Bridging Cloud and HPC towards High Performance Data Analytics for climate science

Donatello Elia[1], F. Antonio[1], C. Palazzo[1], A. D'Anca[1], S. Fiore[2], G. Aloisio[1,3]
[1]Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)
[2]University of Trento
[3]University of Salento

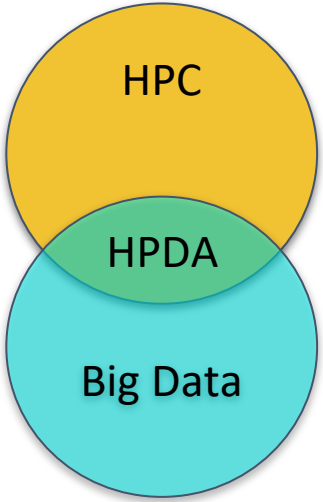
EGI-ACE Lightning Talks: Compute continuum use cases
EGI Conference 2022, 21 September 2022

EGI-ACE receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 101017567.

Convergence of HPC and Big Data Analytics for HPDA

Convergence of **HPC** and **big data analytics** is a key factor for future scientific research and for enabling **HPDA** applications at **extreme-scale**:

- *Support scalable execution models for data **analysis workloads** on top of **HPC infrastructures***
- *Enabling integration of data-driven and compute-driven workloads into a **single workflow** including HPC, analytics and ML components*



Big Data (cloud-based) and HPC software ecosystems have been developed mostly independently

- *Significant gaps in how the two ecosystems are designed (e.g., in terms of deployment, workload, programming models, etc.)*
- *New computing paradigms and software portability across different infrastructures are being addressed by the scientific software community*

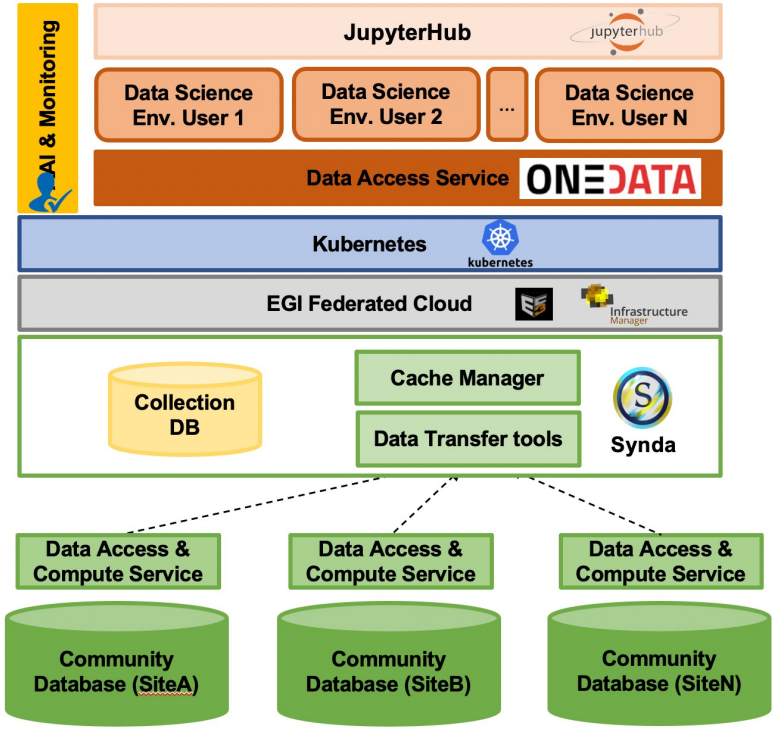
ENES Data Space

*Setup in the EGI-ACE project to deliver an open, scalable and cloud-enabled **data science environment** for **climate data analysis** on top of the **EOSC Compute Platform***

The **ENES *Data Space*** provides a single entry-point to:

- **Datasets** (e.g. CMIP) → most relevant; pre-staged; open
- **Storage** & **Compute** resources → provided by EGI
- **Data Science Software Stack** → to address a wide spectrum of analysis needs (mainly Python-based)
- **Jupyter-based gateway** → to devel/share/(re-)use apps
- **Cloud-Enabled** → SaaS for apps; PaaS for data services

