Research Cloud Data Communities

Bernard Meade[1,2,*], Steven Manos[2], Richard Sinnott[2], Christopher Fluke[1], Dirk van der Knijff[2], Andy Tseng[2]
[1]Swinburne University of Technology
[2] The University of Melbourne
*Corresponding author email: bmeade@unimelb.edu.au

Big Data, big science, the data deluge, these are topics we are hearing about more and more in our research pursuits. Then, through media hype, comes cloud computing, the saviour that is going to resolve our Big Data issues. However, it is difficult to pinpoint exactly what researchers can actually do with data and with clouds, how they get to exactly solve their Big Data problems, and how they get help in using these relatively new tools and infrastructure.

Since the beginning of 2012, the NeCTAR Research Cloud has been running at the University of Melbourne, attracting over 1,650 users from around the country. This has not only provided an unprecedented opportunity for researchers to employ clouds in their research, but it has also given us an opportunity to clearly understand how researchers can more easily solve their Big Data problems. The cloud is now used daily, from running web servers and blog sites, through to hosting virtual laboratories that can automatically create hundreds of servers depending on research demand. Of course, it has also helped us understand that infrastructure isn't everything. There are many other skillsets needed to help researchers from the multitude of disciplines use the cloud effectively.

How can we solve Big Data problems on cloud infrastructure? One of the key aspects are *communities based on research platforms*: Research is built on collaboration, connection and community, and researchers employ platforms daily, whether as bio-imaging platforms, computational platforms or cloud platforms (like DropBox).

There are some important features which enabled this to work.. Firstly, the borders to collaboration are eased, allowing communities to access infrastructure that can be instantly built to be completely open, through to completely closed, all managed securely through (nationally) standardised interfaces. Secondly, it is free and easy to build servers and infrastructure, but it is also cheap to fail, allowing for experimentation not only at a code-level, but at a server or infrastructure level as well. Thirdly, this (virtual) infrastructure can be shared with collaborators, moving the practice of collaboration from sharing papers and code to sharing servers, pre-configured and ready to go. And finally, the underlying infrastructure is built with Big Data in mind, co-located with major data storage infrastructure and high-performance computers, and interconnected with high-speed networks nationally to research instruments.

The research cloud is fundamentally new in that it easily allows communities of researchers, often connected by common geography (research precincts), discipline or long-term established collaborations, to build open, collaborative platforms. These open, sharable, and repeatable platforms encourage coordinated use and development, evolving to common community-oriented methods for Big Data access and data manipulation.

In this paper we discuss in detail critical ingredients in successfully establishing these communities, as well as some outcomes as a result of these communities and their collaboration enabling platforms. We consider astronomy as an exemplar of a research field that has already looked to the cloud as a solution to the ensuing data tsunami.

Keywords: Big Data, cloud computing, virtual infrastructure, virtual machines, platforms, communities, discipline-specific support

Index Terms: Big Data, The cloudscape

Introduction

The research landscape is changing rapidly. More and more, we are being confronted by the "Big Data" revolution. Yet research methodologies are sometimes slow to change and it can seem an almost insurmountable challenge to draw meaningful research from the "data deluge". The timely arrival of cloud computing has been held up as a way for researchers to engage with this new data paradigm, providing a simple, efficient way to adopt Big Data into research activities. But the promise and the reality are often separated by a skills chasm.

The NeCTAR (National eResearch Collaboration Tools and Research – www.nectar.org.au) Research Cloud (RC) was launched in February 2012, with the lead node hosted at the University of Melbourne (NeCTARWeb 2012). By 2014, seven more nodes are expected to come online around the country. Over 1,650 researchers have begun using the RC to underpin their research, with several research groups hosting virtual laboratories directly tackling Big Data problems. From web servers and blog sites, through to ad hoc cluster computing, the RC is in active use across Australia.

Each of these research activities helps us understand better how to use the cloud computing infrastructure to address Big Data challenges. It is clear that the two most significant elements are the combination of community and research platforms. Research is not conducted in isolation, but in collaboration. Connection and collaboration technologies are essential elements in both forming and supporting such communities.

The RC gives researchers an opportunity to change the way they engage with Big Data. It is a new way to work and no doubt this will be challenging for many. But the potential benefits of forming communities with Big Data at the core, connected through research precincts or via disciplines, are enormous. Collaborative platforms that are sharable and repeatable, can be open or tightly secured encourage coordinated use and development, fostering community-orientated methods.

