Integrating 'Big' Geoscience Data into the Petascale National Environmental Research Interoperability Platform (NERDIP): successes and unforseen challenges

Lesley Wyborn
National Computational Infrastructure
Australian National University
Canberra, Australia
lesley.wyborn@anu.edu.au

Benjamin J.K. Evans
National Computational Infrastructure
Australian National University
Canberra, Australia
ben.evans@anu.edu.au

***Abstract*—The Australian Government has begun an initiative to organise publicly funded national data assets and make them accessible for research through the Research Data Services initiative (RDS), which supports over 40 PBytes of multidisciplinary data at eight nodes around Australia. One of these nodes is at the National Computational Infrastructure (NCI) that provides a national comprehensively integrated high performance computing facility. NCI is a partnership between the ANU, the Australian Bureau of Meteorology, Geoscience Australia (GA) and the Australian Commonwealth Science and Industry Research Organisation (CSIRO) and particularly focuses on Earth system sciences. As part of its activity in RDS, NCI has collocated over 10 PBytes of priority research data collections spanning a wide range of disciplines from geosciences, geophysics, environment, climate, weather, and water resources, through to astronomy, bioinformatics, and the social sciences. To facilitate access, maximise reuse and enable integration across the disciplines, data have been built into a platform that NCI has called, the National Environmental Research Data Interoperability Platform (NERDIP). The platform is co-located with the significant HPC resources: a 1.2 PetaFlop supercomputer (Raijin), and a HPC class 3000 core OpenStack cloud system (Tenjin). Combined, they offer unparalleled opportunities for geosciences researchers to undertake innovative Data-intensive Science at scales and resolutions never before attempted, as well as enabling participation in new collaborations in interdisciplinary science. However, compared with other 'Big Data' science disciplines (climate, oceans, weather, astronomy), current geoscience data management practices and data access methods need significant work to be able to scale-up and thus to take advantage of the changes in the global computing landscape. Although the geosciences have many 'Big Data' collections that could be incorporated within NERDIP, they typically comprise heterogeneous files that are distributed over multiple sites and sectors, and it is taking considerable time to aggregate these into large High Performance Data (HPD) sets that are structured to facilitate uptake in HPC environments. Once incorporated into NERDIP, the next challenge is to ensure that researchers are ready to both use modern tools, and to update their working practises so as to process these data effectively. This is an issue in part because the geoscience community has been slow to move to peak-class systems for Data-intensive Science and integrate with the rest of the Earth systems community.**

Keywords—Data-intensive Science, High Performance Data, High Performance Computing, Big Data, Geosciences, Data Platforms.

I. INTRODUCTION

The Education Investment Fund (EIF) Super Science Initiative, announced in the May 2009 by the Australian Government, included AU$50 million for a Research Data Storage Infrastructure (RDSI) project to enhance the development of distributed multi-Petabyte data centres across Australia that support collaborative access to data assets of national significance that have been collected mainly by publicly funded activities from both the research and government sectors. Ultimately, the RDSI project funded storage for 42 PBytes of data collections across eight high capacity nodes (Figure 1) that was configured to:

- Support the classes of access and retention appropriate to the discipline research data held;
- Provide a common access infrastructure to provide a uniform experience for researchers accessing the data held; and
- Provide appropriate specialist access infrastructure hosting tools appropriate to the disciplines related to these collections.

In 2013, RDSI evolved into a new project, the Research Data Services (RDS) project which changed the focus from building infrastructure that stored the multiple individual collections of data to a focus on Data as a Service in support or research communities with petascale data challenges. The emphasis was on shaping and connecting the data access and data sharing services across all eight nodes, to optimise the value of the data, and to facilitate the transition of the research community to the new paradigm of Data-intensive Science. The RDS project advocates a community driven approach (https://www.rds.edu.au/project-overview) to ensure that:

- Researchers will be able to access the data in a consistent manner which will support a general interface, as well as discipline specific access; and
- Researchers will be able to use the consistent interface established/funded by this project for access to data collections at participating institutions and other locations as well as data held at the Nodes

