Getting up to speed with Dask

Aaron Richter
July 2020

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

Saturn Cloud

Saturn Cloud

Bringing together the fastest hardware + OSS

DASK

- Pythonic parallelism
- Rapidly scale PyData

RAPIDS

- Multi-GPU computing
- The future of HPC

PREFECT

- Workflow orchestration
- Flow insight and mgmt

kubernetes

- Fast setup
- Enterprise secure

DASK

Dask

Parallel computing for Python people

Dask

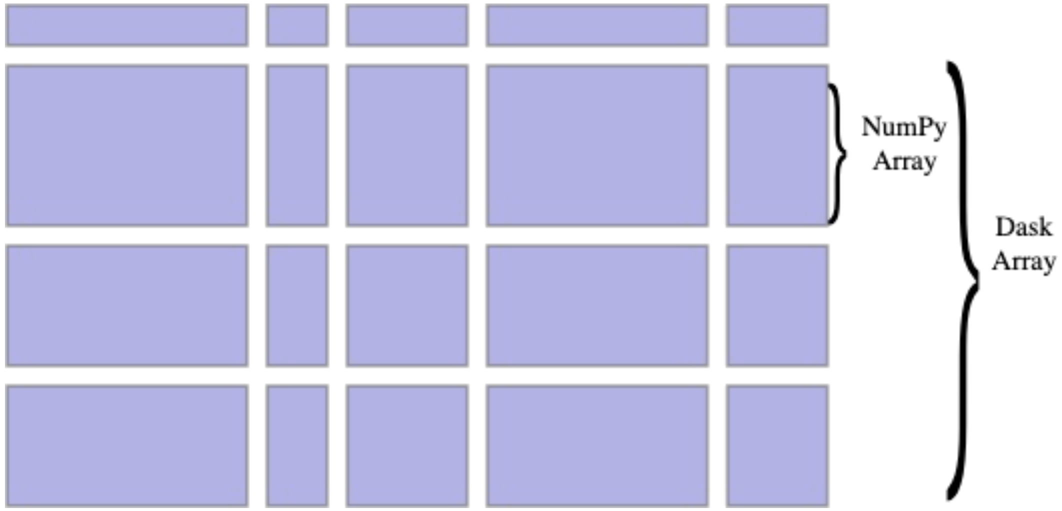
What does it do?

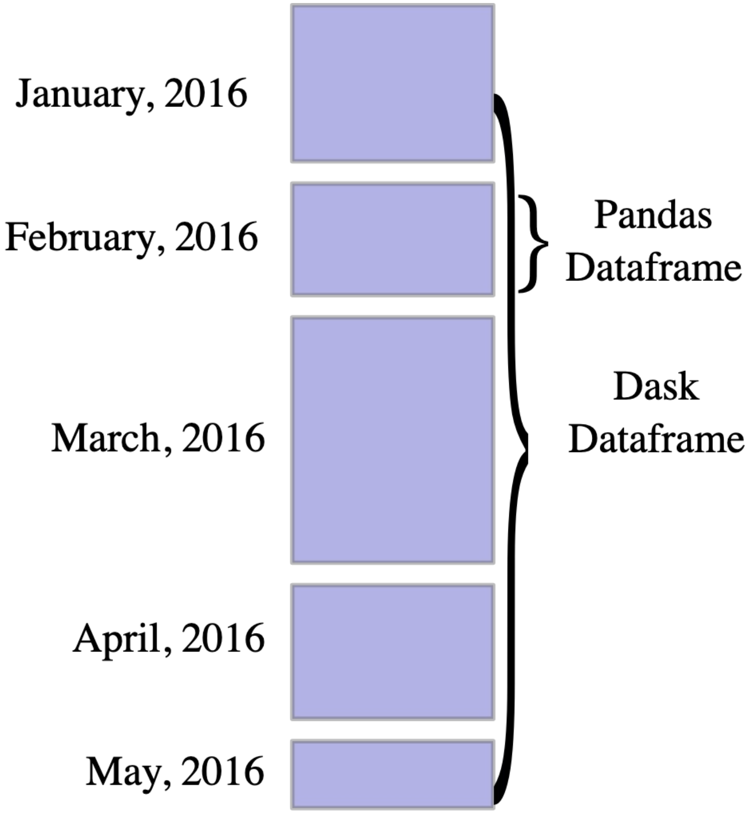
- Parallel machine learning (scikit)
- Parallel dataframes (pandas)
- Parallel arrays (numpy)
- Parallel anything else

https://docs.dask.org/en/latest/user-interfaces.html

What does it do?

