Coiled

Scaling Data Science with Dask

Pavithra Eswaramoorthy

Nov'21

Hi! I'm Pavithra :)

Data Science Evangelist at Coiled

FOSS contributor - Bokeh & Wikimedia

Bachelor's in Computer Science

Internet: @pavithraes

Email: pavithra@coiled.io

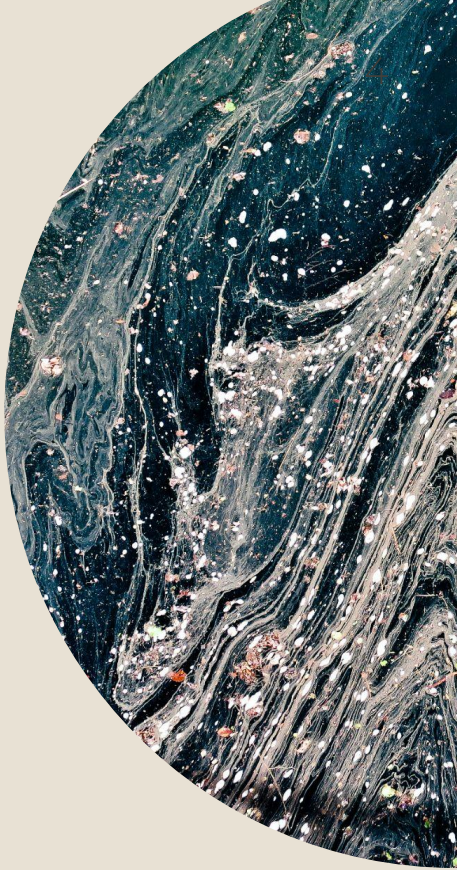
bit.ly/
pyladies-dublin-dask

What’s your experience with Scalable Data Science?

On a scale of 1 - 5

1: What is scalable computing?

5: I understand the challenges and have a working solution



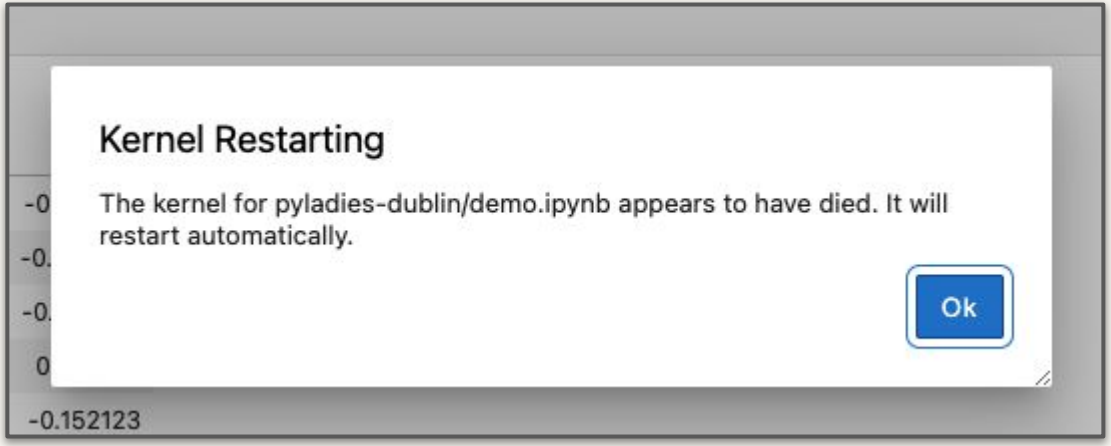
We will discuss:

- PyData Ecosystem - *awesomeness and limitations*
- pandas computations (1 GB)
- Dask - *What? Why? How?*
- Local parallelism (50 GB)
- Distributed cloud computing (1 TB)
- Build intuition around Dask

pandas

1 GB dataset

~50GB dataset:

