



NCI's GSKY - a scalable Geospatial data server

Ben Evans, Pablo Larraondo, Joseph Antony, Jon Smillie, Sean Pringle, Rui Yang, Matthew Sanderson, Chris Allen, Jingbo Wang, and Clare Richards

Australian National University, National Computational Infrastructure, ACTON, Australia (ben.evans@anu.edu.au)

Earth systems, environmental and geophysical datasets are an extremely valuable resource for a wide range of research, government, and industry applications. For researchers analysing, transforming, and integrating these large datasets into their work, the traditional approach has been to either download a relevant part of data and analyse these data subsets in an ad-hoc manner, or to invest significant work into batch processing large data and then store and organise for further analysis. This is now rapidly becoming infeasible due to the amount of storage space and data transformation work that it requires - and out of reach for most end-users that are unfamiliar with how to work with data at this scale. Recent developments in significant data repositories with integrated data processing infrastructure opens the door for new ways of processing data on demand.

The National Computational Infrastructure (NCI), hosted at the Australian National University (ANU), has developed a highly distributed geospatial data server, called GSKY which provides a new capability for high performance data analysis. GSKY is currently being used in some national and international initiatives - providing fast access to programs and tools over the network, and allowing researchers to analyse NCI's multi-petabyte nationally significant research data collections: from satellite data products, climate and weather simulations, and rich geophysics data.

GSKY supports on demand processing of data that allows interactive data exploration presented as 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). GSKY has functionality for specifying how ingested data should be aggregated, transformed and presented. It dynamically and efficiently distributes the requisite computations among computational nodes and thus provides a scalable analysis framework.

GSKY has required improvements in data management practice, ensuring that the data and service meets a new level of quality assurance to help meet data processing performance and end-user application requirements.

In this talk we will be seeking collaborative opportunities to use, improve and further develop GSKY's capability.