Abstract: GeoEDF: A Framework for Designing and Executing Reproducible Geospatial Research Workflows in Science Gateways

Rajesh Kalyanam, Jungha Woo, Lan Zhao, Carol Song
Research Computing
Purdue University
West Lafayette, USA
{rkalyana}@purdue.edu

Jack Smith
Marshall University
West Virginia, USA
smith1106@marshall.edu

ABSTRACT

Central to most geospatial research workflows are a series of data wrangling steps that need to be completed before the necessary data for a particular simulation model is available. For instance, a hydrologist developing a new streamflow prediction model may need to acquire precipitation data, geological data (such as elevation), vegetation and landcover data for their region of interest to effectively model the various operative parameters. On the other hand, an agricultural economist studying the impact of climate change on crop production and downstream economic effects will need to acquire historical production data, various climate parameters (temperature, precipitation), and economic data for agriculture-related activities. Due to the diverse nature of these datasets and their siloed hosting and availability, such data wrangling is often an ad-hoc process that combines purpose-built tools that prioritize convenience and speed over reproducibility. Consequently, researchers end up utilizing a variety of desktop tools and ad-hoc scripts, transitioning to a gateway only for compute intensive tasks or publication of results. This ad-hoc approach not just impacts research reproducibility, but also prevents the effective utilization of high-performance computing (HPC) resources to scale such research workflows.

We previously developed GeoEDF [1], an extensible geospatial data framework that allows researchers to conceive of their workflows as a sequence of abstract data acquisition and processing steps that are implemented as community contributed, containerized, reusable software building blocks. This enables researchers to accomplish the necessary data wrangling as part of an end-to-end workflow rather than as siloed processes. A complementary Python-based workflow engine enables these abstract workflows to be executed entirely in a variety of science gateway environments.

GeoEDF has now been used to develop and execute workflows in hydrology, water quality monitoring, and agricultural economics. It has also been deployed and evaluated in a variety of environments ranging from a well-established science gateway, MyGeoHub [2], as a self-contained Docker container that can be run by CI developers, and as a Jupyter notebook server container that can be integrated into JupyterHub-based gateway deployments. In our presentation we would like to demonstrate the generality of the GeoEDF workflow syntax and semantics, our continuous integration/continuous deployment (CI/CD) community contribution process, diverse research workflows from several domains, and our various deployment options.

Keywords—workflows; geospatial; reproducible; gateway; deployment

REFERENCES

[1] R. Kalyanam, L. Zhao, X. C. Song, V. Merwade, J. Jin, U. Baldos, and J. Smith. 2020. GeoEDF: An Extensible Geospatial Data Framework for FAIR Science. Practice and Experience in Advanced Research Computing. Association for Computing Machinery, New York, NY, USA, 207–214. DOI:https://doi.org/10.1145/3311790.3396631

[2] R. Kalyanam, L. Zhao, C. Song, L. Biehl, D. Kearney, I. L. Kim, J. Shin, N. Villoria, V. Merwade, MyGeoHub—A sustainable and evolving geospatial science gateway, Future Generation Computer Systems, Volume 94, 2019, Pages 820-832,https://doi.org/10.1016/j.future.2018.02.005.