High performance Jupyter: Faster workloads with Dask and RAPIDS

JupyterCon 2020

https://github.com/rikturr/high-performance-jupyter

Hi!

Aaron Richter, PhD

Senior Data Scientist @ Saturn Cloud
Organizer @ PyData Miami

> *I work to make data scientists faster and happier*

aaron@saturncloud.io
rikturr.com
@rikturr

Saturn Cloud

Bringing together the fastest hardware + OSS

DASK

- Pythonic parallelism
- Rapidly scale PyData

RAPIDS

- Multi-GPU computing
- The future of HPC

SaturnCloud

PREFECT

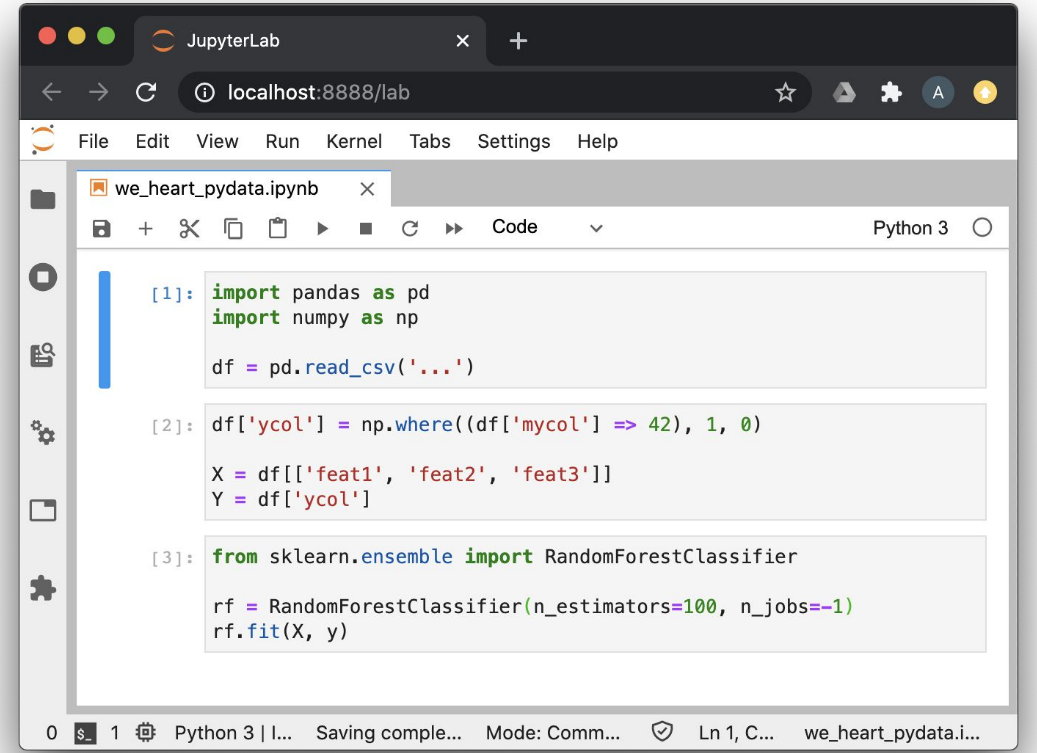
- Workflow orchestration
- Flow insight and mgmt

kubernetes

- Fast setup
- Enterprise secure

Data science with Jupyter

Data science with Jupyter


JupyterLab
localhost:8888/lab
File Edit View Run Kernel Tabs Settings Help
we_heart_pydata.ipynb
Code
Python 3
0 1 Python 3 | I... Saving comple... Mode: Comm... Ln 1, C... we_heart_pydata.i...

```
[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)
```

