Analyse de données scientifiques à l'échelle sur HPC ou dans le Cloud avec Pangeo

Guillaume Eynard-Bontemps 

CNES Computing Center team, Centre National d'Etudes Spatiales
Toulouse, France

Tina Odaka 

Laboratory for Ocean Physics and Satellite Remote Sensing, UMR LOPS, Ifremer, Univ. Brest, CNRS, IRD, IUEM Brest, France

Overview

Introduction & Motivation to Pangeo (Tina Odaka)

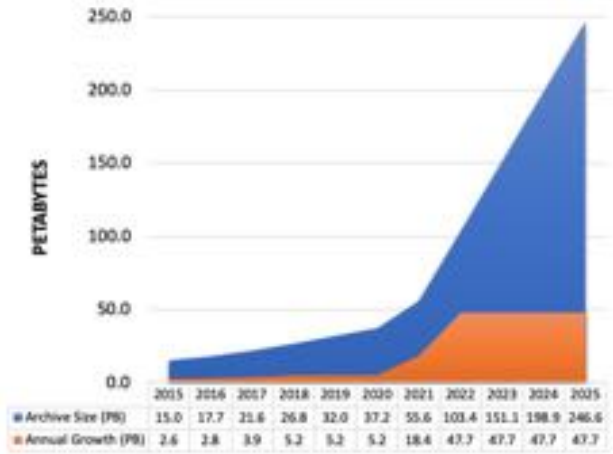
Demonstration on CNES computing platform (Guillaume Eynard-Bontemps)

Introduction and Motivation for PANGEO

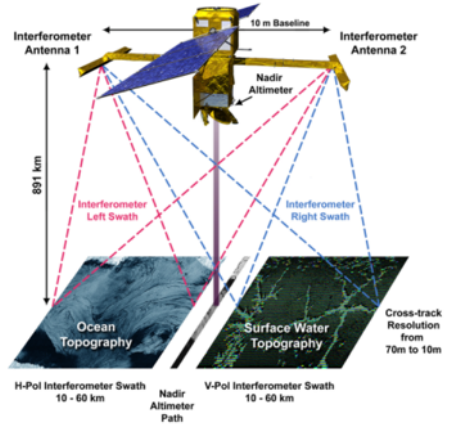
Analysis of Big Geosicence Data (observations and simulations)

- CMIP5 (Coupled Model Intercomparison Project Phase 5) 2PB, CMIP6=18PB

- NASA earthdata cloud evolution :



https://earthdata.nasa.gov/eosdis/cloud-evolution NASA EOSDIS graphic.



SWOT Satellite instruments : https://swot.cnes.fr/fr

Crisis for traditional data analytics workflows

PANGEO to free scientists to explore their data

Dugornay Olivier (2014). **Déploiement d'un profileur Arvor face au rocher du Lion**. Ifremer. https://image.ifremer.fr/data/00378/48926/

Pangeo is a funded *collaboration* dedicated to the advancement of scalable, interactive, easy-to-use data analysis for the climate and weather community.

Alfred P. Sloan FOUNDATION

EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

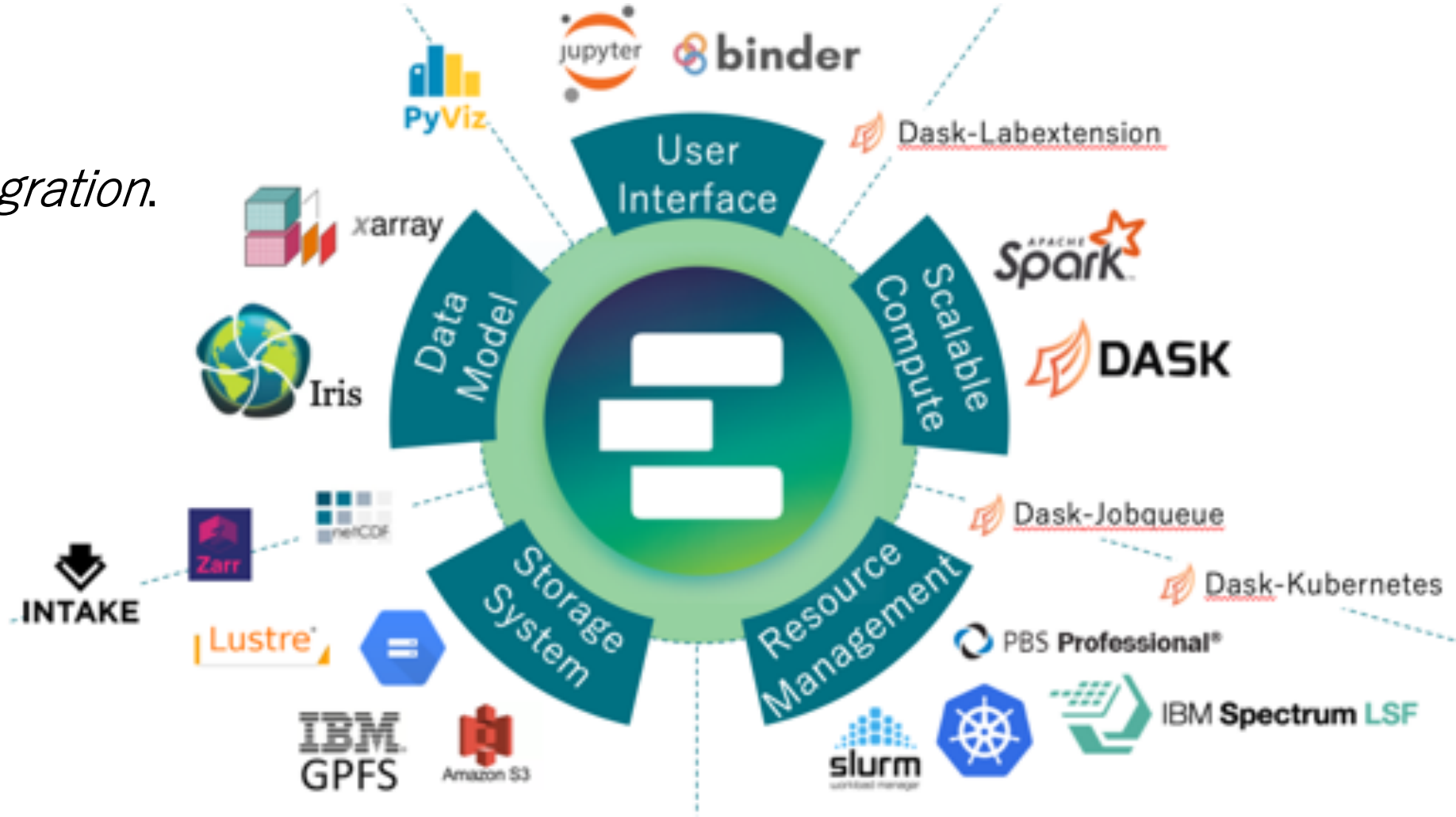
« Pangeo: An Open Source Big Data Climate Science Platform » Ryan Abernathey et al.
https://figshare.com/articles/Pangeo_NSF_Earthcube_Proposal/5361094

