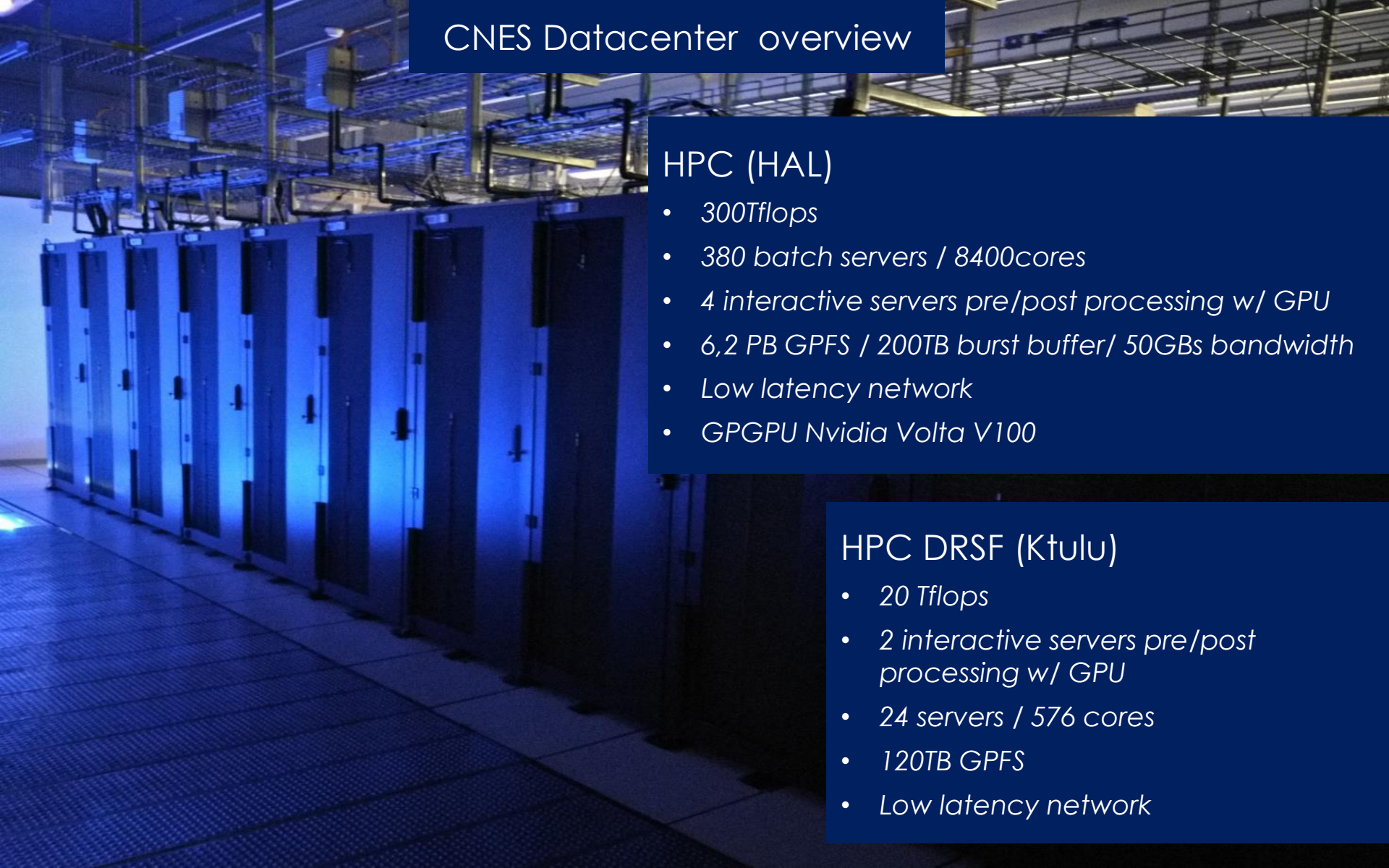
Jupyter, Dask : traitement distribué simple et interactif sur HPC avec l'écosystème Pangeo

JCAD 2018, Guillaume Eynard-Bontemps, communauté Pangeo

cnes
CENTRE NATIONAL D'ÉTUDES SPATIALES



CNES Datacenter overview

HPC (HAL)