Arrays and Dataframes



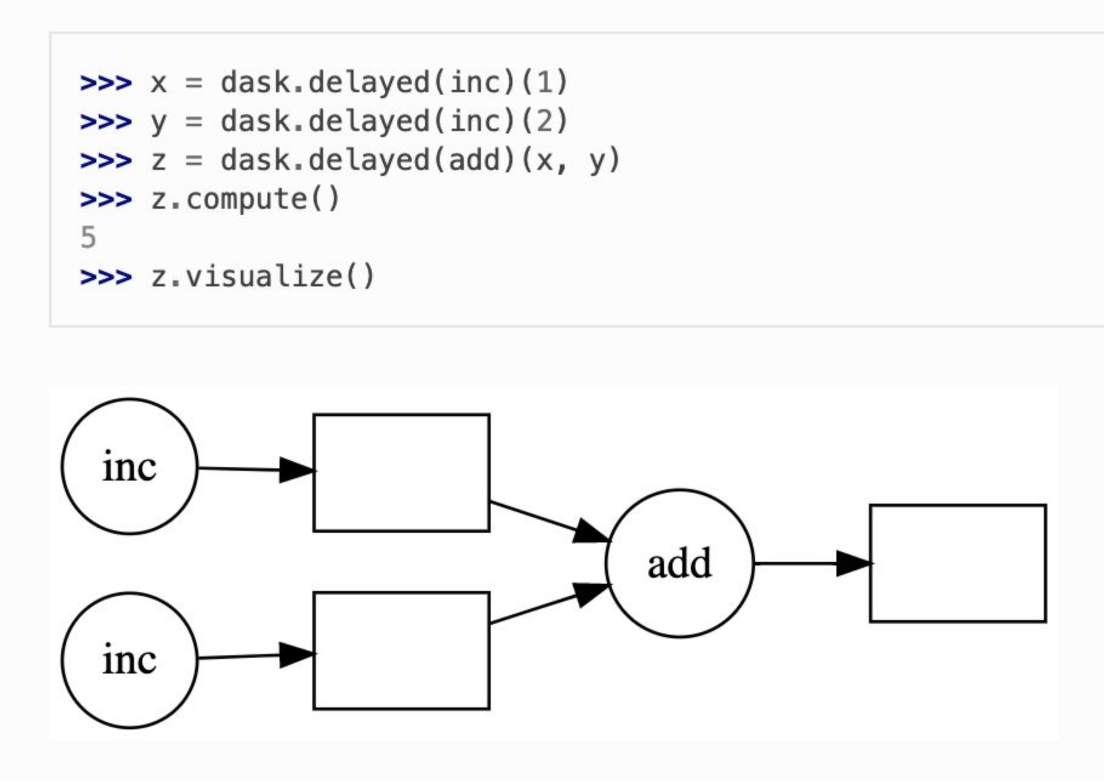


https://docs.dask.org/en/latest/array.html

https://docs.dask.org/en/latest/dataframe.html

What does it do?

Anything else!

