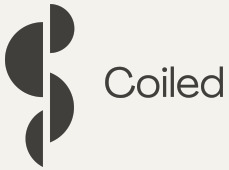


Scaling Data Science with Dask

Pavithra Eswaramoorthy

PyData Delhi, May 2022

Hi! I'm Pavithra :)

Community+OSS at Coiled

Core at Bokeh

Past, OSS Outreach at Wikimedia

Bachelor's in Computer Science

@pavithraes on the internet

hi@pavithraes.me

We will discuss:

- Parallel, Distributed, and Cloud Computing
- What is Dask? How it works?
- Dask DataFrame API (parallelize pandas)
- Dask Delayed API (parallelize general Python)
- Distributed Scheduler and Diagnostic Dashboards
- Best Practices + Resources

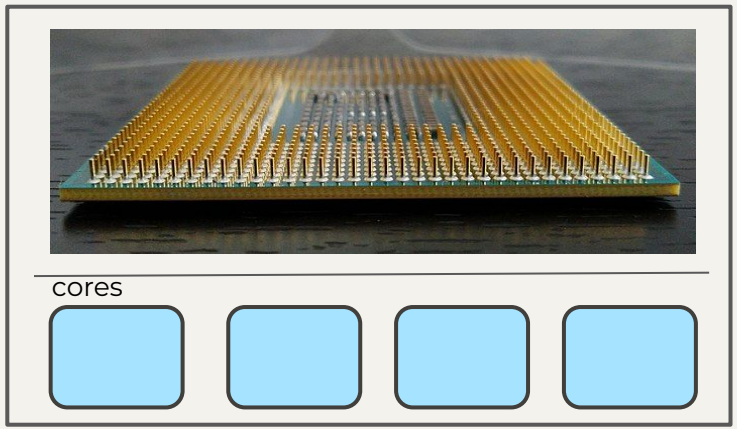
What is Dask?

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.

