Geophysical Research Abstracts
Vol. 19, EGU2017-11098-1, 2017
EGU General Assembly 2017

GSKY: A scalable distributed geospatial data server on the cloud

Pablo Rozas Larraondo, Sean Pringle, Joseph Antony, and Ben Evans

National Computational Infrastructure, Australian National University, Canberra, Australia (Pablo.Larraondo@anu.edu.au)

Earth systems, environmental and geophysical datasets are an extremely valuable sources of information about the state and evolution of the Earth. Being able to combine information coming from different geospatial collections is in increasing demand by the scientific community, and requires managing and manipulating data with different formats and performing operations such as map reprojections, resampling and other transformations. Due to the large data volume inherent in these collections, storing multiple copies of them is unfeasible and so such data manipulation must be performed on-the-fly using efficient, high performance techniques. Ideally this should be performed using a trusted data service and common system libraries to ensure wide use and reproducibility. Recent developments in distributed computing based on dynamic access to significant cloud infrastructure opens the door for such new ways of processing geospatial data on demand.

The National Computational Infrastructure (NCI), hosted at the Australian National University (ANU), has over 10 Petabytes of nationally significant research data collections. Some of these collections, which comprise a variety of observed and modelled geospatial data, are now made available via a highly distributed geospatial data server, called GSKY (pronounced [jee-skee]).

GSKY supports on demand processing of large geospatial data products such as satellite earth observation data as well as numerical weather products, allowing interactive exploration and analysis of the data. It dynamically and efficiently distributes the required computations among cloud nodes providing a scalable analysis framework that can adapt to serve large number of concurrent users.

Typical geospatial workflows handling different file formats and data types, or blending data in different coordinate projections and spatio-temporal resolutions, is handled transparently by GSKY. This is achieved by decoupling the data ingestion and indexing process as an independent service. An indexing service crawls data collections either locally or remotely by extracting, storing and indexing all spatio-temporal metadata associated with each individual record.

GSKY provides the user with the ability of specifying how ingested data should be aggregated, transformed and presented. It presents an OGC standards-compliant interface, allowing ready accessibility for users of the data via Web Map Services (WMS), Web Processing Services (WPS) or raw data arrays using Web Coverage Services (WCS). The presentation will show some cases where we have used this new capability to provide a significant improvement over previous approaches.