International Journal of Multidisciplinary and Current Research

Research Article

ISSN: 2321-3124
Available at: http://ijmcr.com

Cloud Based Framework for Efficient Management of Research Data

Samuel Oluwarotimi Williams, Omisore Mumini Olatunji, and **Obe Olumide Olayinka**

Department of Computer Science, Federal University of Technology Akure, P.M.B. 704, Akure, Nigeria

Accepted 05 December 2013, Available online 01 January 2014, **Vol. 2 (Jan/Feb 2014 issue)**

Abstract

Data Management in the domains of business enterprises and scientific research has posed several challenges which have relatively made it difficult for business managers and scientists to fully explore the potentials inherent in large volume of data. Challenges such as inappropriate storage facilities, lack of adequate data protection mechanism, poor data sharing means, inadequate data analysis tools, and the huge cost associated with the procurement and maintenance of contemporary computing facilities are responsible for the setback experienced in the management of research data. This research therefore proposes a Cloud Based framework with a number of integrative mechanisms for the efficient management of research data. The proposed framework provides a robust secured means for storing, sharing, processing, and analysis of huge amount of data within a short space of time. The framework also provides a mechanism that bills a consumer in accordance to the cloud services subscribed for and the duration for which the cloud service is used.

Keywords: *Cloud Computing; Research Data; Data Management; Cloud Services; Cloud Consumer; PaaS; IaaS; SaaS*

1. Introduction

A huge amount of dataset regarding scientific research has been obtained for experimental purposes over the years. This large collection of data has not been properly managed as a result of the draw backs of the conventional computing tools employed. The size (volume) and complexity of such data is found to be increasing at an alarming rate every now and then. Due to these factors (volume and complexity of the data), there is need to device a new improved means of storing, processing, and managing such data.Over the years, proper preservation and processing of scientific research data has been a major challenge to individual researchers, institutions of higher learning, and research institutes across the globe when conducting a research. Often, only little part of the data used in such research is published which makes it a bit difficult for re-use. Publishers of academic research works have also craved for an efficient way of preserving and processing research data in order to support future academic publications. Proper preservation and processing of scientific research data enables dataset to be shared efficiently, brings about the re-use of research data, provides basis for validating research findings, leads to economic gains in facilitating innovations, and aids effective policy formulation among others.

In recent times, research has identified Cloud Computing (CC) as a new and substantial business model capable of providing efficient services in the domain of research data management [1]. The use of CC offers a potential solution by allowing individuals and organizations to gain access to vast quantities of computing power that is far greater than the quantities of computing resources that the budget of such an individual or organization can acquire and use [2]. Among its numerous advantages, CC allows scaling of resource to meet current research needs with payment being only for the time that the individual/organization 'rents' the resources. CC provides individuals/organizations access to a wide variety of computational resource types either not normally available to them or which would not gain enough utilization to warrant purchasing [3].

This research therefore proposes a Cloud Based Approach for the Management of Research data by individual researchers, institutions of higher learning, research institutes, and publishers of scientific researches across the globe. The framework developed by this research provides a means of preserving large amount research dataset for re-use in the future and also support efficient processing of such dataset in order to meet certain basic research needs and objectives.

The remaining part of this paper is structured as follows: Section 2 presents detailed information on the background of the study; Section 3 presents the proposed

cloud based framework and the methods adopted while Section 4 presents the conclusion which was drawn from the findings of the research.

2. Research Background

This section of the manuscript presents key facts regarding the research topic as discussed in some relevant literature.

2.1 Research Data Management

The management of research data is recognized as one of the most pressing challenges facing higher education and research institutes in most countries. Research data generated by publicly-funded researches are seen as a public good and should be available for verification and re-use. In recognition of this principle, all UK Research Councils require their grant holders to manage and retain their research data for re-use, unless there are specific and valid reasons not to do so. Researchers who have experienced the innovative and transforming potential of data intensive research through data re-use, recombination or meta-analysis, are also calling for data to be as open as possible.

