GeoEDF: A Workflow Framework for Composing and Executing Geospatial Research Workflows

Rajesh Kalyanam, Carol X.Song
Purdue University

Geospatial research workflows often involve data retrieval from various data sources using different protocols, data inspecting, filtering, and preprocessing to prepare the data for use with researchers' own domain-specific code, and finally analysis or simulation carried out on HPC resources. In practice these steps may comprise various scripts, custom-built code, and publicly available libraries that are then carried out on a mix of desktop tools, web-based science gateways, and HPC resources. Consequently, researchers spend an inordinate amount of time wrangling and transferring data between these locations, with the overall reproducibility of the workflow being impacted. The GeoEDF workflow framework addresses these issues, providing a framework for developing reusable building blocks that implement individual workflow steps, and a template for composing these building blocks into end-to-end research workflows that execute entirely on HPC resources.

GeoEDF workflows are conceived as an abstract sequence of data acquisition and processing steps specified in the YAML format. The actual data acquisition and geospatial processing operations are implemented in open-source, reusable Pythonic data "connectors" and "processors". A Python-based GeoEDF workflow engine transforms a YAML GeoEDF workflow into executions of connectors and processors on a HPC system, automating data transfers between steps. The GeoEDF workflow engine is currently deployed in the Jupyter computing environment on the publicly accessible MyGeoHub science gateway. Users can instantiate, execute, and monitor the status of a workflow from this Jupyter environment. The data connectors and processors are managed in GitHub where a continuous integration/continuous deployment process transforms community contributed connectors and processors into Singularity containers that can be executed on HPC resources. In this talk, we will describe the implementation of GeoEDF and demonstrate the use of the GeoEDF workflow engine on MyGeoHub via a proof-of-concept hydrology workflow that retrieves MODIS data for a certain date range and watershed (in ESRI Shapefile format) as HDF files from a NASA Distributed Active Archive Center (DAAC) and returns a new watershed shapefile with per-polygon averages for the user's choice of HDF variables.

Accepted for presentation at UCAR Software Engineering Assembly's 2022 Improving Scientific Software conference (online)
April 4-8, 2022