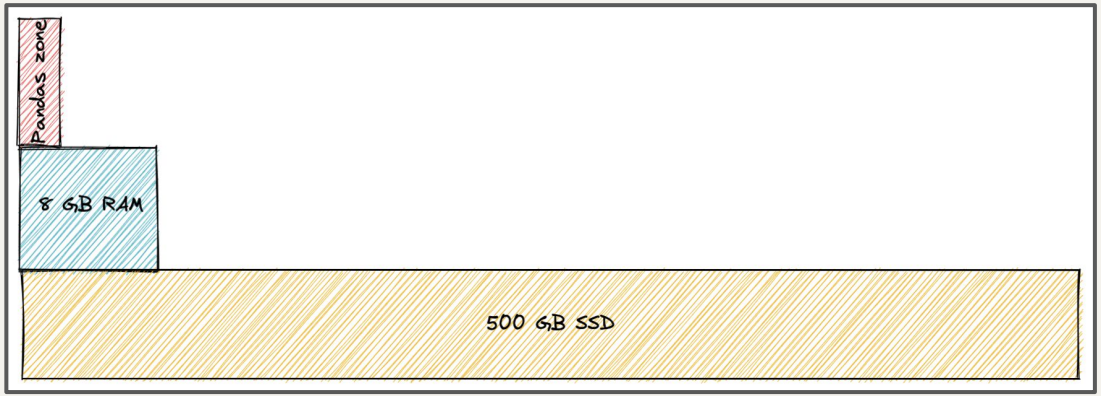


pandas rule of thumb

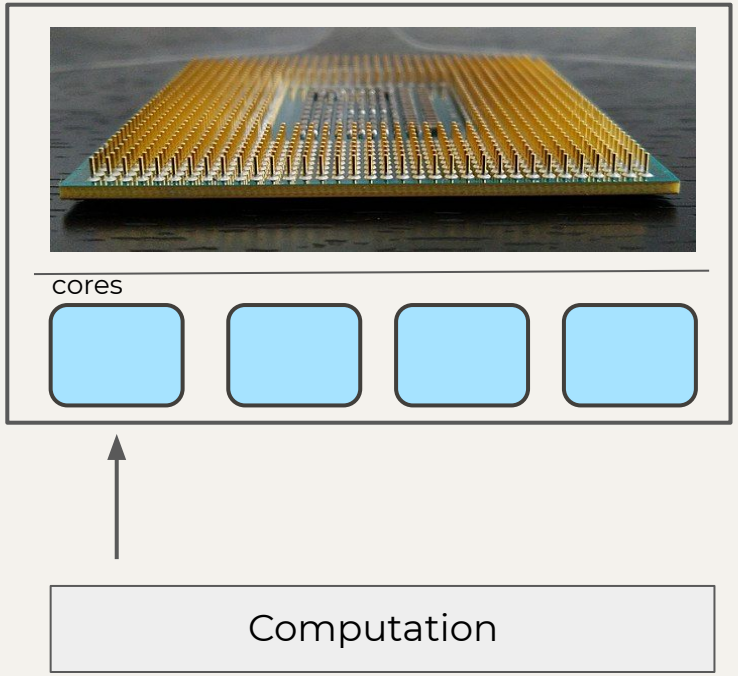
"Have 5 to 10 times as much RAM as the size of your dataset"

~ *Wes McKinney (2017) in* *10 things I hate about pandas*

Let me know if you have an updated rule of thumb!



Single-core computing



Dask

Dask is an open source library for
parallel and distributed computing in Python,
that follows the familiar syntax
of the existing PyData ecosystem.

Dask is an **open source** library for
parallel and distributed computing in Python,
that follows the familiar syntax
of the existing PyData ecosystem.

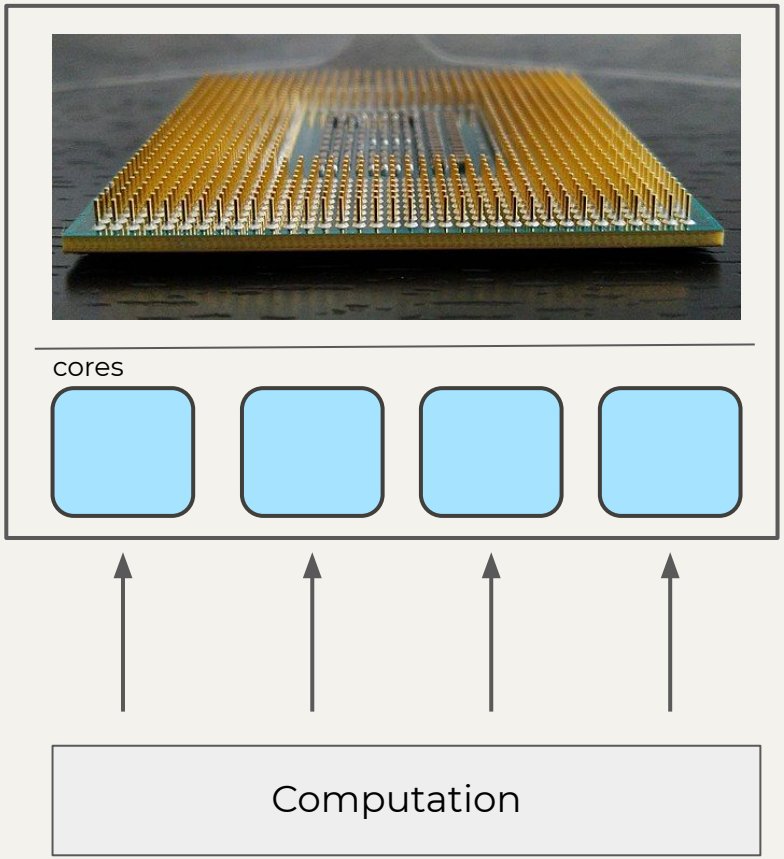
Dask is an open source library for
parallel and distributed computing in Python,
that follows the familiar syntax
of the existing PyData ecosystem.

