Parallel and distributed data science with Dask

Aaron Richter
Deep Learning Adventures meetup
September 2020

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

Saturn Cloud

Hi!

Aaron Richter

Senior Data Scientist @ Saturn Cloud
Organizer @ PyData Miami

> I work to make data scientists faster and happier

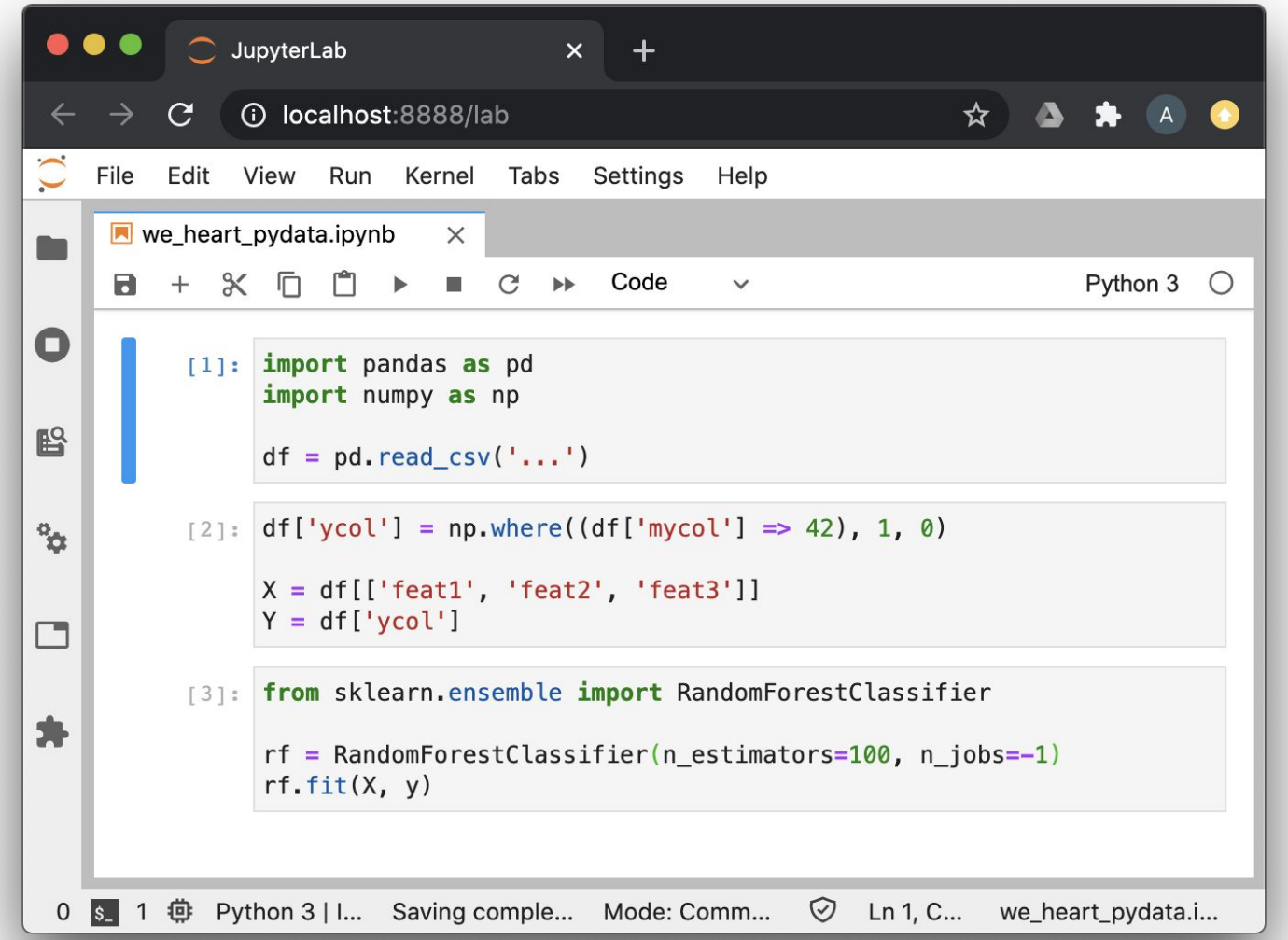
PhD in Machine Learning

aaron@saturncloud.io
rikturr.com
@rikturr

Data science with Python

Saturn Cloud

Data science with Python



Python with “Big Data”