Single-core computing

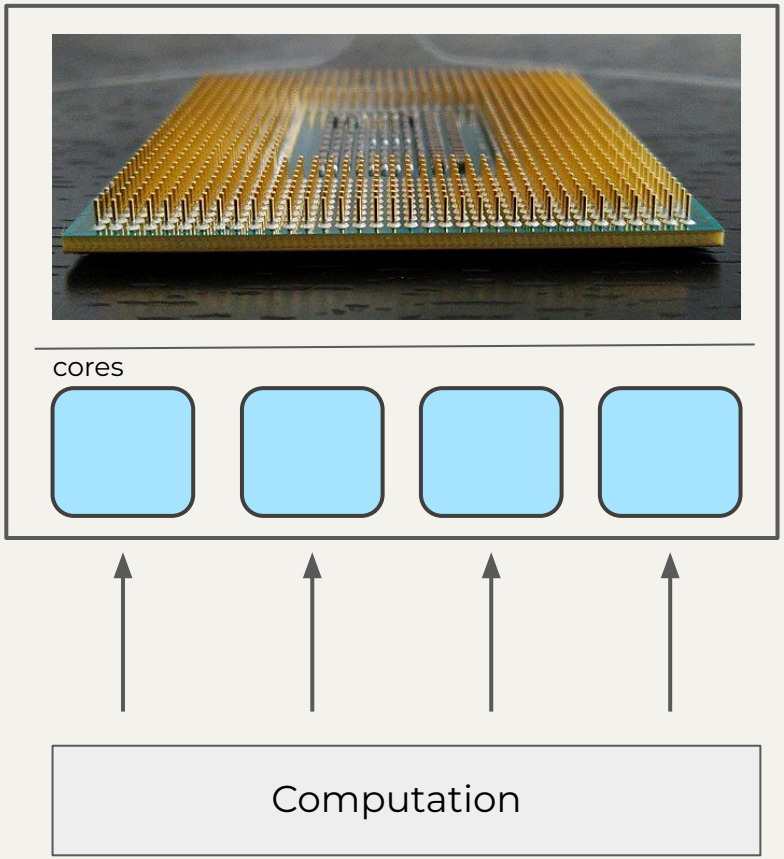


Parallel computing

Working in parallel

Use all CPU cores

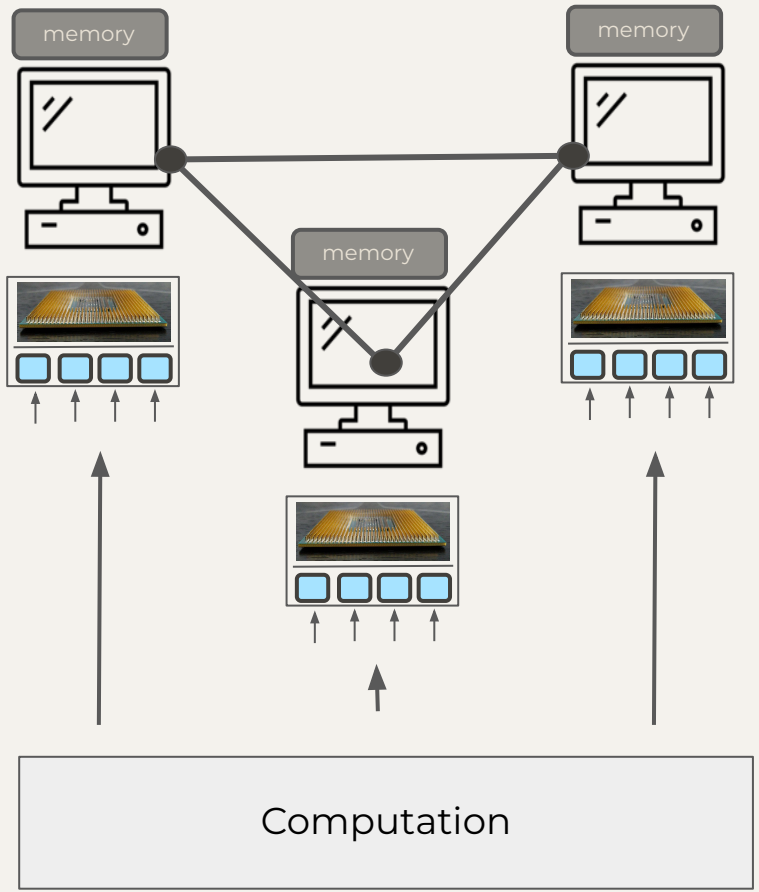
Multiple processes and shared memory



Distributed computing

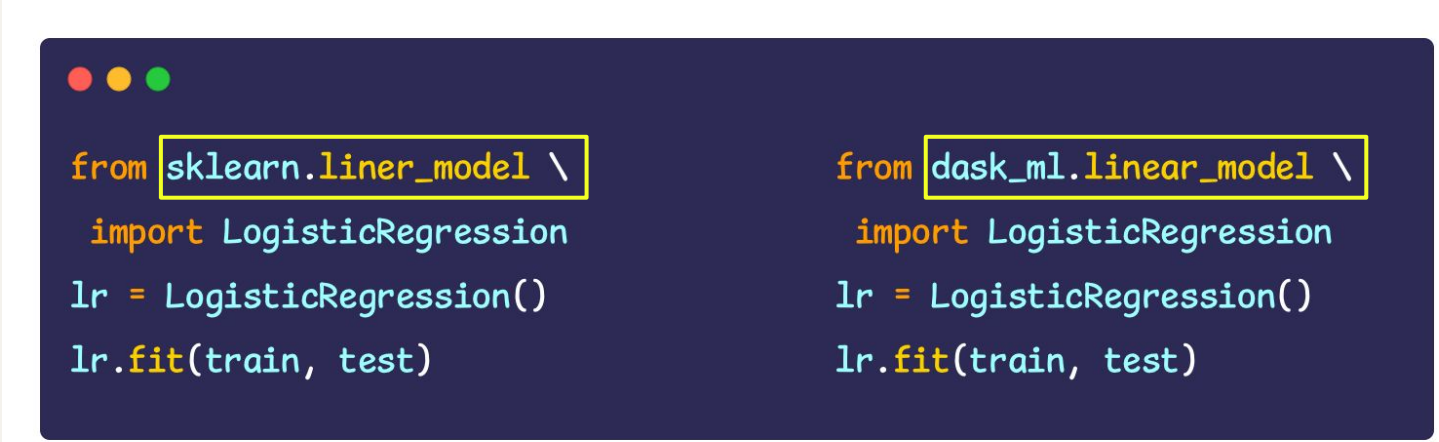
Using groups of machines

Each machine has processors and memory

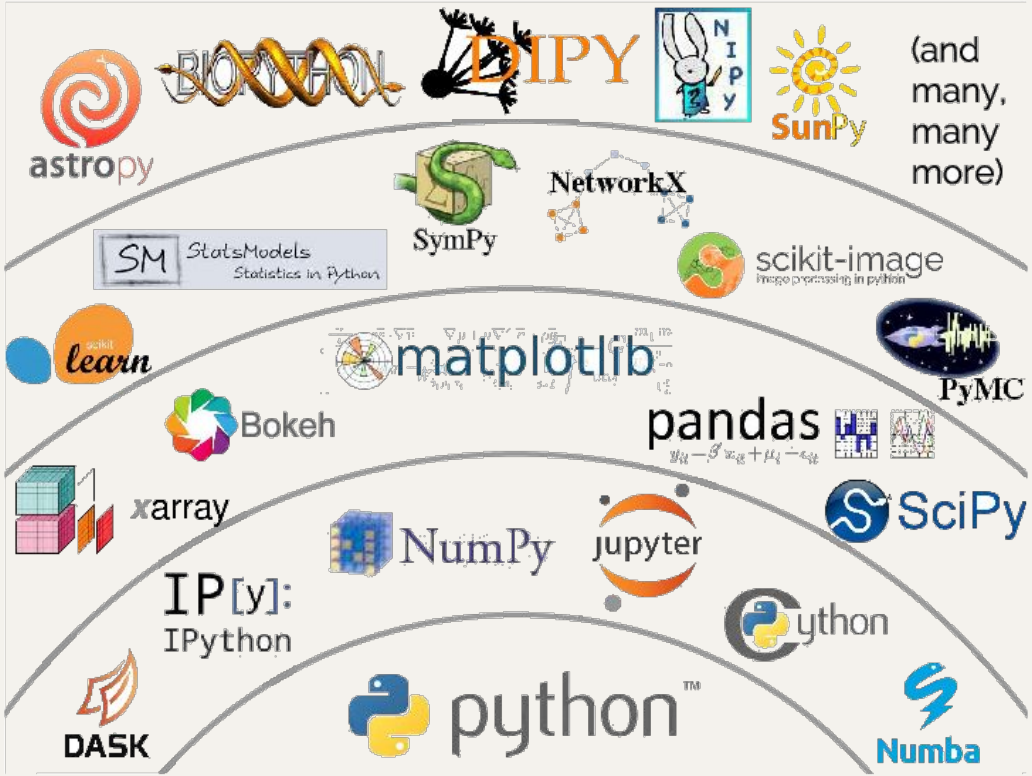


Dask is an open source library for
parallel and distributed computing in Python,
that follows the **familiar syntax**
of the existing PyData ecosystem.


from sklearn.liner_model \
import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
from dask_ml.linear_model \
import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)

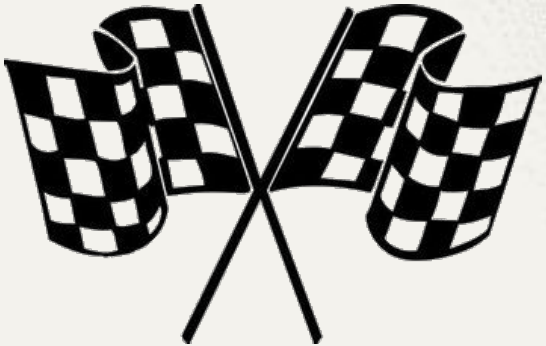

from sklearn.liner_model \
import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)
from dask_ml.linear_model \
import LogisticRegression
lr = LogisticRegression()
lr.fit(train, test)

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



Credit: Jake Vanderplas 2017 PyCon Keynote

What makes Dask special?