In this paper we discuss in detail the specifics of establishing these communities, as well as some of their research outcomes derived from use of collaboration enabling platforms. We start with a general background to the concepts of Big Data and cloud computing, followed by a discussion of the NeCTAR Research Cloud specifically, focusing on those aspects that can benefit data communities, as well as addressing some of the potential risks. Next, we look at Communities and introduce the idea of Virtual Laboratories, highlighting some of the current projects already running on the RC. Using astronomy applications as an example we then discuss cloud computing platforms, followed by a discussion of the relationship between cloud computing and HPC. We also consider the challenges and potential of cloud computing in terms of data management and provenance, as well as the need for effective integration into an institution's IT ecosystem. Finally we discuss what we might expect the research landscape to look like a few years from now.

Background

New research instruments, sensor networks and computer simulations are producing data at an unprecedented rate. Scientific disciplines, such as astronomy, have been dealing with Big Data challenges for several years. However, the value of Big Data is now being recognised across many more "non-traditional" fields, e.g. the humanities and social sciences.

What is Big Data?

Big Data means different things to different people, but the generally accepted concept is that the accumulated data exceeds the capacity of typical or traditional processing means. See Table 1 for some examples of Big Data. The size of data collections stored in services such as Research Data Storage Infrastructure (RDSI) will most likely follow a power distribution, where there are a few very large collections (1PB+) such as the LHC, EBI, etc., more biomedical and imaging DBs on the scale of 100's of TB, and then 1000's of smaller - but equally important datasets - such as survey results - in the order of GB's or MB's. This can mean the data volume exceeds the capacity of local databases, or even local hard-drives, or it may mean the data is accumulating too fast for a desktop computer to process. It can also mean the data required is sourced from a variety of repositories, and is

heterogeneous in nature. In all cases, Big Data means local storage and manipulation is impractical at best, impossible at worst.

Resource	Data volume
Large Hadron Collider (LHC)	1TB/second, 13PB in 2010
Human Genome (e.g. European Bioinformatics Institute (EBI))	100GB/personal human genome, 30,000 human genomes processed in 2011
Research Data Storage Infrastructure (RDSI)	Expected to exceed 100PB

Table 1. Examples of Big Data [source: (Brumfiel 2008; "Data, Data Everywhere | The Economist" 2010; "Another Node Announced for Research 'big Data' Project - Research Data Storage Infrastructure - The University of Queensland, Australia" 2012)]

The best use of these expanding networks is to provide access to remote data stores for researchers. To paraphrase a saying, if the data won't come to the computation, then the computation must go to the data. Indeed, using remote computing with services such as VNC (Virtual Network Computing), researchers are provided with an interface to a virtual desktop that operates very much like the one on the local computer. With the explosion of mobile devices such as smart phones and tablets, the performance of the virtual interface is every bit as good on an iPad as it is on the very latest desktop computer, provided sufficient network bandwidth is available.

What is cloud computing?

Cloud computing offers a way to obtain computing resources on demand, rather than having to commit to potentially unnecessary hardware. It allows an economy of scale to the service provider, and provides consumers with a cost effective way of harnessing the required computing power. For example, by purchasing an amount of computing resource or storage from a cloud provider, a user can ensure that they only pay for what they use, as opposed to a computer under a desk that is paid for whether it is being used or not. The US Federal Government created its "Cloud First" policy to ensure departments investigated the potential of cloud services before investing in IT (Kundra 2011).

Cloud computing is also very attractive for a web service, particularly when the server experiences sporadic loads. For example, a web resource that experiences low usage by students during semester might come under significant strain during exam time. Rather than pay for a high-end computer that can handle the maximum load, and have it sit almost idle for most of the year, a cloud computing hosted virtual web server can exist as a small server costing very little until the demand exceeds a certain level, when additional servers are automatically brought online to cope, instantly balancing the load. This expansion on demand is known as cloud-bursting or elastic cloud as typified by Amazon's Elastic Compute Cloud offerings ("Amazon Elastic Compute Cloud (Amazon EC2), cloud computing Servers" 2013).

Cloud computing also provides an opportunity to test configurations without risk. Launching an instance of a VM typically happens in a matter of minutes and can be terminated just as quickly, making it "cheap to fail". The image can be cloned and modified, launched several or even hundreds of times. It can even provide an ad hoc expansion to an HPC cluster. Images can be used like templates, preconfigured and shared like documents, with links to databases and application already installed, ready to go.