Parallel computing

Working in parallel

Use all CPU cores

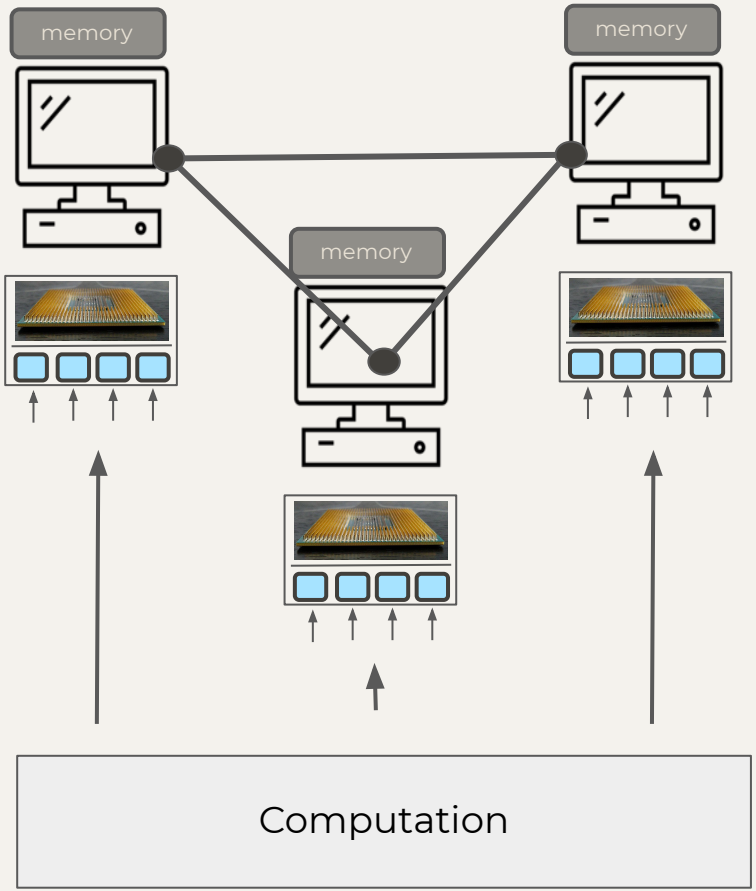
Multiple processes and shared memory



Distributed computing

Using groups of machines

Each machine has processors and memory



Dask is an open source library for parallel and distributed computing in Python, that follows the **familiar syntax** of the existing PyData ecosystem.

```python
from sklearn.liner_model \
 import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
```

```python
from dask_ml.linear_model \
 import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
```

```python
from sklearn.liner_model \
 import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
```

