

Getting up to speed with Dask



Aaron Richter
ChiPy Meetup - April 2021



<https://github.com/rikturr/getting-up-to-speed-with-dask>



Saturn Cloud

Hi!

Aaron Richter



Senior Data Scientist @ Saturn Cloud

- I work to make data scientists faster and happier

PhD in Machine Learning

aaron@saturncloud.io

rikturr.com

[@rikturr](https://twitter.com/rikturr)



Saturn Cloud

Saturn Cloud

Bringing together the fastest hardware + OSS



- Pythonic parallelism
- Rapidly scale PyData

RAPIDS

- Multi-GPU computing
- The future of HPC



- Workflow orchestration
- Flow insight and mgmt



kubernetes

- Fast setup
- Enterprise secure



Dask

Parallel computing for Python people

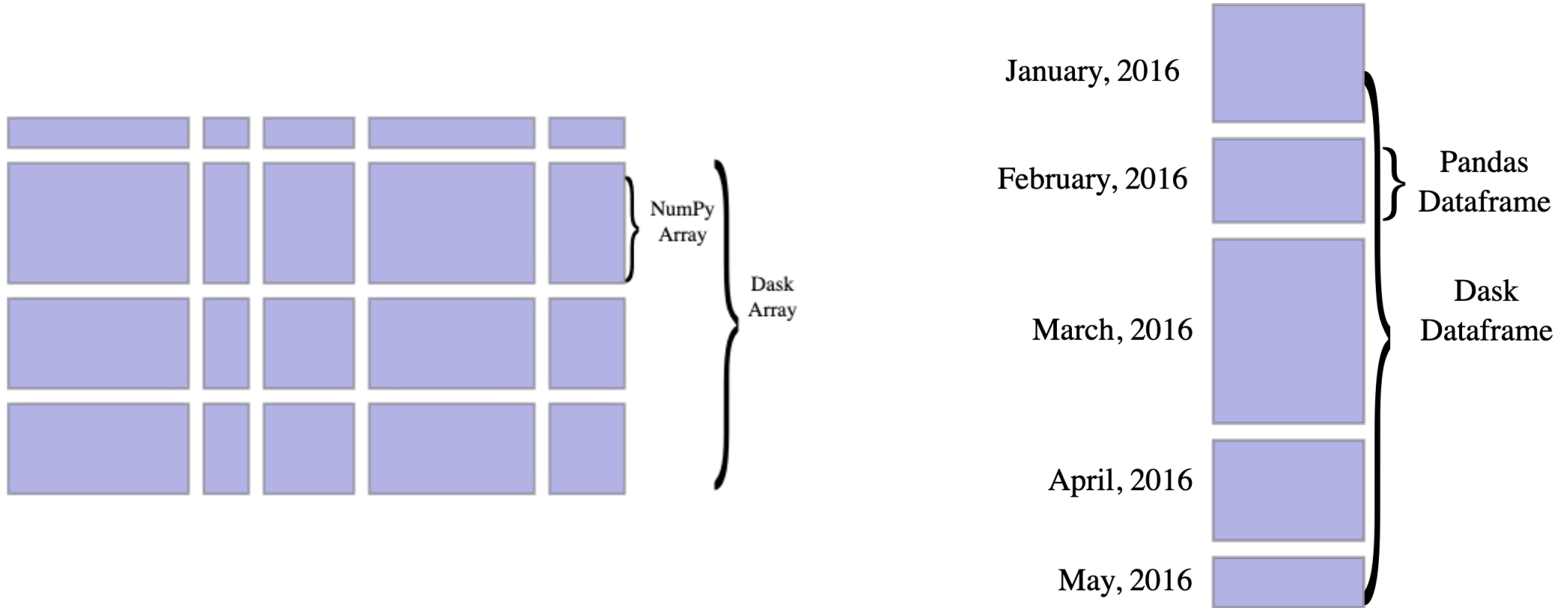
Dask

What does it do?

- Parallel machine learning (scikit)
- Parallel dataframes (pandas)
- Parallel arrays (numpy)
- Parallel anything else

What does it do?

