Virtual Geophysics Laboratory (VGL): scientific workflows exploiting the Cloud

Ryan Fraser[1], Terry Rankine[2], Josh Vote[3], Ben Evans[4], Lesley Wyborn[5]
[1]CSIRO, Perth, Australia, ryan.fraser@csiro.au
[2]CSIRO, Perth, Australia, terry.rankine@csiro.au
[3]CSIRO, Perth, Australia, josh.vote@csiro.au
[4]National Computational Infrastructure, ANU, Canberra, Australia, ben.evans@anu.edu.au
[5]Geoscience Australia, Canberra, Australia, lesley.wyborn@ga.gov.au

INTRODUCTION

The Virtual Geophysics Laboratory (VGL) is a scientific workflow portal that provides geophysicists with access to an integrated environment that exploits eResearch tools and Cloud computing technology. The VGL is a collaboration between the CSIRO, Geoscience Australia (GA) and the National Computational Infrastructure (NCI) facility and has been funded by a grant from the Education Investment Fund of the Federal Government.

The VGL provides scientists with easy agent to exploit multiple technologies provided by eResearch and Cloud in a user driven interface. The VGL was developed in close collaboration with the geophysics user community and, with representatives from GA and ANU, has been deployed directly into their environment

BACKGROUND

Over the last 50 years, Australian geoscientists in academia, in government geological surveys, and in industry have amassed substantial volumes of digital geoscience data. Significant amounts of these data are now available either via the internet or at a nominal cost of transfer. There is estimated to be up to 3 Petabytes of publically funded geoscience data sets in Australia, and the majority of this is held by GA. This volume is growing exponentially as improvements in the capability of instruments have resulted in data being gathered at a greater rate and at higher resolution. In addition to the issue of storing this data appropriately and accessibly, the resolution and size of digital geoscience data sets have reached the point where the subsequent analysis of this data is exceeding the computational infrastructure of most organisations.

In summary most organisations (research, government and industry) have the following problems:

1. Capability to store and dynamically access the data sets internally;
2. Capability to process the data sets to their highest resolution;
3. Inability to provide them online to partners, clients and stakeholders, both nationally and internationally.

For end users, rather than finding and then locally downloading and storing whole data sets for onsite processing, there is a growing case for a shift to new computing paradigms. The VGL provides a distributed system whereby a user can enter an online virtual laboratory that allows seamless access to dynamic user-defined subsets of data which can then be coupled to online software and compute resources.

The VGL provides end users with access to an intuitive, user-centered designed [1] interface. This interface incorporates access to four key elements:

1. eResearch tools such as the Spatial Information Services Stack (SISS) [2]
2. Data storage for input and output datasets on Cloud storage [3]
3. Processing of data via computing Clouds such as NeCTAR[4], NCI and Amazon [5]
4. Registration of the resultant datasets for future use and discovery

Through a workflow portal, a core element of VGL was the establishment of community to utilize and progress the technology forward. Additionally, the patterns and technologies employed are easily repurposed for other use cases beyond the geophysics, for example natural hazards, satellite processing and many other areas that require spatial data discovery and processing.

REFERENCES

1. International Organization for Standardization (ISO), *Human-centered design processes for interactive systems*. ISO 13407: 1999
2. *Spatial Information Services Stack.* Available from http://siss.auscope.org, accessed 30 June 2011
3. *Amazon Simple Storage Service*. Available from http://aws.amazon.com/s3/, accessed 15 June 2011
4. *NeCTAR Research Cloud.* Available from http://www.nectar.unimelb.edu.au/project_components/research_cloud_program, accessed 15 June 2011
5. *Amazon Elastic Compute Cloud (Amazon EC2).* Available from http://aws.amazon.com/ec2/, accessed 16 June 2011

ABOUT THE AUTHOR(S)

Ryan Fraser
Ryan Fraser is a Project Leader within CSIRO's Minerals Down Under flagship. He leads several large projects dealing with the exchange and delivery of spatial information and eResearch tools. He manages projects that focus on enabling the delivery of data in an interoperable manner to the various science domains. He has a software engineering background, has expertise in high performance data and computational technologies, and has primarily been involved in the design and execution of systems to deliver spatial information and the provision of data and computing services to the research community and industry. He leads a large team to deliver technologies to enable data exchange and orchestrate change within the community.

Joshua Vote
Joshua Vote is trained in Computer Science and has been involved in developing a broad spectrum of applications for both government and commercial organisations. His professional interests include user interface design, human computer interaction and generally making software accessible to as many people as possible.

Since joining CSIRO in 2009, Joshua has been involved with the AuScope Grid project and has taken an active role in developing the AuScope Portal and integrating it with the Spatial Information Services Stack (SISS). This has involved supporting various community standards over numerous components of the SISS and weaving them into a single consistent application.

Terry Rankine
Terry Rankine is a Research Group Leader in CSIRO's Earth Science and Resource Engineering Division. He leads the Computational Geoscience group, a capability providing science and technologies to integrate and interpret geoscience data and knowledge in order to understand, quantify, and predict geological processes. In particular, these tools have been applied in the minerals exploration context, with the aim of reducing risks and uncertainties and potentially leading to cheaper, faster discovery. Terry originally studied Computational Chemistry, and has a background in High Performance Computing, data management, data mining, and workflow engines, and various collaboration toolkits. He is currently working with Ryan Fraser and his team, combining those skills into community virtual laboratories, like the VGL.

Ben Evans
Ben Evans is the Head of the ANU Supercomputer Facility at the Australian National University. He leads projects in HPC and Data-Intensive analysis, working with the partners of NCI and the research sector.

Lesley Wyborn
Lesley Wyborn is a Senior Geoscience Advisor at Geoscience Australia and is a member of the Australian Academy of Science National Data in Science Committee, and the Executive Committee of the Earth and Space Science Informatics Focus Group of the American Geophysical Union. She is currently the GA lead on the GA/CSIRO eResearch Collaboration Project and the GA/NCI High Performance Computing Pilot Project.