Scalable Data Science with Dask

COILED

Hi!
I’m Pavithra :)

- Evangelist at Coiled
- FOSS contributor - Bokeh & Wikimedia
- CSE Student, India
- @pavithraes on Twitter
- hi@pavithraes.me

What's your experience with Scalable Data Science?

On a scale of 1 - 5

1: *What is scalable compute?*

5: *I understand the challenges and have a solution that works for me*

We'll talk about

- Big data - What is it?
- Parallel and distributed computing
- Dask for scaling data science
- Coiled - Dask on the cloud

Big data

What is big data?

- Doesn’t fit on your local machine
- Traditional tools and methods fail

What is big data?

- Doesn’t fit on your local machine
- Traditional tools and methods fail

Characteristics:

- Volume
- Velocity
- Variety
- Veracity

COILED

Scalable compute

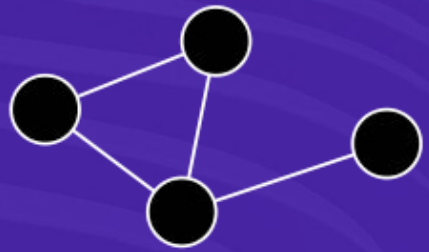
Parallel computing

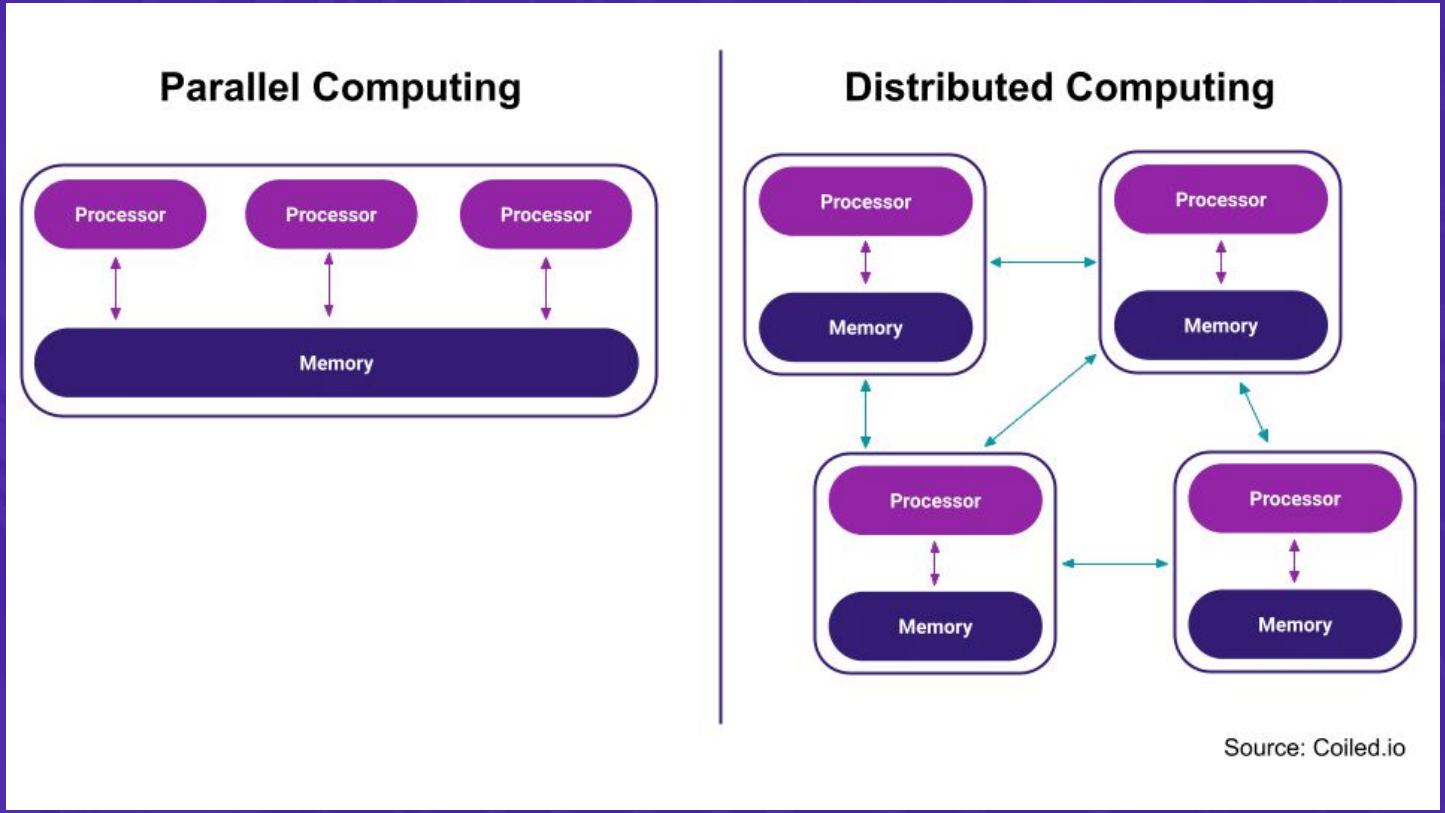
- Working in parallel
- Use all CPU cores
- Multiple processes and shared memory