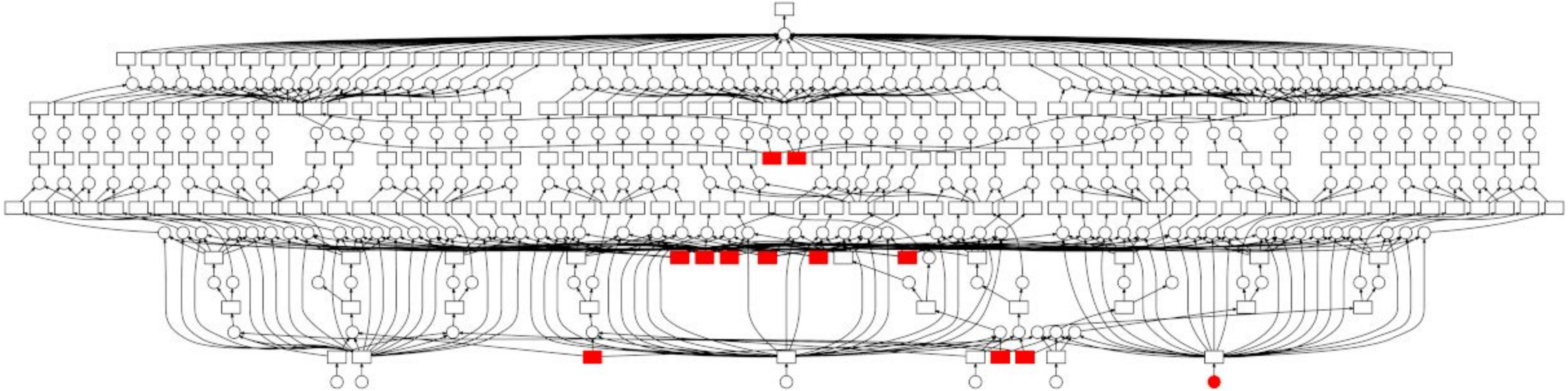


```
>>> x = dask.delayed(inc)(1)
>>> y = dask.delayed(inc)(2)
>>> z = dask.delayed(add)(x, y)
>>> z.compute()
5
>>> z.visualize()
```

https://docs.dask.org/en/latest/delayed.html

What does it do?

Anything else!



https://dask.org/

Dask

Why should I use it?

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

Dask

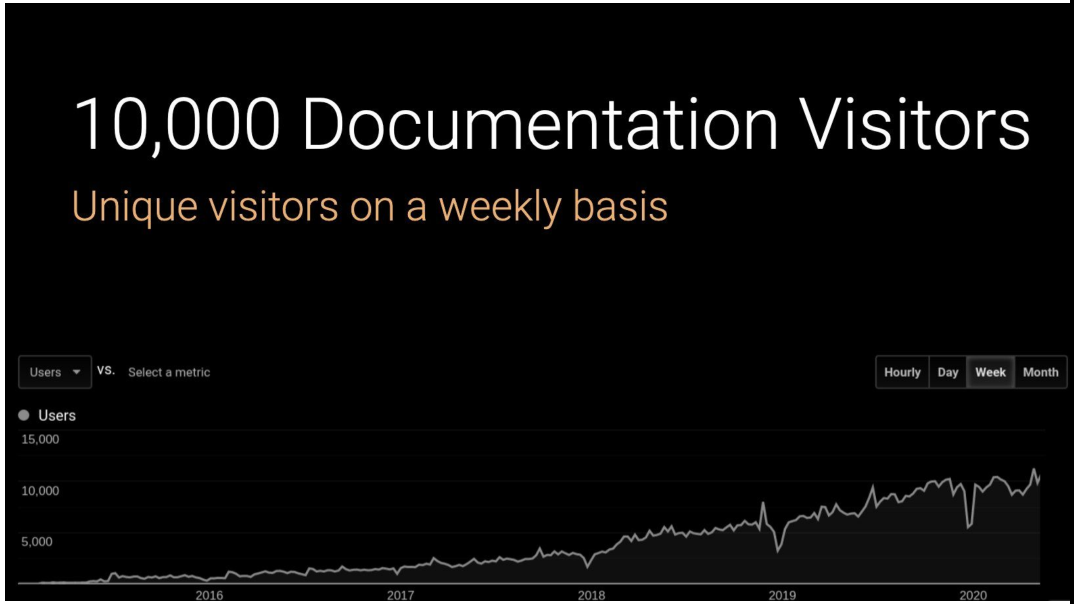
Why should I use it?

- Make your Python faster with a “pip install”
- 2 to 50 times faster than Spark*
- Will bring you fame and fortune**