Welcome to **DASCAR 2021**

Dask APIs

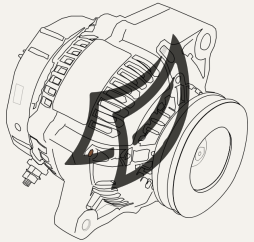
Race cars!
(High-level collections)

`Dask Array`

`Dask DataFrame`

`Dask Bag`

Dask engine
(Low-level collections)



`Dask Delayed`

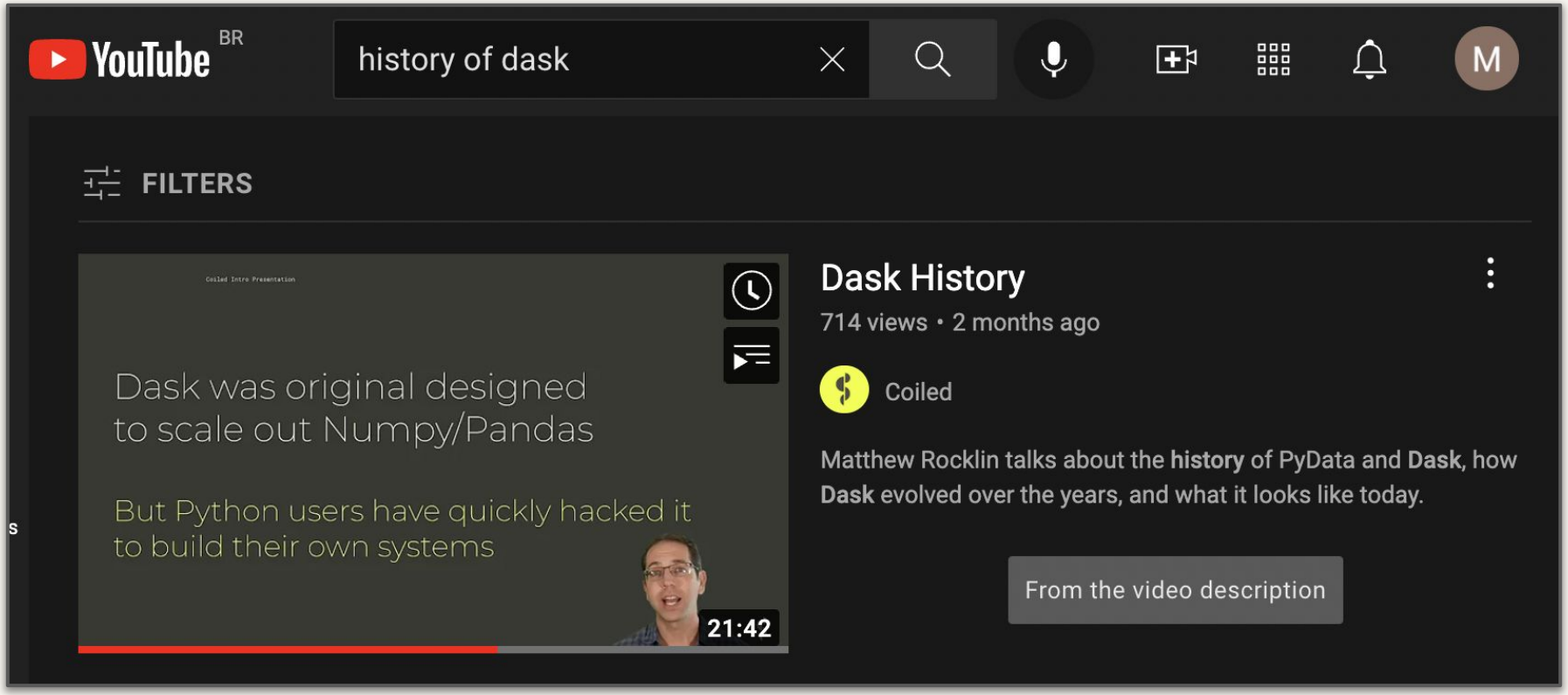
`Futures`

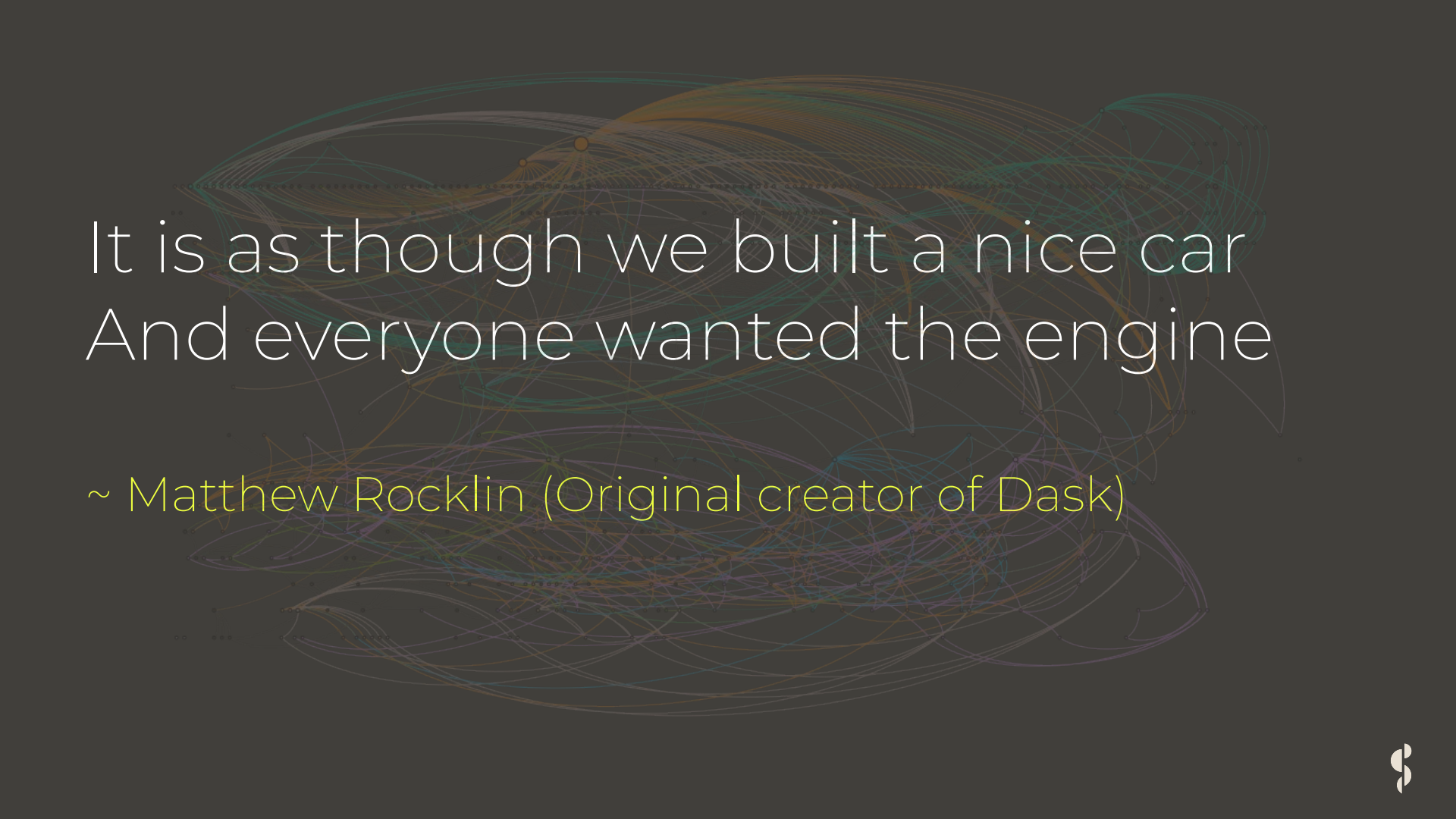
You can use an existing race car.

Dask Array

Dask DataFrame

Dask ML

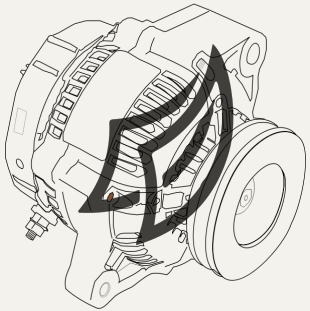

YouTube BR
history of dask
M
FILTERS
Coiled Intro Presentation
Dask was original designed
to scale out Numpy/Pandas
But Python users have quickly hacked it
to build their own systems
21:42
Dask History
714 views • 2 months ago
Coiled
Matthew Rocklin talks about the history of PyData and Dask, how Dask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



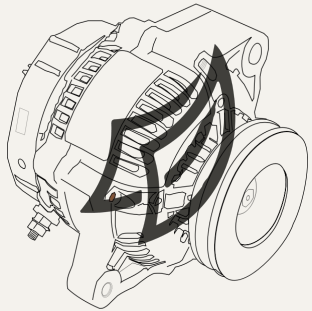
Your own race car!
[Totally optional!]