Several programmes regarding Research Data Management have been developed with a number of important tools, technologies, and services by the UK research council. These programmes are primarily geared towards promoting and supporting Research Data Management among UK higher institutions and research organizations. The programmes also aimed at improving the capability of the institutions in order to manage their research data, provide significant body software, supporting systems, guidance and policies which may be used by other institutions [4].

The need to provide adequate storage for research data and to ensure the best possible exploitation of these assets is recognized as one of the most pressing challenges facing the research communities. Researchers in a number of disciplines have demonstrated the benefits of efficient data processing and sharing. As a result there are vocal calls for making research data as open as possible. These issues touch on those of research integrity. Ensuring research integrity and improving research data management are two closely related challenges facing researchers. Research data management considers the need to have policies and processes governing the retention and preservation of data, and the ability, where necessary, to ensure security and to enable the appropriate sharing and publication of data, fundamental to ensuring research integrity [5]. It is essential, therefore, for research institutes and individual researchers to look into the challenges associated with managing research data and providing lasting solution that will support proper management of such data throughout the research lifecycle.

2.2 Data Flows and Workflows in a Conventional Scientific Research Environment

Scientists mostly feel challenged by the quantity and complexity of data they work with when conducting an intensive research. Often, acquired data must be moved because facilities needed to simulate such data are separate from the analysis facilities available to the researcher. In more complex collaborative activities, data may even be moved to national research centers in other countries to perform resource-intensive processing. A generic model of data flow and workflow in the conventional scientific research environment is shown in Figure 1.

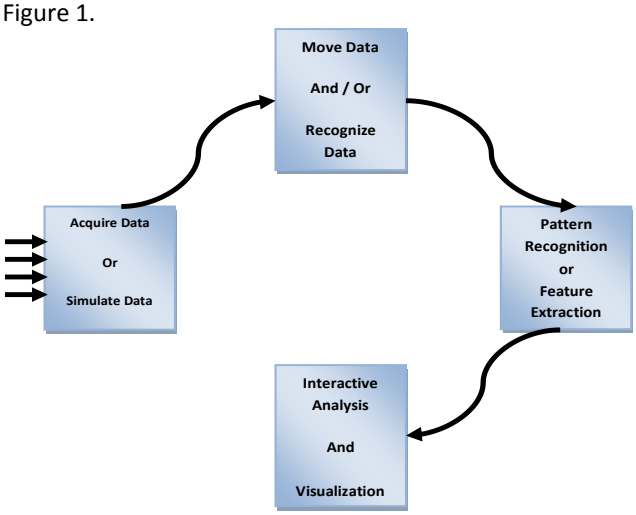

Fig.1 Simple View of Data Flow and Workflow in a Research Environment

Data frequently must be reorganized, for example, the collected dataset must be properly reorganized in order to ease further processing in a given study. Reorganizing a larger amount of research data requires a lot of computing power which may be difficult to achieve with the usual (conventional) computing approach.

Pattern Recognition and Feature Extraction are among the key reasons to taming large volume of datasets that are usually complex to study. In many cases they are simply an automation of the visual searches for patterns and features that can be done by eye on small datasets. However, once the patterns and features have been extracted and stored in a more compact mode, their analysis presents completely new challenges. A framework automating these activities would vastly enhance scientific productivity, particularly in data-intensive research conducted by small and big research teams. Such a framework would also automate the capture of all the steps taken by all participants so that the data provenance was assured. Such assurance becomes vital as small teams evolve into larger teams and then into worldwide collaborating communities. Figure 1 hides the hardware and software components that accomplish its actions while Figure 2 illustrates some of