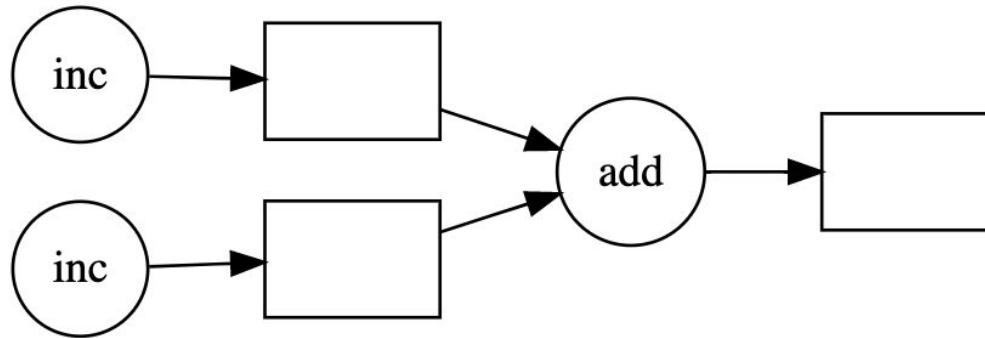
Arrays and Dataframes



What does it do?

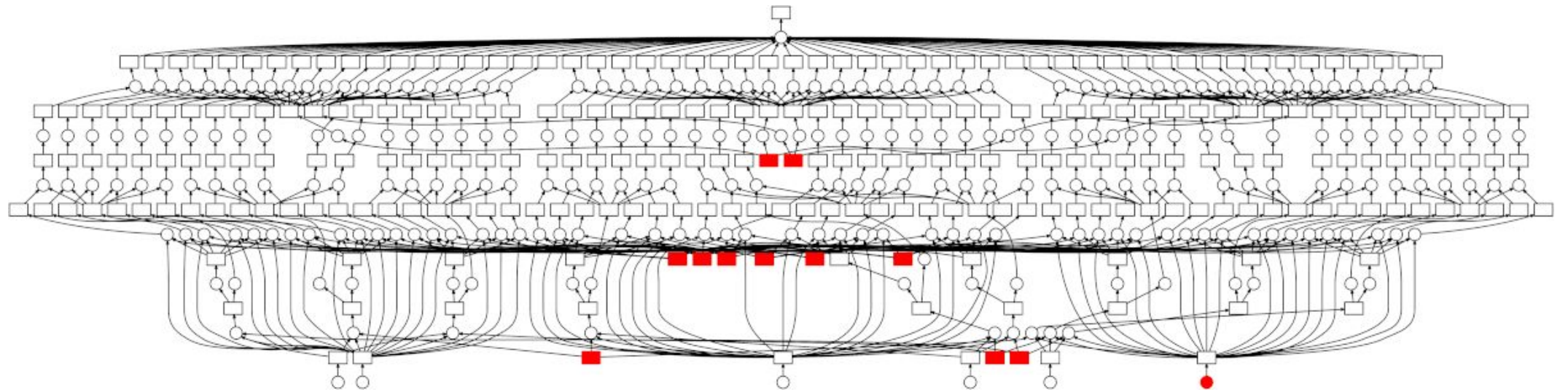
Anything else!

```
>>> x = dask.delayed(inc)(1)
>>> y = dask.delayed(inc)(2)
>>> z = dask.delayed(add)(x, y)
>>> z.compute()
5
>>> z.visualize()
```



What does it do?

Anything else!



Dask

Why should I use it?

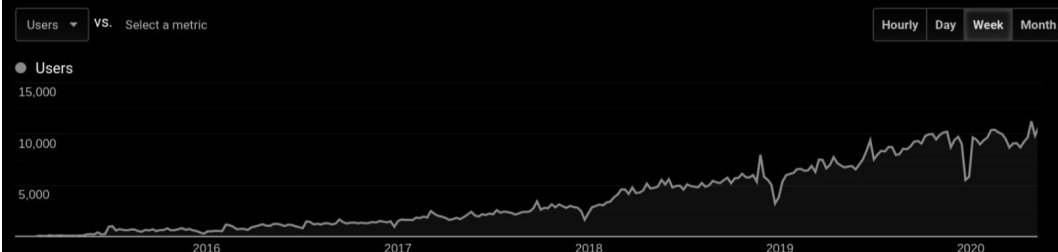
- Python native
- Strong ecosystem (PyData)
- Easily scalable

Dask

Why should I use it?