Big data world

hadoop

APACHE DRILL

APACHE Spark™

cassandra

PyData world

scikit learn

pandas

NumPy

Python + big data!

RAY

DASK

vaex

MODIN

RAPIDS

DASK

Dask

- *Parallel computing for Python people*
- Anaconda, ~2015
- Built in Python; Python API
- Mature, scientific computing communities
- Low-level task library
- High-level libraries for DataFrames, arrays, ML
- Integrates with PyData ecosystem
- Runs on laptop, scales to clusters

https://dask.org/

DASK

Dask

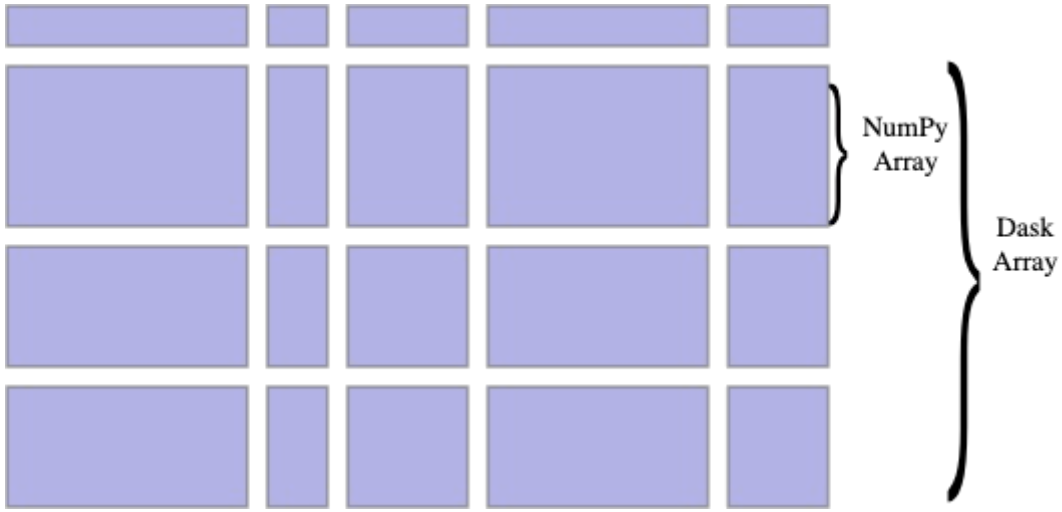
What does it do?

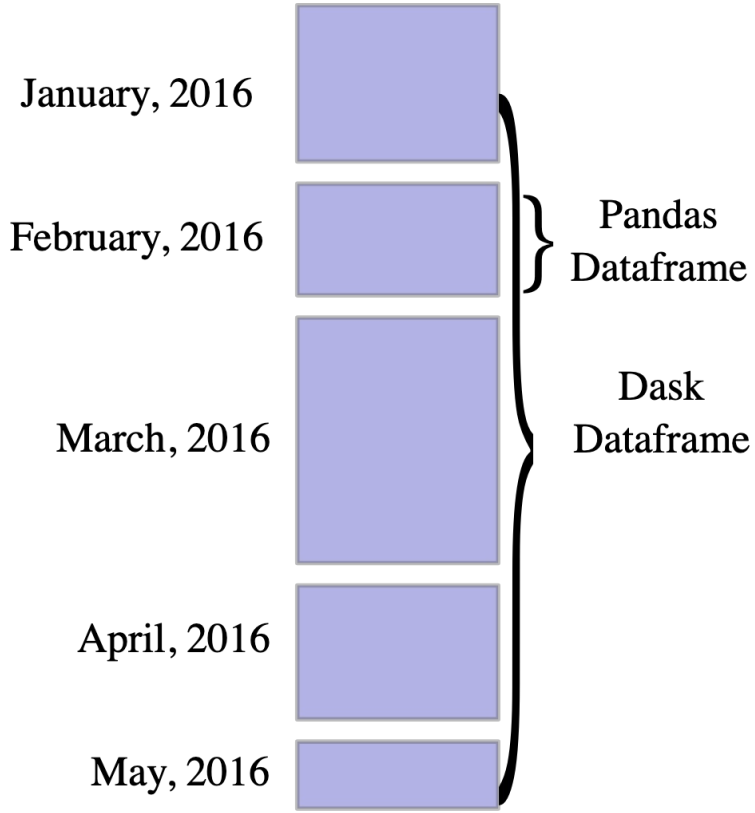
- Parallel machine learning (scikit)
- Parallel dataframes (pandas)
- Parallel arrays (numpy)
- Parallel anything else