→ **Ultimate goal**: promote user's **productivity** and **democratization** of eScience



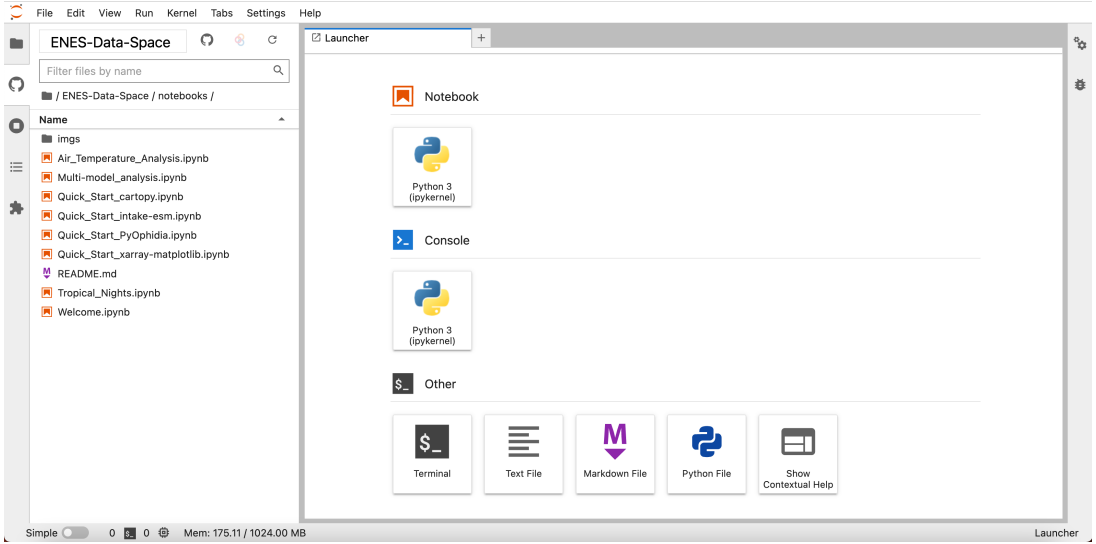
Python-based data science environment

Multi-user **JuptyterHub** instance

- gateway to the whole data space environment

Ready-to-use **JupyterLab** instance equipped with a

- extensible Pythom-based environment (data manipulation, analysis and visualization)
- community-based **parallel analytics frameworks** (e.g., Ophidia)



DASK python jupyter NumPy Ophidia pandas SciPy xarray Cartopy matplotlib

Ophidia High-Performance Data Analytics framework

Ophidia (http://ophidia.cmcc.it) is a CMCC Foundation research project addressing data challenges for eScience

- **HPDA-enabled** framework joining **HPC paradigms** with **scientific data analytics approaches**
- **in-memory** and server-side data analysis exploiting **parallel computing** techniques and **database** approaches
- a **multi-dimensional, array-based, storage model** and **partitioning schema** for scientific data leveraging the **datacube abstraction**
- **PyOphida**: Python bindings for data science applications →

```python
from PyOphidia import cube, client
cube.Cube.setclient(read_env=True)

mycube = cube.Cube.importnc2(
    src_path=input_file,
    measure='tas',
    imp_dim='time',
    ncores=2,
    nfrag=2,
    description="Imported cube"
)
mycube2 = mycube2.reduce(
    operation='max',
    ncores=2,
    description="Reduced cube"
)
data = mycube2.export_array()
var = data['measure'][0]['values'][ : ]
```