- *300Tflops*
- *380 batch servers / 8400cores*
- *4 interactive servers pre/post processing w/ GPU*
- *6,2 PB GPFS / 200TB burst buffer/ 50GBs bandwidth*
- *Low latency network*
- *GPGPU Nvidia Volta V100*

HPC DRSF (Ktulu)

- *20 Tflops*
- *2 interactive servers pre/post processing w/ GPU*
- *24 servers / 576 cores*
- *120TB GPFS*
- *Low latency network*

HPC usecases in CNES

Two main kinds of processing

Numerical simulation (HPC)

- Upstream phase, R&D
- Highly optimized technics
- Fine grain parallelism

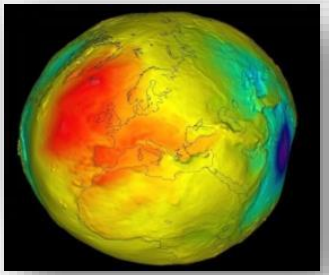
Trends : multiscale, multiphysics

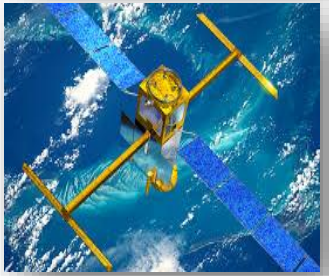
Data Processing (HTC)

- Downstream phase, operation
- Sensors data → scientific data
- Coarse grain parallelism