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

What does it do?

Arrays and Dataframes





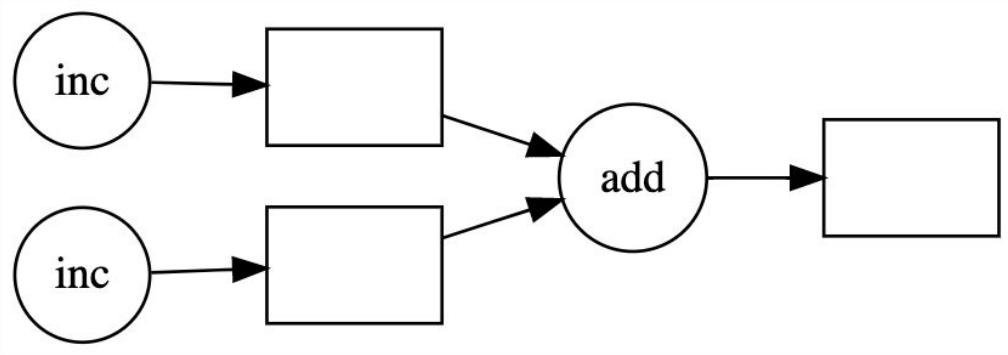
https://docs.dask.org/en/latest/array.html

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

What does it do?

Anything else!

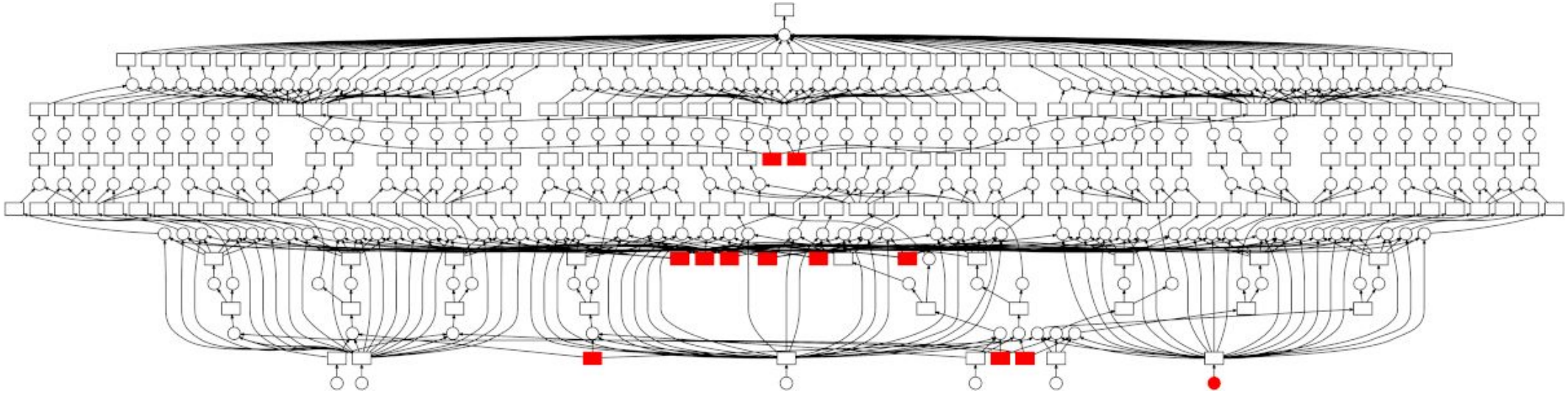
```
>>> x = dask.delayed(inc)(1)
>>> y = dask.delayed(inc)(2)
>>> z = dask.delayed(add)(x, y)
>>> z.compute()
5
>>> z.visualize()
```



