Rebuilding of Data in Cloud Computing

Rabit Ul Islam[1], Jyoti Arora[2]
[1]*M.Tech Student, Desh Bhagat University Mandigobindgarh*
[2]*Assistant Professor, Desh Bhagat University Mandigobindgarh*

Abstract- Cloud computing is the on-demand delivery of computer power , database storage, applications, and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing[1].Cloud computing provides various services in which data storage is the main cloud service. Cloud computing works behind the scene in our day to day activities such as to watch movies, play games, sending mails and listen to music etc, With Cloud computing, we can store, recover and backup data, create new applications, deliver software on demand, host websites and so on. Whenever there is a demand, user can access the services of cloud dynamically via internet[2].

Since the phenomenon of cloud computing was pro-posed, there is an unceasing interest for research across the globe. Cloud computing has been seen as unitary of the technology that poses the next-generation computing revolution and rapidly becomes the hottest topic in the field of IT. This fast move towards Cloud computing has fuelled concerns on a fundamental point for the success of information systems, communication, virtualization, data availability and integrity, public auditing, scientific application, and information security. Therefore, cloud computing research has attracted tremendous interest in recent years. In this paper, we aim to precise the current open challenges and issues of Cloud computing. We have discussed the paper in three-fold: first we discuss the cloud computing architecture and the numerous services it offered. Secondly we highlight several security issues in cloud computing based on its service layer. Then we identify several open challenges from the Cloud computing adoption perspective and its future implications. Finally, we highlight the available platforms in the current era for cloud research and development and other IT resources through a cloud services platform via the internet with pay-as-you-go pricing. Cloud computing provides various services in which data storage is the main cloud service.. In this article CES method concisely gives way for cheaper storage of non-critical data on a Cloud environment, specifically, Amazon's S3. Also, the transfer of files from S3 to RRS allocates free space for critical data in the main storage bucket. The removal of unimportant data from S3 helps in better performance of S3 as it is involved in the computation of several applications that deal with critical information. A cost comparison of S3 and RRS led to the derivation of the CES method, which minimizes capital costs involved in data storage. We have also been able to prove the difference between costs with the help of a graph curve between the S3 and RRS storage. The CES method is restricted to non-critical data and old file versions for small scale organizations. An effort can be made to reduce redundancy of critical information as well. This would increase performance and efficiency of Cloud services. Also, even though the costs have been reduced, we have not been able to test whether the data can be reproduced from the RRS bucket. Hence this can be an area of research to reproduce old file versions when necessary. This form of storage need not be restricted only to Amazon's S3 and RRS. We could also aim at generalizing the solution for all Cloud environments by creating two types of storage options, one for critical information, and another cheaper storage for non-critical information. This would cut down non- critical redundant data and save the storage space available in Clouds.

Keywords- Cloud Computing; CES; RSS Storage, S3.

I. INTRODUCTION

The study and research carried out during in this article investigates how to minimize redundant information on the Cloud and further store it in a cost effective manner. There are several organizations that intend to move their businesses to the Cloud. This implies that most of the data storage would be done on the Cloud. The solution we intend to propose, aims at enhancing data restoration by firstly sorting the data to differentiate between critical and non-critical data. The critical data is stored in a storage bucket that is more reliable than the bucket used to store non-critical data.

Hence, we intend to provide two types of storage buckets, where the non-critical information is not as expensive. It is up-to the user to decide what information is critical and what is non-critical. In this article, we aim at reducing the amount of non-critical information and redundancy associated with such data before storing these files. This will add value to the data being stored by ensuring that the storage space and investments made are utilized optimally. Existing data storage techniques are analyzed to derive a new solution in storing versioned data in order to manage costs and avoid having to pay a heavy price in storing non critical data in a Cloud.

