Research Management with Cloud Storage

**University of Surrey**

**Department of Computing**

**2011 - 2012**

Giang Tran Hoang

This project is effort to improve current solutions of Research Data Management products by adding the power of Cloud Storage.

# Abstract:

(To be written in the end)

Table of Contents

[Abstract: 1](#_Toc322280345)

[I. Introduction: 1](#_Toc322280346)

[1. Overview of the project: 1](#_Toc322280347)

[2. Aim and Objectives: 1](#_Toc322280348)

[2.1. Aim of the project: 1](#_Toc322280349)

[2.2. Objectives: 1](#_Toc322280350)

[3. Structure of the Report: 1](#_Toc322280351)

[II. Literature Reviews: 1](#_Toc322280352)

[1. Data Curation: 1](#_Toc322280353)

[1.1. Data sharing methods: 1](#_Toc322280354)

[1.2. File formats and Standards: 1](#_Toc322280355)

[1.3. Access Restriction: 1](#_Toc322280356)

[1.4. Back-up and Replication: 1](#_Toc322280357)

[1.5. Metadata: 1](#_Toc322280358)

[1.6. Archiving: 1](#_Toc322280359)

[2. Cloud Storage: 1](#_Toc322280360)

[2.1. Cloud Computing: 1](#_Toc322280361)

[2.2. Definition of Cloud Storage: 1](#_Toc322280362)

[2.3. Advantage of Cloud Storage: 1](#_Toc322280363)

[2.4. Disadvantage of Cloud Storage: 1](#_Toc322280364)

[III. Planning: 1](#_Toc322280365)

[1. General project plan: 1](#_Toc322280366)

[1.1. Project miles stone 1](#_Toc322280367)

[1.2. Project Objectives: 1](#_Toc322280368)

[1.2.1. General requirements during the projects: 1](#_Toc322280369)

[1.2.2. Achievable Objectives: 1](#_Toc322280370)

[1.2.3. Time-permitting Objectives: 1](#_Toc322280371)

[2. Software Development plan: 1](#_Toc322280372)

[2.1. Choosing a development methodology: 1](#_Toc322280373)

[2.2. Development plan following Prototype methodology: 1](#_Toc322280374)

[IV. Analysis: 1](#_Toc322280376)

[1. Current Existed Technologies: 1](#_Toc322280377)

[2. Technology Choices: 1](#_Toc322280379)

[3. Product requirements: 1](#_Toc322280380)

[V. Storage Connectors: 1](#_Toc322280381)

[1. Prototype 1: 1](#_Toc322280382)

[1.1. Objectives: 1](#_Toc322280383)

[1.2. Designs: 1](#_Toc322280384)

[1.3. Implementation: 1](#_Toc322280385)

[1.4. Testing: 1](#_Toc322280386)

[1.5. Planning for Prototype 2: 1](#_Toc322280387)

[2. Prototype 2: 1](#_Toc322280388)

[VI. DataVerse Connector: 1](#_Toc322280389)

[1. Prototype 1: 1](#_Toc322280390)

[1.1. Objectives: 1](#_Toc322280391)

[1.2. Designs: 1](#_Toc322280392)

[1.3. Implementation: 1](#_Toc322280393)

[1.4. Testing: 1](#_Toc322280394)

[1.5. Planning for Prototype 2: 1](#_Toc322280395)

[VII. Final Product Evaluation: 1](#_Toc322280396)

[1. Test plans: 1](#_Toc322280397)

[2. Test Results: 1](#_Toc322280398)

[3. Feedbacks: 1](#_Toc322280399)

[VIII. Future plan 1](#_Toc322280400)

[IX. Discussion and Conclusion: 1](#_Toc322280401)

[X. References: 1](#_Toc322280402)

# Introduction:

## Overview of the project:

Science is the most important factor for evolution and development in human history and scientific research activities is the most vital. Nowadays, research activities are shifting more and more to the fourth paradigm: data exploration (Hey et. al, 2009), and the common research cycle is now data capture and curation then data analysis and visualization. However, the tools for all those activities are now very limited and that makes the whole process very dreadful. There are needs for a good application helping managing all those research data, providing access of the needed data to the correct people while maintaining the intellectual integrity. Many challenges have been addressed, and I, with this project will make an attempt to tackle one of them: the storage of that massive amount of data with respect to its special nature.