Cloud computing is typically built on big infrastructure, and is therefore ready to handle Big Data. The high-speed interconnects provide excellent access to data stored either adjacent to the compute resource, or via multi gigabit links to other parts of the country, or even the world.

NeCTAR Research Cloud

The NeCTAR Research Cloud was launched in February 2012 and in its first fifteen months of operation has seen over 1,650 research individuals and more than 110 projects sign on. Berriman et. al. (2010; 2013) provide an excellent summary of cloud computing in scientific workflows when comparing commercial clouds such as Amazon and institutional HPC facilities. However, the NeCTAR RC blazes a new trail for research communities. Rather than weighing the cost benefits of internal resources versus commercial cloud providers, it aims to weigh the value of research opportunities and outcomes against the cost of purchasing and supporting institutional facilities.

There are many directly measurable benefits of cloud computing, and these become even more obvious in the NeCTAR RC context. Initial outlay of capital, operational costs of maintaining space for equipment, power, cooling are easily measured. However, the most significant benefits stem from the fact that hitherto impractical research activities become viable. Many researchers confronted by Big Data are finding new ways to engage with their data, and ultimately produce valuable new research.

Community benefits of the Research Cloud

There are many benefits to using the RC as opposed to deploying your own infrastructure. Understanding the value of communities around Big Data is key to successfully utilising the RC to extend and enhance collaborations.

1. *Borders to collaboration are eased.* Communities need to be able to share resources, and research collaborations are often national if not global in nature. Fast and efficient sharing of resources, either as infrastructure or information, is essential to the success of these teams. Having the ability to create "instant" computing resources as required, and having full control over the access to that resource, allows researchers to work together no matter where they are in the world, in a secure environment, and to make their work available to a global audience as necessary.
2. *Free and easy.* The NeCTAR RC is free for Australian researchers, allowing them to build virtual servers and infrastructure as required to facilitate their research. This has the benefit of allowing for experimentation, with servers able to be launched and terminated with ease and without penalty. Moving beyond code testing, researchers can now test servers and services in ways that were simply impractical, impossible or simply too expensive before.
3. *Sharing infrastructure.* Perhaps the most exciting aspect of virtual infrastructure is that it can be shared between/across collaborations. In the past, sharing code, systems and results between remote collaborators, writing papers together over long distances has been non-trivial. To develop/integrate code from multiple sources often required researchers to be physically co-located. With RC, virtual servers can be connected to from anywhere in the world by multiple people concurrently. What's more, the actual virtual machine (VM) itself can be shared, cloned and archived. Others can extend the research activity by launching a copy of a preconfigured VM, running simulations or data interrogations with their own parameters. And this can happen in a matter of mere minutes (Hiden et al. 2013).
4. *Big Data Infrastructure.* Today's data centres are built with the capacity to handle Big Data. Physical machines are packed closely together with extremely fast interconnects between them. These racks of machines are in turn connected to high-speed Internet backbones, giving the very best speeds available to other facilities. This greatly exceeds the capability of a typical desktop computer. For many researchers working with Big Data, the proximity of the data to the processing facility is a necessity.

Risks of the Research Cloud

As with any new technology, there can be significant risks associated with early or insufficiently planned adoption. Cloud computing in general and RC specifically is not a panacea to Big Data difficulties. It is important that institutions and researchers consider their own application before employing the RC for their research (Canon 2011).

1. *Ethical considerations.* Many datasets have strict use controls that limit the way data can be distributed. In some cases, this may preclude storing or transferring the data via public networks.
2. *Security management*. Like any server operating online, there is an onus on the operator to ensure the system cannot be easily compromised and exploited. For many researchers, this will mean employing a system administrator to maintain their servers. The lack of financial barrier to entry may tempt cash strapped researchers without sufficient experience to try to manage their own server, which may result in their systems being compromised. There is also the chance that data stored online might be compromised if the hosts security prevention measures are overcome. In recent years, even high-security services such as those used at financial institutions, have been shown to be not immune to breaches, so it is reasonable to expect that successful attacks will happen for services running on the RC, either through unpatched exploits in the system or inadequate security measures on the VMs themselves.
3. *Network dependence.* While many researchers are already dependent on the presence of a robust network, for cloud computing it is imperative. Large institutions such as the University of Melbourne have high-bandwidth and high-quality network services. However, it is essential that researchers consider the stability of their own environment before committing themselves and their research to the RC. Fortunately, most Universities and research institutions around Australia have excellent network infrastructure, and connectivity to the wider community via broadband networks like the NBN (National Broadband Network) ensures that the reliability and bandwidth of networks will only get better.
4. *Sustainability and technical capabilities.* It is hard to predict the impact of some of the challenges relating to the long-term sustainability of cloud services. At this time, Government funding for the NeCTAR project is uncertain beyond 2014, and the potential for the service to be fully funded by research institutions independently is by no means certain. Sustainability also relies on the persistence of technical capabilities of those creating, operating and maintaining VMs and Virtual Laboratories (VLs). There is a risk that without adequate documentation, once systems are put in place, the processes for establishing new or improved services could be lost.

