Distributed computing in Python

Petr Janík, 2023

Motivation

- 😊 It's easy to develop POC ML algorithms locally on small data.
- 😓 It's hard to put them to production and/or run them on large data.

Solution

Distributed computing

Project intro

 Poetry

 FastAPI

Dask.distributed

A lightweight library for distributed computing in Python.

How to install

```
poetry add dask distributed
```

Ray

Open-source unified compute framework that makes it easy to scale AI and Python workloads.

How to install

```
poetry add "ray[default]"
```

Ray Client

Connects to a remote Ray cluster.

```python
ray.init("ray://localhost:10001")


@ray.remote
def square_number_ray(num: int) -> int:
    return square_number(num)



future = square_number_ray.remote(2)
result = ray.get(future)  # 4
```

Ray Jobs

The recommended way to run a job on a Ray cluster.

Jobs CLI

```
ray job submit --address ray://localhost:10001 -- python script.py
```

Python SDK:

```python
client = JobSubmissionClient("ray://localhost:10001")
job_id = client.submit_job(
    entrypoint="python script.py",
)
```

https://docs.ray.io/en/latest/cluster/running-applications/job-submission/sdk.html

Thank you

Code + Slides: https://github.com/petr7555/dask-ray-demo