Cloud storage, which is described as a perfect solution for storing data with high volume but require high availability seems to be an interesting road to follow. It has the definition as providing “the illusion of inﬁnite computing resources” (Armbrust et. al, 2009) is a perfect fit for the massive amount of research data need managing. The data will grow over time very fast, and the transparently expandable properties of cloud storage will keep the system running smoothly. Also, cloud storage provides great accessibility to the resource, as long as the connections are alive. This is vital for research data management, as the point of such system is for scientific results to be shared to the right people. Moreover, it can be improved with many additional functions to make it customized for a specific type of data. Those features could be redundancy, archiving, and multi-layered authentication or customized metadata. All of these would be very useful for scientist and researchers to manage their projects.

With that in mind, I really want to see the applications for this purpose make use of that technology. This project is an effort to make that happens. There are some existed products helping research society in managing their results, but hardly any of them making use of cloud storage. Most of them just took a disk drive in the application server for storing files, which will slow down the performance of the file operations. The main software application to be done here will be a connector to the cloud storage system(s). This software will put the files into one or many cloud storage systems, depending on the usage of the data. It will maintain the mapping between the files in each research results and theirs actual storing place as well as store the necessary metadata. The application is hoped to be very flexible, and more storage systems can be added to the software based on the provided framework.

Then, another part of this project will be a sample connector for the recommended research data management software: DataVerse. This is an application made by Harvard University and it has quite a lot of good features and is attracting much attention. It is a massive product, I would say, and is one of the most complete solutions in this field. Learning about this product will definitely help me in understanding the nature of research data, and how to manage it efficiently. Also, for such complication of this product, making a working connector to make use of my storage connector will be a very strong proof for the usability of my creations. Moreover, University of Surrey is now in need of a good system for managing its science projects, and my project if succeeds will provide a very good choice for it. A complete and stable system for managing their precious data, with improved storage functions will be ready to use. This is the big reason I follow this path instead of making an application to manage research data from scratch.

## Aim and Objectives:

### Aim of the project:

The main aim of this project is to create an application helping integrating the cloud storage technologies into research data management software. The expected result will be an application being able to connect any existed data managing products to a transparent and stable pool of cloud storage. The cloud storage functionality may consist of services from many cloud storage systems, which could be on-site (E.G: OpenStack’s SWIFT), or an online cloud provider (E.G: Amazon S3).

This feature will be customized for the nature of research data. Not all functions provided by the cloud storage technologies will be used, but the ones which help the most with research data will be utilized. For example, the extra metadata fields provided by SWIFT will be very useful for storing information about the files. But the function of fake-directory by adding path for files is not useful as the users do not care where the files are store, as long as they can retrieve the files as they sent when they need.

Currently, there may be not many complete solutions for the scientific data storage problem, but there is still a fair amount of effort for this. In the UK, there are some projects sponsored by JISC ([www.jisc.ac.uk](http://www.jisc.ac.uk)) researching on a meta-framework for research data management. There is also the Data Curation Centre ([www.dcc.ac.uk](http://www.dcc.ac.uk)) which addresses a lot of features, tasks and general knowledge specified for research data. There are also documents from all over the world giving guidance for managing such data. All those sources will be exploited to gain understandings about the challenges as well as the solutions for this project. I will do my best to gain as much information as possible to provide a good explanation on the problem and the approach to make the storage connector suitable for this type of data.

The connector to the storage will be planned to be able to use many cloud technologies from different providers. In able to do this, I will need to make something like a template, which can be applied to any cloud services to make a corresponding connector. One or two sample connectors will be made, depending on the time permitted. One cloud solution will be definitely use, and that is SWIFT from OpenStack. This service is very new and still on the development, but it has very great potential. It is open-source, providing data redundancy, quick mapping, easy to install and expand and has built-in RESTful functionality. The other important part of this storage side will be a manager for all those connector. This will decide which files go to which storage systems, and maintaining an efficient mapping and look-up mechanism to retrieve the files as quick as possible when required.

Alongside with that, another part will be the connector to one of the most sophisticated application in the field, which is DataVerse. This is more like a demo how to use the cloud connector. I will try my best to make the documentation as complete as I can, which will help any future application of the cloud connector easier. The manual of how to set up the connector, how to make use of the classes and how to integrate it into the application will be documented. This plus the javadoc of the software itself will make sure any future use of the connector very applicable. Moreover, as I said, with this decision, after the project, there will be a nice system ready to be used by any party in need of a complete solution of research data management with nice storage. I may not use the latest code of DataVerse but use a stable version of it instead, but the demonstration value of the connector should still remain.