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

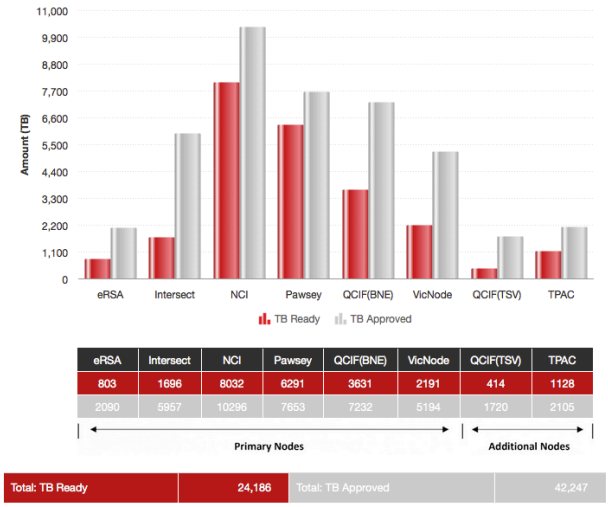

Figure 1: Volumes data approved and ingested on the Australia RDS nodes as of 1 September 2015 [1]

The RDS project recognised nine initial research domains as having petascale challenges: Earth Systems Science; Life Sciences (genomics); Medical and Health; Image Publishing Services; Astronomy; Culture and Community; Marine Sciences; Terrestrial Systems; and the Geosciences. Each of the eight nodes specialised around at least one of these domains, and collaborated with other nodes working on the same domain. In a separate process, the RDS project identified nationally significant data collections which were deemed of lasting value and importance and could be used to support new and innovative ways of research and provided funding to support their development at the most appropriate node.

The largest of these RDS funded nodes is at the National Computational Infrastructure (NCI) at the Australian National University (ANU), which has co-located and targeted 10+ PBytes of priority national and international data assets that mainly focused on the earth systems science domain including climate, weather, geosciences, environment, water resources, but also included astronomy, bioinformatics and the social sciences (http://nci.org.au/data-collections/data-collections/). The data at NCI are accessible using an integrated HPC-HPD environment, and through adopting and harmonising community standard data services, will provide unparalleled opportunities for the Australian Researchers.

The RDS funding enabled opportunistic co-location of geoscience data with other large earth systems data collections on NCI RDS node, including additional satellite remote sensing data, elevation and depth, geodesy, geophysics, geochemistry, geological data, and 3D geological/geophysical models. To standardise access and enable integration across the disciplines, the data collections have been assembled into what we have called the National Environmental Research Interoperability Data Platform (NERDIP), which is described below.

This approach has allowed integration with, as well as comparison and contrast, on the 'Big Data' methods used across the domains and how each links data to tools and compute. Overall in the geosciences, access to 'Big Data' in peak-class computing environments like large clouds and HPC are not familiar, and the current methods, practices and ambitions of geoscientists are not fully ready take advantage of this. However, others, such as the climate, oceans and weather communities who are already familiar with such infrastructure and high performance computing methods and techniques are ready to make new advances by accessing these additional rich data sources. By joining a shared computational platform, the geosciences community is now better able to trial new methods of analysis, adopting from the other communities wherever possible, and can now undertake cross-domain integrative science and at scales and resolutions not hitherto achievable.

This paper is about highlighting some of these 'Big Data' issues for the geosciences. Critical to this analysis is a review of recent changes in global computing environments and data management practices. The obvious barriers to enabling access to geosciences are heterogeneous data and scarcity of tools and people skills: the unforseen ones are the lingering impact of legacy practices and standards.

II. Changes in the Global Computational Landscape

In modern computation, there is a new strong emphasis on data management/handling to be fully integrated within workflows and systems. This strong movement includes data acquisition, compute output, data analysis, updating, sharing, accessing, and archiving.

To gain full advantage of the computing power available, the processing, people and tools need to be brought to the data. In particular, the requirement for Earth systems modelling to implement more realistic and higher fidelity physics, integrated with reference data through forcings or data assimilation methods, and over longer time-scales, has meant that the computational power has dramatically increased, and the volume and quality of data either required or generated for this work has been critical to progressing the science. This has prompted a need for the 'Big Data' collections to be re-organised as **High Performance Data** (HPD): a term defined by Evans et al. as data that are carefully prepared, standardised and structured so that it can be used in Data-Intensive Science on HPC [2]. HPD offers an opportunity to take 'Big Data' collections and aggregate multiple heterogeneous files into self-describing data arrays and data cubes that homogenize data on regional/continental scales.