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
- Pairs with RAPIDS for GPU acceleration
- Will bring you fame and fortune*

https://www.saturncloud.io/s/supercharging-hyperparameter-tuning-with-dask/

https://www.saturncloud.io/s/random-forest-on-gpus-2000x-faster-than-apache-spark/

*YMMV

RAPIDS

- *GPU accelerated data science*
- NVIDIA, ~2018
- Built in C++(CUDA), Python; Python API
- Large dev team, support from NVIDIA
- Native DataFrames, arrays, ML, graph, streaming, spatial
- Integrates with PyData ecosystem
- Scales to clusters with Dask integration

https://rapids.ai/

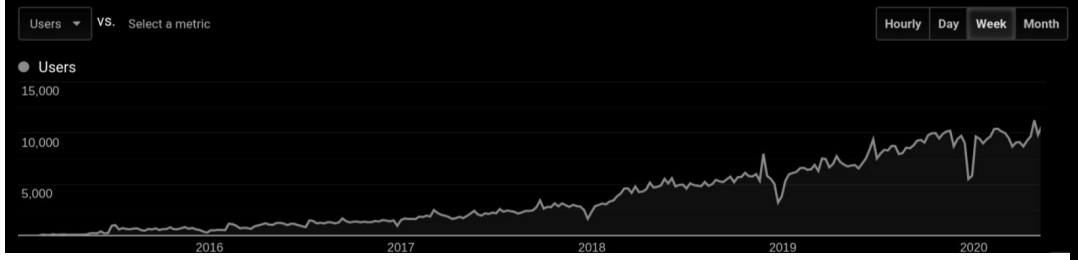
RAPIDS

Dask

Why should I use it?

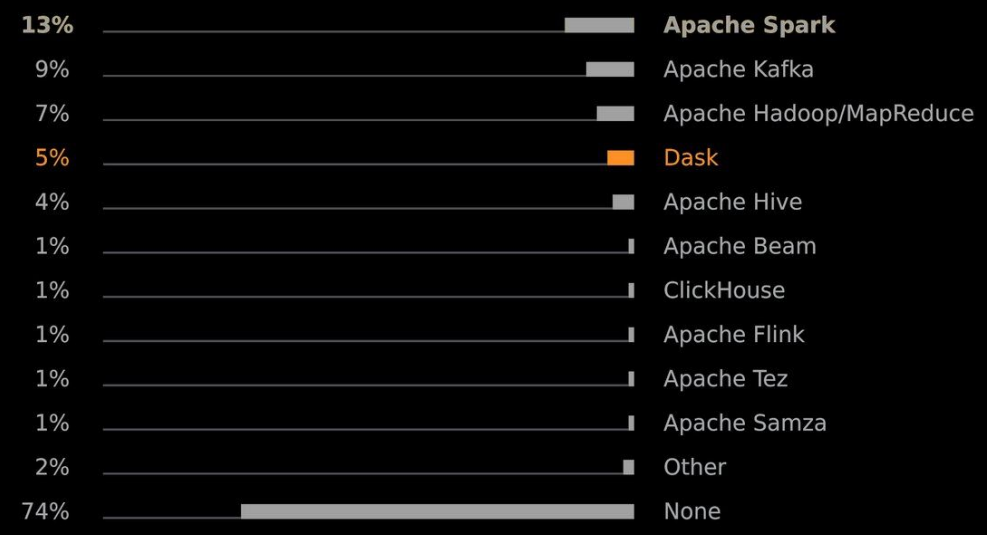
10,000 Documentation Visitors

Unique visitors on a weekly basis



5% of Python developers

(among those who take the Python survey)