Pangeo is a *community* that promotes open, reproducible, and scalable science.

. . . just to name a few.

Pangeo is a '*platform*' that is deployable on HPC or cloud.

Pangeo is about *integration*.

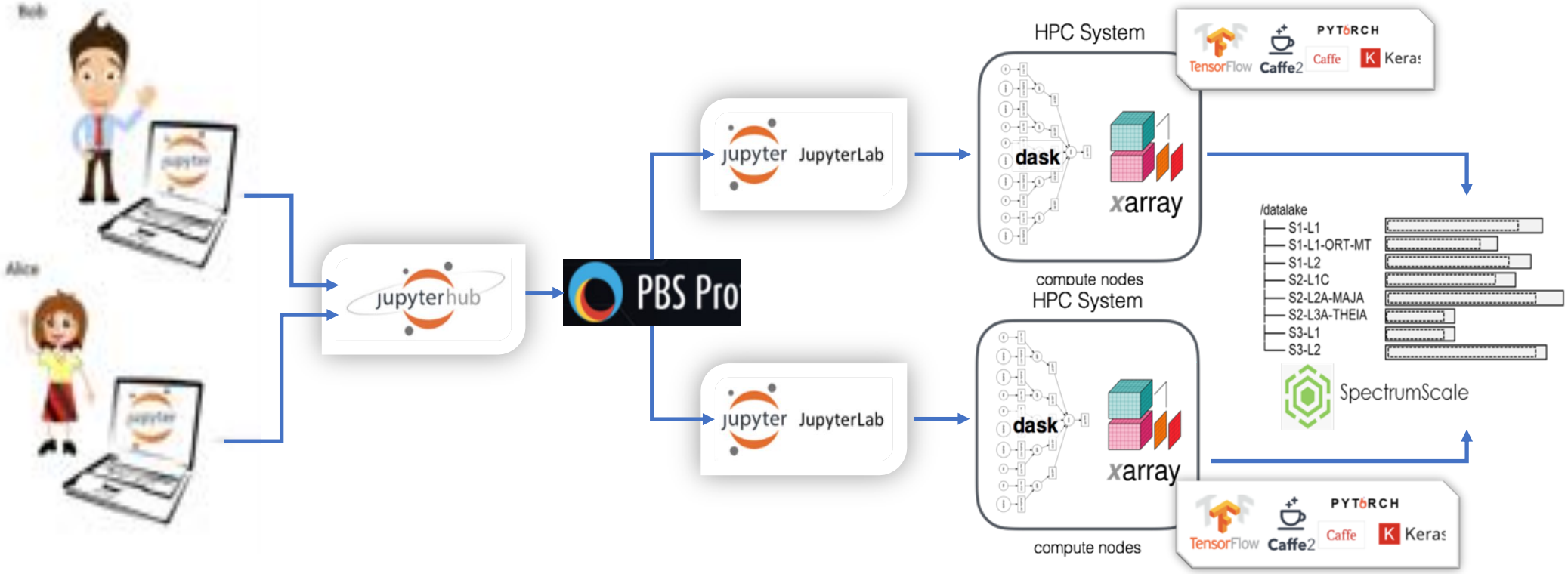


What is behind parallel computing in Pangeo

- Distributed computing = Dask
  - Dask worker clusters:
    - HPC: Dask jobqueue: on HPC, spawns Dask workers via job schedulers (e.g. PBS Pro, Slurm)
    - Cloud: Kubernetes and Helm integrations
    - Hadoop: Dask-Yarn
  - Workers process computations on each ‘chunked’ data.
  - Dask Scheduler orchestrates tasks among workers.

- Distributed IO , Zarr and other cloud ready formats
  - Directory tree format, each zarr directory is made of many files, each contains one chunk of data.
  - Each Dask workers can read /write only the necessary chunk of data.

Demonstration on CNES HPC



Try it with Pangeo Binder: http://gallery.pangeo.io/

Conclusion

Pangeo makes it possible to explore big-data geoscience using HPC or cloud in an interactive manner.

Easy to try

https://github.com/pangeo-data

Active community

https://discourse.pangeo.io

Start from here

http://pangeo.io/about.html