COILED

Distributed computing

- Using groups of machines
- Each machine has processors and memory

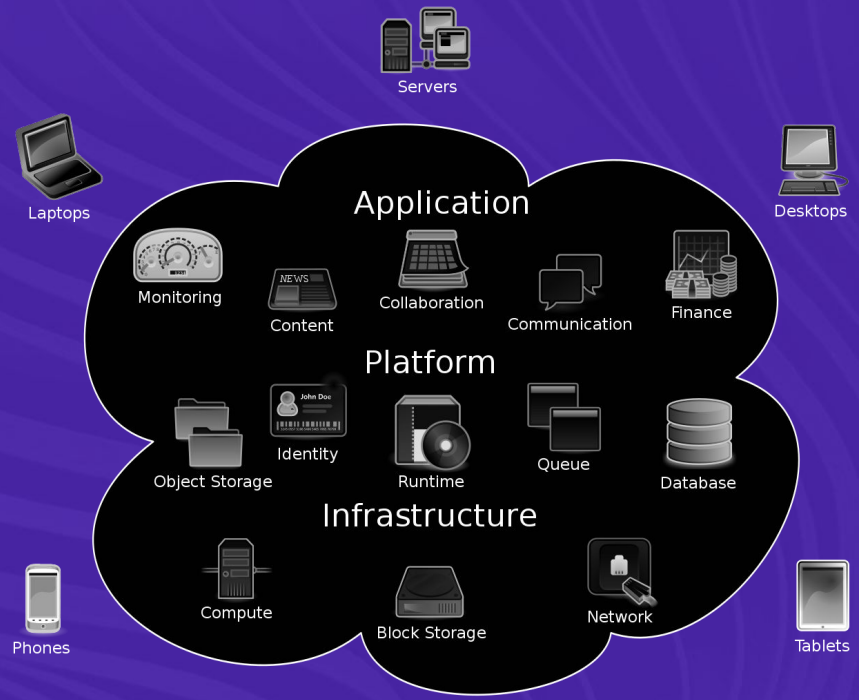




Source: [Coiled.io]

Cloud computing

- Using cloud resources
- AWS, Azure, GCP
- Lots of storage and computational power