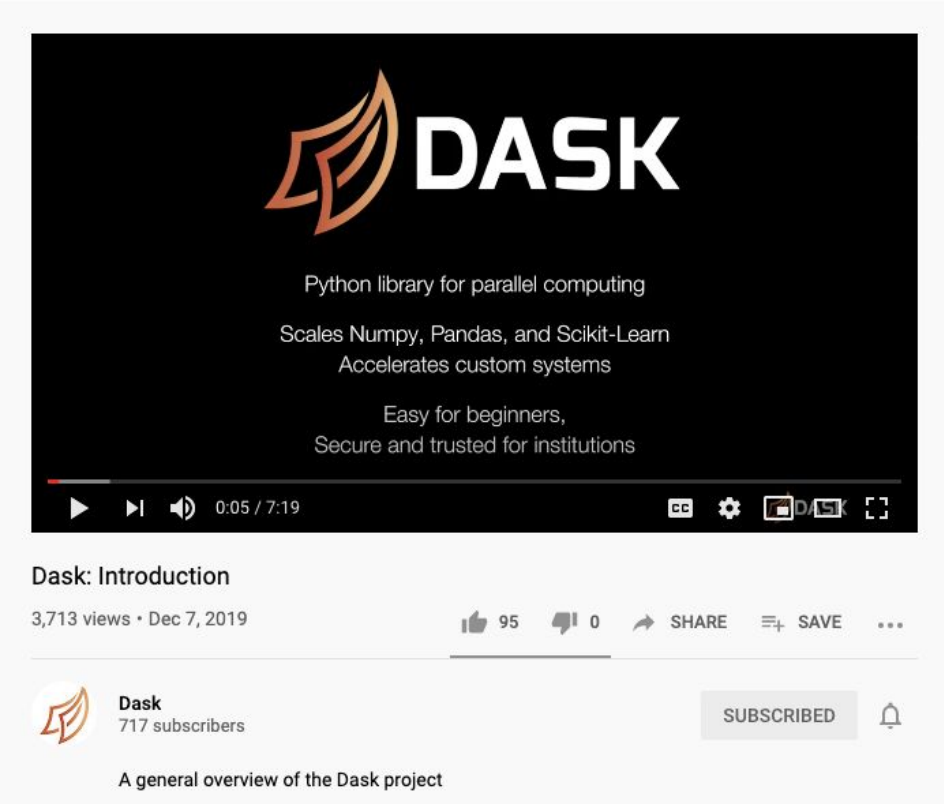


https://www.jetbrains.com/lp/python-developers-survey-2019/

DASK

How do I get started?

