PyData

From your laptop to all resource that you need

@PyDataVenice #PyDataGlobal #2023

Alessandra Bilardi

Data & Automation Specialist

- Coderdojo member
- PyData Venezia member
- AWS User Group Venezia member

alessandra.bilardi@gmail.com

@abilardi

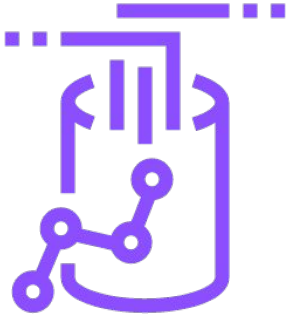
bilardi

Alessandra Bilardi

Data & Automation Specialist

- Coderdojo member
- PyData Venezia member
- AWS User Group Venezia member

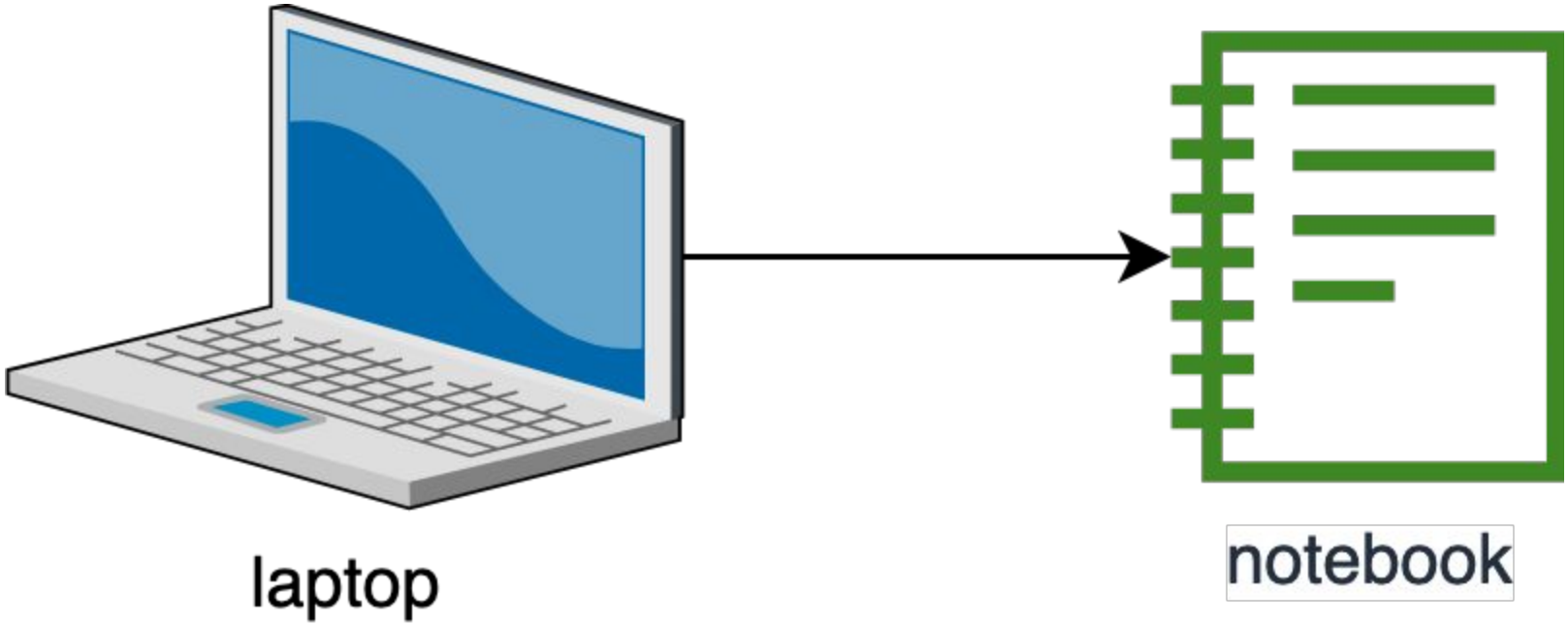
alessandra.bilardi@gmail.com
@abilardi
bilardi