Ophidia

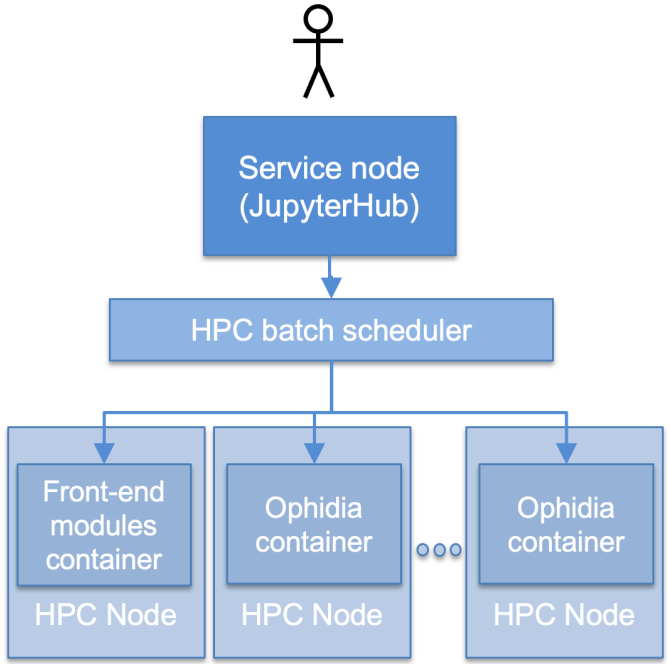
cmcc Centro Euro-Mediterraneo sui Cambiamenti Climatici

S. Fiore, D. Elia, C. Palazzo, F. Antonio, A. D'Anca, I. Foster, G. Aloisio, "Towards High Performance Data Analytics for Climate Change", ISC High Performance 2019, LNCS Springer, 2019

HPC ENES Pilot overview

Exploit **HPC resources** for HPDA in **climate applications**:

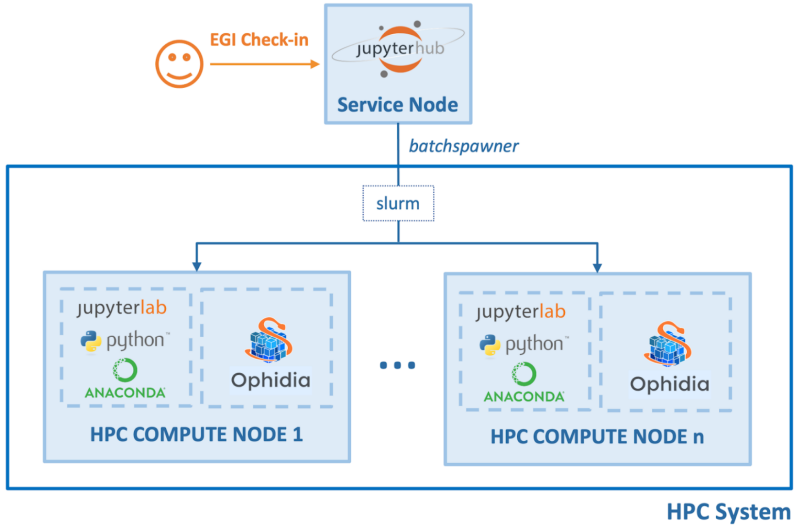
- Initial PoC Data Science environment for **data analytics** and **visualization** on top of a **HPC infrastructure** developed in the context of EGI-ACE
  - Focus on subset of data space software stack: *Ophidia, Jupyter, Python Data Science libraries*
- Explore how **to simplify** the **deployment** of HPDA services over different infrastructures
- **Software containers** for supporting **transparent** portability and deployment
  - From Cloud (K8s) to HPC (schedulers)
- *Leading to novel service models:* ***HPC as a Service***



Results from the HPC Pilot PoC & future plans

*Initial PoC setup on the **HPC resources** from the **EGI Federation (Tubitak)**:*

- *different usage scenarios evaluated*
- ***non-privileged container-based solutions** for deployment of the environment on HPC (i.e., udocker)*
- ***Jupyter** as gateway to the HPC resources (i.e., customized batchspawner by INFN)*
- *integrate federated solutions for user **AAA** (e.g., EGI Check-in) to run on HPC*



Next steps:

- ***transparent integration** of **HPC infrastructures** in the general data space*
- *better support for **multi-node execution***
- *stronger integration with **federated storage resources** (e.g., DataHub)*

Thank you!

Ophidia website	*http://ophidia.cmcc.it*
ENES Data Space:	*https://enesdataspace.vm.fedcloud.eu/*
EGI-ACE	*https://www.egi.eu/projects/egi-ace/*
IS-ENES	*https://is.enes.org/*

These activities are supported in part by EGI-ACE and IS-ENES3 projects:

- ***EGI-ACE*** *receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101017567*
- ***IS-ENES3*** *is a project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824084*