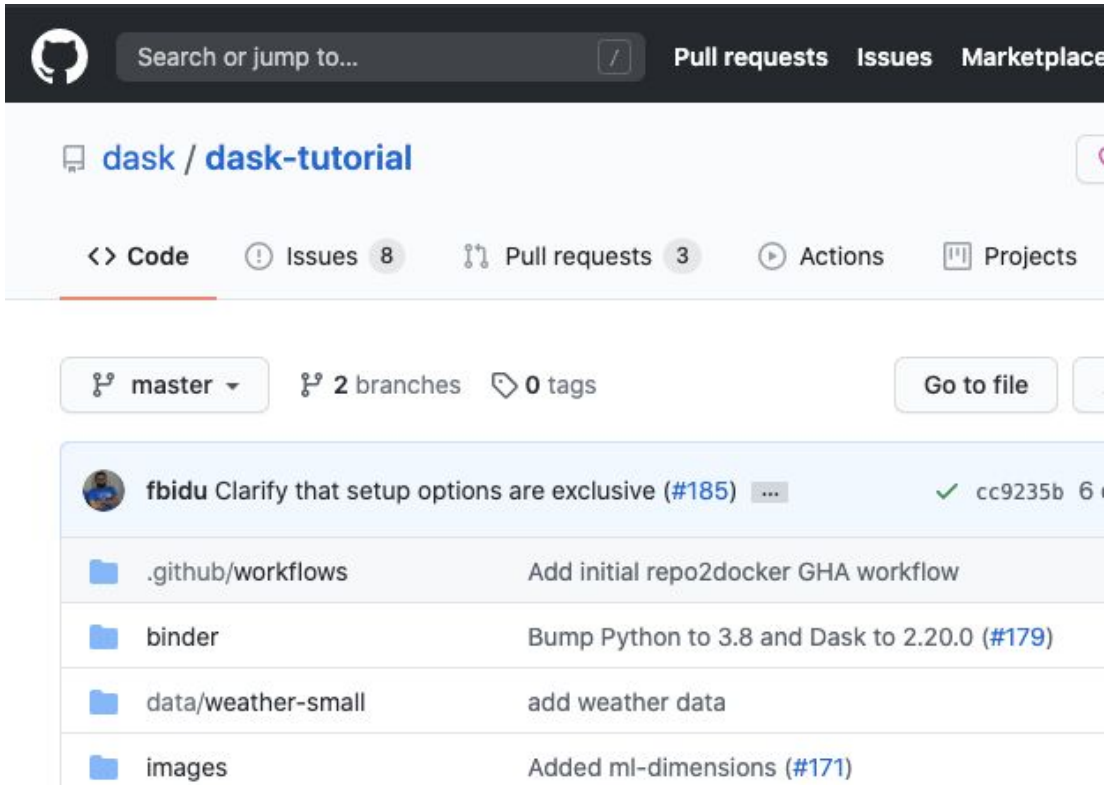
Videos



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

How do I get started?