II. WORK DONE

The advent of Cloud computing leaves us with a fact that the data that is believed to be "ours" is not really physically stored on our personal computer. In fact, it is stored on a remote server and is available to us virtually [l]. The storage space used by this data on the remote server is not free. The methodology to put forth this article uses pre-existing solutions to store data on remote servers. More details will be provided in the later chapters of this document. Several research papers talk about how data can be stored on Cloud systems today. This shift from local storage to Cloud storage is taking place in order to minimize capital costs involved in storage of data. Cloud computing possesses a number of benefits with regard to storage and costs. Firstly, the costs involved in data storage are cut by a noticeable difference. Also, the time and efforts involved in maintenance of the data also is minimized. In the further parts of this section, we shall give more insight on how we can reach the solution we intend to propose.

The concept of "Data redundancy" is studied in order to understand the importance of data replication. Data can be sectioned based on the frequency of usage. Thus, we aim at partitioning the data stored on the Cloud as critical and non-critical data. Once this is done, our next step would focus on reducing costs involved in storing data. This is done by attempting to bring down the redundancy of data that is not in use for a noticeable amount of time, hence the name "non-critical data." Several Cloud storage types are studied in the further chapters to concede how data is stored. Below mentioned are a few examples of storage servers dedicated in storing data in their servers remotely, and thus cut capital costs involved in data storage.

Amazon's Simple Storage Service (S3) provides data storage for users on the internet. This is studied in detail to understand how file systems are stored on the Amazon Cloud. To add, Rack space is also a used for storage. Both Amazon and Rack space provide for data storage and have inbuilt tools to recover data at the time of a disaster. Redundant information on a virtual server aids in data recovery. In our article, we aim at using these concepts to derive a better understanding of redundancy levels.

This would further help in partitioning critical information from non-critical data. A scenario for small scale and medium scale organizations is taken to get a pronounced view of our area of research. A company would like to store all the old files even though these files have not been used for a long time. Such data can be categorized as non- critical data. Although this sort of data can be categorized as non- critical, allocation of a significant amount of memory for future reference of a company's records. We would research on these lines and thereby reduce the costs involved in current mechanisms of storage of data on the Cloud.

There are critical considerations made with time and cost while this article is being conceived. Care is taken to ensure that the problem in question will be analyzed to deduce an effective solution, with technical justification and an instance to implement this solution to its best capabilities in the real world scenario, keeping in mind the time span taken to deliver the same. Versioning is another feature which is used in order to derive a conclusion to store old non-critical file systems in a cost effective way.

III. LITERATURE SURVEY

Cloud computing always is tagged with Storage, security and cost concerns in the recent past. In a nutshell, most researches on Cloud computing today can broadly be classified in two types: Storage and security of Clouds and its computation. Cloud storage on the whole, talks about the issue of outsourced storage of data. Ateniese et al. first termed a model called provable data possession (PDP), this let a client verify his data stored at non trusted server and proved that the data in possession by the server could not be used by it [2]. In order to carry out the process, they took help of RSA-based homomorphism tags to audit outsourced data and ensure that it is available at all times. But they failed to keep in mind, the dynamic data storage technique, which says that upon sudden changes in data, they need to be updated and stored in the place they allocate.This needs to happen dynamically in order to make sure that the data being used is always updated and current [3]. Nevertheless, their later work proposed a slightly dynamic edition of the same PDP scheme [3] [4]. Further, an almost identical scenario was presented where data storage was partial and dynamic and happened within a Cloud [5].

At present, research on storage issues of Cloud computing still looks into minimizing the redundancy associated with data. There could be several data chunks that do not require replication. Redundant data makes it easy for Clouds to be able to beavailable with data at any given point. But the question here is, how much of data needs to be redundant or replicated and what type of data can do without replication. Since the storage cost of data is billed based on space usage, measured in GB, the user requires optimal storage options. Reduced redundancy would also increase the efficiency of performance and computation of data that reside on Cloud systems.