*stay tuned **YMMV

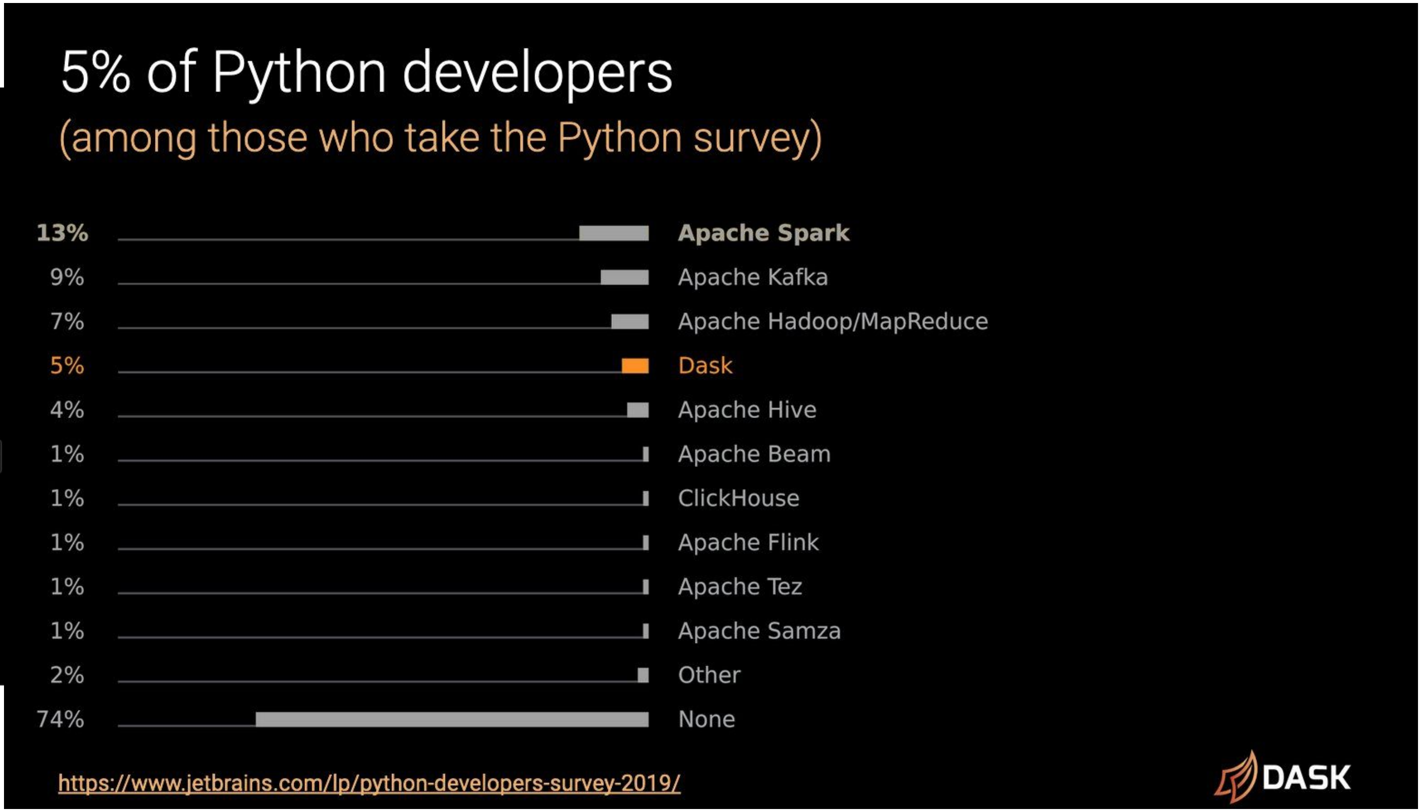
Dask

Why should I use it?



