Pangeo : une plateforme communautaire pour le traitement de données scientifique à l'échelle

2019-10-24, Guillaume Eynard-Bontemps, CNES/Pangeo

cnes
CENTRE NATIONAL D'ÉTUDES SPATIALES

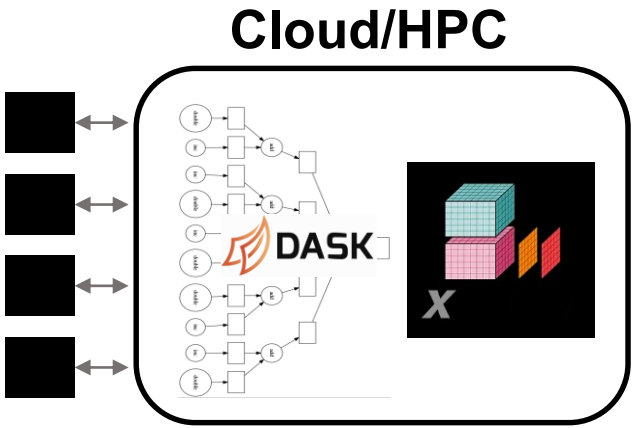
Pangeo c'est quoi ?

A community platform
for Big Data geoscience

- ✦Une communauté internationale ouverte
- ✦Un écosystème logiciel open source
- ✦Une infrastructure open source

Pangeo ecosystem

Données, prêtes à l'analyse, stockées et cataloguées sur un système de stockage distribué accessible à l'échelle globale (p. ex. S3, GCS)



Jupyter pour un accès interactif sur des systèmes distants.

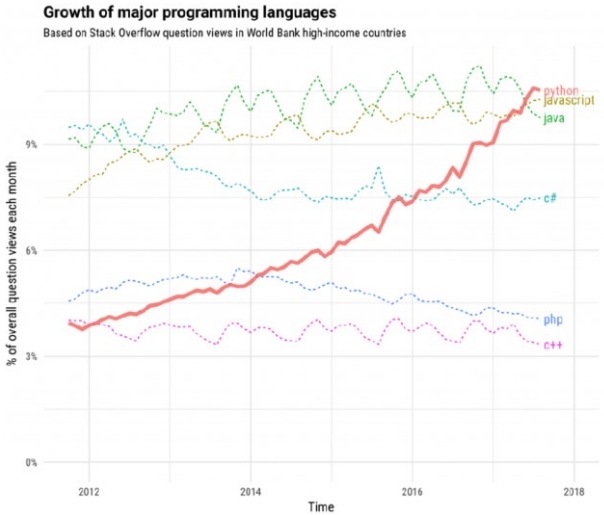
Xarray fournit des structures de données et une interface intuitive pour interagir avec les ensembles de données.

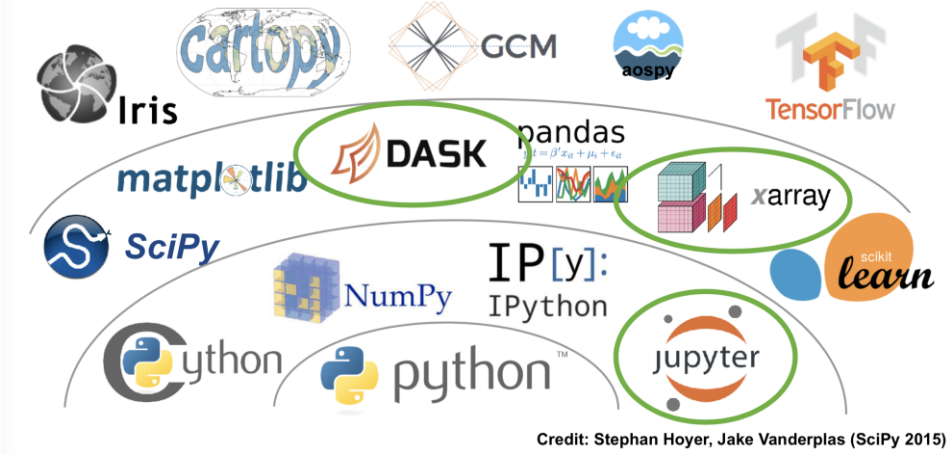
Système de calcul parallèle construit sur Kubernetes ou HPC. Dask dit aux nœuds ce qu'ils doivent faire.

- Un déploiement ouvert sur le cloud (Jupyterhub) hub.pangeo.io
- Nombreux déploiements sur cluster/HPC : https://pangeo.io/deployments.html
- Dont les systèmes HPC du NCAR et de la NASA, et le cluster HAL du CNES.

Pourquoi Ouvert ?

- Communauté scientifique, utilisatrice d'outils libres (écosystème Python)
- Communauté internationale, besoin d'échanges simplifiés
- Être le plus inclusif possible et bénéficier de toute contribution :
  - Un maximum de discussions : clarifier les besoins, les problèmes
  - Un maximum de collaboration : faire évoluer les logiciels, les documentations, les pratiques…

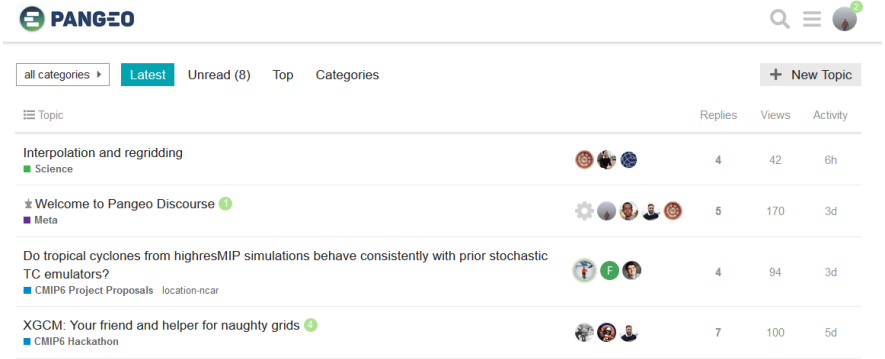




Credit: Stephan Hoyer, Jake Vanderplas (SciPy 2015)