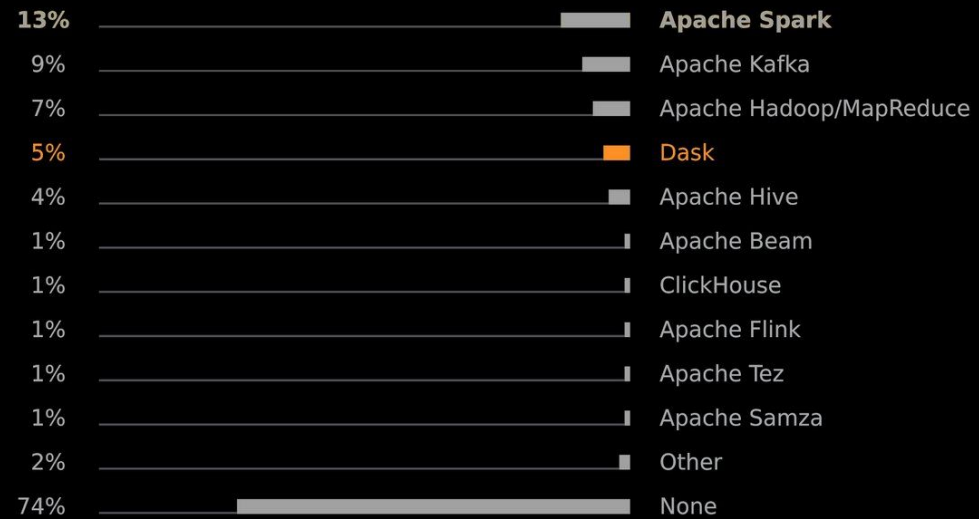
10,000 Documentation Visitors

Unique visitors on a weekly basis



5% of Python developers

(among those who take the Python survey)

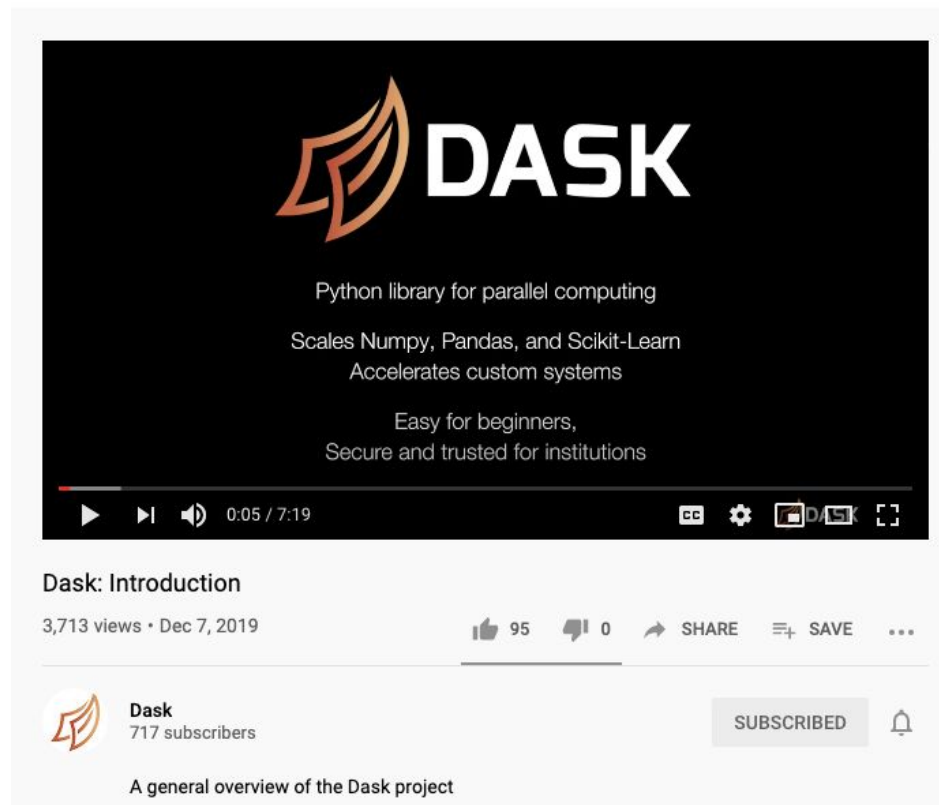


<https://www.jetbrains.com/lp/python-developers-survey-2019/>



How do I get started?