Source: Wikimedia Commons

COILED

Dask

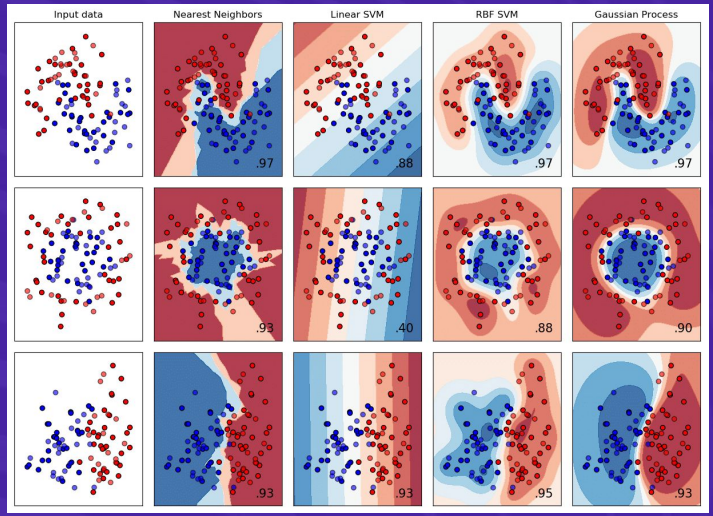
COILED

Dask

Library for parallel and distributed computing in Python

Dask

Makes it easy to scale-up your workflows to use all cores in your local machine

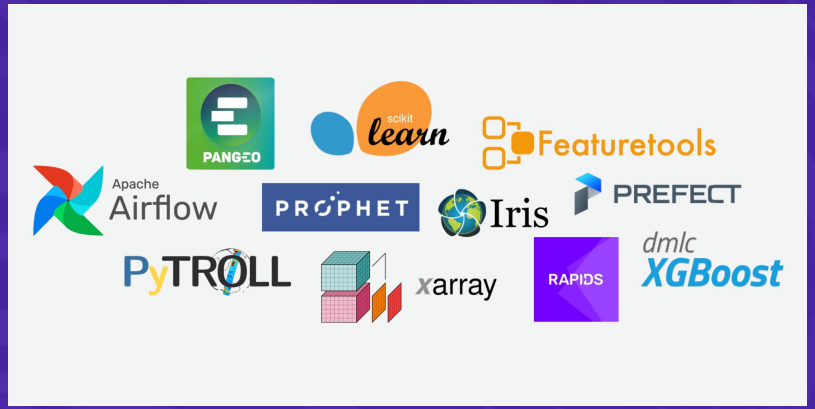


COILED

Dask

Provides a distributed computing framework

Powers tools like RAPIDS, Airflow, PyTorch, and more!



COILED

Dask features

Familiar API

Resembles normal pandas, NumPy, scikit-learn code

```
import pandas as pd
df = pd.read_csv("data_taxi/yellow_tripdata_2019-01.csv")
df.groupby("passenger_count").tip_amount.mean()
```

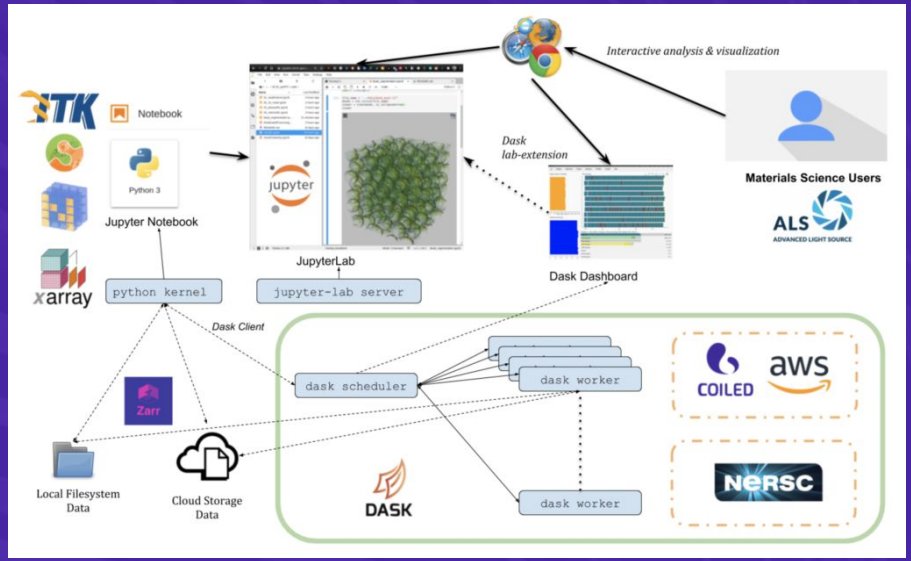
```
import dask.dataframe as dd
df = dd.read_csv("data_taxi/yellow_tripdata_2019-*.csv")
mean_amount = df.groupby("passenger_count").tip_amount.mean()
mean_amount.compute()
```

COILED

Dask features

Flexible

Local machine, on-prem, on the cloud, anywhere.