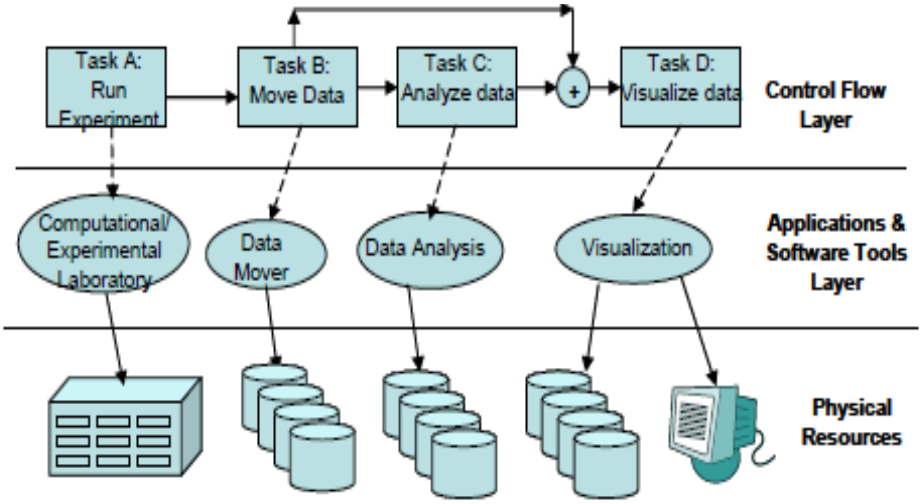

Fig. 2 Detailed View of Data Flow and Workflow in a Research Environment

the hidden components that accomplish the multiple data related actions performed in many experiments and simulations using the conventional computing approach.
In Figure 2, a framework that shows the workflow created in a scientific investigation process, with its three basics layers: control flow, applications/software tools, and physical computer hardware, is presented. The top layer illustrates the control activities; the middle layer presents the software components; while the bottom layer shows the physical resources needed for the activities.

2.3 Meeting the Challenges of Data Management

Institutions today have more means than ever to collect research data from researchers that are affiliated to them. The traditional methods, such as face-to-face meetings, exhibitions and bought-in lists are still as relevant as ever. In addition, the new communication channels, primarily Email and the Internet, have opened up new opportunities for research institutes to acquire new set of research data in a convenient manner. But along with these opportunities come the threats, and in pulling this data together from so many disparate sources, the main challenge is to retain control of the data, and to put in place institution-wide procedures that will keep such data accurate and up-to-date.

A number of database approaches have been applied to the management of research data in recent times. However, database experts across the globe still encounter some challenges when it comes to accessing such stored data. Also, the effects of poor data collection and database management are felt in such institutions. Surprisingly, many institutions still choose to ignore the problem, and have no visibility of how much money poor data quality is costing them [6]. It is reported in literature that the majority of database professionals seek advice on the subject of data quality preservation. And it as well assumed that managers of data are aware, to a certain degree, of the errors that creep into large collection of research data over time. It has been reported in literature that the challenges of data management is gradually exceeding the conventional computing power. Hence, the need to adopt CC into research data management in order to efficiently meets these data management problems.

2.4 Cloud Computing

CC is a model that enables ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storages, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. CC provides computation means, software, data access, and storage resources without requiring cloud users to know the location and other details of the computing infrastructure [7].

The concept of CC is rapidly gaining grounds in various fields of human endeavor due to its enhancing features which are: On-demand self-service; Broad network access; Resource pooling; Rapid elasticity; and Measured services. Basically, CC has three service models (Software as a Service, Platform as a Service, and Infrastructure as a Service) and four deployment models (Private Cloud, Community Cloud, Public Cloud, and Hybrid Cloud) [8]. Security is considered a key requirement for CC consolidation as a robust and feasible multipurpose solution [9]. This viewpoint is shared by many distinct groups, including academia researchers [10][11], business decision makers [12] and government organizations [13][14]. Fortunately, enhanced security measures have been built into Cloud based services in recent times.

In general, the amount of data relating to both scientific and business activities is growing at an exponential rate thereby requiring larger storing and processing infrastructure. This large amount of data needs proper storage, processing and analysis techniques in order to discover hidden patterns and new knowledge.

CC platforms are believed to provide a more efficient and robust mechanism for individuals and organizations to collaborate, share and process data with ease [15].