Trends : data volume explosion





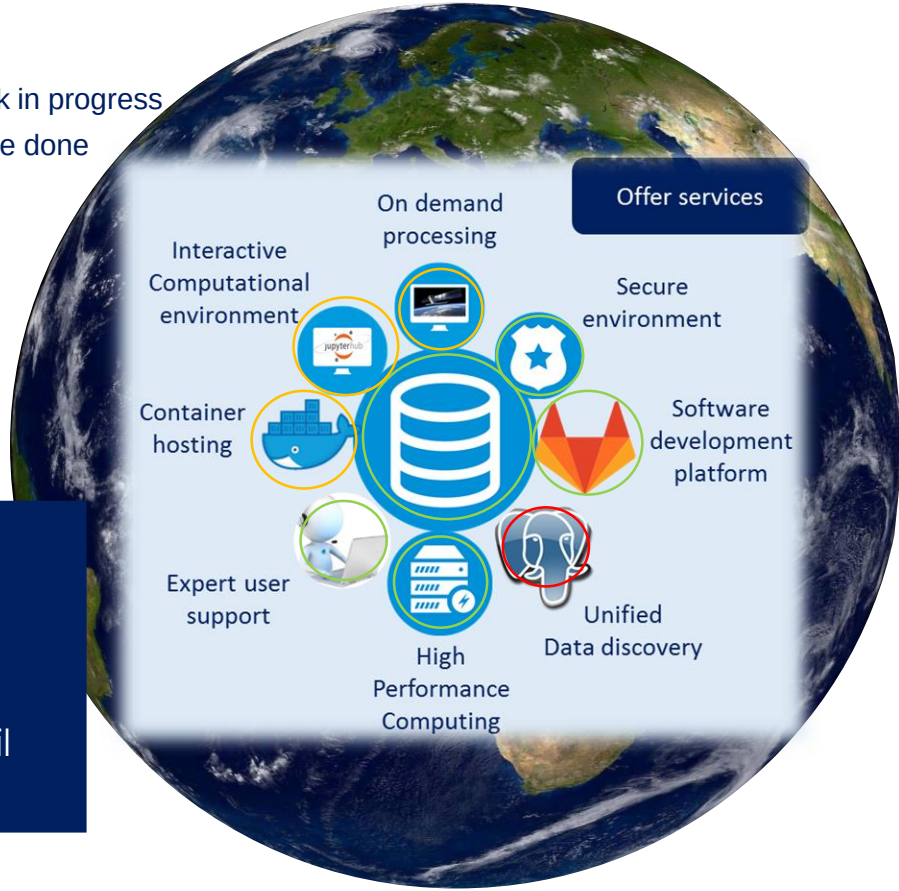
datalake : approche de haut niveau

Objectifs

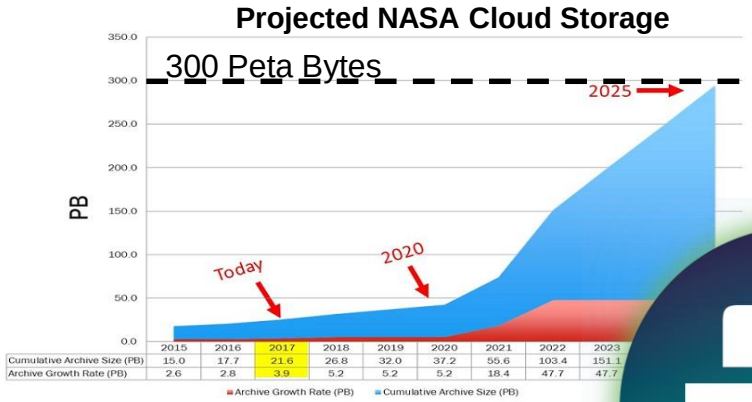
- Casser les silos de données
- Simplifier l'accès aux données
  - Entrée (Sentinel, Landsat, AUX, etc.)
  - Produite (Theia, CFOSat, etc.)
- Performance
- Offre de services complète

Challenges

- Fournir un environnement de développement standard, simple, flexible et performant (!)
- Favoriser l'adoption des nouveaux paradigmes de développement parallèle (Dask, Spark, ..), le travail des données en mémoire (vs orienté fichier)



Pangeo goals and motivation



Problems

- Data volume crisis in (geo)sciences
- Software multiplication, non reproducibility
- Many copies of the same datasets
- Local vs HPC vs Cloud
- Technology gap: industry vs academia

Mission

To cultivate an ecosystem in which the next generation of open-source analysis tools for the geosciences can be developed, distributed, and sustained.

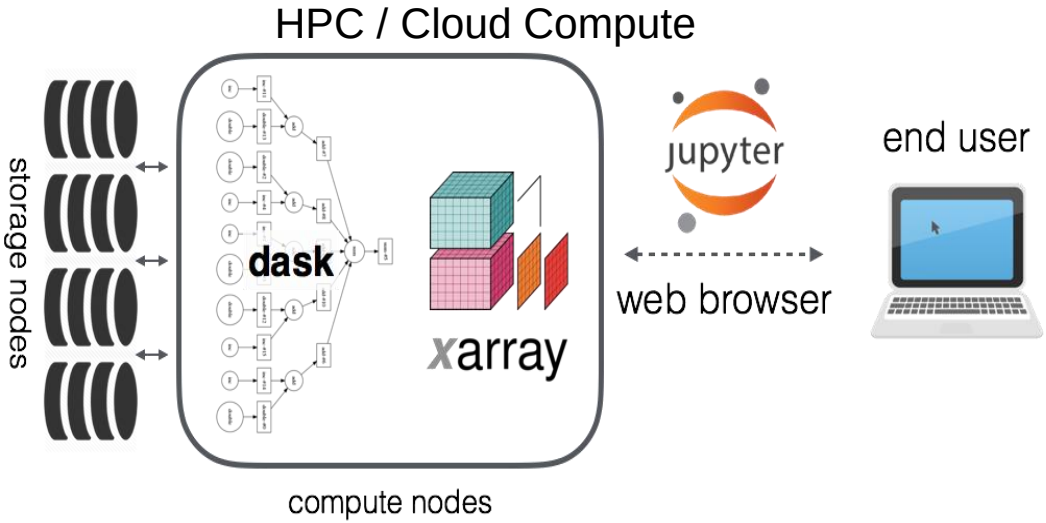
Goals/vision

- Foster collaboration around the open source Scientific Python ecosystem:
  - open and collaborative development
  - Welcoming and inclusive culture
- Support the development with domain-specific (geo)science and transverse packages
- Improve scalability of these tools to handle gigabytes to petabytescale datasets