Compared to the deeply computational community with climate, ocean and weather, the overall transformation and uptake by traditional geoscientists has been slower. To take advantage of these systems, the traditional geoscientist needs to move away from a local desktop approach of simple file/service discovery and data download for processing, and move towards processing through well-managed facilities with data repositories that are collocated with HPC or cloud systems. Once organized for HPD, it is then possible for these data cubes to be accessed via community standard interfaces (such OGC) to allow a broader community access - such as through using public/private clouds through to mobile devices (smart phones, tablets).

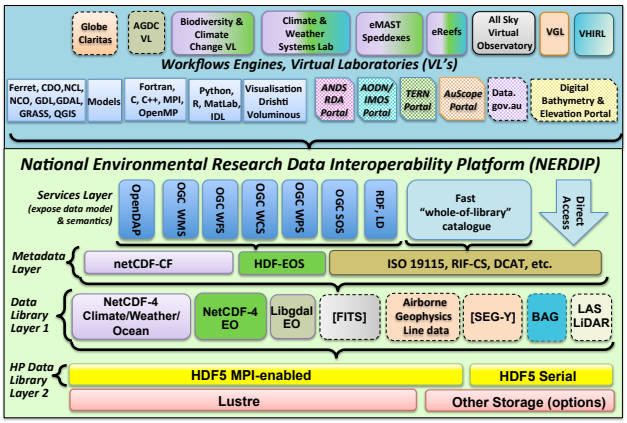

Figure 2 Architecture of the NCI NERDIP

III. The NCI NERDIP: A Platform for High Performance Data Collections

With additional funding through the RDS program, the NCI has organised data into a 10+ PBytes research data collection repository that is collocated with significant HPC and cloud resources. The NCI repository manages nationally significant data collections categorised as 1) Earth system sciences, climate and weather model data assets and products, 2) Earth and marine observations and products, 3) geosciences, 4) terrestrial ecosystem, 5) water management and hydrology, and 6) astronomy, social science and biosciences. Collectively these collections span the lithosphere, crust, biosphere, hydrosphere, troposphere, and stratosphere and they comprise one of the largest collections of Earth and Environmental data in the world at a single site. The data are largely sourced from the NCI partners (who include the custodians of many of the national scientific records), major research communities, and collaborating overseas organisations. The data are formally organised into the National Environmental Research Data Interoperability Platform or **NERDIP** (Figure 2).

At NCI, NERDIP is a key component of an integrated High Performance Computing and Data (HPC-HPD) environment, which comprises a 1.2 PetaFlop supercomputer (Raijin), a HPC class 3000 core OpenStack cloud system (Tenjin) and several highly connected large-scale high-bandwidth Lustre filesystem. Researchers have the choice of logging on to utilise the HPC infrastructure for these data collections, or accessing the data via standards-based web services. This integrated HPC-HPD environment is designed for users from high performance research that specifically need such an infrastructure to support their operations at NCI. But it is was also designed for scale-out; to allow access by individual researchers for small scale, local research projects by dynamic extraction of minor data volumes and then processing in the cloud.

The NERDIP provides a new holistic data platform within which the data are organized and made accessible, that can be utilised for multiple use cases for modelling, analysis and visualization. Data collection metadata can be searched in the NCI data catalogue (http://geonetwork.nci.org.au/) to allow easy discovery of data collections and datasets, as well as the deeper level search of data attributes. The data are fully accessible for both primary and secondary use – either on the NCI filesystems for those needing highly performing access (and hence direct access via NCI) or via data services for broader research community access.

Work on the integrated data platform was started just over 18 months ago. It is early days yet, but it is apparent that some data communities have easily moved into these HPC/HPD environments, whilst for others it has been a challenge. Those communities that are already in the 'Big Data' community and have a history of interaction with HPC computing (such as climate, weather, and ocean) have adapted readily. Getting the geoscience data into a HPD form suitable for HPC has not been easy as the current method, practices and ambitions of geoscientists are not fully ready take advantage of this and some old approaches inhibit the transition required to make full use of the potential. It is not just a matter of transferring and scaling up existing work practices: these have to be transformed into a new paradigm that involves:

- Creating cohesive, value-added data sets (data cubes, data arrays) that can be accessed within realistic time frames; and
- New tools and applications that can analyse these large scale data sets.

A. *Integrating Geoscience Data into Petascale Computing environments*

Now that the larger, cohesive data sets are being assembled on NERDIP it is clear that there will be new opportunities for geoscientists to transform the way that they do their research. Much larger data volumes can be analysed in single passes, and if required, the data can be analysed to very high resolutions. Because of the larger compute capacity available, it is no longer necessary for geoscience data from larger data sets to be averaged, subsampled, tiled, or gridded to suboptimal scales: the data can be analysed to varying resolutions at continental or even global scales.