One of the key issues within the research industry is the storage of research data during an experimental phase and over the entire life of the research. Several companies have developed servers for the storage and processing of research data. However, these servers often require local infrastructure and maintenance within the organization that is using them - tending to utilize either central (accessible to all team members over the WAN) or local (accessible to team members over the LAN) connectivity. Nevertheless, there are still issues with data interoperation surrounding this, i.e. complete mapping between different formats is not possible due to the use of proprietary extensions. There is however, currently, a drive to overcome these constraints and move towards a standardized Cloud based platforms [16].

3. Proposed Cloud Based Framework

The architecture of the proposed Cloud Based Framework for the management of research data is shown in Figure 3. The left part of Figure 3 shows a set of research data collected from different researchers across the world. The data collected by each researcher is labeled Research $Data_I$.

Where **I** represent **I**th researcher's dataset and I = 1, 2, 3, . . ., n. On the right side of Figure 3, is the Cloud Based Infrastructure which contains a *Cloud Client Interface* (Client Computer), a *Control Node*, *Storage Facility*, *Network Infrastructure*, and a set *Application Servers* for proper management of research data.

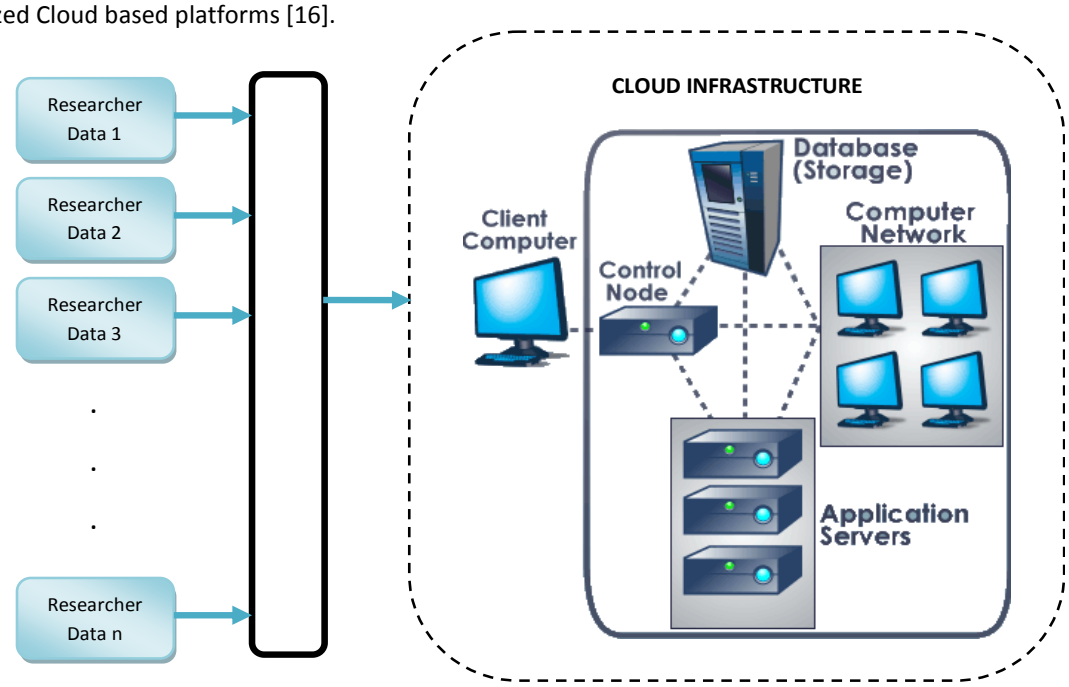

Fig. 3 Cloud Based Framework for Research Data Management

The left part of Figure 3 shows a set of research data collected from different researchers across the world. The data collected by each researcher is labeled Research $Data_I$.
Where **I** represent **I**th researcher's dataset and I = 1, 2, 3, . . ., n. On the right side of Figure 3, is the Cloud Based Infrastructure which contains a *Cloud Client Interface* (Client Computer), a *Control Node*, *Storage Facility*, *Network Infrastructure*, and a set *Application Servers* for proper management of research data.