Communities

Research communities are the backbone of research. The communities can form around disciplines, institutions, and even methodologies. Communities provide support and form the basis of the peer-review system. The ‘dude who knows about computers’ is often your PhD student or a postdoc.

In the era of Big Data, communities can also form around datasets and data collection resources and methods. Because the value of the data goes beyond the initial collection motivation, further research based on a dataset or collection of sets is brought about by community awareness. This potential for reuse of data for entirely new research is a key ingredient to justifying expenditure on high-end resources, rather than myriad low-end resources.

Virtual Laboratories

The NeCTAR RC is aiding the formation of data communities with the VL concept. A VL, also known as a remote laboratory, is an online resource that provides remote access to data collection and analysis tools, and/or data archives. A VL will typically allow resources to be used in very much the same way as if they were stored locally, however, the potential for collaboration is greatly enhanced. Access to the VL is no longer determined by proximity to the computation or the data collection equipment. Processing the data is equally simplified. Table 2 shows some of the current RC Virtual Laboratories.

Virtual Laboratory	Purpose
Virtual Geophysics Laboratory	Scientific workflow portal for Geophysicists
Virtual Genomics Laboratory	"Sequence-oriented" genome-related molecular bio-sciences
Marine Virtual Laboratory	Marine and ocean-climate science
The All Sky Virtual Observatory	"Hardware, tools and services to bring together data from radio telescopes, optical telescopes and supercomputers, covering all parts of the southern sky, under a Virtual Observatory"
Climate and Weather Science Laboratory	"Support an intrinsically complex Earth-System Simulator that allows scientists to simulate and analyze climate and weather phenomena."
Humanities Networked Infrastructure (HuNI)	Unlocking and uniting Australia's cultural data
Characterisation Virtual Laboratory	Research environments for exploring inner space

Table 2. NeCTAR Research Cloud Virtual Laboratories [source: (NeCTARWeb 2012)

Cloud computing platforms for astronomy

As an example of the way RC can support scientific communities, we look to a field where Big Data is already a reality: astronomy. For astronomers, the challenge of coping with new telescopes such as the Square Kilometer Array (SKA) is a real and present concern. While network bandwidths are increasing, astronomers are loath to forego their traditional approach to interrogating data. However, in the next few years, even with significant expansion of bandwidth, the networks will be overwhelmed by the appropriately nicknamed "data tsunami" (Berriman & Groom 2011). Table 3 shows some examples for Big Data in astronomy.

Resource	Data Volumes
Sloan Digital Sky Survey	357 million unique objects, 15.7TB FITS images, 26.8TB Other data objects, 18TB catalogs
Large Synoptic Survey Telescope	Will capture 20TBs/night, 60PBs over ten years
Australian Square Kilometer Array Pathfinder	72Tb/second raw data stream, enough to fill 120 million Blu-Ray discs/day
Square Kilometer Array	~1EB/day (2x global daily internet traffic, 100x Large Hadron Collider data collection)

Table 3. Examples of Big Data in Astronomy [source: ("SDSS Data Release 7" 2013), ("LSST Data Management | LSST" 2013), ("CSIRO Launches the ASKAP Telescope – and a New Chapter for Radio Astronomy Begins" 2012), ("Amazing Facts - SKA Telescope" 2013)]

Like many disciplines, researchers in astronomy have been confronting the problem of working with datasets that are simply too large to transfer. The Big Data challenge is currently met by remotely processing data using collocated HPC facilities, such as the International Virtual Observatory Alliance (IVOA) ("International Virtual Observatory Alliance" 2013). However, HPC resources are

not always a viable solution for many researchers. For one thing, there is a significant learning curve in developing suitable code to run efficiently on such systems.