High performance data science 🚀

DASK

RAPIDS

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

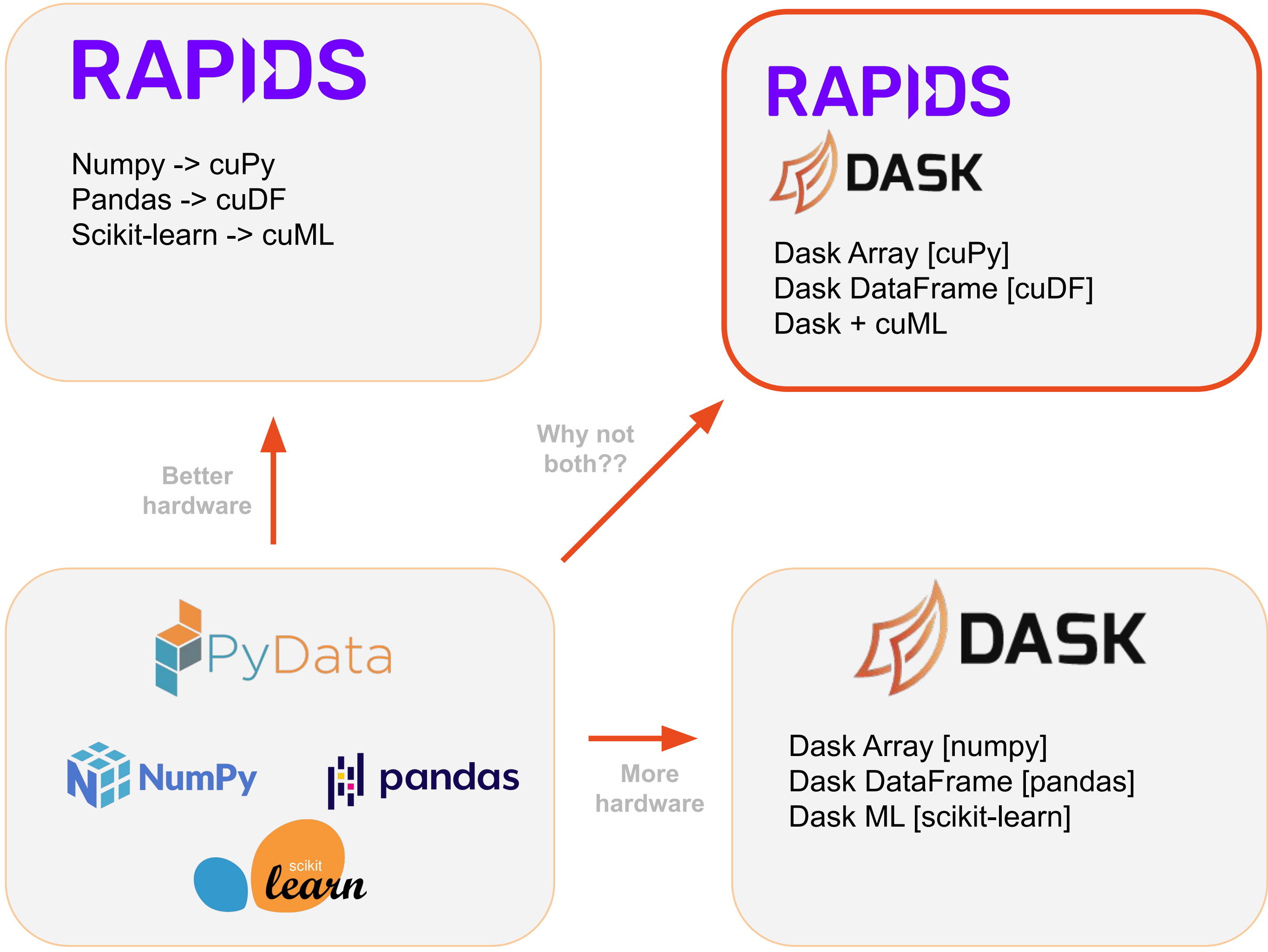
https://dask.org/

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



Single GPU
Data <16GB

100x faster than CPU

RAPIDS

Numpy -> cuPy
Pandas -> cuDF
Scikit-learn -> cuML

RAPIDS
DASK

Dask Array [cuPy]
Dask DataFrame [cuDF]
Dask + cuML

1000s GPUs
Data TB++

100x faster + “Big data”

Single machine
Data <200GB

Slow computation

PyData

NumPy
pandas
scikit learn

DASK

Dask Array [numpy]
Dask DataFrame [pandas]
Dask ML [scikit-learn]

1000s machines
Data TB++

Horizontal scaling

Code time!

https://github.com/rikturr/high-performance-jupyter