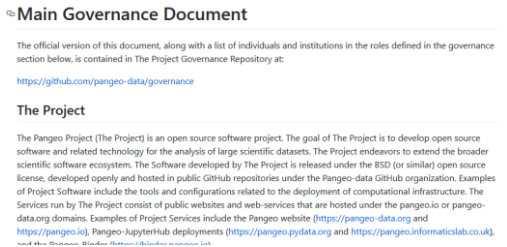
Nos outils

- Github : https://github.com/pangeo-data
  - Gestion de code
  - CI : Travis, Circle-ci,
  - Gestion documentaire (ReadTheDoc)
  - Discussions ouvertes : https://github.com/pangeo-data/pangeo/issues
  - Mais pas facile pour tout le monde
- Discourse (new), discussion plus accessible que github : https://discourse.pangeo.io/
- Blog Medium : présentation de résultats et évènements : https://medium.com/pangeo



Notre organisation et gouvernance

- Groupe international
- Réunions hebdomadaires de synchro
  - Horaires compliqués à trouver
- Rencontres annuelles
- Groupes de travail et d'intérêt
  - Education, Machine Learning, data model…
- Steering commitee
  - https://github.com/pangeo-data/governance
  - Discussions stratégiques, orientations



Main Governance Document

The official version of this document, along with a list of individuals and institutions in the roles defined in the governance section below, is contained in The Project Governance Repository at:

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

The Project

The Pangeo Project (The Project) is an open source software project. The goal of The Project is to develop open source software and related technology for the analysis of large scientific datasets. The Project endeavors to extend the broader scientific software ecosystem. The Software developed by The Project is released under the BSD (or similar) open source license, developed openly and hosted in public GitHub repositories under the Pangeo-data GitHub organization. Examples of Project Software include the tools and configurations related to the deployment of computational infrastructure. The Services run by The Project consist of public websites and web-services that are hosted under the pangeo.io or pangeo-data.org domains. Examples of Project Services include the Pangeo website (https://pangeo-data.org and https://pangeo.io), Pangeo-JupyterHub deployments (https://pangeo.pydata.org and https://pangeo.informaticslab.co.uk),

Partnering with Cloud Providers

Ryan Abernathey Follow

Sep 25 · 6 min read

Ryan Abernathey and Joe Hamman cowrote this blog post following discussion with the Pangeo Steering Council.

The basic ingredients for a Pangeo deployment are a fast parallel storage system, scalable high-performance compute nodes, access to the internet, and software which makes all these elements work together for an amazing interactive data-analysis experience.

Valorisation

- Articles de blog, Conférences, Rencontres, Formations… Evangélisation.
- Mise à disposition de plateformes pour tests : http://pangeo.io/, https://binder.pangeo.io/
- Site web http://pangeo.io/ et documentation
- Information ouverte via github, discourse… http://pangeo.io/meeting-notes.html
- Tutoriels/binder : https://github.com/pangeo-data/pangeo-tutorial, https://github.com/pangeo-data/pangeo-ocean-examples ...
- Charte graphique : https://github.com/pangeo-data/branding
- Publications scientifiques : http://pangeo.io/publications.html
  - Pas facile à récupérer : pour l'instant en mode volontaire.
- Adoption par les institutions ou reconnaissances dans d'autres communautés

Pérennité du projet

- Aux US, financement sur appel à projet pour les chercheurs
  - Des potentiels problèmes RH ?
- Briques logicielles indépendantes, avec des contributeurs et une communauté différents
  - Peu de chance que tout l'écosystème disparaisse
- Des sociétés ou organisations de taille comme Anaconda, Nvidia, ou même Google
- Nombreux contributeurs ou membres de la communauté.
  - Mais finalement pas tant d'acteurs tout le temps actifs.
- Tout est disponible en ligne et utilisable sous licence permissive
  - Possibilité de maintenir les outils si abandon de la communauté
- Au final pérennité bien supérieure à un développement interne

Pourquoi participer ? De quelle manière ?

- http://matthewrocklin.com/blog/work/2018/08/21/institutional-open-source
- Participer à l'orientation du projet, définition des besoins : discussions ouvertes
- Remonter les problèmes, améliorer la maintenance ou la doc : issues github
- Ne pas redévelopper nos propres solutions : réduire les coûts de développement et de maintenance
  - Contribuer aux projets ouverts plutôt que de développer du code spécifique au dessus ou à la place : contribution github
  - Bénéficier d'une grande force de développeurs, et de fonctionnalités annexes
- Bénéficier de l'expertise et de retours d'expériences des membres de la communauté : discussions ouvertes
  - Récemment, discussion gestionnaire de workflow

Comment participer à Pangeo ?

http://pangeo.io

- Utiliser et contribuer à Xarray, Dask, Zarr, Jupyterhub…
- Accéder et utiliser une plateforme Pangeo sur un cluster HPC (au CNES) ou sur des ressources cloud (http://pangeo.io/deployments.html)
- Parlez-nous !
  - https://github.com/pangeo-data/pangeo/issues - plutôt dev
  - https://discourse.pangeo.io/ - plutôt science et discussions ouvertes (nouveau)

PANGEO