

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

A screenshot of the JupyterLab web interface in a browser. The browser address bar shows "localhost:8888/lab". The JupyterLab window has a menu bar with "File", "Edit", "View", "Run", "Kernel", "Tabs", "Settings", and "Help". Below the menu is a toolbar with icons for file operations and execution. The main area shows a code editor with a file named "we_heart_pydata.ipynb". The code is written in Python 3 and consists of three cells. The first cell imports pandas and numpy, and reads a CSV file. The second cell uses numpy.where to create a new column 'ycol' based on the value of 'mycol', and extracts features 'feat1', 'feat2', and 'feat3' into a matrix X. The third cell imports RandomForestClassifier from sklearn.ensemble and creates a classifier object with 100 estimators, fitting it to the data X and y.

```
[1]: import pandas as pd
import numpy as np

df = pd.read_csv('...')

[2]: df['ycol'] = np.where((df['mycol'] >= 42), 1, 0)

X = df[['feat1', 'feat2', 'feat3']]
Y = df['ycol']

[3]: from sklearn.ensemble import RandomForestClassifier

rf = RandomForestClassifier(n_estimators=100, n_jobs=-1)
rf.fit(X, y)
```



Python with “Big Data”

Big data world



PyData world



Python + big data!





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/>



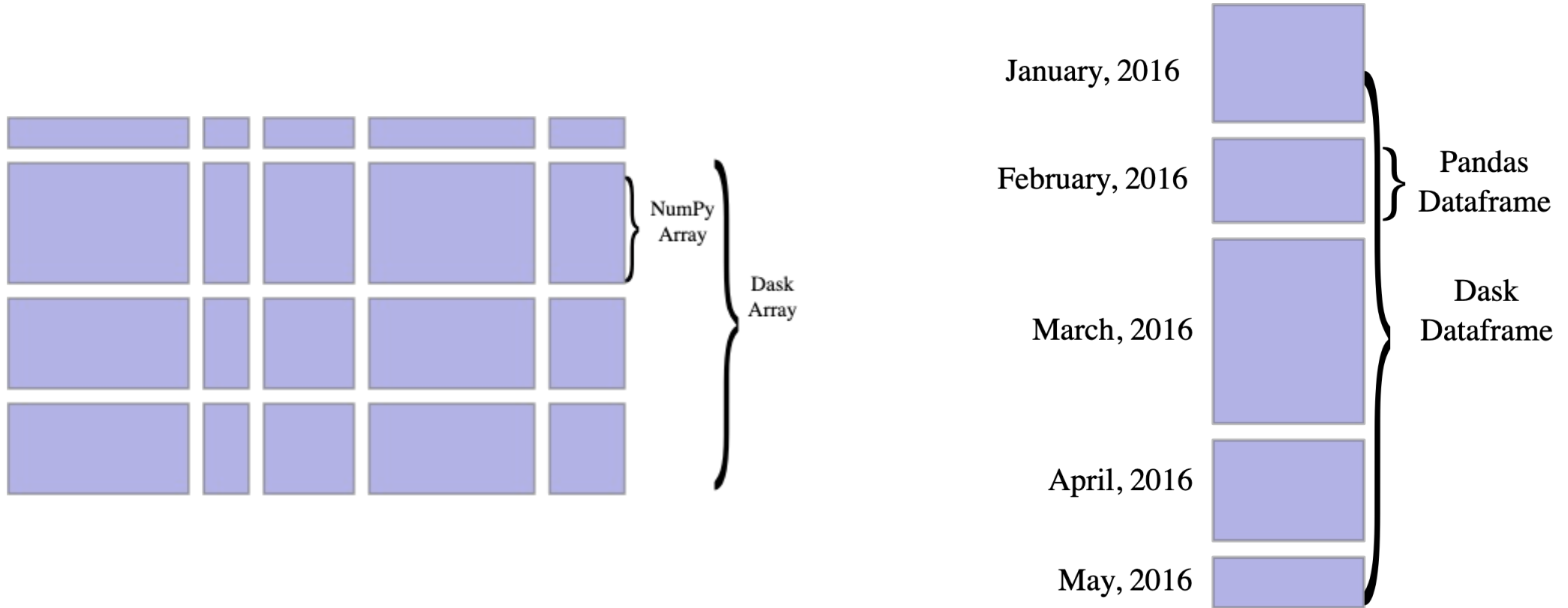
Dask

What does it do?