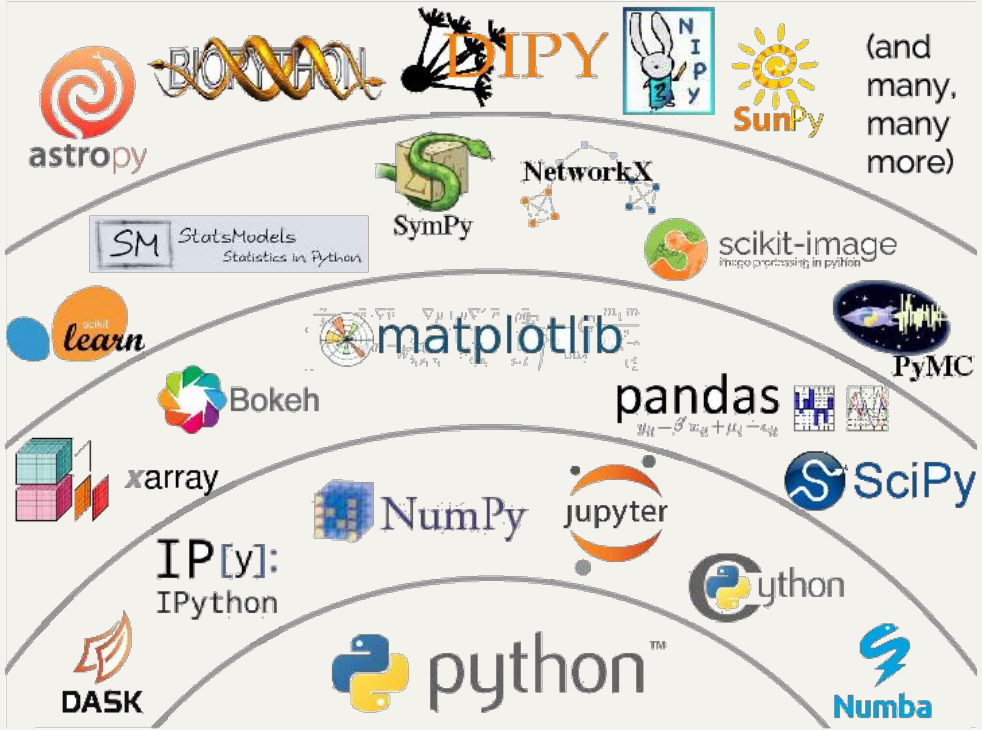
```python
from dask_ml.linear_model \
 import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
```

Dask is an open source library for parallel and distributed computing in Python, that follows the familiar syntax of the existing **PyData ecosystem**.

Python Ecosystem

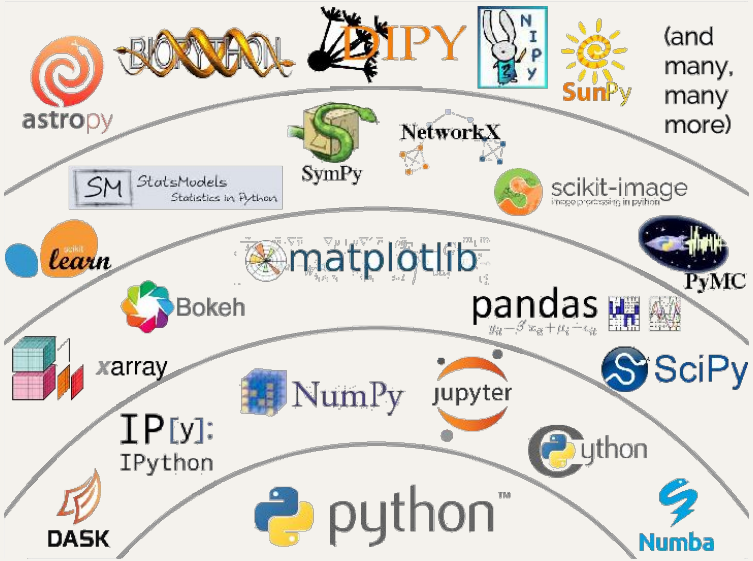
Python ecosystem is great

But.... it did not scale.