10,000 Documentation Visitors

Unique visitors on a weekly basis



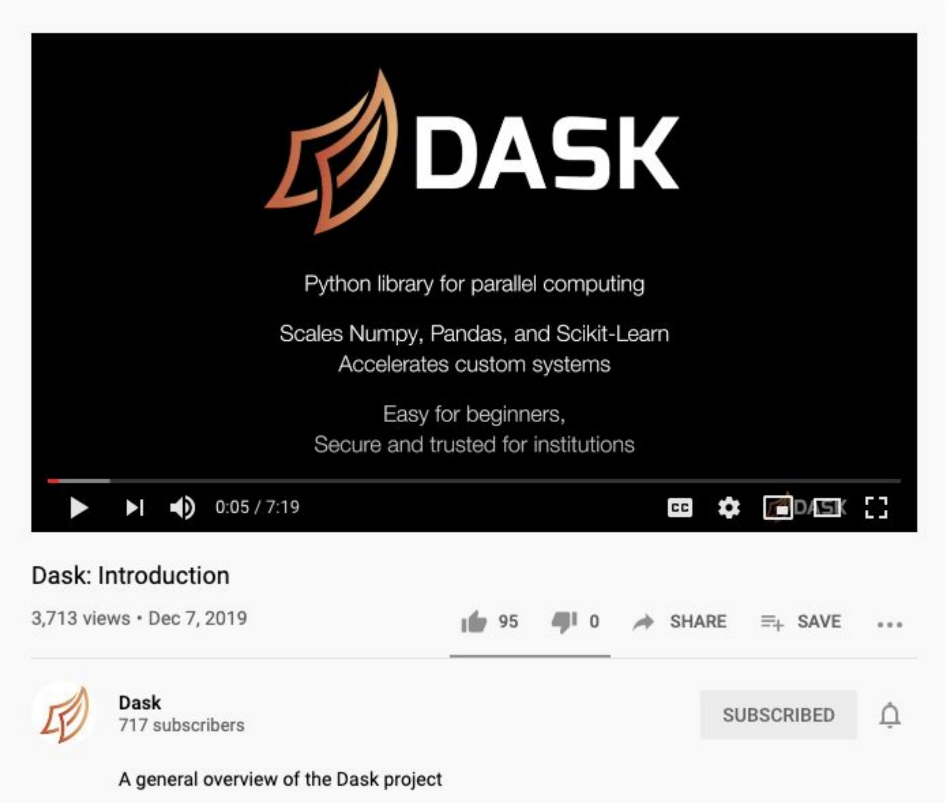
5% of Python developers

(among those who take the Python survey)

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

How do I get started?

Videos



DASK

Python library for parallel computing

Scales Numpy, Pandas, and Scikit-Learn
Accelerates custom systems

Easy for beginners,
Secure and trusted for institutions

Dask: Introduction

3,713 views • Dec 7, 2019

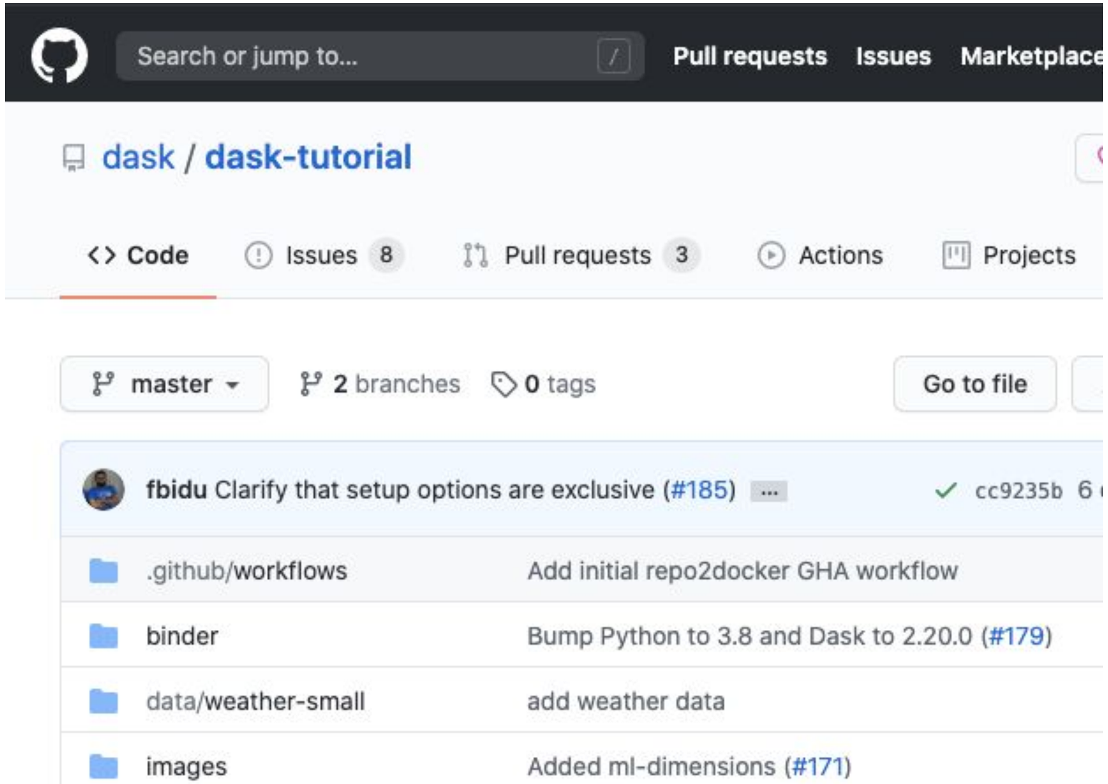
Dask
717 subscribers

A general overview of the Dask project

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

How do I get started?