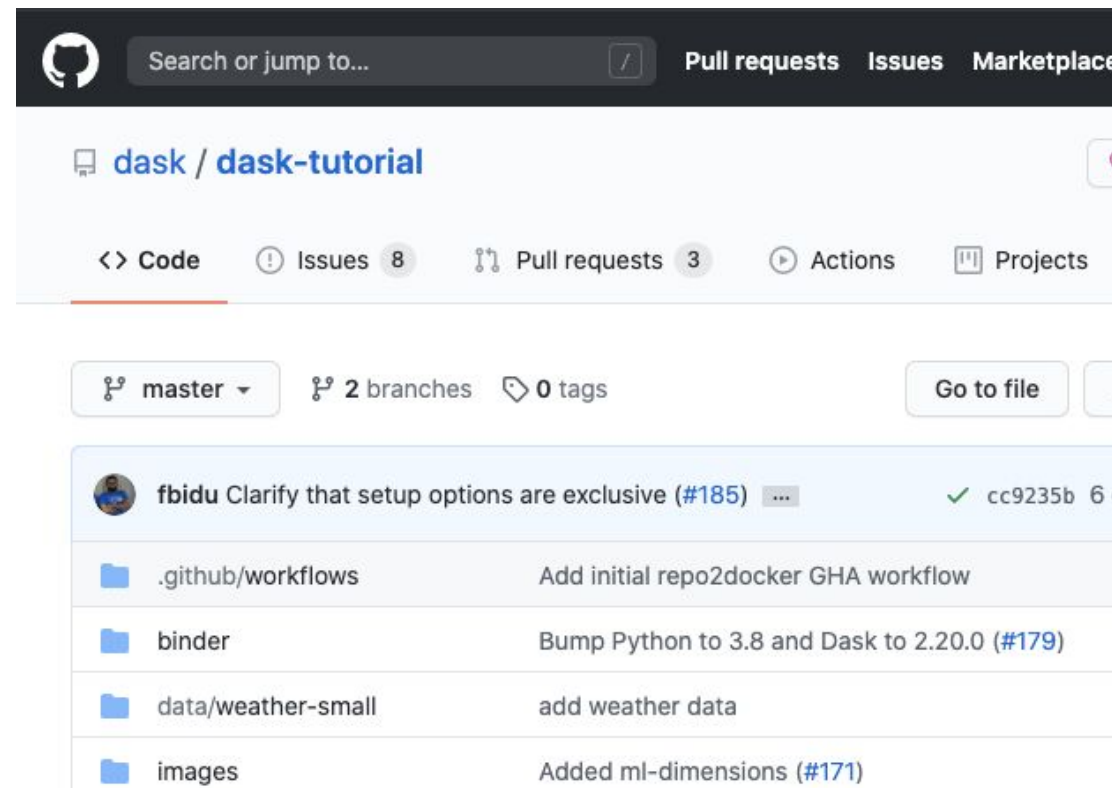
Videos



https://www.youtube.com/watch?v=nnndxbr_Xq4

How do I get started?