A. Data Storage

Increasing costs involved in the purchase of memory makes data storage a challenge faced by almost every organization or individual today, since most of the companies are now considering the option of moving all their data into the Cloud. We use the various advantages of Cloud computing to store data on the Cloud, so that the data is easily available to its users. Replication and redundancy of this data is to ensure that data is easily accessed by users virtually. The reason for opting to store data on a Cloud would be to improve business by bringing down capital costs involved in purchasing memory for data storage. Several companies initially invest on infrastructure in order to carry out their businesses. With the influx of Cloud computing, these costs are cut by a huge margin. More on infrastructure of Cloud computing will be discussed in the further chapters. The cost of storing data online is much cheaper than storing it on local memory. Also, the user/company would pay only for how much data they intend to store. Scalability and usability is improved upon moving data storage to the Cloud.

Documents can be shared among several users at the same time with the help of replication. Data replication will be explained more in detail in the further parts of this chapter. Data on the Clouds reside as data servers. Here, the user is provided with several means to recover from data loss, therefore making this type of storage beneficial. Data recovery is easily done in such type of storage.

Consider a scenario where a company has encountered an unrecoverable deletion of a whole data server. It is required to resolve this problem in a timely manner, such that operations could be resumed. The data that has been lost could be critical. It could take up to an hour for operations to resume, depending on the complexity of the problem. For some users, one hour of unavailable data could mean huge loss of business and for others, even an entire week of unavailable data would not alter in operations involving the data. The first and foremost solution here would be to make the lost data available in another data server in a timely manner, which means that the lost data needs to be replicated or moved to another destination from where its users can use the data. Hence, Data redundancy becomes necessary in such circumstances. The next challenge is to develop routines that can recover the data and also move the data from its primary location to another data center. This ensures its availability even when the disaster occurs.The characteristics of Cloud computing has methods to recover data that has been lost during a disaster [6]. For example, the Amazon S3 synchronizes user's work regularly while keeping a mirror of the production environment we work upon. This is an efficient way to recover data at the time of a disaster. Amazon S3 also has tools that help with data encryption.

Cloud computing accommodates for data storage in a cost effective manner, to minimize the costs involved with infrastructure, application processes and other businesses. The data services running on Clouds offer many solutions especially for small scale businesses. These solutions include data replication by file back-ups, done by making disk images of the processes and environments they work with. Even the data transfers in and out of Clouds are cost effective. The customers here decide on how much storage space they require, computation of processes and bandwidth usage. They pay only for how much usage they require. Along with storage system, Clouds also provide the customers with a file system to organize their data within the Cloud [7]. In the further part of this chapter, we would study more about the popular types of Cloud storage available today.

B. Storage Types

There are several data storage types on Clouds presently. All the storage centers target the same goal, i.e. to guarantee easy availability and accessibility of large scale data storage for IT administrators, developers and several clients [8]. Data storage services generally provide many benefits such 'as low costs, data availability and dynamic sharing of these resources among several end users [9].

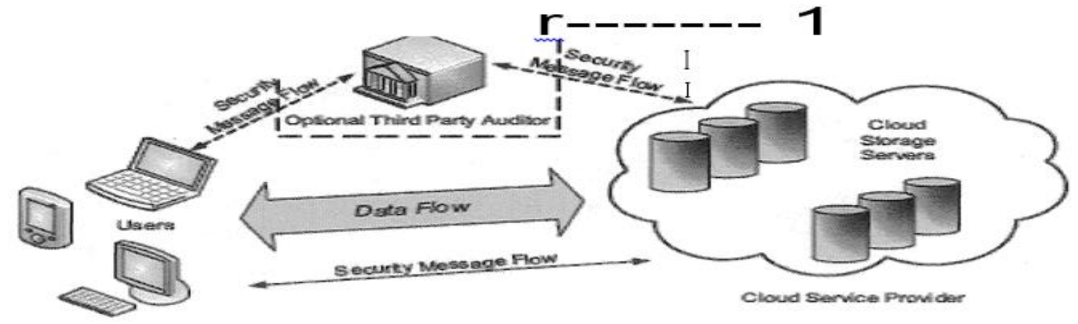


Fig.1: Cloud Storage architecture [10].