A model adopted by international facilities like the CyberSKA in Canada ("CyberSKA: Authorized Application Tokens" 2012; Willis 2011) or "OneSpaceNet" from the National Institute of Information and Communications Technology, Japan (NICT) (Morikawa et al. 2010) is that of a portal to a remote processing facility. Similarly in the case of the RC, the portal is created in virtual machine hosted in the Cloud. This portal already has links into both HPC and data storage facilities, often with the two connected with very high-speed interconnects. The user can submit requests via the portal to the HPC system, often with preconfigured widgets, which in turn draws on the data from the connected store, either adjacent to the facility or from wherever it is located on the globe. Only the results of the processing are sent to the researcher (see Figure 1). This methodology has also been adopted by the Canadian Astronomy Data Centre in the form of the Canadian Advanced Network for Astronomical Research (CANFAR) (Ball 2012).

The collaborative potential of this approach is for several researchers to work together to determine the parameters of the request, with the results distributed to each researcher simultaneously. In the case of astronomy, these results could be ultra-high resolution images automatically displayed on remote tiled display walls. Being able to observe the images and discuss the results in real-time, would allow the researchers to refine the parameters and resubmit their query. For astronomers interested in real-time quality control of terabyte-scale radio astronomy data before the raw data gets erased, this may allow for essential refinement of parameters and result in a significantly better scientific outcome.

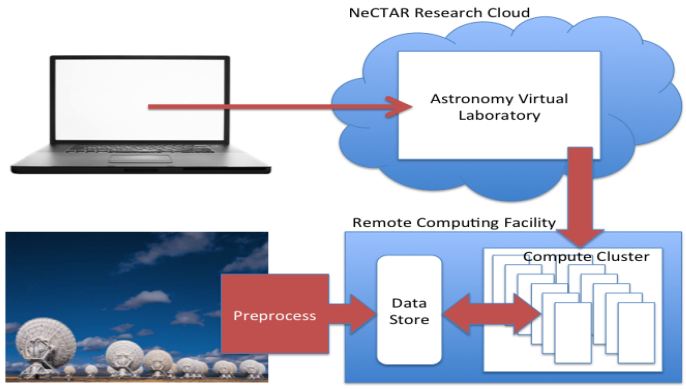

Figure 1: Model for an astronomy virtual laboratory

Recent results from the Space Telescope Science Institute (STScI) shows that more papers are being produced using archived data than from new data ("HST Publication Statistics" 2013; Berriman & Groom 2011). This means the value of the stored data has tipped from validation of research to maximising the scientific return of captured data. These massive datasets can therefore become the core of a research community. The reuse of data increases the potential of research instruments and aids in the justification of expenditure.

Forming communities around data and data-generating instruments, such as telescopes and HPC clusters, is easily facilitated using the Research Cloud. For example, a research group investigating a particular data set,can produce a VM with all their code and links to the dataset in place. This VM can be stored along with data for both provenance and sharing. Another group wishing to extend the original research could clone the VM and conduct new investigations, furthering the original research.

Archiving and provenance as ends in themselves are also better served using VMs that can be backed up and transferred at will. As technology advances, out-dated equipment is typically decommissioned, sometimes to the detriment of being able to reproduce the original environment of the research. With a VM, the virtual environment in which experiments and data were created can be persisted, however this also has challenges that must be overcome, e.g. for how long should they be stored?

Research Cloud and HPC

HPC can be seen as the forerunner to cloud computing. Rather than utilising local desktop computation resources, HPC allowed users to take advantage of available compute cycles on a massive remote resource. cloud computing achieves a similar outcome. Both HPC systems and cloud computing are based on clusters of computers interconnected by some high-speed network, often managed by a dedicated additional (head) node.

Cloud computing and HPC differ in that HPC systems are predominantly task based whereas cloud computing is more often characterized as Infrastructure as a service (IaaS). On HPC systems, users submit tasks to a queuing system, which then allocates resources to the task as they become available. User tasks all run in the same software environment. cloud computing on the other hand allows the users to develop VMs with their chosen software environment, which they then submit to an allocation system that allocates them the resources they need.

The major differences are that on HPC systems, users are guaranteed exclusive access to the allocated resources for a limited time and sharing is accomplished by having tasks wait on a queue until resources become available, while in the Cloud resources are shared by being oversubscribed, but VMs are allowed to be persistent. This leads to the two systems having different best use situations. HPC, as the name implies, is most suited to well defined and bounded computational problems, whilst Cloud is most suited to ongoing continuous loads. Cloud systems also have the capability to add VMs in a dynamic fashion to cope with varying demand in a way that HPC systems find difficult, and this makes them suited to many collaborative activities where demand is hard to predict (Cohen et al. 2013; Suresh, Ezhilchelvan, and Watson 2013).