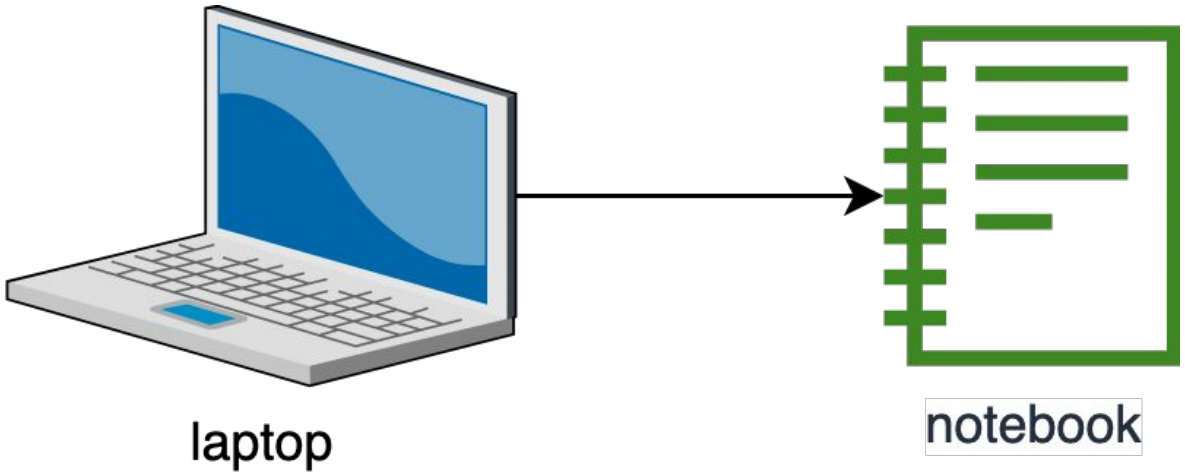
Agenda

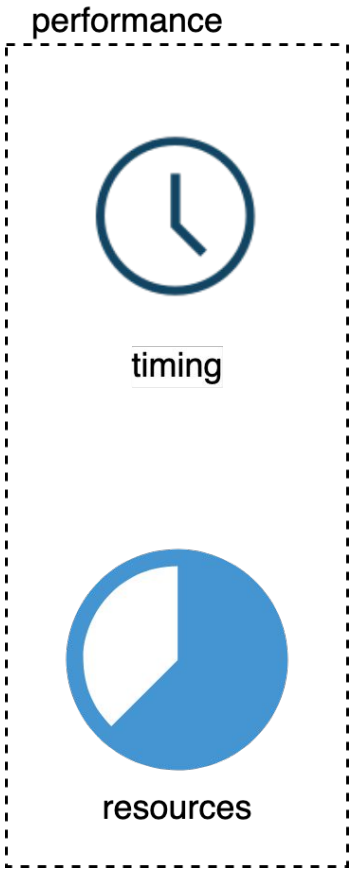
Goal
Solution
Cloud
Demo
Best practices

Goal

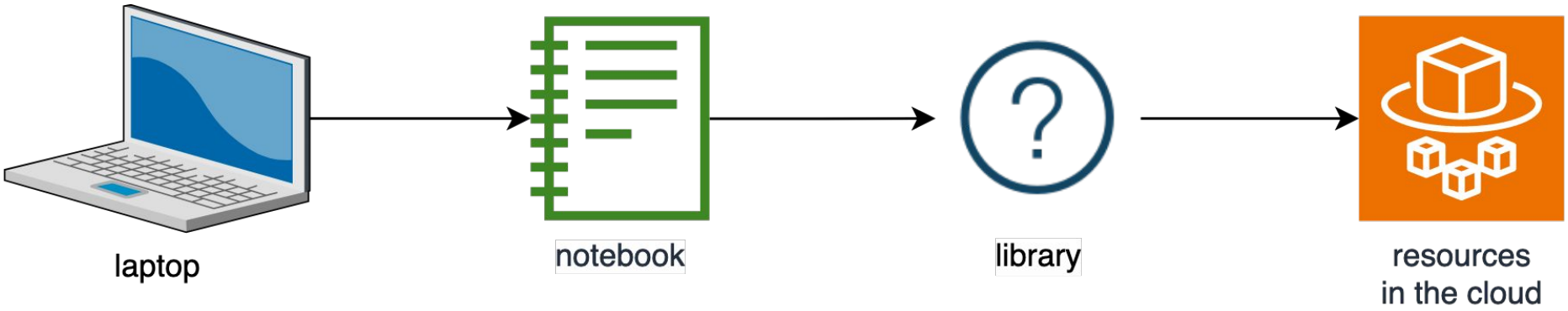


Goal





Solution

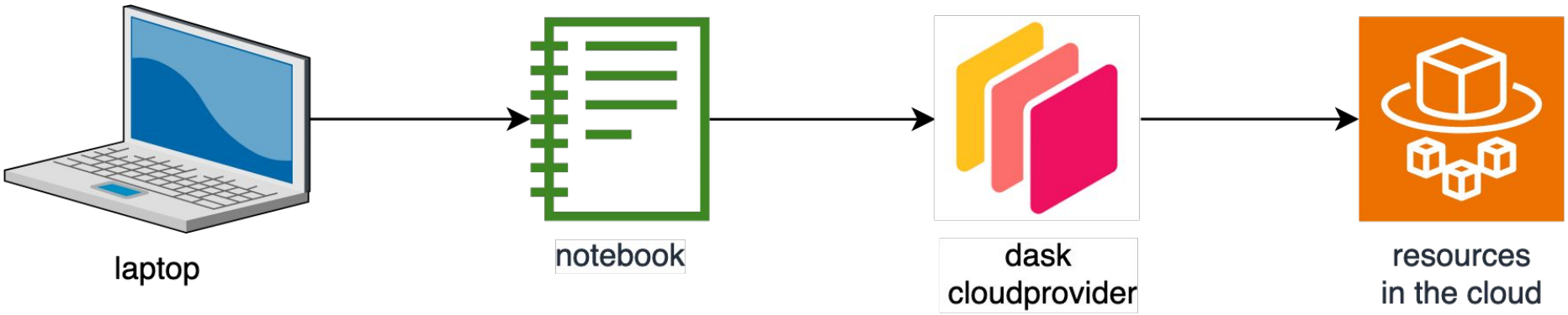


may

open source

with you

Solution



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

PyData

Dask

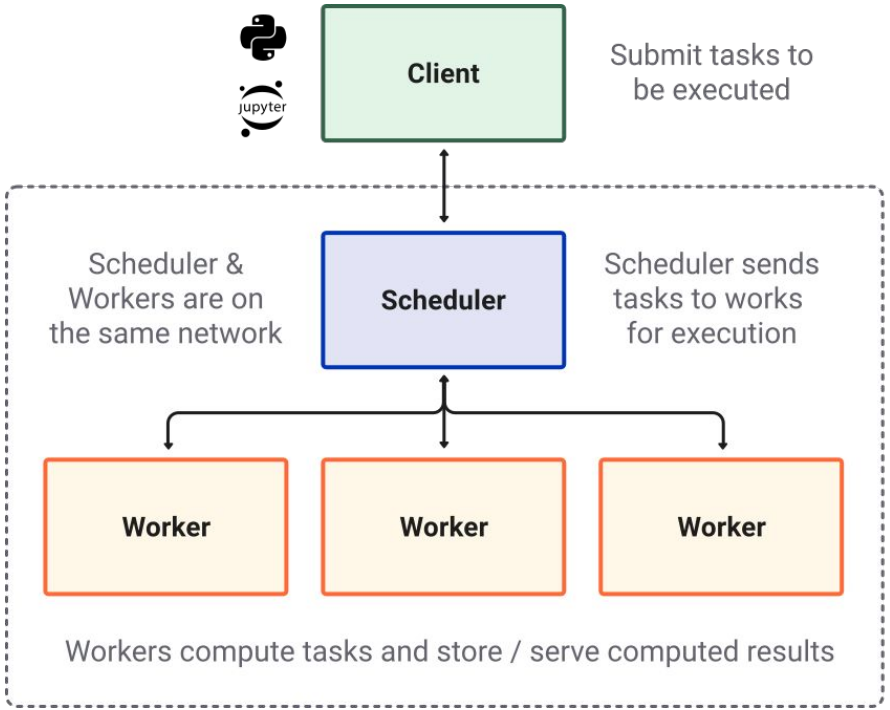
Fundamentals

- Dataframe
- Array
- Bag
- Delayed
- Futures
- Distributed

ML integrations

- Sklearn & Joblib
- XGBoost & LightGBM
- PyTorch
- Keras and Tensorflow

Dask Cluster



https://tutorial.dask.org/00_overview.html#Dask-Cluster

scale

the python tools

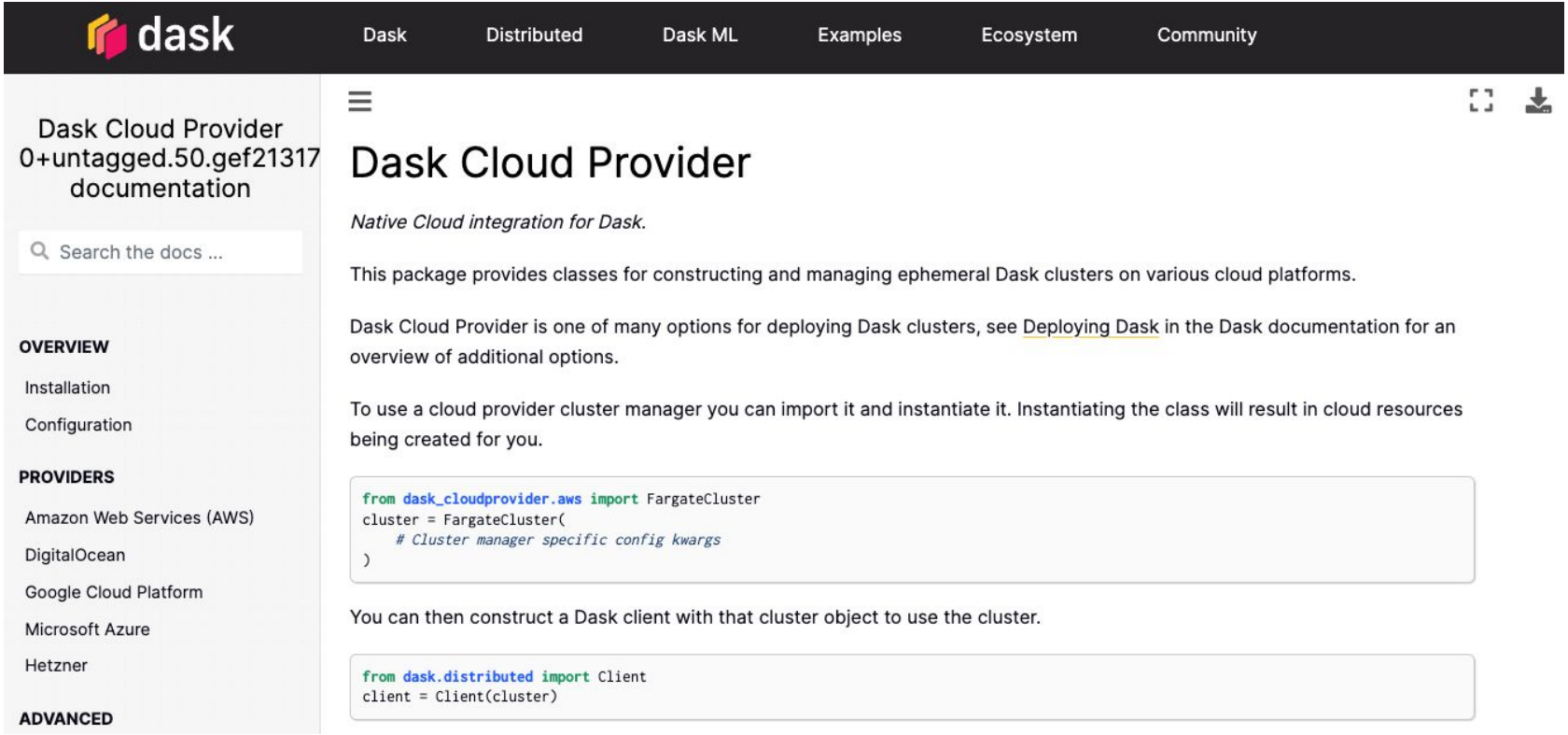
you love

Cloud

Top 10 Cloud Provider 2023

PyData

Dask Cloud Provider



PyData

Today

Goal:
to run our code on