### Objectives:

* Background research on data management
  + What are the challenges of it?
  + Metadata for scientific projects/files
  + What helpful features should be considered?
* Technologies understanding
  + Current applications for research data management
  + Cloud storage technologies
  + Web application technologies
  + Choose the specific technologies to use
* Analysis and design the application
  + Decide functions to be implemented
  + Design the structure of the application
  + UML diagram
* Develop the application with a suitable methodology
  + Evaluate and choose a development methodology
  + Keep the documentation following the chosen one
* Write a detailed and concise report
* Maintain efficient project management
  + Setting repositories, project workspace
  + Create the documents, keep them updated
  + Keep the weekly logs

## Structure of the Report:

The structure of the report will follow the flow of a software development methodology, plus all the normal parts of any scientific report. Because the software will be split into two part, the report will have the two chapter covering the development cycle of each part. Each part will have its own test phase. The integration tests and evaluation will be put in a separate chapter.

* Chapter I gives a quick overview for the projects and the report itself.
* Chapter II shows the research results have been done on all the required aspect of the problem to be solved and the application.
* Chapter III covers all the planning decision with milestones and expected time line.
* Chapter IV contains the technologies, the structures and other design aspect of the application.
* Chapter V follows the development cycles of the cloud storage connector: this is the main part of the application.
* Chapter VI follows the development cycles of the application connector: a sample demo how to use the storage connector.
* Chapter VII reports the results of the integration tests between two parts in to a complete system.
* Chapter VIII provides the future plans for more development, usage and user support.
* Chapter IX provides the final evaluation of the projects as a whole.
* Chapter X is the references list.

# Literature Reviews:

## Data Curation:

On this scope of this project, data is digital research data which is any kind of data produced during the research processes, and can be stored digitally. Data curation as defined by DCC’s definition is the actions maintaining, preserving and adding value to digital research data throughout its lifecycle. The processes of this cycle are (from DCC’s definition):

* “**Conceptualise:** conceive and plan the creation of digital objects, including data capture methods and storage options.
* **Create:** produce digital objects and assign administrative, descriptive, structural and technical archival metadata.
* **Access and use:** ensure that designated users can easily access digital objects on a day-to-day basis. Some digital objects may be publicly available, whilst others may be password protected.
* **Appraise and select:** evaluate digital objects and select those requiring long-term curation and preservation. Adhere to documented guidance, policies and legal requirements.
* **Dispose:** rid systems of digital objects not selected for long-term curation and preservation. Documented guidance, policies and legal requirements may require the secure destruction of these objects.
* **Ingest**: transfer digital objects to an archive, trusted digital repository, data centre or similar, again adhering to documented guidance, policies and legal requirements.
* **Preservation action:** undertake actions to ensure the long-term preservation and retention of the authoritative nature of digital objects.
* **Reappraise:** return digital objects that fail validation procedures for further appraisal and reselection.
* **Store:** keep the data in a secure manner as outlined by relevant standards.
* **Access and reuse:** ensure that data are accessible to designated users for first time use and reuse. Some material may be publicly available, whilst other data may be password protected.
* **Transform:** create new digital objects from the original, for example, by migration into a different form.”

These processes are almost all of the requested functions for a good research data management system. They could form a very good evaluating framework for any existed products or upcoming projects concerning the same problem.

The Australian National Data Service (ANDS) also recommends a similar structure of functionality for a data management tools. These processes are grouped up two three big categories (from the ANU Data Management manual):

* Data Organisation
* Data Administration
* Data Archiving and Sharing

Most of the aspects of the first two functions are handled by the front-end data management application and will be used as the template for evaluating existed products in the next part of the report. The functions relating to data archiving and sharing concern directly to the storage of the files, so they will be explained in more details and from that I will point out the requirements for a storage solution for research data.

### Data sharing methods:

The manual from ANDS states that there are currently three popular methods for sharing research data:

* “Email request – Interested researchers email and request the dataset. This is the most common way that data is shared.
* Website – Researchers place datasets on their website that anyone can download.
* Archiving – Researchers place their dataset in an archive.”

But it is also said that the most preferred options is archiving, due to its usefulness in both data preservation and dissemination. Archives also often have search functionality integrated, making it usually indexed by search engines thus more probably to be found by fellow researchers. My though about this is that a combination of a central web service managing all the archives of the data may be the ideal system. The author may still have a customized website, showing his research projects, but all the file storage will be manage by that application. It will take care of the file location, indexing and providing search utility so that the archives reach their optimum availability. That is the application I am looking for and trying to integrate the cloud storage feature to make the archiving function the most efficient and fastest.