`Dask Delayed`

`Futures`

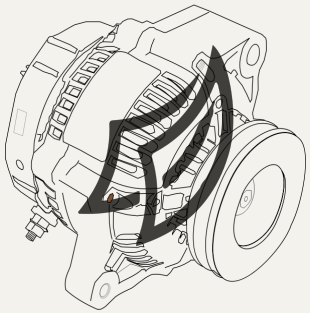
You can build your own custom distributed execution race car

Dask DataFrame



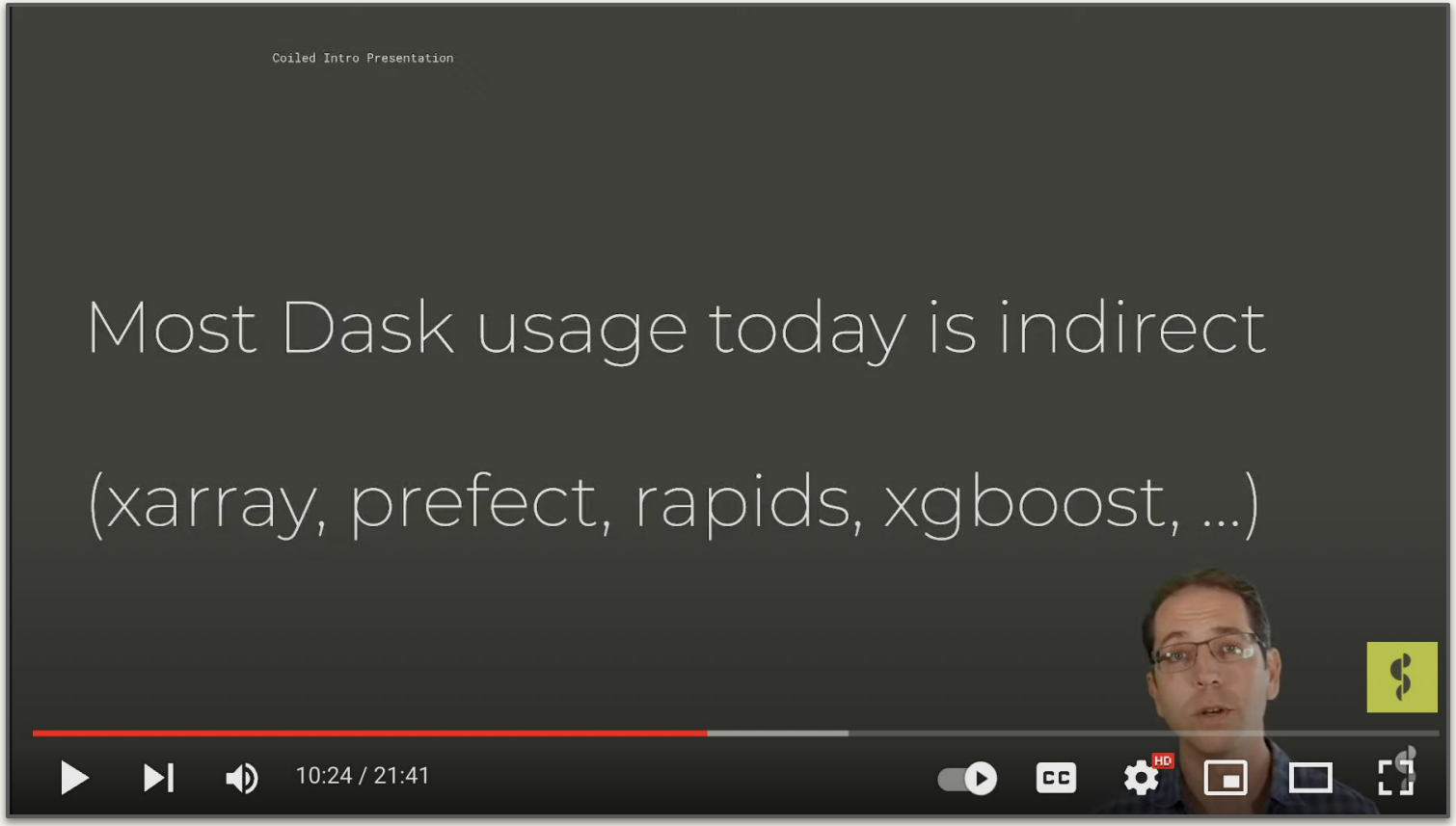
+ pandas

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

Dask Array

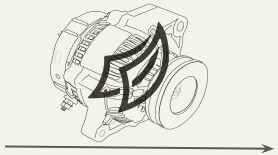
Dask DataFrame

Dask Bag

Indirect Dask users

Dask Delayed

Futures



RAPIDS

dmlc XGBoost

pandas

scikit learn

PyTorch

PREFECT

Who uses Dask?

