National Science Data Fabric (NSDF) Distinguished Speaker Series

https://nsdf-fabric.github.io/2022/03/25/nsdf-seminar-series.html

Pangeo Forge: Crowdsourcing Analysis Ready Data in the Cloud

Speaker: Ryan Abernathey, Columbia University, Department of Earth and Environmental Science and Lamont Doherty Earth Observatory
Date: April 28, 2022 12:30PM ET
How to join: https://utah.zoom.us/j/99659937052

Abstract: Analysis-ready, cloud optimized (ARCO) scientific data is essential for scalable big data analytics in the cloud. ARCO can massively accelerate statistical analysis, visualization, and machine learning workflows on large-scale scientific datasets. However, most scientific data is distributed in archival formats that are not optimized for large-scale analysis. Pangeo Forge (https://pangeo-forge.org/) is an open-source framework for data Extraction, Transformation, and Loading (ETL) of scientific data. The goal of Pangeo Forge is to make it easy to extract data from traditional data archives and deposit it in cloud object storage in ARCO format.

Pangeo Forge is made of two main components:
- Pangeo Forge Recipes: an open source Python package, which allows you to create and run ETL pipelines (“recipes”) and run them on your own computer.
- Pangeo Forge Cloud: a cloud-based automation framework which executes these recipes in the cloud from code stored in GitHub and deposits the data into cloud object storage.

By storing data recipes in version-controlled GitHub repositories, we can maintain perfect provenance information from archival repository to ARCO copy. Using Pangeo Forge, we are collaboratively populating a petabyte-scale library of open ARCO climate data distributed across multiple cloud storage services, including Open Storage Network.

Pangeo Forge is inspired directly by Conda Forge, a community-led collection of recipes for building conda packages. We hope that Pangeo Forge can eventually play the same role for datasets, encouraging open, interdisciplinary collaboration around data curation.

Bio: https://github.com/rabernat#short-biography