Data management and provenance

As research outcomes becoming more varied and versatile, data management becomes a crucial component of research when dealing with massive datasets. It is essential for research institutions to establish relevant policies and services in order to address these ever-increasing Big Data challenges (Turilli et al. 2013).

Reliability is particularly fundamental when it comes to managing high volumes of research data. Transferring the research data to a trusted cloud environment, that has been set up specifically to accommodate researcher's needs, dramatically reduces the risks of their valuable data being lost or stolen, at the same time lowering the time and resources needed compared to managing data stored in different locations.

For instance, in 2011, the Higher Education Funding Council for England ("Higher Education Funding Council for England (HEFCE)" 2013) has announced the availability of the Universities Modernisation Fund (UMF) to assist UK universities and colleges to take the advantage of the new cloud computing technology to provide more efficient cloud-based services that can be utilised and shared by all research communities.

Three key areas were identified in the UMF initiative:

- *Infrastructure as a Service* (IaaS) offers access to virtual servers, data storage and high-performance computation;
- *Platform as a Service* (PaaS) provides virtual tools for researchers to develop and host individual customised applications; and
- *Software as a Service* (SaaS) enables the users to publish their applications online for easy public access.

Another critical element of cloud-based data management is the data provenance. Data provenance is important because not only does it identify the source or origin of the data, but also ensures its integrity and quality as well. The Open Provenance Model ("The Open Provenance Model" 2013), for example, is a community-driven model providing guidelines on how to allow provenance information to be exchanged between systems which in turn enables developers to build and share tools that operate on the same agreed provenance model.

Cloud computing enables data to be stored and accessed from the very same shared, remote environment as software and computation power. It empowers researchers with a greater control in what they could do with their research data better than they could have imagined which leads to a more productive research experience.

Cloud as part of an effective institutional IT ecosystem

Sustainable research communities need a good base to be built upon. To tune this base of services to meet the needs of academics is often seen as too challenging. This is understandable as the Research IT environment is quite complex. The customers come from diverse disciplines, each with their own tools, data formats, experience levels and expectations of quality and price ('but DropBox is free?'). Users are geographically dispersed, academics consume collaborations, not services, yet we provide services. Innovation is occurring at breakneck speed elsewhere on the internet, injecting free and easy to use services direct to academics. So, what is the role of the Research Cloud and the institution more broadly in that environment? It is to complement the evolving continuum of services that are provided by local, departmental, faculty, state and national levels, as well as the myriad of other service providers.

However, the final hurdle often remains the incompatibility of the traditional IT helpdesk with researchers. The problem here is that "The very first assumption about an IT helpdesk is that the researcher will know that IT can help them with their problem." The mapping of research problem to IT problem is often the biggest hurdle. This is where growing communities is imperative. They can enable researchers to identify their IT problem more clearly and in context of their discipline, and thereby begin a course of action to solve their problem.

To meaningfully support data communities, IT services need to be made up of a few things to be effective:

1. Good communications, helping researchers understand the benefits in a way that are adapted to discipline-specific audiences and skill levels
2. Community & connection & trainer knowledge
3. Flexible underlying (technical) services that give users *full control* – Academics are very self-sufficient, so enabling them to take ownership and control of their services is key (e.g NeCTAR dashboard).

Discussion

It seems inevitable that cloud computing will become standard practice, even to the point of overtaking the typical desktop computer. Laptops, tablets and even mobile phones now provide our typical access to the network resources and this will only increase, probably to the point of rendering a local, "anchored to the desk" PC redundant. Our work activities are also shifting to cloud platforms, such as online email, web browsing, journal access and office suites such as Google Docs or MS Office 365. We are already using many cloud platforms and in the future, the seamless integration of these environments will possible (Fransham et al. 2010; Armstrong et al. 2010).

In the next few years, e-Research will have evolved to simply being Research. Researchers will expect a high-bandwidth, "always there" network with simple and efficient access via devices they carry on their person. The data collected for their research will be entirely managed in datacenters across the globe and will be accessible by others in their research community, and beyond. They will also have access to data collected by others, with little difference in procedure between newly collected data and archived material.