### File formats and Standards:

This is one of the most important aspects to be considered when implementing a storage system, especially when it serves a specific purpose. The ANDS’s guide for Data Management Planning suggests the requirement to specify the file formats used for the data in each research. Abrams (2007) states that: “Format can be used by curation managers as an important organizing principle for their digital collections.” The format of the files if being standardized will help not only with the storing/archiving tasks, but also with the searching, research reproducing as well. Some formats like tabular ones support on-the-fly data extraction and analyse, which is very useful in research environment.

Any systems will have to consider which file formats it will support, document that decision in an authority and inclusive manner. The decision must consider a wide range of platform-independent tools for content creation validation, modification and rendering (Abrams, 2007). The decision must be also robust and persistent enough so that the content created will be usable in the long time-span.

At this moment, the most widely used file formats classification is MIME (Multipurpose Internet Mail Extensions) media type registry. It uses a system of two-level identifiers. For example, a JPEG images is defined as “image/jpeg”. This has been used and considered very useful in the world wide web, especially in the mail-centric applications. However, the information about the format of the files using this registry is at minimal level, and the classification of the files sometimes is not sufficient for scientific purposes. Also, there are many file formats classified as more than one type, which may cause management problem. There are various efforts to build a more suitable file format specification system. The most standardized and recommended system is probably Open Archival Information System (OAIS – ISO 14721). Many popular research centers in this field, including the DCC, Harvard University, are promoting the use of OAIS in digital curation and preservation.

The tasks attracting the most interest in this topic are not just making use of a customary classification, but also creating standard file formats for each kind of research data. These formats are aimed to be preservation-friendly, which can persist in long-term systems. The new formats are created hoping to include many attributes being conductive to the conservation of the data. For example, JPEG 200 (ISO 15444-1) is a format that “supports calibrated colour management, built-in error resilience features, provision for embedding arbitrary metadata in XML or other forms, and a flexible internal structure of nestable units, or “boxes”.” (Abrams, 2007). Those new formats, if applied to the data curation cycle, will help the managerial operation of identification, validation, characterization and assessment of the files.

However fancy and promising those new formats may sound most of the current application and software library still supports MIME as the main registry. The effort when creating any application will use MIME as the normal system, the OAIS will only be as a bonus, time-permitting features.

### Access Restriction:

For most of the files when created and during it life-cycle, the access rights to it is usually set to one of these (according to the ANU manual):

* Unrestricted– Anyone can download.
* Registered– Users must give their name and affiliation so the data owner can track who is using their data.
* Requested– Users must submit a request outlining how they will use the data.
* Closed– No access (i.e., confidential data)

The access restriction can be even more specific to part of the data in a file. For example, a tabular file has 5 columns and 500 rows, the file is supposed to be “unrestricted”, everyone should be able to get the data from the file. However, the last column of the table contains sensitive data which should be accessed by only registered users. Therefore, when an unregistered user wants to capture the data from the file, the last column must be eliminated. This use-case poses more requirements for the information about the file itself to be saved, but is not necessarily required. This feature can be considered as non-essential and may be implemented if time permits only.

The authentication scheme for this feature is usually implemented in the front-end applications and the storage system may not need any authentication at all. But knowing how this authentication is handled will help the implementation of the connectors between those two. There is no point in trying to ask the storage for a file while the front-end authentication failed, and depending on the access rights of the files, they can be arranged for faster access.

### Back-up and Replication:

For such precious data as scientific results, relying for the durability of the hard drive is simply unacceptable. There are needs for multiple redundant copies in a stable and credible storage pool. The ANDS manual states this as “probably the single most important item on this list”, and also suggests an automated back-up process.

Luckily, as we are trying to corporate cloud storage into this, and redundancy is a usual feature of cloud storage. Each of cloud storage system has its own replication scheme, and this is ready to use, we may not have to concern about this. However, if we really want to have even more redundancy, we can make it happen. The software design is expected to have many cloud systems under management of a central storage manager. For data with the need of extra special warranty, we can have the storage manager to back it up across the cloud systems. This way, the data will gain a great redundancy, but will also pose more complication in the operation of the storage manager. It will now have to decide which copy of the file across the cloud systems will be downloaded to be given to the user, as well as have to maintain the integrity among the copies of the file when it is deleted or modified.