3D microstructure interactive image analysis and visualization system architecture.
Source: Article presented at Super Computing 2020

Dask users in retail

Walmart

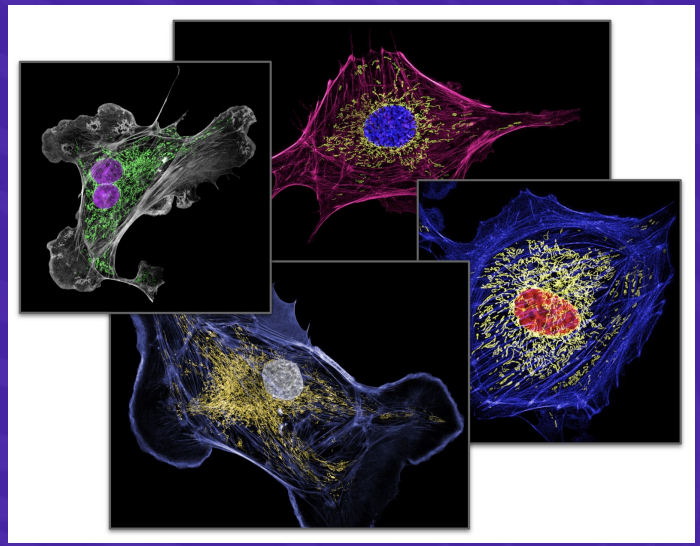
- demand forecasting
- 500M+ store combinations
- 100x speedup from RAPIDS and Dask

COILED

Dask users in life science

High resolution, 4-dimensional, cellular imagery

- Harvard Medical School
- Howard Hughes Medical Institute
- Chan Zuckerberg Initiative
- UC Berkeley Advanced Bioimaging Center