The geosciences clearly have many 'Big Data' sets generated by co-locating hundreds if not thousands of multiple, individual data sets from copious numbers of data collection campaigns, but for uptake in HPC environments, these need to be pre-processed (e.g., calibrated, levelled, geo-located, etc.) and then aggregated into analysis ready data cubes and data arrays. The larger the number of individual data sets that are to be merged, and the longer the time frame they were collected over, the greater the chance will be for incompatible formats, varying survey parameters (line spacing, survey heights, sampling frequencies) and different standards, vocabularies and ontologies.

Such heterogeneous data sets are increasingly difficult to dynamically access within realistic time frames and there are new requirements for data to be transformed and aggregated into HPD data sets that comprise cohesive, nationally calibrated data arrays or data cubes, that are self describing, and comply with international standards that enable programmatic access to data.

B. *Incompatible data formats in the geosciences*

Incompatible data formats have been a key issue in building geoscience data sets into NERDIP. For example, NCI hosts nearly 2.0 PBytes of Earth Observation data, but there is not universal agreement as to how the data should be stored and accessed. The Landsat community has traditionally used data in raster formats, such as GeoTIFF, and then structured data collections to facilitate file downloads for local processing. In contrast, other Earth Observation communities on the NCI NERDIP (e.g., MODIS, VIIRS, AVHRR) have moved to NetCDF-CF formats that enable in-situ access via OGC-compliant data services. But ultimately, the choice of formats is not the decision of the NCI users alone, as there is an additional 700 TB of Earth Observation data on the other Australian RDS nodes that need to be taken into account. Further, Earth Observation data on the NCI node are already being used in international collaborations such as the Earth Systems Grid Federation (ESGF) and the European Union (EU) EarthServer project. Hence considerable care and broad consultation will be needed in deciding what formats Earth Observation data will be stored on NERDIP and how they will be accessed/delivered.

Likewise data formats are a key issue when trying to merge data from differing domains. This became apparent when NCI participated with CSIRO and Geoscience Australia in the development of the Virtual Hazards, Impact and Risk Laboratory (VHIRL). The aim of VHIRL was to advance natural hazard and risk modelling by accessing a collection of open community standards-based web services for both data selection and then processing of selected data. VHIRL had a clear dependency on the accessibility and availability of data suitable for hazard and risk modelling, in particular digital elevation, bathymetry and geoscience data collections. Although these collections are co-located on the NCI platform, incompatibility of standards limited the uptake of VHIRL for several codes. For example, the ANUGA Hydrodynamic Modelling code widely used to model the impacts of hazards such as tsunamis and storm surges, requires merged grids that comprise a single data file which describes the elevation at the coastal boundary and is created by merging bathymetry and topography data. Suitable LiDAR data from the elevation community was in LAS format, whilst data from the bathymetry community was in ESRI Grid, CARIS or ASCII.

A similar problem exists in geophysics where data were made available in ER Mapper or Intrepid formats, which are mainly accessible to commercial software packages: few open source programs and libraries can consume these formats. The geophysics community at NCI is now working on transforming all of the geophysics data sets into NetCDF-CF format so as to enable standardisation across this collection and to support better integration with other data communities at NCI, in particular the terrestrial and Earth Observation communities.

Data formats are a key stumbling block to the geosciences moving forward in modern HPC/HPD environments. Until now, users have been able to use differing formats to suit their end application. The new Data-intensive research paradigm will not allow the flexibility researchers previously had and they will no longer have the freedom to reformat and store major data collections to suit their own needs. The cost of storage on these major platforms is such that multiple copies of PByte-scale data sets in different formats are untenable.

However, to enable within- and cross-domain integration, data from many communities will need to be migrated to better standards that provide programmable interfaces through data services under a common model. Of all the formats available, NetCDF4/HDF5 is proving the most promising. This is largely due to the careful attention to self-describing data standards, enabling for the HPC community the realisation of layered software stacks that enable the data model to be exposed to network access, and the strong adoption across multiple communities in the earth systems sciences domains.

As the format and standards issues are not going to be resolved in the short term, one of the challenges for NCI has been to support existing domain specific techniques and methods, while carefully preparing the underlying infrastructure for the transition needed for the next generation of interdisciplinary, Data-intensive Science.