In the framework shown in Figure 3, the researcher is viewed as the Cloud consumer since he appears to be in need of the services offered by the CC based platform. The Cloud consumer in this respect could be an individual researcher or organization that maintains a business relationship with the Cloud provider and as well uses the services offered by the Cloud provider. Usually, a list of services offered by each Cloud provider is compiled and made available in form of catalog from each Cloud provider. The Cloud consumer browses through this catalog to subscribe for appropriate service(s), sets up service contracts with the Cloud provider, and eventually uses the service. The Cloud consumer will be billed for the service provisioned, and needs to arrange payments accordingly.

Service Level Agreements (SLAs) consists of terms regarding the Quality of Service (QoS), Security provision, and Remedies for performance failure (Fault Tolerance Mechanism) which usually hold between the Cloud consumer and Cloud service provider. The Cloud provider often list in the SLAs a set of promises explicitly not made to consumers, i.e. limitations, and obligations that Cloud

consumers must accept while the Cloud consumer needs SLAs to specify the technical performance requirements that would meet their research needs. A Cloud consumer can freely choose a Cloud provider with better pricing and more favorable terms.

The 'pay per use' funding model adopted by many Cloud-based services provider makes it an appealing option for researchers who wish to store and/or process data over a short time period and have limited or no existing infrastructure to perform the activity. However, maintaining the data in the cloud beyond the project lifecycle is highly feasible. Researchers are often updated about changes in the prices of Cloud services and this would enable the researchers to either retain or migrate their data to a more favorable Cloud provider. Guidelines for removing data once the retention period has expired are also made available. The charges paid by a Cloud consumer to the Cloud provider is dependent upon the services requested while the activities and usage scenarios vary among Cloud consumers. Figure 4 presents some example Cloud services available to a Cloud consumer [13].

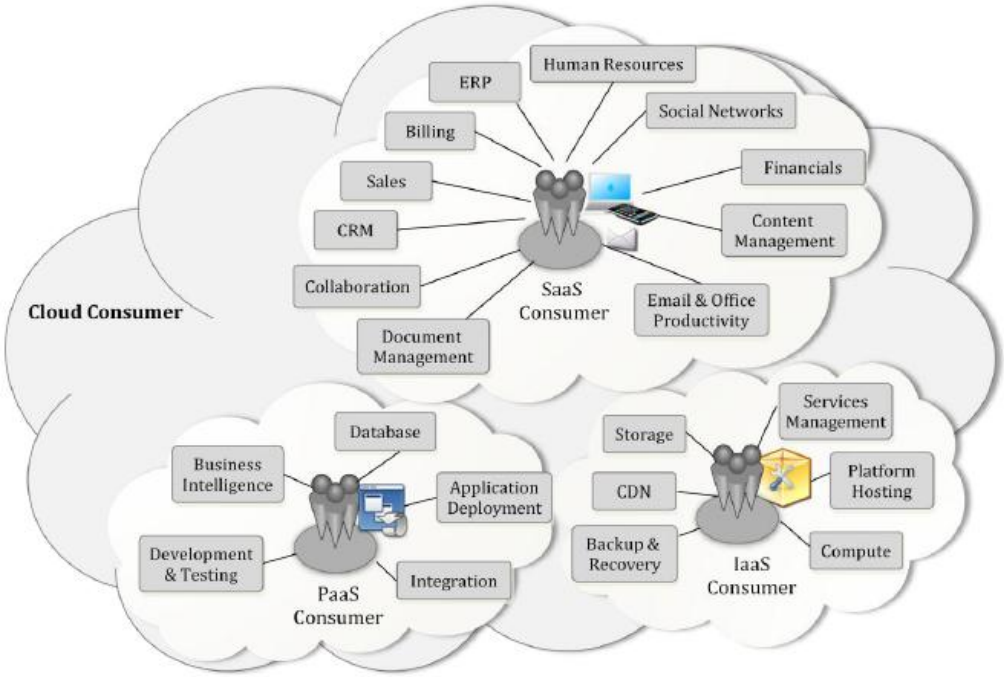

Fig. 4 Services Available to a Cloud Consumer

3.1 Cloud Based Services

Software **a**s **a** **S**ervice (SaaS)**:** these are applications in the Cloud that could be accessible via a network by cloud consumers. SaaS consumers are billed based on; the number of end users; the time of use; the network bandwidth consumed; and the amount of data stored or duration of stored data.