Tutorial



dask / dask-tutorial

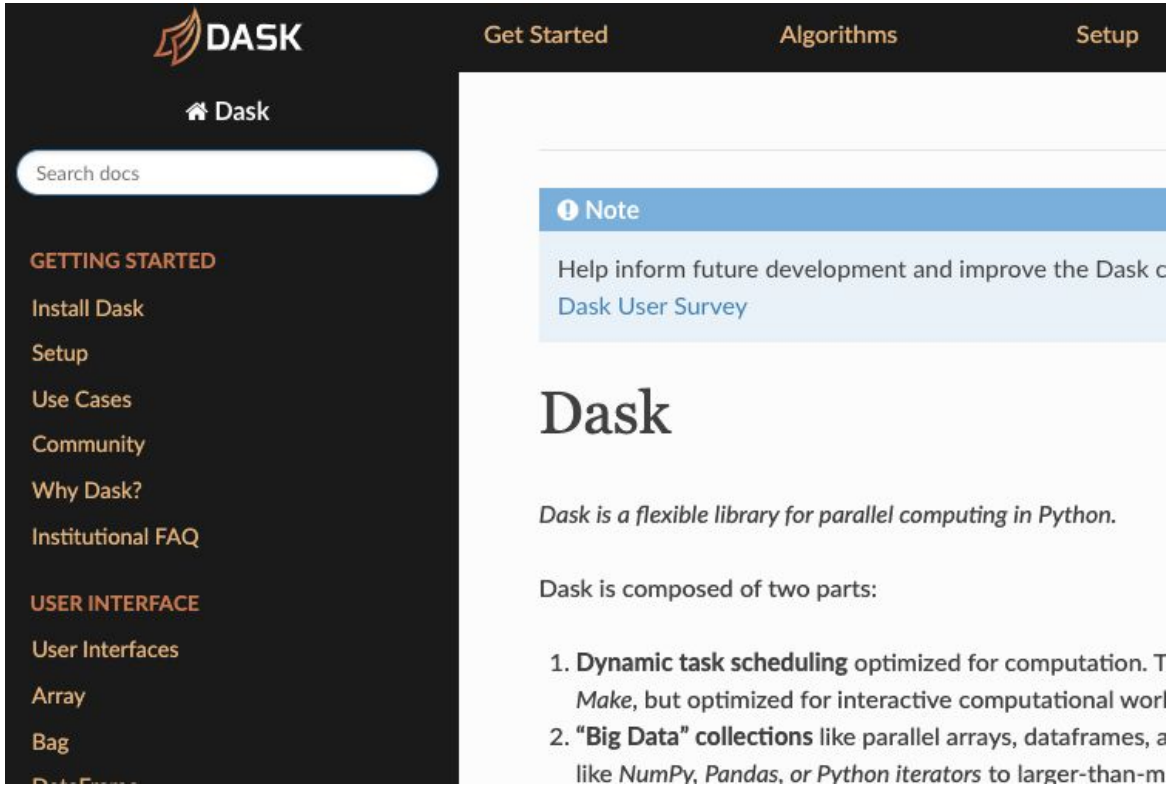
fbidu Clarify that setup options are exclusive (#185) ... ✓ cc9235b 6

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

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

How do I get started?

Docs



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)

https://www.saturncloud.io/s/should-i-use-dask/

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 Computing

What's coming next?

Exciting stuff!

- High level graph optimization
- Scheduler performance
- Chan Zuckerberg Initiative life science grant

Jacob Tomlinson
@_JacobTomlinson

I just finished giving this talk. Here's the tweet version. #SciPy2020 1/

Jacob Tomlinson @_JacobTomlinson · Jul 8
In 20 minutes I'll be giving the @dask_dev update on the #SciPy2020 Maintainers Track.

https://threadreaderapp.com/thread/1280885850914553856.html

Get involved!

Thriving community of open source contributors

- https://docs.dask.org/en/latest/develop.html
- Do you use Dask now?
    - Take the survey: https://dask.org/survey

Thank you!

aaron@saturncloud.io
https://rikturr.com
@rikturr

Saturn Cloud