Credit: Jake Vanderplas 2017 PyCon Keynote

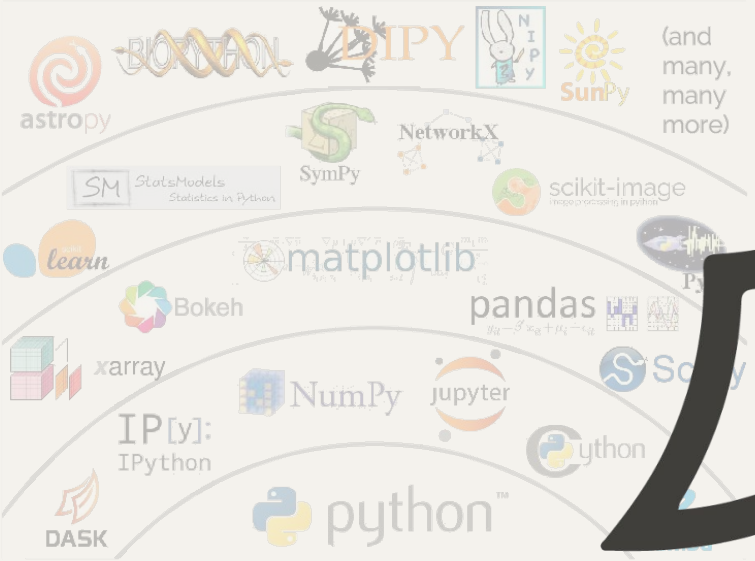
Python tools



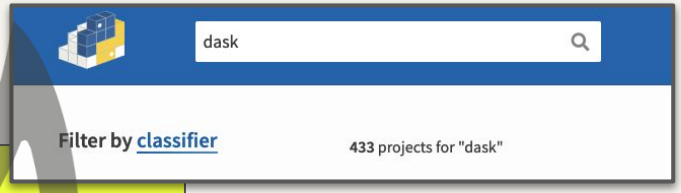
Scalable tools

Python tools

Scalable tools



The Python Ecosystem today is Scalable



PREFECT | RAPIDS | ...

dmlc XGBoost | PyTorch | OPTUNA

SQL | scikit learn | xarray | scikit-image image processing in python | Realtime

Asyncio | Web Servers | Numba | NumPy | pandas | HPC Networks | CUDA Python

Dask (local)

50GB dataset

What about a **1TB** dataset?

My hard drive only has 500GB

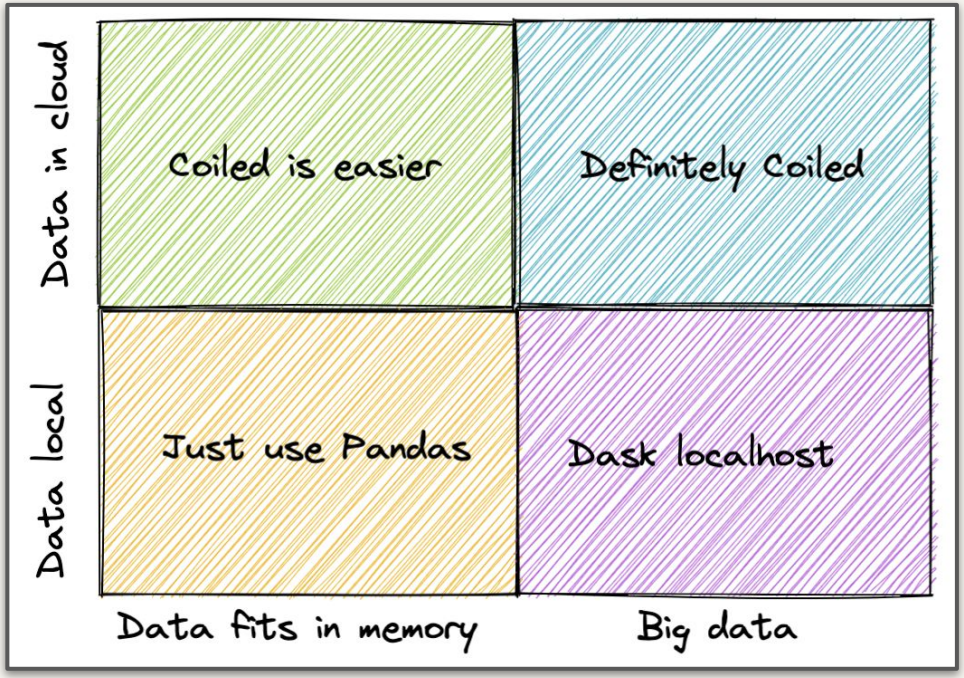
Dask local wouldn’t work for this dataset

Need a bigger computer or… the cloud?

Dask (cloud)

1TB dataset

pandas vs Dask rules of thumb



What makes Dask special?

Welcome to **DASCAR 2021**

Dask APIs

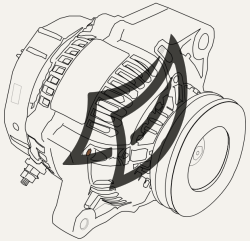
Race cars!