### Metadata:

This part is definitely the most important feature to consider when working with storing data, especially research data. Metadata is defined generally as data about data, but to express the importance of it to scientific research data, we will need more in-depth analyse. Some characters of metadata stated by ANDS are: metadata can actually applied to anything, metadata is data and metadata generally has little value on its own. It contained extra information for any object (in this topic, it could be a research project, a study, or a file, etc…) which help increase the usefulness of the object itself in different ways.

Day (2005) suggest that it is more useful nowadays to “define metadata in relation to its use, chiefly the functions that it is intended to support.” There are many classification scheme for metadata. The ANDS’s guide for metadata suggests the classification of metadata for scientific related data as:

* Provenance metadata: This relates to the origin of the data, and ranges from the human to the highly technical.
* Rights and access metadata:  This provides information about access and usage rules.
* Structural metadata: This provides fundamental information for a person or a computer to read the data.
* Preservation metadata:  This builds on the history from the Provenance and Rights metadata, but also includes information to help build a sense of trust in the data, and allow for the data to be used long into the future.

Metadata is becoming essential in today’s digital world, and in the case of research data management, it is even more crucial. The amount of data created in the scientific projects are booming very fast, as data-intensive is the new paradigm of science. Moreover, those results tend to be shared by scientists, as a basic operation of the data curation cycle. Also, there are requirements of the data while being shared still need to keep their originality and intellectual properties intact. Without the help of good metadata, those data will simple unmanageable and the sharing purpose cannot be fulfilled.

Metadata first important use is to discovery and retrieve resources. Most of the current search engines can only look up for the appearance of the key word(s). If there is no metadata, the search will only return the text files containing the keywords, but it does not mean the same thing as the text file is about the topic you want. Moreover, all other files with different formats like image or video will not be retrieved by the search as it cannot get any information about the content of those media. The core elements providing information for the search to work are the metadata values. In the case of scientific results, the metadata will also help narrow down the scope of the search. This is very important as it is very difficult to find the close match for the intended search in such massive world of knowledge where everything can be related to each other in one or another way.

To have good search functionality, good metadata scheme been able to describe and classify objects is a key. For research data management, Campbell (2007) promotes the use of IEEE Standard for Learning Object Metadata (LOM) (documented at <http://ieeeltsc.org/wg12LOM/>). This is an open standard specifying the syntax and semantics of metadata directly useful in learning/researching environment. This is built upon many IEEE standards and provides a complete “tree” structure of conceptual metadata and each element has specified data types and value spaces.

The next important role of metadata is to manage the vast and growing pool of data created by researchers. To use any dataset, its structure, description and source must be known, and only metadata can contain such information. This is the main purpose of administrative metadata. This, for example, records the process of production and curation of the data as well as the intellectual property rights going with it.

Metadata also play a crucial part in archiving the data. For the long-term usage, archiving is a must, but in order to use that archived data, it will have enough credibility and provides a high level of trust. This can be satisfied with help of highly standardised metadata. Ballegooie and Duff (2006) suggests the following information is required to be captured by archival metadata: “a) information about the origin or provenance of the documents, e.g. who created accumulated and/or used them; b) their documentary form, e.g. diaries; c) their filing structure, e.g. chronological; d) their relationships with other records; and e) the ways they can be used, e.g. restrictions on the use or reproduction of the records”. This is a very nice guidance for me to consider when implement the archival functionality of the storage system.

The last important role for metadata to play is to help with the data sharing and reuse. Research domains are where data tends to be shared and used by other researchers, and there is a need to maintain contextual and other information about the data for it to be reproduced, interpreted and analysed by others.

Metadata can be stored in many different ways. It can be stored in a database and maps the values to the objects like any normal information. It can also be attached to the files as extra properties fields or even stored in a separate text file (E.G: the manifest file). In any case, a standard scheme of which metadata values need to be store is the key and also the complication of building metadata.

### Archiving:

As I stated, archiving is a very good practice for data wishing to be used in long-term. The ANU manual describes archiving as being able to help data avoid being lost, forgotten or unusable, and provide it with access control and security. This is a good feature to consider in any data management plan, and there are several aspects need preparing for this action:

* The general good practice for archiving is only doing it to final state data with high credibility and value.
* As stated above, good archival metadata is required to ensure the value of the data in long-term future.
* The data will likely to be reused, so file formats must be chosen carefully, making sure that it will not be obsolete and not lose quality.
* The time and cost for archival can be unexpected. The power to do this task need to be calculated to make sure it does not interfere with other tasks. It is a better practice to do archiving continuously than wait for a peek time.

