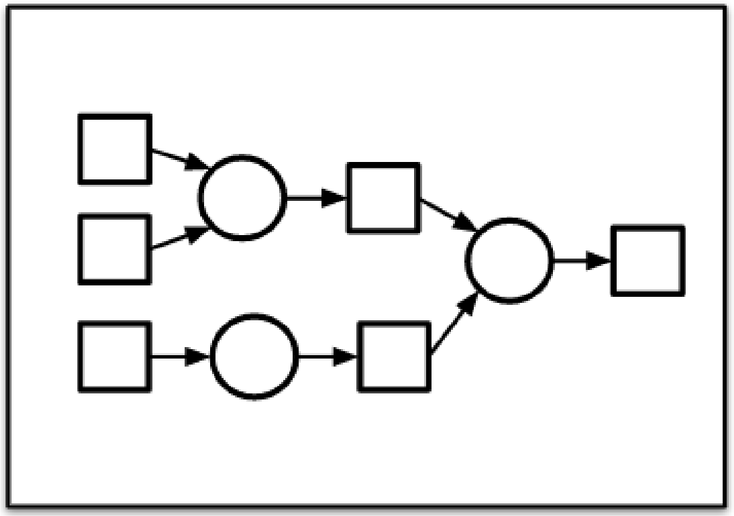
→

Task Graph

→

Schedulers
(execute task graphs)

- Dask Array
- Dask DataFrame
- Dask Bag
- Dask Delayed
- Futures



- **Single-machine**
  (threads, processes, synchronous)
- **Distributed**

DASK-ML

01.

Machine Learning Integration

Integrates with popular machine learning libraries like scikit-learn, XGBoost, TensorFlow, and PyTorch to parallelize and scale machine learning workflows.

02.

Ease of Use

Offers a familiar scikit-learn-like interface, making it easy for users to transition existing workflows to distributed environments.

THANKYOU!

Any Questions?