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

What does it do?

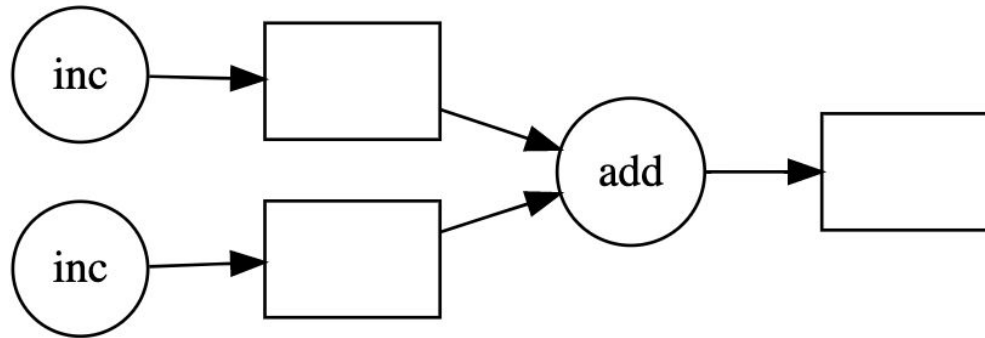
Arrays and Dataframes



What does it do?

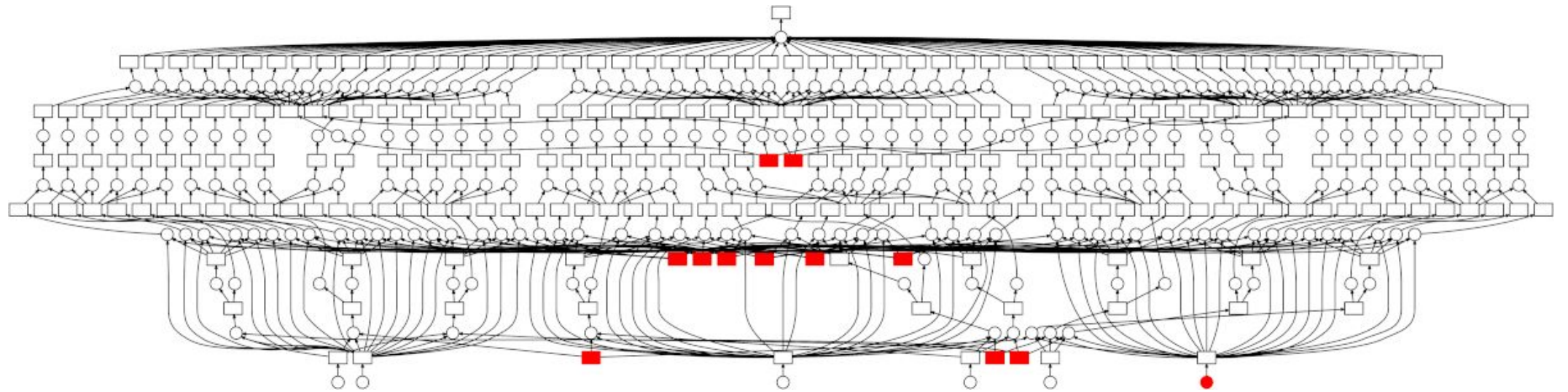
Anything else!

```
>>> 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?

- 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/>

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/>

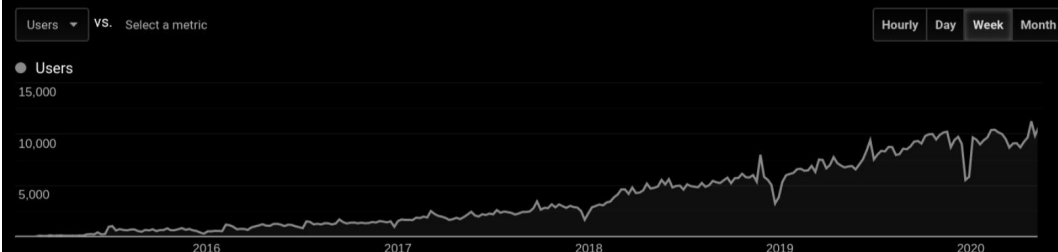
RAPIDS

Dask

Why should I use it?

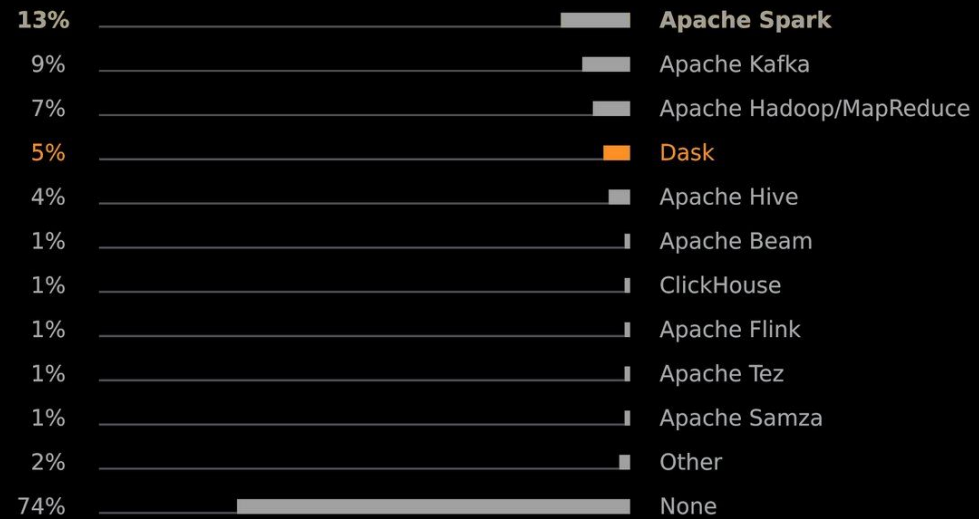
10,000 Documentation Visitors

Unique visitors on a weekly basis



5% of Python developers