Tutorial



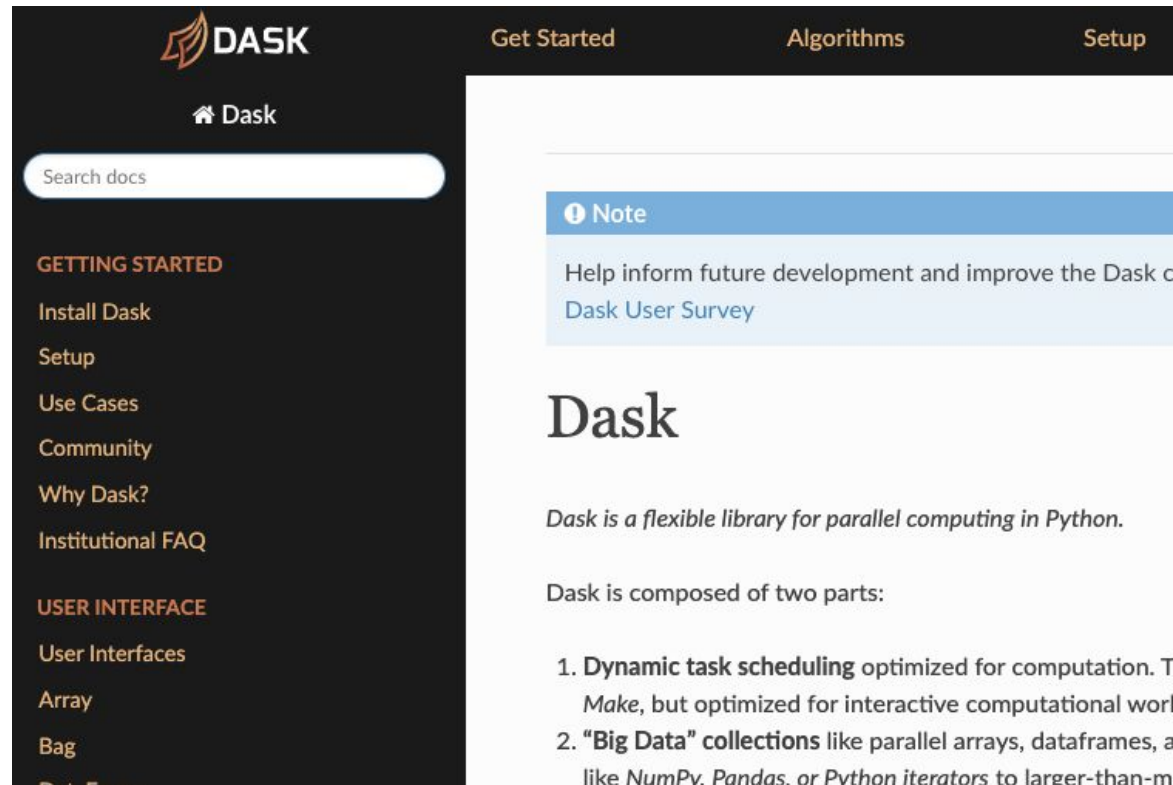
The screenshot shows the GitHub interface for the `dask / dask-tutorial` repository. At the top, there's a search bar and navigation links for Pull requests, Issues, and Marketplace. Below the repository name, there are tabs for Code, Issues (8), Pull requests (3), Actions, and Projects. The 'Code' tab is selected. Underneath, it shows the 'master' branch with 2 branches and 0 tags. A 'Go to file' button is visible. The main content area displays a list of files and folders with their commit messages:

File/Folder	Commit Message
<code>.github/workflows</code>	Add initial repo2docker GHA workflow
<code>binder</code>	Bump Python to 3.8 and Dask to 2.20.0 (#179)
<code>data/weather-small</code>	add weather data
<code>images</code>	Added ml-dimensions (#171)

<https://github.com/dask/dask-tutorial>

How do I get started?

Docs



The screenshot shows the Dask documentation website. The top navigation bar includes the Dask logo and links for 'Get Started', 'Algorithms', and 'Setup'. A left sidebar contains a search bar and a list of navigation links under 'GETTING STARTED' and 'USER INTERFACE'. The main content area features a 'Note' box, the title 'Dask', a description of Dask as a flexible library for parallel computing in Python, and a list of its two main components: dynamic task scheduling and 'Big Data' collections.

DASK

Get Started Algorithms Setup

🏠 Dask

Search docs

GETTING STARTED

- Install Dask
- Setup
- Use Cases
- Community
- Why Dask?
- Institutional FAQ

USER INTERFACE

- User Interfaces
- Array
- Bag
- DataFrames

Note

Help inform future development and improve the Dask c
[Dask User Survey](#)

Dask

Dask is a flexible library for parallel computing in Python.

Dask is composed of two parts:

1. **Dynamic task scheduling** optimized for computation. T
Make, but optimized for interactive computational worl
2. **"Big Data" collections** like parallel arrays, dataframes, a
like NumPy, Pandas, or Python iterators to larger-than-m

<https://docs.dask.org/en/latest/>

How do I get started?

Key concepts

- Task graph
- Lazy execution
- Parallel objects are “normal” objects under the hood

Code time!



<https://github.com/rikturr/getting-up-to-speed-with-dask>

Dask

Should I use it now?

- Use pandas (numpy) until you can't
- Then use Dask on your laptop
- Then try a big machine in the cloud
- *Then* go for clusters (in the cloud)

Dask

Running on a cluster

- Dask runs on most cluster/HPC platforms
 - Hadoop/YARN, Kubernetes, SLURM, etc.
- Rent your machines! (AWS, Azure, GCP)
- Managed solutions like Saturn Cloud, Coiled

Get involved!

Thriving community of open source contributors

- <https://docs.dask.org/en/latest/develop.html>



Thank you!

aaron@saturncloud.io

<https://rikturr.com>



@rikturr