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

Let’s jump in!

Documentation: https://docs.dask.org

GitHub: https://github.com/dask/dask

Discourse: https://dask.discourse.group/

Slack: https://dask.slack.com/

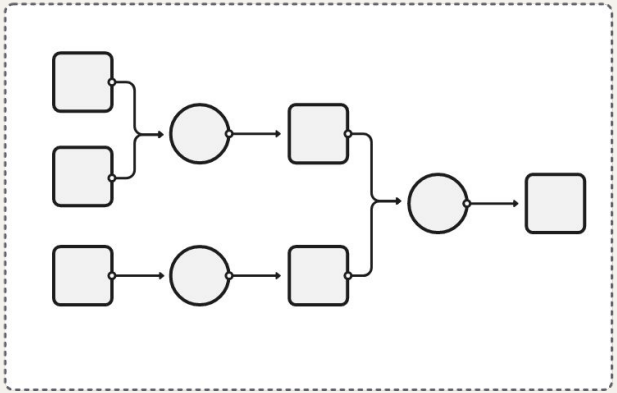
How does Dask work?

Collections
(create task graphs)

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

→

Task Graph



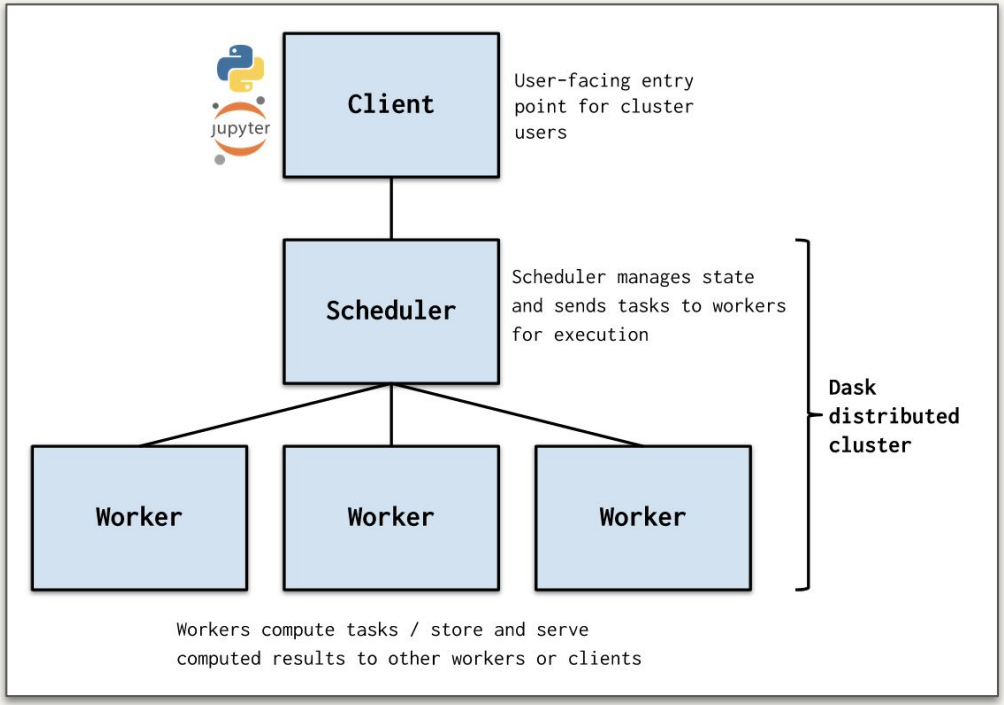
→

Schedulers
(execute task graphs)

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

Source: docs.dask.org

Dask Jargon: Client, Scheduler and Workers



Note: *The Scheduler and the Workers are on the same network, they could live in your laptop or on a separate cluster.*

When to use Dask?