When a researcher needs data to drive their research or to support their hypotheses, they will be able to access relevant Big Data stores almost instantly. Where archived data lacks appropriate information, researchers will be able to collect new data from remote facilities, contributing to these online datasets. Research students will be able to complete their research degrees using nothing but archived data. Research communities will collectively decide on the use of limited access facilities such as telescopes, capturing datasets that will satisfy the largest number of research activities. All collected data will also be available to citizen scientists, who in turn will be able to work with research communities to aid the research endeavours.

Big Data and cloud computing will underpin the majority of research activities in the next few years. Whether as primary methods of supporting new research or as supplement, both Big Data and cloud computing will become so ingrained in research methodology and computing in general, that like the "e" in e-Research, they will simply merge into the term, "Research".

Conclusion

Big Data and cloud computing have already begun to change the research landscape. Researchers have begun to embrace both in an effort to continue to produce cutting edge research. Big facilities like the Pathfinder projects for the Square Kilometre Array and the Large Hadron Collider produce Big Data, but Big Data can also come from sensor networks and crowd-sourced repositories. The volume of data being captured often provides a resource well beyond the original purpose, and it heralds a new way of thinking for many researchers. New skills are needed and this is where communities and the associated platforms are critical to success.

Over the next few years, cloud computing services like NeCTAR RC will prove key to the development of research data communities. With six nodes online by mid-2014, NeCTAR RC will represent a crucial computation resource for a wide variety of projects. Virtual Laboratories from numerous disciplines will exist, with dozens of communities forming around these resources. Communities will develop platforms that will be able to cross disciplines, and make the use of Big Data a natural extension of research activity.

The next few years will provide an opportunity to observe and understand how cloud computing and Big Data changes how researchers work. The combination of community and research platforms will enable far greater collaboration and in turn, better research outcomes. The reuse of platforms and Big Data datasets will be made possible by the ability of cloud computing proliferate customized VMs throughout a research community.

This future will not be without challenges of its own. It is imperative the due diligence be paid to issues such as security and skills development, as well as improving the stability of the underpinning technology. As more research finds its way into the cloud, frailties of the system will be exposed, and will need to be addressed decisively.

While these risks exist and need to be attended, the potential benefits are enormous. The simple fact that Big Data offers such a rich opportunity for research, and is reusable in ways beyond the original purpose, justifies the effort to capture and retain this scale of information. Research communities that form in precincts, around disciplines or even around Big Data, can create collaborative platforms that are shareable and repeatable. Adept users can manage their VMs fully, creating open systems for the wider community to use, or highly secured systems to protect valuable or sensitive data.

The future of cloud computing is all but assured, growing with the same inexorability as the Internet itself has over the last decade. Provided we understand this growth and the opportunities it presents, it can only serve to enrich research as we know it.

References

Amazing Facts - SKA Telescope. (2013). Retrieved January 17th, 2013 from http://www.skatelescope.org/media-outreach/fun-stuff/facts-figures/.

Amazon Elastic Compute Cloud (Amazon EC2), cloud computing Servers. (2013). Retrieved January 17. http://aws.amazon.com/ec2/.

Another Node Announced for Research 'big Data' Project - Research Data Storage Infrastructure - The University of Queensland, Australia. (2012).Retrieved from http://www.rdsi.uq.edu.au/nodes-announced.

Armstrong, P., A. Agarwal, A. Bishop, A. Charbonneau, R. Desmarais, K. Fransham, N. Hill, I. Gable, S. Gaudet, and S. Goliath. (2010). Cloud Scheduler: a Resource Manager for Distributed Compute Clouds. arXiv Preprint arXiv:1007.0050. Retrieved from http://arxiv.org/abs/1007.0050.

Ball, N. M. (2012). Astroinformatics, cloud computing, and New Science at the Canadian Astronomy Data Centre. In *American Astronomical Society Meeting Abstracts* 219.219:145.11. American Astronomical Society Meeting Abstracts.

Berriman, Bruce, and Steven L. Groom. (201). How Will Astronomy Archives Survive the Data Tsunami? In *Communications of the ACM* 54 (12) (December 1): 52. doi:10.1145/2043174.2043190.

Berriman, G. B., E. Deelman, G. Juve, M. Rynge, and J. S. Vöckler. 2013. The Application of cloud computing to Scientific Workflows: a Study of Cost and Performance. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371(1983). Retrieved from http://rsta.royalsocietypublishing.org/content/371/1983/20120066.short.