EarthCube
TRANSFORMING GEOSCIENCES RESEARCH
NSF

Pangeo ecosystem

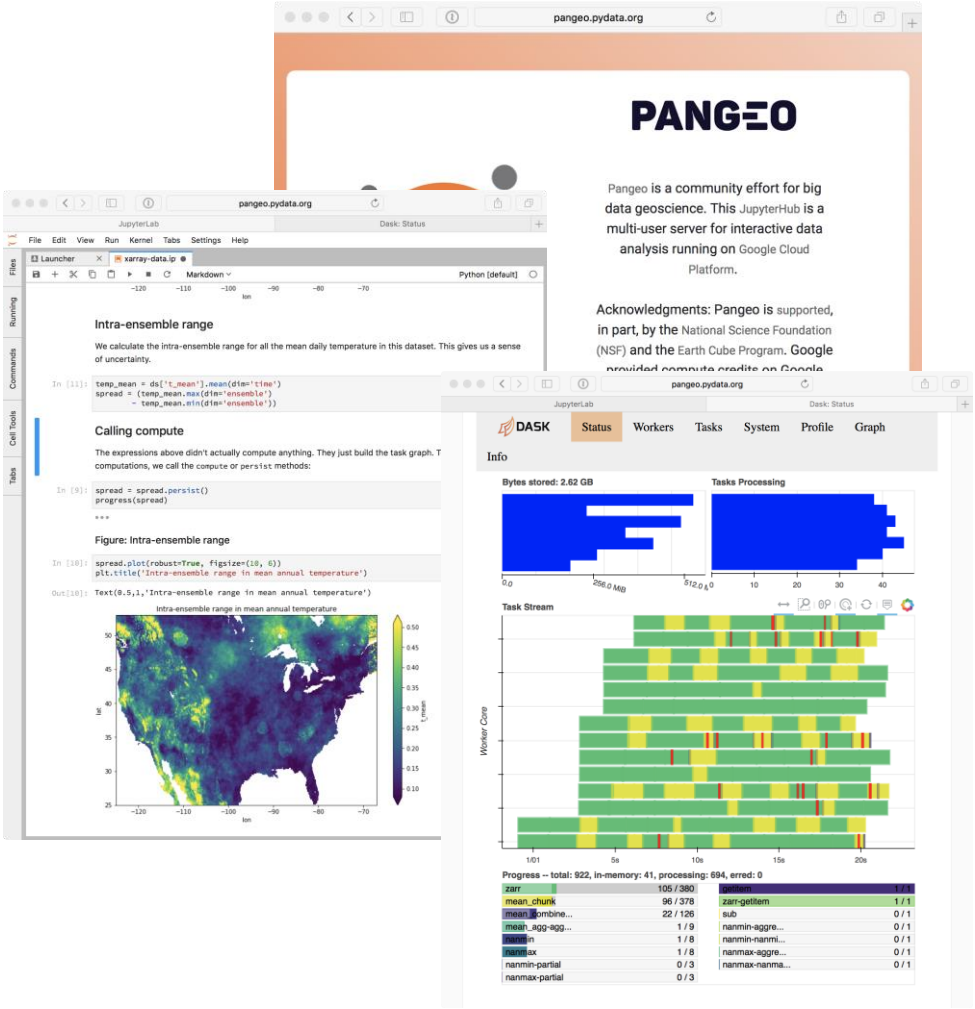
- Set of tools that will facilitate science at all scales
- Platform agnostic
- The core of the Pangeo ecosystem includes:
  - **Xarray** (data-model and toolkit for working with N-dimensional labeled arrays)
  - **Dask** (parallel computing)
  - **Jupyter** (interactive computing)
- Extensible: Series of 3rd party packages that build on top of core libraries
- Flexible: Individual components may be swapped in/out



Examples of 3rd party packages in the Pangeo Ecosystem:

- Data discovery
- Regridding and GIS
- Vector calculus
- Signal processing
- Thermodynamics