## Cloud Storage:

### Cloud Computing:

There are a lot of different definitions for Cloud Computing. In general, my thought about it is a type of service where the computing resources provided to the users in a pay-per-usage basis. The architecture of the system is hidden to users. They all form a large pool of easily usable and accessible virtualized resources and can be configured to suite different cases of loads for optimal utilization (Vaquero et al., 2009).

The National Institute of Standards and Technology define cloud computing as: “…a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” (NIST 2009).

Armbrust et al. (2009) states that the service to be considered to be cloud computing need to have at least three of properties: “(1) the illusion of inﬁnite computing resources; (2) the elimination of an up- front commitment by cloud users; and (3) the ability to pay for use as needed ...”.

An example for the cloud computing services is the Elastic Compute Cloud (EC2) provided by Amazon. This is a service allow users to have “virtual machines” (VM). The user sign up for the service and will have access to an isolated fully function machines through the internet. The actual resources for the whole system of VM is hidden from the users, but as long as the promised minimum computing power by Amazon is satisfied, the user will be happy with the service. The user can install his own software on his VM, can run any application for any purpose he want on it, and Amazon will make sure all of his data, software and settings are stored and executed nicely.

Cloud computing can sub-divided into small forms in different ways, one common division is based on the services provided as cloud computing (Wu et al. 2010):

* Infrastructure as a Service (IaaS):
* Platform as a Service (PaaS),
* Storage as a Service (StaaS),
* Software as a Service (SaaS),

### Definition of Cloud Storage:

Cloud Storage is a type of cloud computing where the users are provided a virtual personal storage space. The files will be stored in a cloud system consist of many storage server and accessed through the internet. There will be some gate-way or client or web services made by the provider to make sure the user can access to the services. The storage structure (folder/directories) can be implemented to help the user organize their files easier, and the cloud service providers (CSP) will make sure the file location appear to the user persist over time while the actual place of the files in the system is not necessary static. It can be moved between hard drives, clusters or servers, but everything is transparent to the user. The whole cloud can be formed by many machine clusters in many locations. The files stored in the cloud can be redundantly stored so that in case the failure occurs, the files could be recovered without having to notice the users. This option will raise the reliability of the cloud storage but may cost more for the user.

This service can be very cost effective for the users as they do not need to build an expensive local storage system. They will still have the secured redundancy for the files, the availability for access everywhere with network connection and the guaranteed bandwidth. They do not have to worry about cost for maintaining, updating and fixing the system. With the distribution of multiple copies of files, when an error occurs, the data will still be accessible. This feature is very hard to gain in a small low/mid-range storage system. Also, the users may have “infinite” amount of storage which can be expand as they need (as long as they can pay for it) but may only have to pay for what they use (Wu et al., 2010). All the above is the ideal properties for a cloud storage service.

### Advantage of Cloud Storage:

Chen et al. (2011) states that most of the current distributed storage designs are storage area network (SAN) or Network Attached Storage (NAS), but they are not robust enough for the level of wide area network (WAN). There are several benefits of Cloud computing over such on-site storage system:

* Cost effectiveness: This will be definitely cheaper compare to the on-site data center with the same level of quality, which is very appealing to any business manager as cutting cost is one of the most important tasks for them. Using cloud storage help saving the capital not just for the hardware but also the software and the human resource to runs the systems (Wu et al., 2010). During the time like this, where the shadow of the economic crisis of 2009 is still there, this advantage of cloud storage and cloud computing in general is too good to miss. Moreover, with the fact that cloud storage is becoming more popular, there will be more service providers coming to the field and the cost for this service will be more competitive. Some groups like OpenStack even provide their platform as open source, which will make the implementation of a cloud storage system very easy and will drag the price down.
* Maintaining: The client will no longer have to worry about maintaining the data centre, updating the software or training employees to take care of that system. All the hassle is now on the side of the CSPs. The cost for maintaining the hardware and software for storing data is reduced or eliminated.
* Reliability: All the cloud services usually claim to have the reliability of 99.99%, due to their multiple redundancy storage pools. Any failure will not be seen by the user as it is automatically recover through the copy of the files in deferent locations of the cloud. The sample scenario for any planned upgrades or outages.
* Availability: This is one of the best bits of cloud computing. As long as the client can connect to the network, they will have access to their files. They can download, modify and save the files using any devices, anywhere on the go and any time in a day. This also makes sharing the data with other parties much easier. The file owners just need to setup the authentication rules for the files and provide the others with correct information and they will be able to access the files through an interface.
* Elasticity: This is the properties that make the cloud storage a much better choice compare to other traditional solutions. As the need of the user grows, the can simply add (and pay for) more storage capacity into his account. This will be handle quick and simple by the CSPs and the added resources can be use right away, no disruption happens.
* Pay-per-use cost model: The user only have to pay the usage they actually need. They do not have to invest capital for the resource they may need for back-ups or future use. This will maximize the effectiveness of their capitals and help avoiding the risk of keeping unused resource.
* The interfaces for the cloud storage services are usually very simple and easy to use.