Does your data fit in memory?

Yes: Use pandas or NumPy.

No : Dask might be able to help.

Do your computations take for ever?

Yes: Dask might be able to help.

No : Awesome.

Do you have embarrassingly parallelizable code?

Yes: Dask might be able to help.

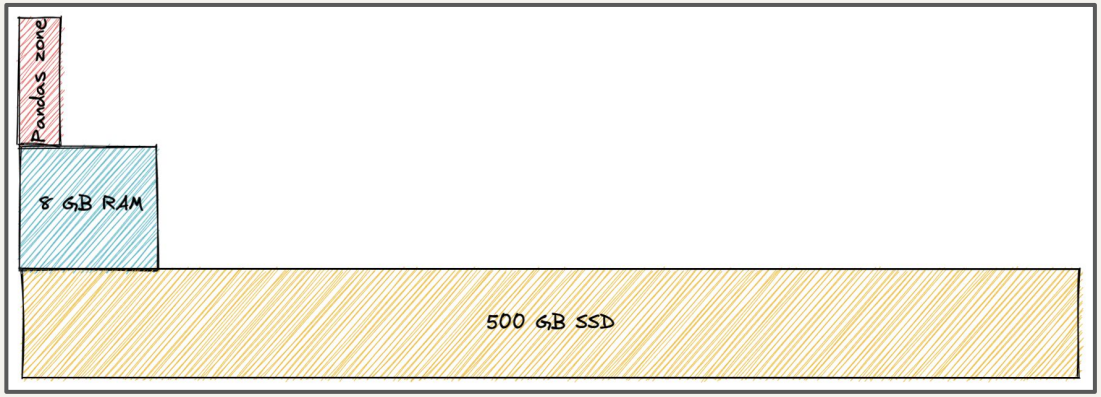
No?: If you are not sure here are some examples.

No: I'm sorry, although Dask might have some hope for you.

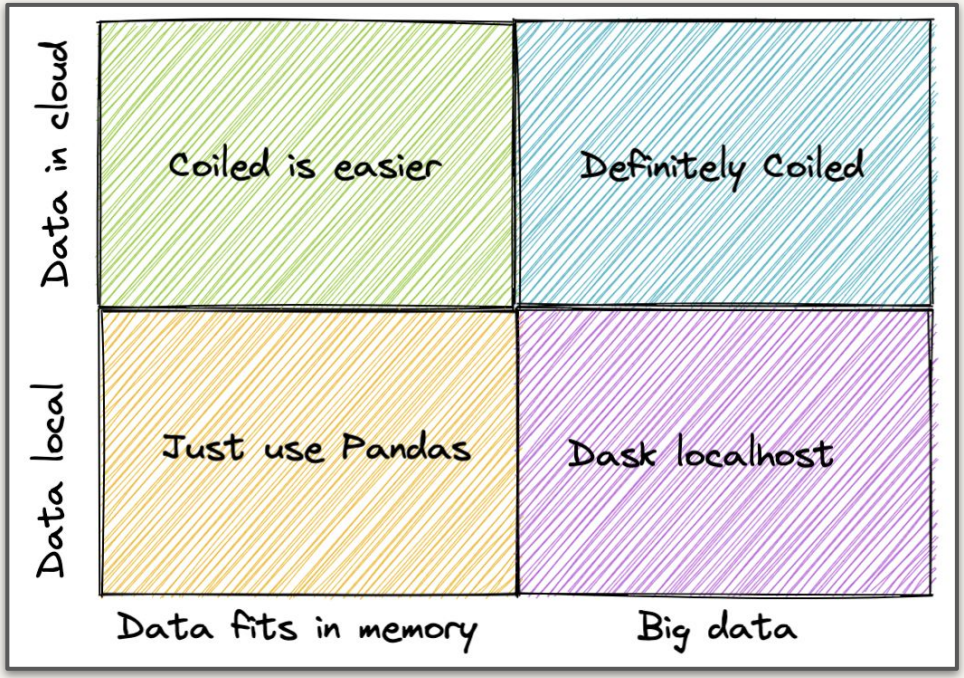
pandas rule of thumb

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

~ *Wes McKinney, creator of pandas (2017)*



pandas vs Dask rules of thumb



bit.ly/
pydata-delhi-dask

Any Questions?