Pangeo public cloud deployment



pangeo.pydata.org

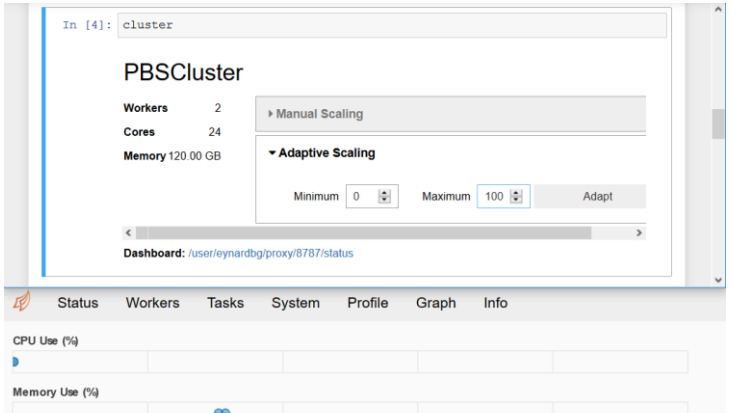
pangeo.binder.io

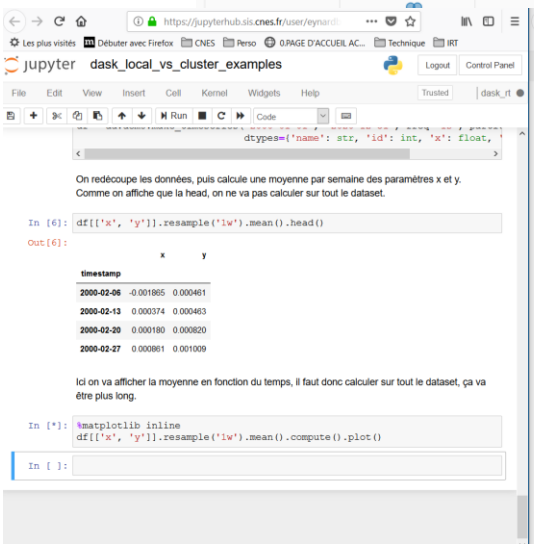
JupyterHub/Binder running on the Google Cloud

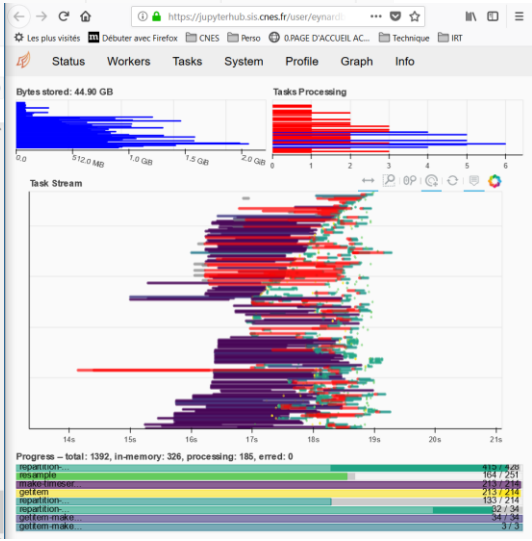
- Kubernetes for both Jupyter and Dask-distributed
  - Dask-kubernetes
- Exploring/evaluating:
  - Cloud storage
  - User environment customization
  - Data discovery
- Kubernetes Helm-chart (github.com/pangeo-data/helm-chart)
- Clones of our deployment have been made on AWS and Azure.

Pangeo at CNES

- JupyterHub and notebooks for interactive computing
  - Hub on a VM with qsub access
  - Batchspawner, Wrapspawner
- dask.distributed: parallel workers across many HPC nodes
- Xarray for computational toolkit and I/O
- New tool for deploying dask clusters on HPC: **dask-jobqueue**
  - Start a cluster from a notebook
  - Interactive (or not) distributed computing
  - Auto scaling capabilities