Cloud consumers of **P**latform **a**s **a** **S**ervice (PaaS) can employ the tools and execution resources provided by Cloud providers to develop, test, deploy and manage the applications hosted in a Cloud environment. PaaS consumers are mostly application developers who design and implement application software, application testers who run and test applications, application deplorers who publish applications into the cloud, and application administrators who configure and monitor application performance on a platform. PaaS consumers can be billed according to, processing, database storage and network resources consumed by the PaaS application, and the duration of the platform usage.

Consumers of **I**nfrastructure **a**s **a** **S**ervice (IaaS) have access to virtual computers, network-accessible storage, network infrastructure components, and other fundamental computing resources on which they can deploy and run to access these computing resources, and are billed according to the amount or duration of the resources consumed, such as CPU hours used by virtual computers, volume and duration of data stored, network bandwidth consumed, number of IP addresses used for certain intervals.

3.2 On-Demand Pricing Mechanism

The On-Demand pricing technique adopted by Cloud service providers have made it flexible and easier for Cloud consumers to subscribe to their services. For instance, assuming a Cloud consumer wishes to subscribe for the Storage service (**Ss**) of a given Cloud provider for a given time interval ($\mathbf{T_i}$), then, the consumer is expected to pay in return the following charge:

$$CSsC = CPUS_0 * \sum_{i=0}^{n} T_i \quad . \quad . \quad . \quad . \quad . \quad (1)$$

Where **CSsC** denotes the Cloud Storage Service Charge, $\mathbf{T_i}$ represents the time over which the service was accessed by the consumer while $\mathbf{CPUS_0}$ is the charge per unit storage as stated by the Cloud service provider.

If the consumer desires to access the Processing service (**Ps**) rendered by the provider over a period of time ($\mathbf{T_i}$), then, the consumer is also expected to pay the following charges:

$$CPsC = CPUP_0 * \sum_{i=0}^{n} T_i \quad . \quad . \quad . \quad . \quad . \qquad (2)$$

Invariably, **CPsC** denotes the Cloud Processing Service Charge, $\mathbf{T_i}$ represents the time for which the service was used by the consumer while $\mathbf{CPUP_0}$ is the charge per unit processing as stated by the Cloud service provider.

Then, we can equally determine the total amount that a Cloud service consumer is expected to pay when subscribed for a number of services as follows:

$$CSCC = CPUS_0 * \sum_{i=1}^{n} T_i + CPUP_0 * \sum_{i=1}^{n} T_i + \ldots + N \qquad (3)$$

$$CSCC = CSsC + CPsC + \ldots + N \quad . \quad . \quad . \quad . \qquad (4)$$

where **N** represents the $\mathbf{N^{th}}$ Cloud service charge.

In summary, the Cloud service provider charges the customer a cost ($\mathbf{CPUS_0}$ and $\mathbf{CPUP_0}$) to use the computing resource for an interval **i**. The interval **i** is chosen by the provider, and it represents the minimum period (typically an hour) for which a resource may be used, and for which the consumer will be charged. Aside the efficient pricing mechanism that the proposed framework offers, it also provides a substantial storage technique which helps to guarantee the proper storage of research data irrespective of its volume.

Conclusion

The challenge of building dependable, available and scalable data management systems capable of serving a large volume of data to a considerable number of users has confronted the data management research community as well as large Internet enterprises. The growing popularity of Cloud computing, the resulting shift of a large number of Internet applications to the Cloud, and the quest towards providing efficient data management services in the Cloud, has opened up the challenge for designing a robust Cloud based data management system. Going by these challenges, this research proposes a sustainable Cloud based platform for the management of research data. The proposed framework will provide an enabling environment for proper management of research & business data. Aside its convenient pricing mechanism, the framework also has enhancing data storage and processing tools which in turn refines the overall performance of the proposed Cloud based framework. However, this research has only looked at the application of Cloud computing concept in the domain of data management in scientific researches but the initiative can as well be extended to other domain in the future.