With the nature of research data as being in massive amount, growing quickly, require high availability and sharing ability, those advantages of cloud computing are very tempting. The appliance of cloud storage into management of this kind of data will definitely improve the values of the system.

### Disadvantage of Cloud Storage:

Having such good potential but cloud storage and cloud computing in general still has many difficulties in getting their shares in the market. This is due to some concerns about “is the cloud ready for you?”, not just “are you ready for the cloud?” (Edwards, 2011). Some of the arguments put as the advantages of cloud computing are (Wu et al., 2010):

* Integrity: Putting a file in that massive pool of storage and then taking it out when required sounds like a simple task. However, with such large systems and with such high expectation of speed and redundancy, errors can happens easily. The users just have one requirement which is getting the files back as they were sent, but this will take lots of effort from the developer to produce such integrity.
* Security: The users hand their precious data to the CSPs, and they can only trust them. The physical places of the files are out of reach, and they can never be sure is there any intrusions or attacks on their data unless the CSPs inform them (this is another matter of trust). Moreover, as the architecture of cloud computing is not fully stable, there are still many unknown aspects of the security in the cloud. As the evaluation for the cloud security is not easy to prove, the user may feel insecure and abandon the technology.
* Resource shortage: The cloud providers will have to make sure the system has sufficient power of hardware, processing and network through put to support the expected size of the user usages. They promised that the files will have many copies across their machines, and that will take them a great deal of resources which they need to use the most effectively as possible.
* Replication: This feature is essential for the cloud storage, but also gives a big challenge. The replication speed will need to be enough to keep the transparency and efficiency high, but also should not interfere with the file transferring speed, which is more visible to the user.
* Pricing arguments: Cloud providers are still business parties and they will try to get the most profit as always. One thing they usually consider is putting the features like multiple redundancy and back-ups, or extra security options into a higher price plans. This is like pushing to cost of managing the risk all to the customer and might not sound convincing at all.

Among all those arguments, security seems to be the biggest concern these days. As the user losing their controls over the actual storing location of the files, they will need the providers to prove the security of the system in order to gain the trust. This is very hard as there are no concise scoring mechanism to evaluation such new technologies. One suggestion is to only use cloud storage for low-value data and still have important data under control, but this restrict so much the potential advantage of cloud storage.

# Planning:

## General project plan:

### Project miles stone

I have an expected time line for the projects, keeping track of the most important milestones:

* Have all the project management tools and initial documents set-up by 20/10/2011
* Have all the background research requiring for software design done by 01/01/2012
* Finish the initial system design by 23/01/2012
* The first working code item done by 14/02/2012
* A stable working version done by 23/03/2012
* The final version done by 20/04/2012, with all the integration tests with results
* The final report is finished and submitted on time, 07/05/2012

### Project Objectives:

### General requirements during the projects:

* Making 20 hours available in the time table every week
* 500 hours in total
* Keeping weekly logs on MS OneNote
* Keep track of all changes of plans and discuss with the supervisor
* Submit all the deliverables on time

### Achievable Objectives:

* Finish all the research on research data management framework and cloud storage technologies
* Making a storage connector for SWIFT of OpenStack
* Making a meta storage manager
* Improving a stable build of DataVerse, build a working instance making use of the storage manager
* Testing the whole system, including performance indices.

### Time-permitting Objectives:

* Build another storage connector for a different cloud storage provider (Amazon S3).
* Implement the features for the storage manager to choose the files to be archived or distributed to a different storage connector
* Do more testing on expanding the storage systems. Test the performance, disruption during the expansion.