COILED

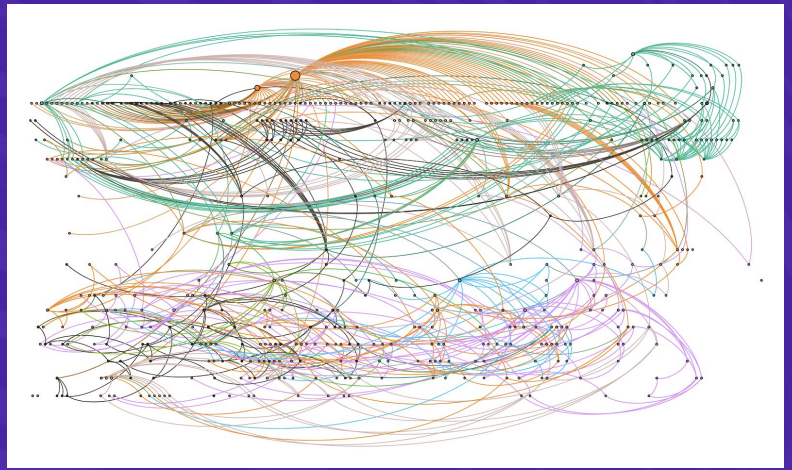
Dask users in finance

Capital One

- ETL and ML pipeline speedup

Barclays

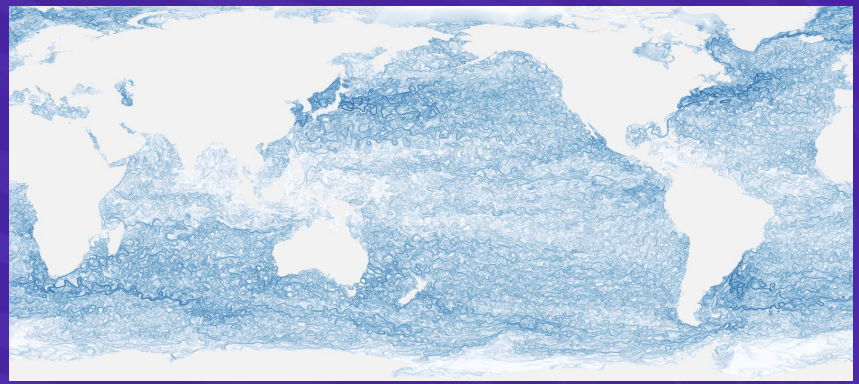
- Financial system modelling



COILED

Dask users in Geo

Farallon Institute , Los Alamos National Labs



- Climate Science
- Energy
- Hydrology
- Meteorology
- Satellite Imaging

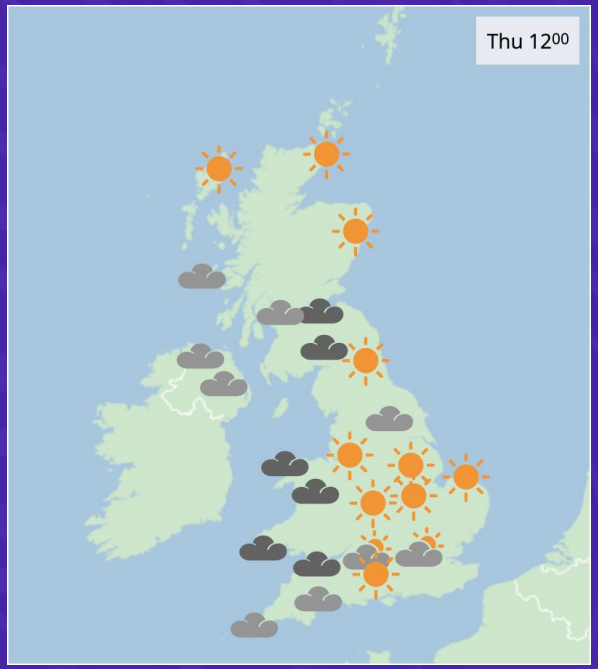
COILED

Dask users

Many many more!

NASA

Software Libraries



COILED

Shout out to Matthew Rocklin, and the entire Dask community for this material!

