Keywords:
#Pangeo, #geoscience, #openscience, #bigdataanalytics, #FAIRdata, #softwareprinciples, #cloudcomputing, #EGIFoundation

Pangeo: Fostering Open Science Collaboration in Big Geoscience Data Analytics

Empowering the European Research Community with Scalable Infrastructure, FAIR Data Practices, and Cross-Disciplinary Synergies

Anne Fouilloux

Senior Research Engineer at Simula Research Laboratory

"The Pangeo ecosystem is increasingly used in different contexts, like bioimaging or earth observation, with the potential to become a 'reference' open science gateway able to leverage infrastructures and data providers for various scientific applications."

"The Pangeo ecosystem is increasingly used in different contexts, like bio imaging or Earth Observation, with the potential to become a 'reference' open science gateway able to leverage infrastructures and data providers for various scientific data driven applications".

The Project Involved

PANGEO

A community platform for Big Data geoscience

Pangeo is a global community-driven initiative that fosters collaboration and scalability in big geoscience data analytics. Scientists, developers, and research engineers unite efforts to develop and enhance software and infrastructure, and address significant challenges in Big Data Geoscience research within this open ecosystem.

The Challenge

The members of the European Pangeo community didn't have a shared deployment where anyone interested in big data could learn, share and exchange knowledge and best practices for delivering efficient Pangeo deployments and for practising FAIR data and software principles, and Open Science in general with big data.

The challenges that Pangeo addressed was how to deploy and use a Pangeo enabled infrastructure on a public cloud and underline the benefits for the European community, onboarding European researchers on this Pangeo EOSC infrastructure.

EOSC service or tool used

Pangeo used several services to enable future deployment, e.g. DaskHub and Kubernetes cluster backend on EOSC leveraging the infrastructure of the EGI Federation.

The Pangeo EOSC JupyterHub deployment capitalises on the EGI Check-in for user registration. It uses the EGI Cloud Compute and cloud-based EGI Online Storage for distributing computational tasks across a scalable compute platform and storing intermediate results generated by users.

To enable future deployments of Pangeo@EOSC, Infrastructure Manager (IM) Dashboard enables future Pangeo deployments across a broad range of private and public cloud providers, including EGI Federated Cloud, OpenNebula, OpenStack, AWS, GCP and more.

The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020 - Grant Agreement Number 101017536

EOSC Future

EOSC in practice story #37

Keywords:
#Pangeo, #geoscience, #openscience, #bigdataanalytics, #FAIRdata, #softwareprinciples, #cloudcomputing, #EGIFoundation

Useful tips & tricks

"Pangeo Training Infrastructure as a Service (PTIaaS)" is the training developed by the European Pangeo community; it can be requested online and offers a dedicated, scalable environment for workshops, deploying a specialised Pangeo JupyterHub instance accessible on cloud platforms or dedicated servers.

Customizable and free of charge, it supports hands-on exercises, collaborative work, and interactive learning, promoting the adoption of Pangeo within the geoscience community.

The Research Community

Pangeo ecosystem is increasingly used in different contexts, like bioimaging or earth observation, with the potential to become a "reference" open science gateway able to leverage infrastructures and data providers for various scientific applications.

Benefits and impact

The availability of the newly developed infrastructure has allowed the strengthening of the Pangeo community in Europe, "Pangeo@Europe", granting it a shared deployment where scientists and/or technologists can exchange know-how and provide feedback.

Such an environment can speed up the learning process: in fact, users have a chance to learn with real-world examples to access, analyse, visualise and share data, Jupyter Notebooks and best practices for making their research work FAIR (Findable, Accessible, Interoperable, Reusable).

In this way, Pangeo has been able to train over 100 researchers to use the infrastructure, thus helping them work in a truly open and collaborative science environment.

Useful material related to this story

Pangeo.io

Why do I need EOSC?

Pangeo users need EOSC-provided solutions to get easier and faster access to data.

Thanks to the services provided by the EGI Federation through EOSC, Pangeo is working towards the improvement of its deployment facilitating Open Science practices through serving a Binder instance with a Dask gateway, and providing a common approach to spatial data analysis, independently of data and infrastructure providers.

Across disciplines

The deployment of Pangeo on EOSC facilitates cross-disciplinary research by offering a scalable and adaptable platform. Pangeo's seamless integration with EOSC ensures flexibility across diverse infrastructures, mitigating concerns of a vendor lock-in.

Additionally, Pangeo's collaboration on Analysis Ready Cloud Optimised (ARCO) data production and the use of Zarr as a common data format create synergies with other disciplines such as bioimaging, exemplifying how cross-disciplinary efforts can be capitalised upon within the Pangeo and EOSC framework.

Limitations and future improvement

Deployment of a Binder instance with a Dask gateway is underway, and we are working to enhance Pangeo@EOSC.

Our goal is to streamline the creation of ARCO data and develop a user-friendly data catalog, enabling users to seamlessly create and share data as soon as it is generated.

Liked this #EOSCinPractice story?

Follow @EOSCFuture for more!