## Software Development plan:

This section will provide the planned timeline for the development process of the application. There are two parts: one for analysing the possible development methodologies and choosing a suitable one, and the second is the milestone and deadline for the development.

### Choosing a development methodology:

(to be written)

### Development plan following Prototype methodology:

# Analysis:

## Current Existed Technologies:

## Technology Choices:

## Product requirements:

# Storage Connectors:

## Prototype 1:

### Objectives:

### Designs:

### Implementation:

### Testing:

### Planning for Prototype 2:

## Prototype 2:

# DataVerse Connector:

## Prototype 1:

### Objectives:

### Designs:

### Implementation:

### Testing:

### Planning for Prototype 2:

# Final Product Evaluation:

## Test plans:

## Test Results:

## Feedbacks:

# Future plan

# Discussion and Conclusion:

# References:

**Abrams, S.** (October 2007), "File Formats", DCC Digital Curation Manual, S.Ross, M.Day (eds), Retrieved 18/12/2011, from <http://www.dcc.ac.uk/resource/curation-manual/chapters/file-formats>

ANDS’s Data Management Planning guide (2011), available online at: <http://ands.org.au/guides/data-management-planning-awareness.html>, last access 17/03/2012.

ANDS’s Metadata guide (2011), available online at: <http://ands.org.au/guides/metadata-working.html>, last access 17/03/2012.

ANU Data Management Manual (2010), available online at [http://ilp.anu.edu.au/dm/ANU\_DM\_Manual\_v10.09.17 -63\_2010 -09-17.pdf](http://ilp.anu.edu.au/dm/ANU_DM_Manual_v10.09.17%20-63_2010%20-09-17.pdf), last access 17/03/2012.

**Armbrust, M., Fox, A., Griffith, R., Joseph, A.D.** and **Katz, R.** (2009). *Above the Clouds: A Berkeley View of Cloud Computing*. Berkeley, CA: UC Berkeley Reliable Adaptive Distributed Systems Laboratory.

**Ballegooie, M.V.** and **Duff, W.** (May 2006), "Archival Metadata", DCC Digital Curation Manual, S.Ross, M.Day (eds), Retrieved 21/01/2012, from <http://www.dcc.ac.uk/resource/curation-manual/chapters/archival-metadata>

**Campbell, L.** (April 2007), "Learning Object Metadata (LOM)", DCC Digital Curation Manual, S.Ross, M.Day (eds), Retrieved 22/01/2012, from <http://www.dcc.ac.uk/resource/curation-manual/chapters/learning-object-metadata>

**Chen, Y., Ku, W., Feng, J., Liu, P.** and **Su, Z..** (2011). Secure Distributed Data Storage in Cloud Computing. In: Buyya, R., Broberg, J., and Goscinski, A. *Cloud Computing: Principles and Paradigms*. New Jersey: Wiley. p221-p248.

**Davenhall, C.** (2011), " Scientific Metadata", DCC Digital Curation Manual, J. Davidson, S. Ross, M. Day (eds), Retrieved 22/01/2012 from <http://www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata>

**Day M,** (November 2005), "Metadata", DCC Digital Curation Manual, S.Ross, M.Day (eds), Retrieved 18/12/2011, from <http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/>

DCC’s definition of Data Curration, available at <http://www.dcc.ac.uk/digital-curation/what-digital-curation>, last visit 17/03/2012.

**Edwards, M. J.,** (2011) Is The Cloud Ready For You?‟ Bluesource White Paper

**Hey, T., Tansley, S.** and **Tolle, K** (2009). *The Fourth Paradigm*. Redmond: Microsoft Research.

ISO 14721, 2003, Space data and information transfer systems – Open archival information system – Reference model, pre-print available <http://public.ccsds.org/publications/archive/650x0b1.pdf> last access 17/03/2012

**NIST 2009**, The NIST Definition of Cloud Computing, viewed 12/12/2011, <<http://www.nist.gov/itl/cloud/upload/cloud-def-v15.pdf>>.

**Vaquero, L. M.**, **Rodero Merino, L., Caceres, J.,** and **Lindner**, **M.**(2009) *A break in the clouds: Towards a cloud deﬁnition*, SIGCOMM Computer Communications Review, 39:50-55.

**Wu, J, Ping, L, Ge, X, Wang, Y**  and **Fu, J** (2010), 'Cloud Storage as the Infrastructure of Cloud Computing', 2010 International Conference on Intelligent Computing and Cognitive Informatics, pp. 380-383