(among those who take the Python survey)

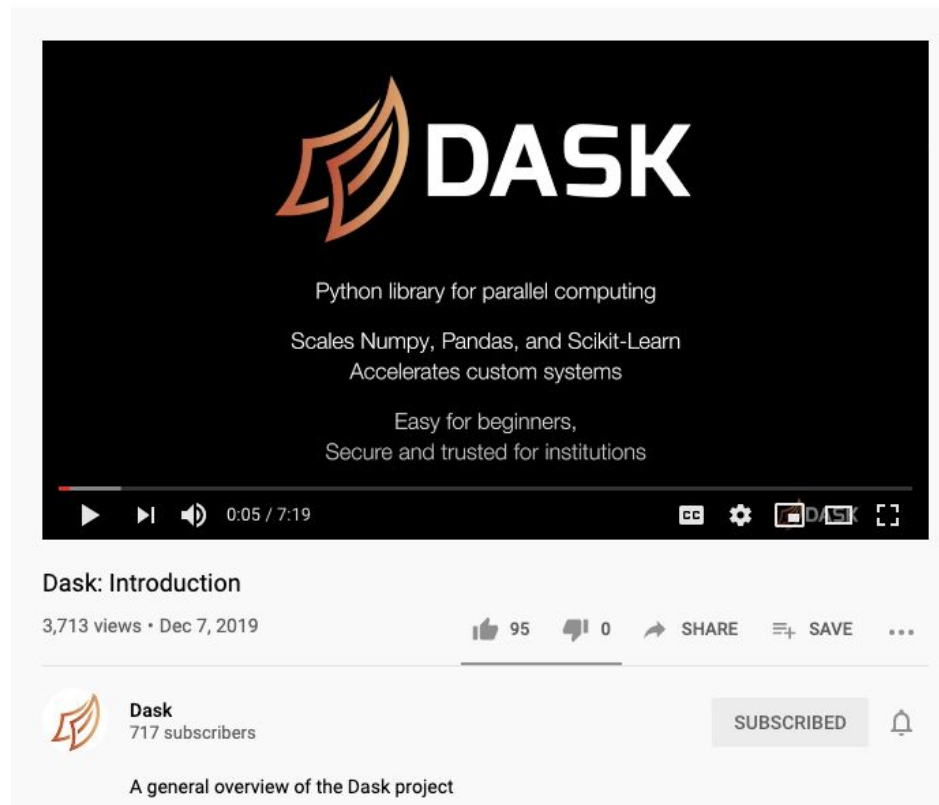


<https://www.jetbrains.com/lp/python-developers-survey-2019/>



How do I get started?

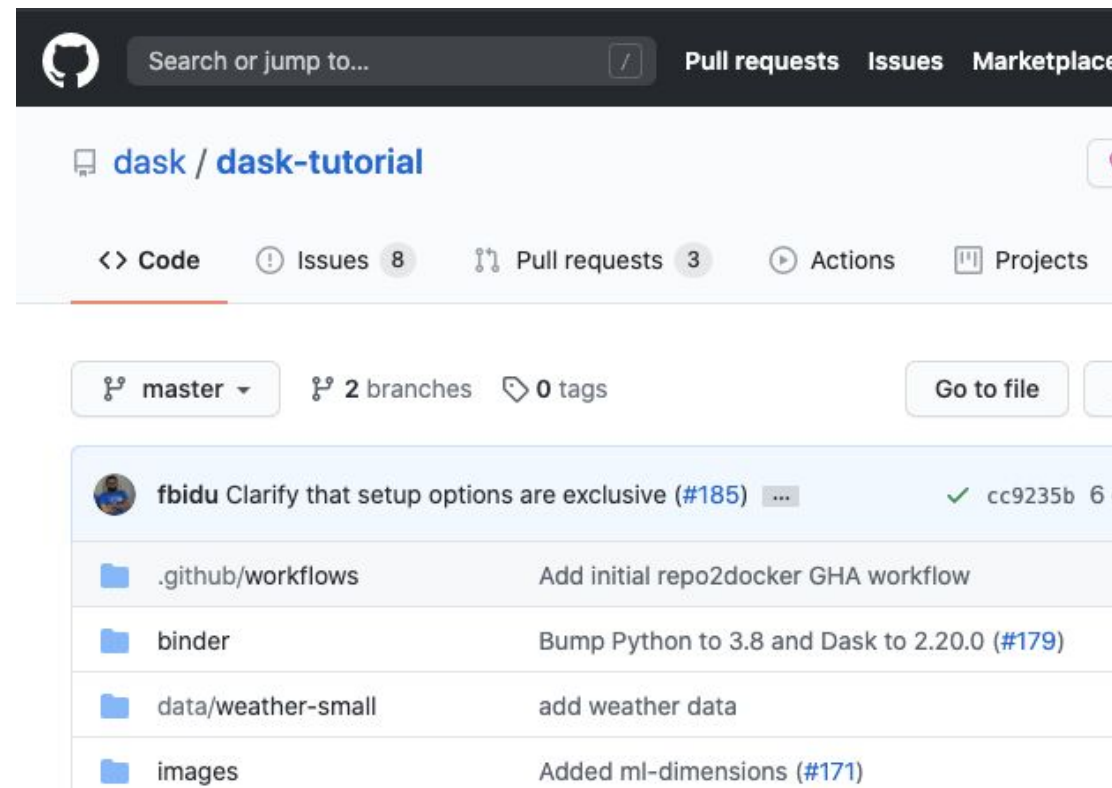
Videos



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

How do I get started?

Tutorial



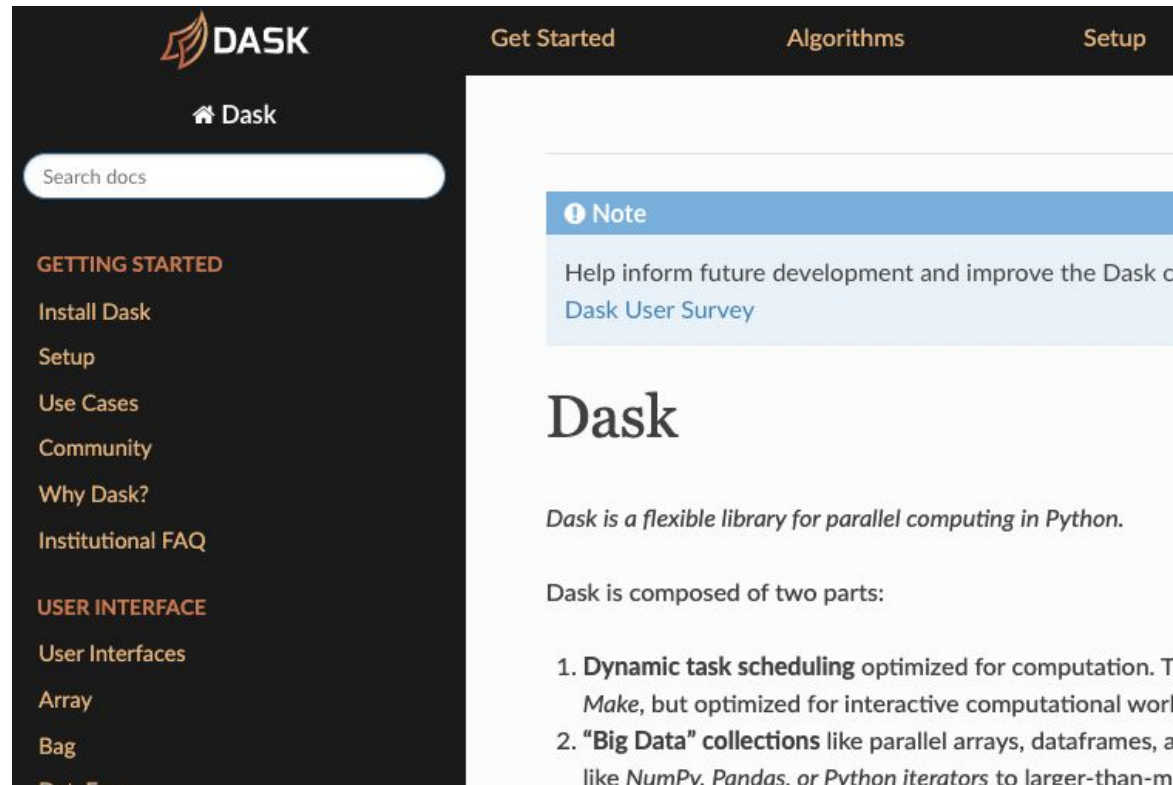
The screenshot shows the GitHub interface for the `dask / dask-tutorial` repository. At the top, there's a search bar and navigation links for Pull requests, Issues, and Marketplace. Below the repository name, there are tabs for Code, Issues (8), Pull requests (3), Actions, and Projects. The 'Code' tab 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>

Larger case study

The screenshot shows a YouTube video player with a video titled "End-to-End Data Science & Machine Learning on Saturn Cloud". The video content displays a web interface for Saturn Cloud, which is a platform for data science and machine learning. The interface features a workflow diagram with the following components:

- Data ingest** (represented by a gear icon)
- Analyze** (represented by a magnifying glass icon)
- Dashboard** (represented by a bar chart icon)
- Model** (represented by a brain icon)
- REST API** (represented by a server icon)

Arrows indicate the flow from Data ingest to both Analyze and Model, and from both Analyze and Model to the Dashboard and REST API respectively.

Below the workflow diagram, there are three sections highlighting Saturn's capabilities:

- Saturn ETL**: Easy cluster computing, 1.6 billion records, ML training datasets.
- Saturn EDA**: 200x faster machine learning, Hyperparameter tuning, Random forest, XGBoost.
- Saturn Deploy**: Low-friction deployments, Panel dashboard, Model serving API.

The video player interface includes a progress bar at the bottom showing the video is at 0:48 / 18:41. The video has 92 views and was uploaded on Aug 11, 2020. The channel is "Saturn Cloud" with 42 subscribers. The video is marked as "SUBSCRIBED".

<https://youtu.be/SgXS1bB4Hik>

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

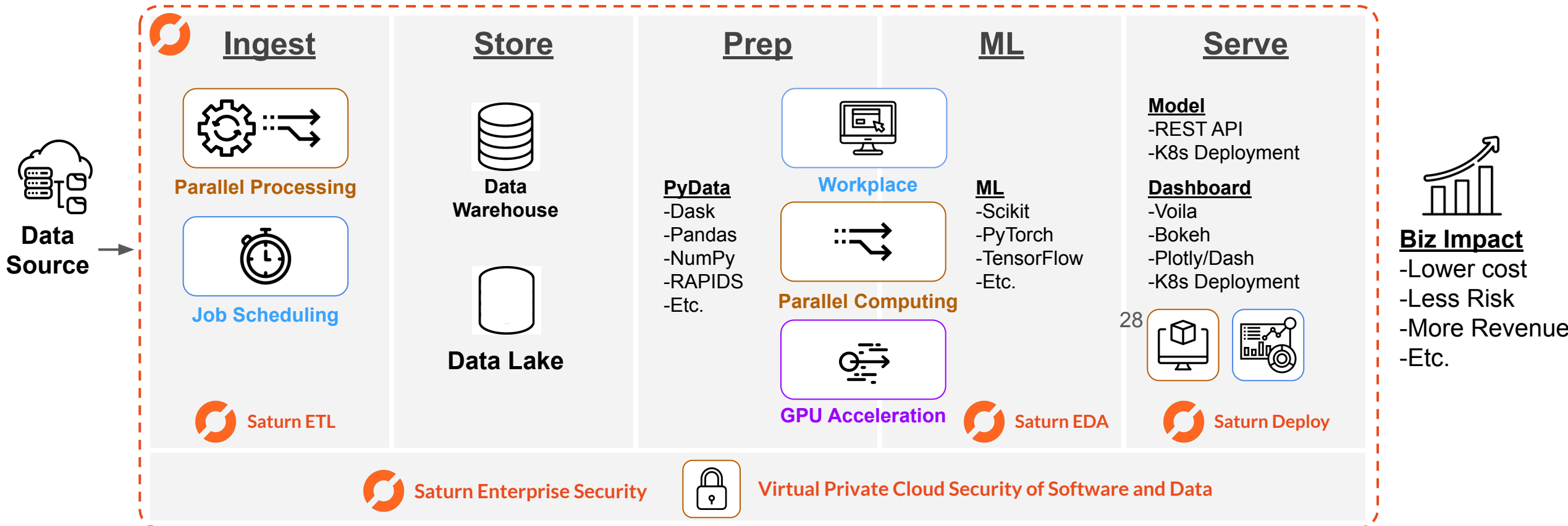
- <https://docs.dask.org/en/latest/develop.html>
- Do you use Dask now?
 - Take the survey: <https://dask.org/survey>



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**





Thank you!

aaron@saturncloud.io

<https://rikturr.com>



@rikturr