Dask Array

Dask DataFrame

Dask Bag

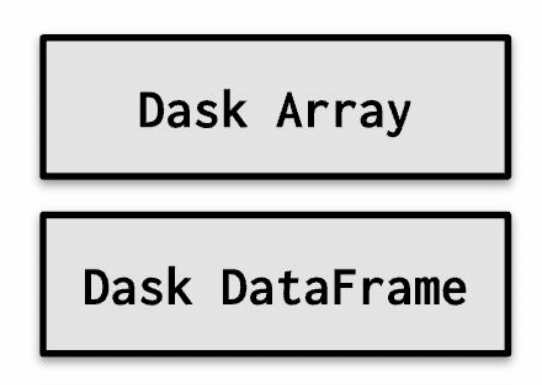
Dask engine



Dask Delayed

Futures

You can use an existing race car.

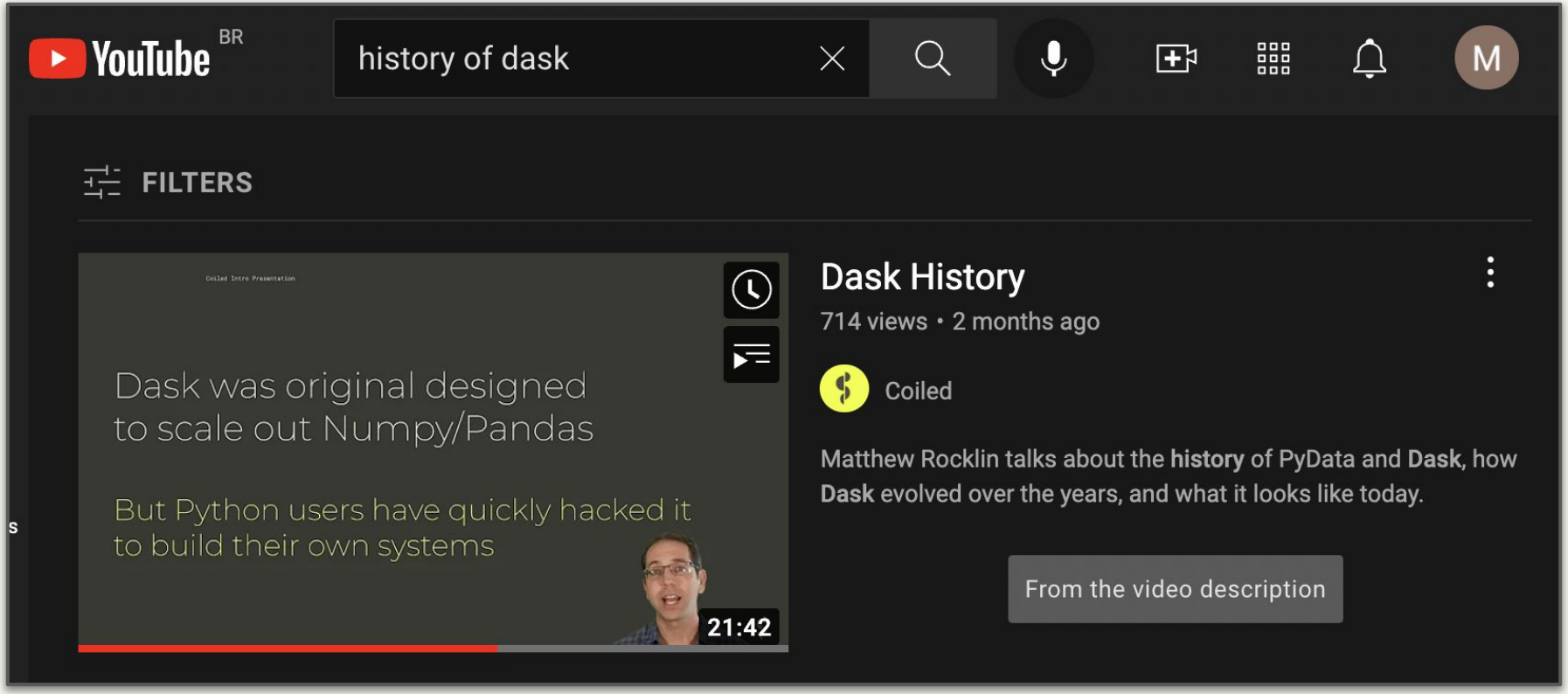


You can use an existing race car.