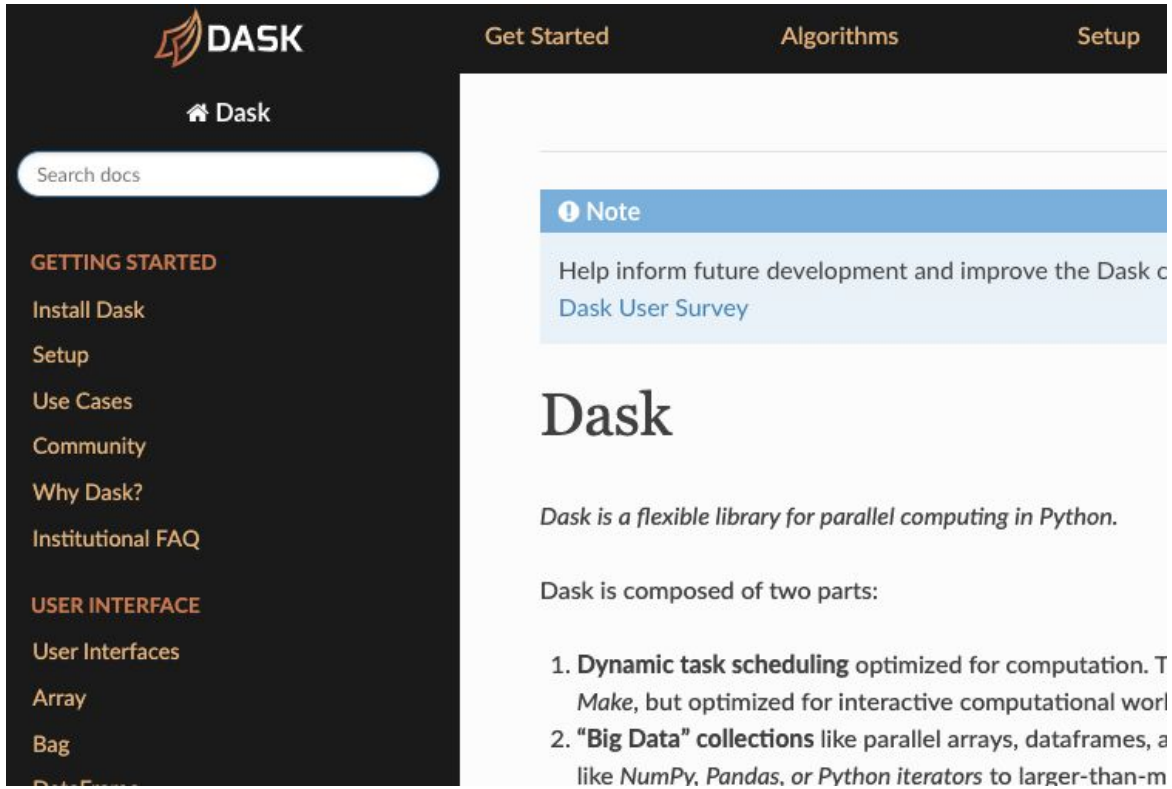
Tutorial



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

How do I get started?

Docs



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

How do I get started?

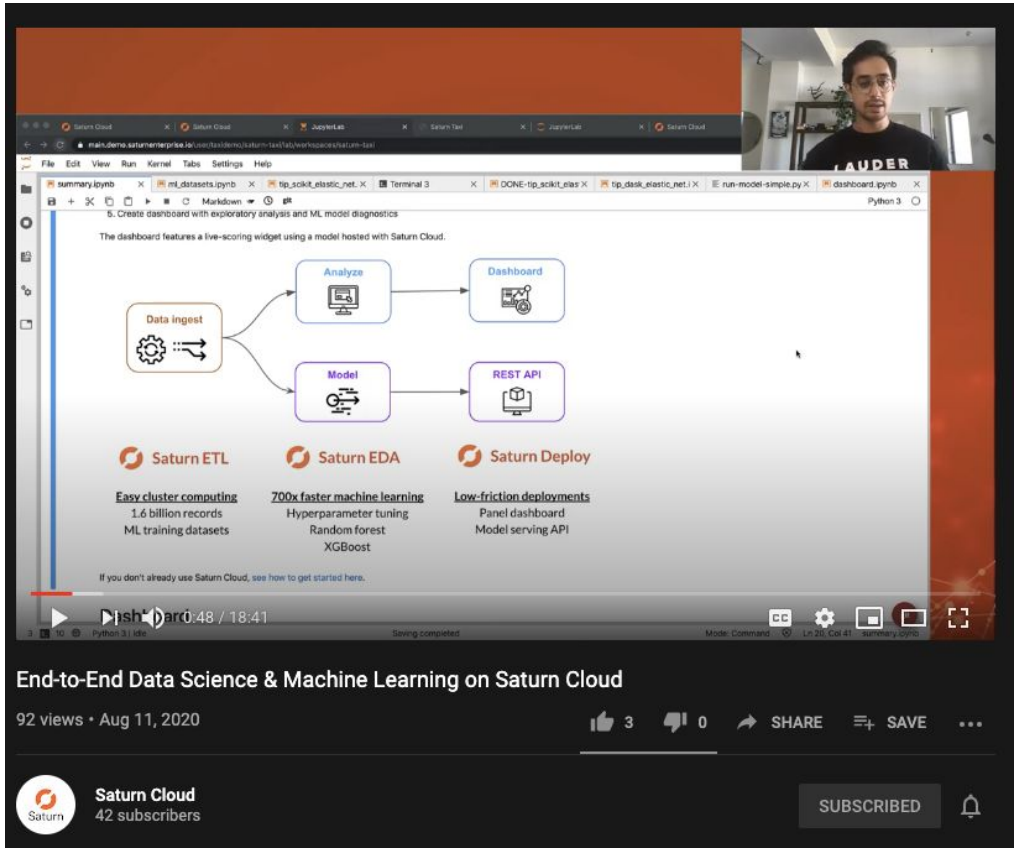
Key concepts

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

Code time!