References

[1]. Samuel, O. W., Omisore, M. O., Ojokoh, B. A., and Atajeromavwo, E.J. (2013) "Enhanced Cloud Based Model for Healthcare Delivery Organizations in Developing Countries", International Journal of Computer Applications (IJCA) Vol. 74, No. 2, pp. 0975-8887.

[2]. Armbrust M, Fox A, Griffith R, Joseph AD, Katz R, Konwinski A, Lee G, Patterson D, Rabkin A, Stoica I, Zaharia M (2010) A view of Cloud computing. Communications of ACM 53(4): 50–58

[3]. Glenis, V., McGough, A.S., Kutija, V., Kilsby, C. and Woodman, S. (2010) Flood Modelling for Cities using Cloud Computing. Journal of Cloud Computing: Advances, Systems and Applications 2013, 2:7.

[4]. Jensen J, Downing R, Waddington S, Hedges M, Zhang J, Knight G (2011) Kindura – Federating Data Clouds for Archiving Proc. Int'l Symp. on Grids and Clouds 2011. Academia Sinica, Taipei, Taiwan, http://pos.sissa.it/archive/conferences/133/039/ISGC%202011%20&%20OGF%2031_039.pdf. Accessed 13 June 2013

[5]. EPSRC Policy Framework on Research Data (2011). http://www.legislation.gov.uk/ukpga/2000/36/contents. Accessed 15 Oct 2012.

[6]. R. Baburajan, "The rising Cloud Storage Market Opportunity Strengthens Vendors," info TECH, August 24, 2011, http://it.tmcnet.com/channels/cloud-storage/articles/211183-rising-cloud-storage-market-opportunity-strengthens-vendors.htm. Date Accessed: 04-02-2013-04.

[7]. Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M (2009) Above the clouds: a Berkeley view of cloud computing In: Tech. Rep. UCB/EECS-2009-28. EECS Department, University of California, Berkeley. http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[8]. Mell, P. and Grance, T. (2011), The NIST Definition of Cloud Computing. National Institute of Standards and Technology (NIST), U.S Department of Commerce, Special Publication 800-145.

[9]. IDC (2009) Cloud Computing 2010 – An IDC Update. slideshare.net/JorFigOr/cloud-computing-2010-an-idc-update

[10]. Armbrust M, Fox A, Griffith R, Joseph AD, Katz RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M (2009) Above the Clouds: A Berkeley View of Cloud Computing. Technical Report UCB/EECS-2009-28, University of California at Berkeley, eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html

[11]. Rimal BP, Choi E, Lumb I (2009) A Taxonomy and, Survey of Cloud Computing Systems. In: Fifth International Joint Conference on INC, IMS and IDC, NCM '09, CPS. pp 44–51.

[12]. Shankland S (2009) HP's Hurd dings cloud computing, IBM. CNET News

[13]. Catteddu D, Hogben G (2009) Benefits, risks and recommendations for information security. Tech. rep., European Network and Information Security Agency,enisa.europa.eu/act/rm/files/deliverables/cloudcomputing-risk-assessment

[14]. CSA (2009) Security Guidance for Critical Areas of Focus in Cloud Computing. Tech. rep., Cloud Security Alliance

[15]. Ludescher, T., Feilhauer, T., Brezany, P. (2013), Cloud-Based Code Execution Framework for scientific problem solving environments. Journal of Cloud Computing, Advances, Systems and Applications 2013, 2:11.

[16]. Beach, T.H., Rana, O.M., Rezgui, Y., and Parashar, M. (2013), Cloud computing for the architecture, engineering & construction sector: requirements, prototype & experience. Journal of Cloud Computing, Advances, Systems and Applications 2013, 2:8