Dask Array

Dask DataFrame

Dask ML

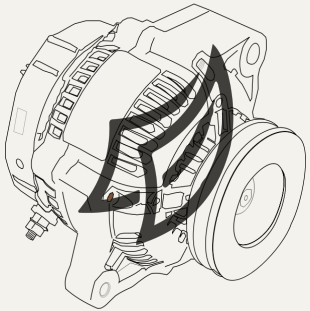

YouTube BR
history of dask
M
FILTERS
Coiled Intro Presentation
Dask was original designed
to scale out Numpy/Pandas
But Python users have quickly hacked it
to build their own systems
21:42
Dask History
714 views • 2 months ago
Coiled
Matthew Rocklin talks about the history of PyData and Dask, how Dask evolved over the years, and what it looks like today.
From the video description

It is as though we built a nice car
And everyone wanted the engine

~ Matthew Rocklin (Original creator of Dask)

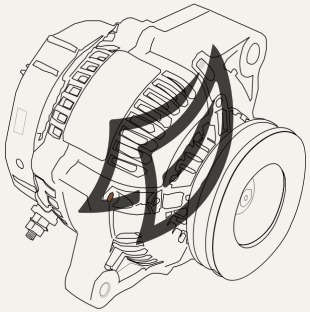
You can build your own custom distributed execution race car

Dask DataFrame



+ pandas

You can build your own custom distributed execution race car

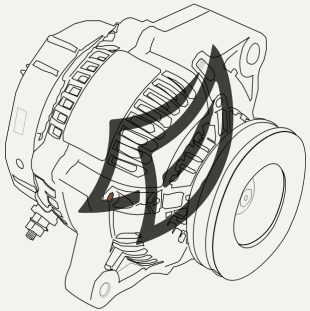


Dask Delayed

Futures

Your own race car!
[Totally optional!]

You can build your own custom distributed execution race car



+

PREFECT

Prefect

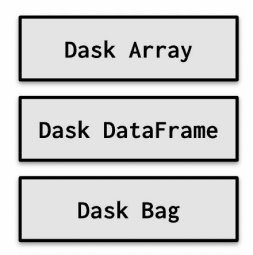
Coiled Intro Presentation

Most Dask usage today is indirect

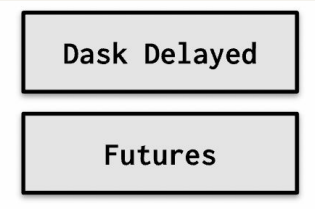
(xarray, prefect, rapids, xgboost, ...)

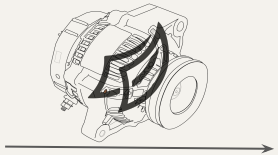
10:24 / 21:41

Direct Dask users



Indirect Dask users





RAPIDS

pandas

PyTorch

dmlc XGBoost

scikit learn

PREFECT

Large Scale ETL

Financial Services | Research | Technology

BARCLAYS
Capital One
zebra MEDICAL VISION
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
PREFECT
ACCURE
RenaissanceRe
Steppingblocks

Machine Learning

Technology | Research | Financial Services

zebra MEDICAL VISION
Steppingblocks
ACCURE
TRUMID
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
Capital One
BARCLAYS
U.S. AIR FORCE
SIDEWALK LABS
datarevenue
NCAR

Large Scale EDA

Life Sciences | Technology | Geophysical Sciences | Government | Research

University of Oxford
Berkeley UNIVERSITY OF CALIFORNIA
NASA
Met Office
FARALLON INSTITUTE
zebra MEDICAL VISION
HARVARD MEDICAL SCHOOL
Chan Zuckerberg Initiative
Los Alamos NATIONAL LABORATORY
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON
Brookhaven National Laboratory
THE WORLD BANK
Steppingblocks
Walmart+
GRUBHUB
BlueYonder
ACCURE
TRUMID
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

Forecasting

Retail | Technology

Walmart+
GRUBHUB
BlueYonder
TRUMID
ACCURE

Image Processing

Life Sciences | Technology | Geophysical Sciences | Government | Research

HARVARD MEDICAL SCHOOL
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH
Berkeley UNIVERSITY OF CALIFORNIA
Los Alamos NATIONAL LABORATORY
Brookhaven National Laboratory
NASA
Chan Zuckerberg Initiative
zebra MEDICAL VISION
FARALLON INSTITUTE
Met Office
WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

Geospatial

Financial Services

THE WORLD BANK

Complex Algorithms

Financial Services

D E Shaw & Co

Want to learn more?

dask.org

@CoiledHQ