Berriman, G. B., G. Juve, E. Deelman, M. Regelson, and P. Plavchan. (2010). The Application of cloud computing to Astronomy: A Study of Cost and Performance. In *e-Science Workshops, 2010 Sixth IEEE International Conference, 1–7.* Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5693133.

Brumfiel, Geoff. (2008). LHC by the Numbers. In *Nature*, September 9. doi:10.1038/news.2008.1085. Retrieved from http://www.nature.com/doifinder/10.1038/news.2008.1085.

Canon, S. (2011, June). Debunking some common misconceptions of science in the cloud. In Raicu, I Beckman, P and Foster, I (Chairs), *ScienceCloud2011, 2nd Workshop on Scientific Cloud Computing*. Workshop co-located with the ACM HPDC 2011 (High Performance Distributed Computing), San Jose, California, June 8th 2011. Retrieved from http://datasys.cs.iit.edu/events/ScienceCloud2011/

Cohen, J., I. Filippis, M. Woodbridge, D. Bauer, N. C. Hong, M. Jackson, S. Butcher, D. Colling, J. Darlington, and B. Fuchs. (2013). RAPPORT: Running Scientific High-performance Computing Applications on the Cloud. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1983). Retrieved from http://rsta.royalsocietypublishing.org/content/371/1983/20120073.short.

CSIRO Launches the ASKAP Telescope – and a New Chapter for Radio Astronomy Begins. (2012).Retrieved from http://theconversation.edu.au/csiro-launches-the-askap-telescope-and-a-new-chapter-for-radio-astronomy-begins-9991.

CyberSKA: Authorized Application Tokens. (2012) Retrieved February 10, 2013 from. http://www.cyberska.org/pg/oauth/catalogue?offset=0.

Data, Data Everywhere (2010) In| *The Economis*t. Retrieved from http://www.economist.com/node/15557443.

Fransham, K., A. Agarwal, P. Armstrong, A. Bishop, A. Charbonneau, R. Desmarais, N. Hill, et al. (2010). Research Computing in a Distributed Cloud Environment. In *Journal of Physics: Conference Series* 256 (1): 012003.

Hiden, H., S. Woodman, P. Watson, and J. Cala. 2013. Developing Cloud Applications Using the e-Science Central Platform. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 371 (1983). Retrieved from http://rsta.royalsocietypublishing.org/content/371/1983/20120085.short.

Higher Education Funding Council for England (HEFCE). (2013). Retrieved January 17, 2013 from http://www.hefce.ac.uk/.

HST Publication Statistics. (2013). Retrieved January 17, 2013 from. http://archive.stsci.edu/hst/bibliography/pubstat.html.

International Virtual Observatory Alliance. (2013). Retrieved January 17, 2013 from http://www.ivoa.net/.

Kundra, V. (2011). Federal cloud computing Strategy. White House,[Chief Information Officers Council] Retrieved from http://www.theresearchpedia.com/sites/default/files/Federal%20Cloud%20Computing%20Strategy.pdf.

LSST Data Management | LSST. (2013). Retrieved January 17, 2013 from. http://www.lsst.org/lsst/science/concept_data.

Morikawa, Y., K. T. Murata, S. Watari, H. Kato, K. Yamamoto, S. Inoue, K. Tsubouchi, et al. (2010). A Science Cloud: OneSpaceNet. In *AGU Fall Meeting Abstracts*. December: D5.

NeCTARWeb. (2012). Home | NeCTAR. Retrieved from http://www.nectar.org.au/home.

SDSS Data Release 7. (2013). Retrieved January 17, 2013 from. http://www.sdss.org/dr7/.

Suresh, Visalakshmi, Paul Ezhilchelvan, and Paul Watson. (2013). Scalable and Responsive Event Processing in the Cloud. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 371. doi:10.1098/rsta.2012.0095. Retrieved from http://rsta.royalsocietypublishing.org/content/371/1983/20120095.abstract.

The Open Provenance Model. (2013). Retrieved January 17, 2013 from. http://openprovenance.org/.

Turilli, Matteo, David Wallom, Chris Williams, Steve Gough, Neal Curran, Richard Tarrant, Dan Bretherton, et al. (2013). Flexible Services for the Support of Research. In *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Science.* 371.

doi:10.1098/rsta.2012.0067. Retrieved from http://rsta.royalsocietypublishing.org/content/371/1983/20120067.abstract.

Willis, A.G. (2011). The Canadian CyberSKA Project. Presented at the 19th Annual Meeting of Astronomy and Astrophysics, May 24, Aveiro, Portugal.