Coiled

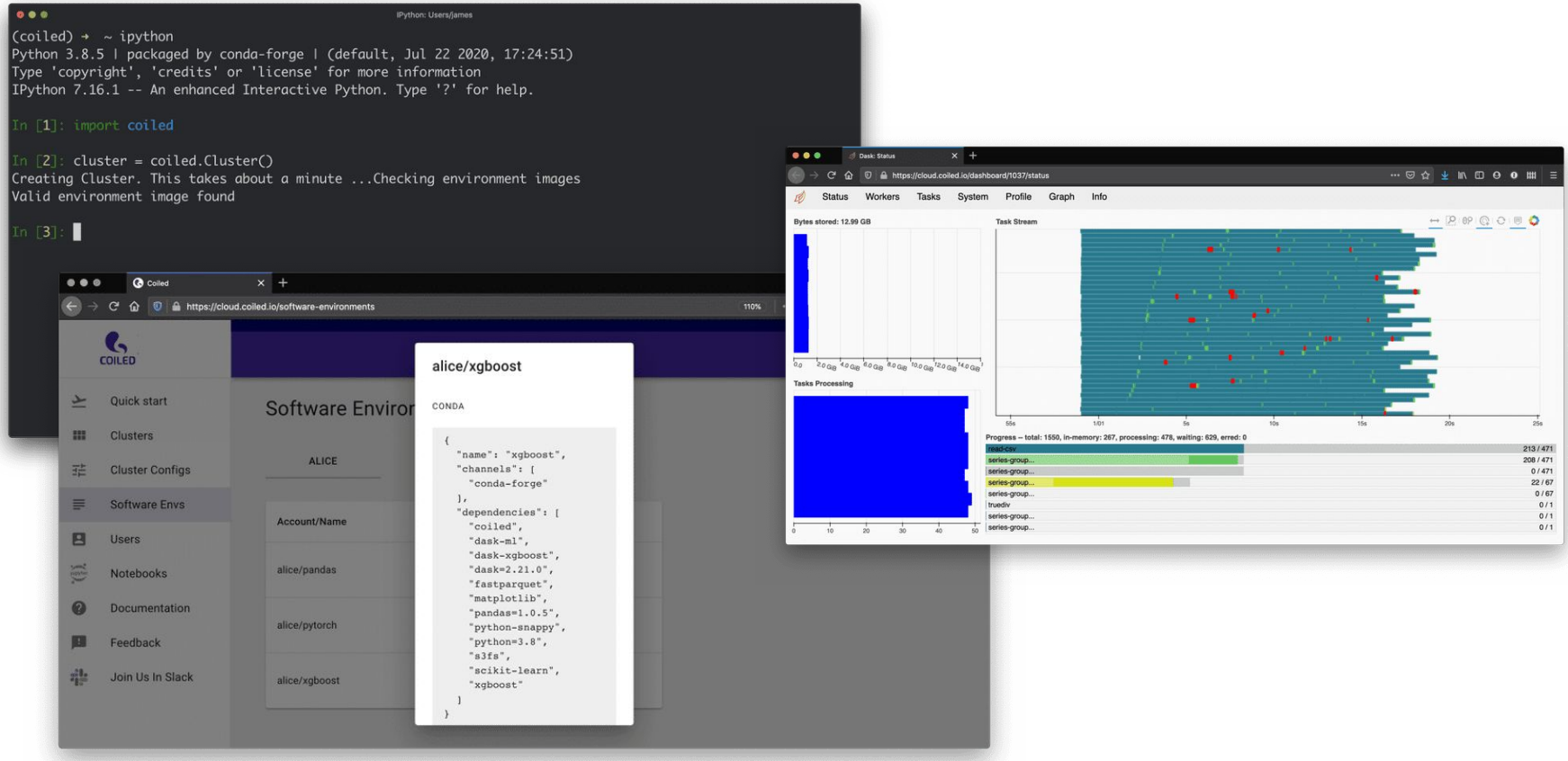
COILED

Built by Dask maintainers, contributors, and enthusiasts.

Open source culture is at the heart of Coiled.

Scalable computing
has some challenges

- Security concerns
- Managing software environments
- Cost optimization


IPython: Users/james
(coiled) ➜ ~ ipython
Python 3.8.5 | packaged by conda-forge | (default, Jul 22 2020, 17:24:51)
Type 'copyright', 'credits' or 'license' for more information
IPython 7.16.1 -- An enhanced Interactive Python. Type '?' for help.
In [1]: import coiled
In [2]: cluster = coiled.Cluster()
Creating Cluster. This takes about a minute ...Checking environment images
Valid environment image found
In [3]:
Coiled
https://cloud.coiled.io/software-environments
110%
COILED
Quick start
Clusters
Cluster Configs
Software Envs
Users
Notebooks
Documentation
Feedback
Join Us In Slack
Software Enviror
ALICE
Account/Name
alice/pandas
alice/pytorch
alice/xgboost
alice/xgboost
CONDA
{
"name": "xgboost",
"channels": [
"conda-forge"
],
"dependencies": [
"coiled",
"dask-ml",
"dask-xgboost",
"dask=2.21.0",
"fastparquet",
"matplotlib",
"pandas=1.0.5",
"python-snappy",
"python=3.8",
"s3fs",
"scikit-learn",
"xgboost"
]
}
Dask: Status
https://cloud.coiled.io/dashboard/1037/status
Status
Workers
Tasks
System
Profile
Graph
Info
Bytes stored: 12.99 GB
Task Stream
Tasks Processing
Progress -- total: 1550, in-memory: 267, processing: 478, waiting: 629, erred: 0
read-csv
213 / 471
series-group...
208 / 471
series-group...
0 / 471
series-group...
22 / 67
series-group...
0 / 67
truediv
0 / 1
series-group...
0 / 1
series-group...
0 / 1

Coiled tackles these challenges for you.

welcome.coiled.io

Thank you!

Slides and notebook at:
bit.ly/pyladies-berlin-dask

COILED