EGU23-6960, updated on 24 Feb 2024
https://doi.org/10.5194/egusphere-egu23-6960
EGU General Assembly 2023

Remote Sensing Deployable Analysis environmenT

Pranav Chandramouli, Francesco Nattino, Meiert Grootes, Ou Ku, Fakhereh Alidoost, and Yifat Dzigan
Netherlands eScience Center, Amsterdam, Netherlands (p.chandramouli@esciencecenter.nl)

Remote-sensing (RS) and Earth observation (EO) data have become crucial in areas ranging from science to policy, with their use expanding beyond the 'usual' fields of geosciences to encompass 'green' life sciences, agriculture, and even social sciences. Within this context, the RS-DAT project has developed and made available a readily deployable framework enabling researchers to scale their analysis of EO and RS data on HPC systems and associated storage resources. Building on and expanding the established tool stack of the Pangeo Community, the framework integrates tools to access, retrieve, explore, and process geospatial data, addressing common needs identified in the EO domain. On the computing side RS-DAT leverages Jupyter (Python), which provides users a web-based interface to access (remote) computational resources, and Dask, which enables to scale analysis and workflows to large computing systems. Both Jupyter and Dask are well-established tools in the Pangeo community and can be deployed in several ways and on different infrastructures. RS-DAT provides an easy-to-use deployment framework for two targets: the generic case of SLURM-based HPC systems (for example, Dutch Supercomputer Snellius/Spider) which offer flexibility in computational resources; and the special case of an ansible-based cloud-computing infrastructure (Surf Research Cloud (SRC)) which is more straight-forward for the user but less flexible. Both these frameworks enable the easy scale-up of workflows, using HPCs, to access, manipulate and process large-scale datasets as commonly found in EO. On the data access and storage side RS-DAT integrates two python packages, STAC2dCache and dCacheFS, which were developed to facilitate data retrieval from online STAC catalogs (STAC2dCache) and its storage on the HPC system or local mass storage, specifically dCache. This ensures efficient computation for large-scale analyses where data retrieval and handling can cause significant bottlenecks. User-defined input/output to Zarr file format is also supported within the framework. We present an application of the tools developed to the calculation of leaf-spring indices for North America using the Daymet dataset at a 1km resolution for 42 years (~940 GiB, completed in under 5 hours using 60 cores on the Dutch supercomputing system) and look forward to on-going work integrating both deployment targets in the case of the Dutch HPC ecosystem.