C. *New Software and Methods for the Geosciences in HPC/HPD environments.*

In recent years, with the ending of Moore's Law, computer hardware trends are towards increased parallelisation and larger memory. Although HPC and HPD can meet the high computational demand, there is still much more work to fully introduce and adopt the software and data standards for the geosciences so that they can fully exploit them. Cloud computing has lowered the barrier to scalable computing and the adoption rate is now improving.

The geosciences have for many years relied on proprietary software, particularly in the geophysics and spatial communities. In the main, the commercial community has been relatively slow to develop parallel computing methods in their codes for HPC platforms, particularly for introducing distributed memory algorithms. Rewriting to use HPC systems is a non-trivial and requires both scientific and specialised computational science skills to know the algorithms and where scalability through parallel methods can improve the performance of the code.

An alternative, potentially faster to adopt, solution, is for developers of codes and users to join in the growing global community groups that are developing Open Source software and reliable trusted software stacks designed for scalable computing environments that tackle similar geoscientific research problems.

It is also the case that the scientists running these large computations need to come to terms with CPU, memory and other resource requirements; know how to access their data; and to understand how to organise larger bodies of work to take advantage of the additional capabilities available in these environments.

In reality, operating in the new HPC environments requires a new generation of quantitative geoscientists that are more computer literate than ever before: on the use of the models, data analysis techniques, and data management awareness. This change is proving difficult for the more qualitative, descriptive areas of the geosciences.

IV. CONCLUSIONS

The next generation of large-scale Australian and International Data-intensive Earth and environmental interdisciplinary research requires the geosciences to be fully included as a core component to help researchers uncover the new fundamental insights locked in their data. NCI's NERDIP was designed at inception as a standards-based multi-purpose data platform to allow data from multiple disciplines to be combined with HPC to support the new and more complex workflows of Data-intensive Science.

It has been relatively easy to integrate data from those communities where the standards utilised are modern and capable of enabling programmatic access to the data (climate, oceans, weather). The NERDIP platform has enabled access via a variety of methods, including multiple virtual laboratories, online tools, data portals, and direct programmatic access.

We are now making the challenging step of taking the geosciences into HPD/HPC and hence preparing for the next vital component of interdisciplinary, Data-intensive Science. The data sets are generally very heterogeneous, fragmented, and the uptake of self-describing standards that enable programmatic access to data is limited. The geoscience community has also required transitions from older processes, data formats and some retooling of their applications to begin to take advantage of the new infrastructure: it was unforeseen how complex these required transitions were going to be. While progress has been slower, steady advances have still been made within areas of the geosciences, in particular geophysics.

Shared international collaborative efforts continue to be needed to realise the full potential of the new Data-intensive environments in geoinformatics - such as the Global Earth Observation System of Systems (GEOSS), Coupled Model Intercomparison Project (CMIP), ESGF, OneGeology, the Belmont Forum, the Oceans Data Interoperability Platform (ODIP) and EarthCube, which in reality are cross domain computational ecosystems. No one organisation, no one group, no one country has the required resources or the expertise. Due to the size and interdependence of the domain data, the Earth systems data have to be built as global facilities based around national institutions and international communities using international data standards. Those who cannot conform to standards will be left on their own, and unable to enter data into the shared systems suitable for globally transparent scientific research.

ACKNOWLEDGMENT

The authors wish to acknowledge funding from the Australian Government Department of Education, through the National Collaboration Research Infrastructure Strategy (NCRIS) and the Education Investment Fund (EIF) Super Science Initiatives through the National Computational Infrastructure (NCI), Research Data Storage Infrastructure (RDSI) and Research Data Services Projects.

REFERENCES

[1] Nationally significant collections stored at RDS nodes. Accessed 1 September 2015 https://www.rds.edu.au/nationally-significant-collections-stored-nodes

[2] B. Evans, L. Wyborn, T. Pugh. C.Allen, J. Antony, K. Gohar, D. Porter, J. Smillie, C. Trenham, J. Wang, A. Ip, G. Bell, "The NCI High Performance Computing and High Performance Data Platform to Support the Analysis of Petascale Environmental Data Collections" in Environmental Software Systems, Infrastructures, Services, Applications, I R. Denzer, R.M. Argent, G Schinak, J.Hrebicek, Eds., FIP AICT 448, pp. 569–577, 2015.