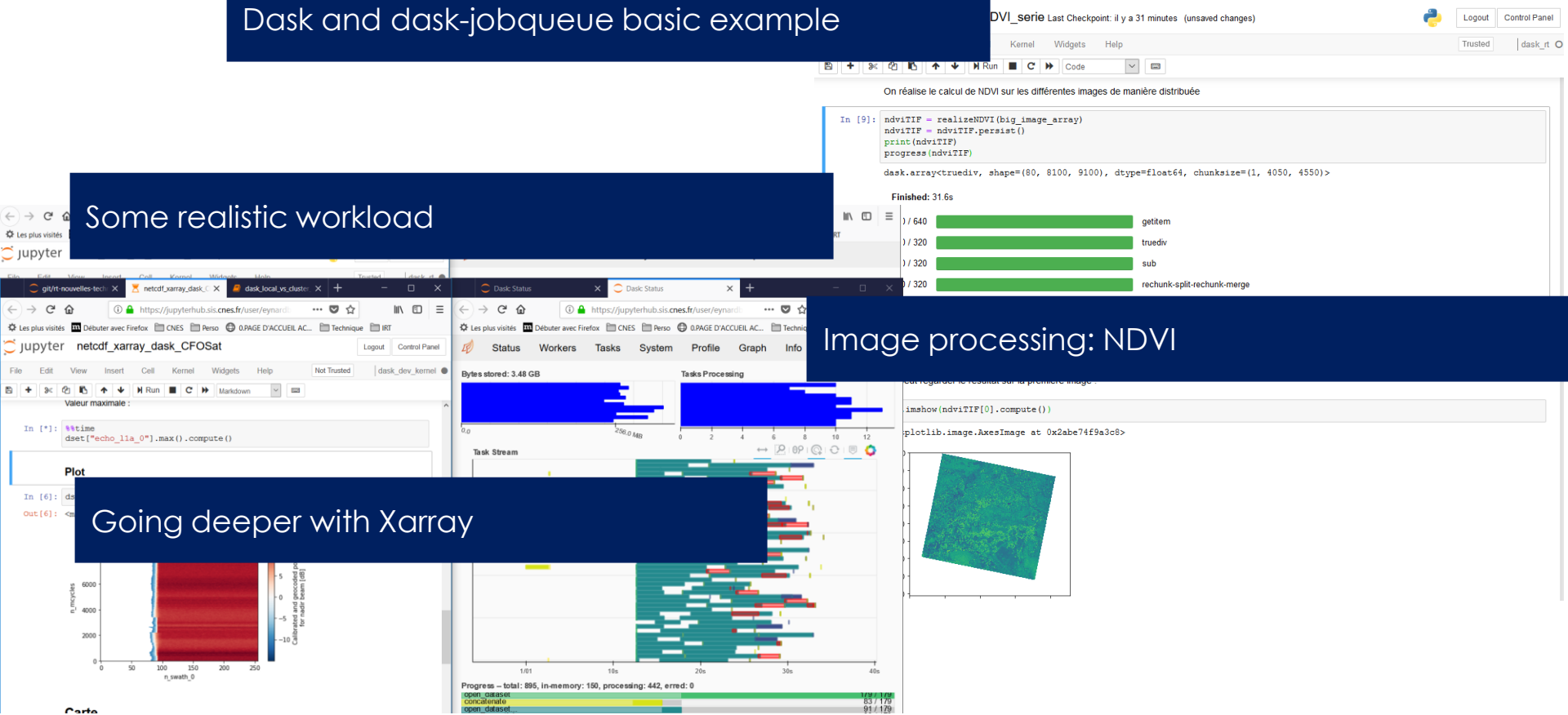
Demos @ CNES

Dask and dask-jobqueue basic example

Some realistic workload

Image processing: NDVI

Going deeper with Xarray



Pangeo Cloud use

Geoscience use cases:
http://pangeo.io/use_cases/index.html

Astronomy with GAIA catalog:

https://github.com/pangeo-data/pangeo/issues/255#issuecomment-427186915

Image processing and visualization

https://medium.com/pangeo/cloud-native-geoprocessing-of-earth-observation-satellite-data-with-pangeo-997692d91ca2

Conclusion, Useful links

Conclusions

- Pangeo ecosystem greatlty facilitates distributed computing and data analysis at scale
- It changes ways of doing it too
- Non monolithic platform built on top of existing Scientific Python stack and new related packages
- Community is always here to help

Next steps

- Broaden users and use cases at CNES
- Work in cooperation with others (Ongoing with Ifremer and CLS on SWOT aval data processing)

Pangeo wesite and discussions:

https://pangeo.io

https://github.com/pangeo-data/pangeo/issues

https://medium.com/pangeo

Pangeo Example + Binder:

https://github.com/pangeo-data/pangeo-example-notebooks

http://binder.pangeo.io/v2/gh/pangeo-data/pangeo-example-notebooks/master

Dask jobqueue:

https://github.com/dask/dask-examples

Dask simple examples:

https://github.com/dask/dask-examples

My email

Guillaume.Eynard-Bontemps@cnes.fr