resources in the cloud

Solution:
to use AWS
and dask-cloudprovider

- our code with dask
  - for local
- configuration of tools
- our code modified
  - for cloud

Code

Configuration

PyData

Configuration

AWS account:
creation & configuration

Laptop side:
configure AWS
use AWS in our code

- AWS configuration
  - create account
  - create credentials
- on our laptop
  - install aws-cli
  - configure credentials
- in our code using libraries
  - boto3
  - dask-cloudprovider

Code

PyData

Results

Goal:

to run our code on
resources in the cloud

Solution:

use AWS
and dask-cloudprovider

- on Laptop
  - workers: 2
  - iters: 500
  - time: 271 seconds
- on AWS, in the cloud
  - workers: 4
  - iters: 500
  - time: 84 seconds

PyData

Best practices

AWS console access:
don’t use root user !

Dask cluster:
delete the cluster !

- AWS
  - use multi-factor authentication (MFA)
  - create your user with only the permissions that you need
- Laptop
  - update AWS credentials
  - create an environment
- Dask
  - use only the resources that you need
  - remember to delete the cluster for saving costs

may

data power

with you

Links

Support

- AWS getting started, https://docs.aws.amazon.com/accounts/latest/reference/welcome-first-time-user.html
- AWS free support, https://discord.gg/N97snpCQ
- Dask Overview, https://tutorial.dask.org/00_overview.html
- Dask Cloud Provider, https://cloudprovider.dask.org/en/latest/

References

- https://ml.dask.org/joblib.html
- https://gist.github.com/jacobtomlinson/ee5ba79228e42bcc9975faf0179c3d1a

PyData

Thanks for listening.

@PyDataVenice #PyDataGlobal #2023