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

Larger case study



https://youtu.be/SgXSIbB4Hik

https://github.com/saturncloud/saturn-cloud-examples/tree/main/taxi_demo

More ML with Dask

- Distributed inference/scoring: ParallelPostFit
- Tune deep learning models (via Skorch, SciKeras)
- Case study: The Future of Computer Vision with AI Pioneer Senseye

Unofficial guide to accelerating Python

- Use “traditional” PyData tools on your laptop until you can’t
- Then, use Dask *on your laptop*
  - RAPIDS if you have a GPU
- Then, get a bigger machine in the cloud
- Then, use a Dask cluster in the cloud
- ***Use the best tool for each workload!***

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

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

Get involved!

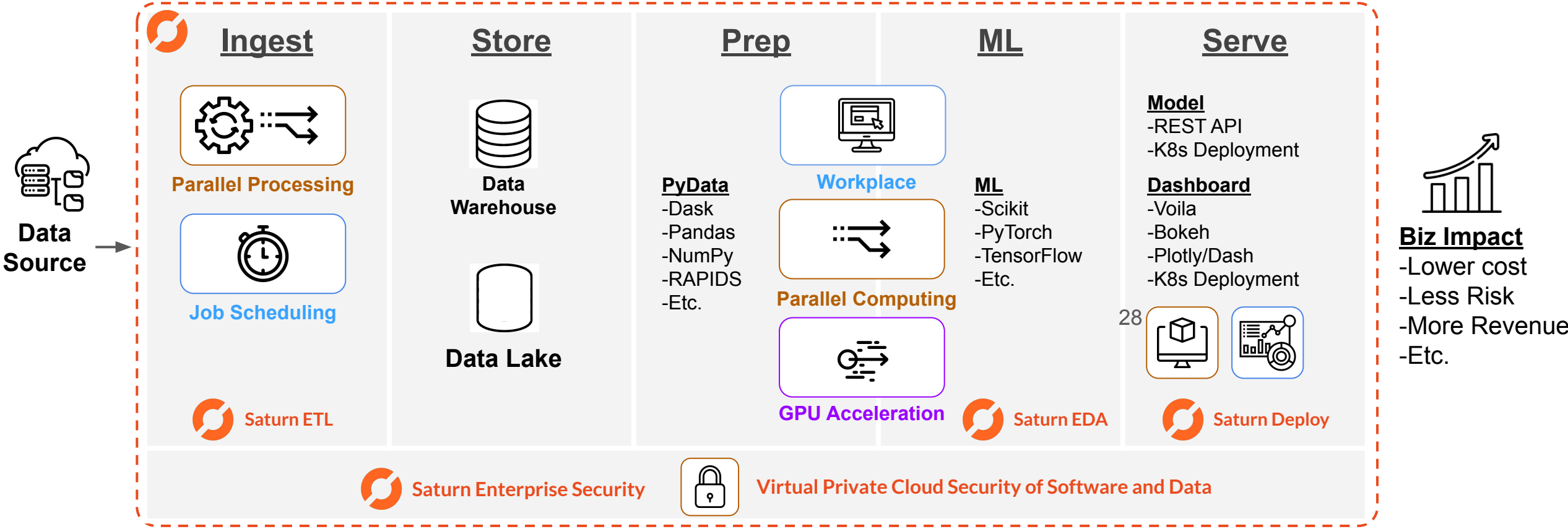
Thriving community of open source contributors

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

Saturn Cloud

Saturn Cloud

Saturn enables end-to-end DS and ML in Python




BETTER LAPTOP
BIG
MACHINE ON AWS
BIG MACHINE
ON SATURN CLOUD
DASK CLUSTER
ON SATURN CLOUD
GPU DASK CLUSTER
ON SATURN CLOUD
imgflip.com

Thank you!

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

Saturn Cloud