Among the several Cloud storage centers, Diomede, Google Docs, Google Cloud picker, Sun Cloud and Cloud loop are good examples of Cloud storage options that co- exist with Amazon's Simple Storage Service.

C. Diomede

This type of Cloud storage is "energy-efficient", dealing with data back-up, file archival and hosts file for data storage services. It can be fused with several applications using

"Restoration Hardware (REST)," "Simple Object Access protocol (SOAP)" and Microsoft programming interfaces [8]. The storage costs for Diomede is considerably less when compared to other Clouds, but the interface with other applications is still under development.

D. Google Docs

"Google provides this service." The functionality of Google docs is not restricted to only word and spreadsheet processing. The data stored here can directly be linked to the Google Apps engine, which is an environment for Cloud based applications. The storage is free up to 1GB and addition storage would cost $0.25 per GB [11].

The above Cloud storage options are constantly improving to keep up with contemporary Cloud storage services. Small scale businesses, especially the ones that have limited budgets are all shifting towards the Cloud for business and enterprise computation as well as storage. The benefits of these would be seen in the later parts of this article. But among various reasons for this sudden shift, the main ones would be disaster recovery and reduced expenses. The company would be dispensed from purchasing expensive hardware for storage. They simply store data at various data centers on the Cloud. This constantly takes a back up of all the data to recover from data unavailability. The redundancy stored on a Cloud is basically information that can be made available at various data centers at any given time. This involves having to dedicate a substantial amount of the organization's budget.

E. Data Redundancy

Data redundancy in simple terms can be defined as the duplication of data in order to make the same data available at all times to its users should one of the sources become unavailable. It is a means through which fault tolerance is done. Here, the same information is saved in several locations at the same time. When the source data is being changed or manipulated, the information in all the other location also subsequently changes. The information that resides in the duplicate locations need to be updated when the original data is updated. This is essential since the same information is stored at different locations and is being used by different users. So, any change in one, must cohere to the other duplicate data as well, or a data miss-match would occur.

The advantages of Data redundancy are many. Most importantly, data unavailability can lead to computational malfunction, especially to people working with real-time data involved in banks and share markets. The data is very crucial and needs to be updated and available at any given point in time. When data loss occurs, redundant data could be used while the original data is being updated. This makes data available at all times at different data stations. Should the data become unavailable at one source, the redundant data from another source is provided as back up. Also, since there are several users making use of the same data at one time, redundant data makes the data more accessible, thus increasing performance and efficiency of computation. The main disadvantage of redundant data is the utilization of space. Since the same information is present at several locations, there is too much memory consumed in storing duplicate data. This eventually leads to high costs involved in data protection and data maintenance related issues.

IV. TECHNOLOGY APPLIED

The system makes use of various tools and technologies to support the execution of the solution we intend to deliver. We have made the best use of Windows, Linux Debian lenny, VMware, FUSE API, Amazon AWS console and GPG.

F. System Construction

The system required to show the file system upload onto the Amazon S3' s Cloud is a Dell Studio 1537 Laptop computer running on a Linux Debian Lenny operating system. The FUSE API is used to mount the files onto the Cloud over an Amazon AWS console. Boto is also used for interface with S3. GNU Privacy Guard (GPG) is software used to provide encrypted data on the Cloud. Documents are prepared with Microsoft office 2007.

G. FUSEAPI

Fuse (File system Use space) is a simple library Application Programming Interface (API). This package is used to mount files securely over the Amazon Cloud. It provides efficient interface at kernel. The installation of fuse on Linux systems is a simple and free process.. The file descriptor then is used to mount the actual file system [17].

Amazon's Web Services (AWS) and S3- Simple Storage Service (S3) is a Cloud storage option designed to deliver simple data storage service to its users. The developers benefit in various ways with S3, as it is hassle free and provides for efficient web-scale computation. Its interface is not complicated due to which, storage and retrieval of huge data chunks from any remote server is possible. This makes for highly scalable and dependable infrastructure, while providing speed and security [18].

V. RESULTS AND DISCUSSIONS

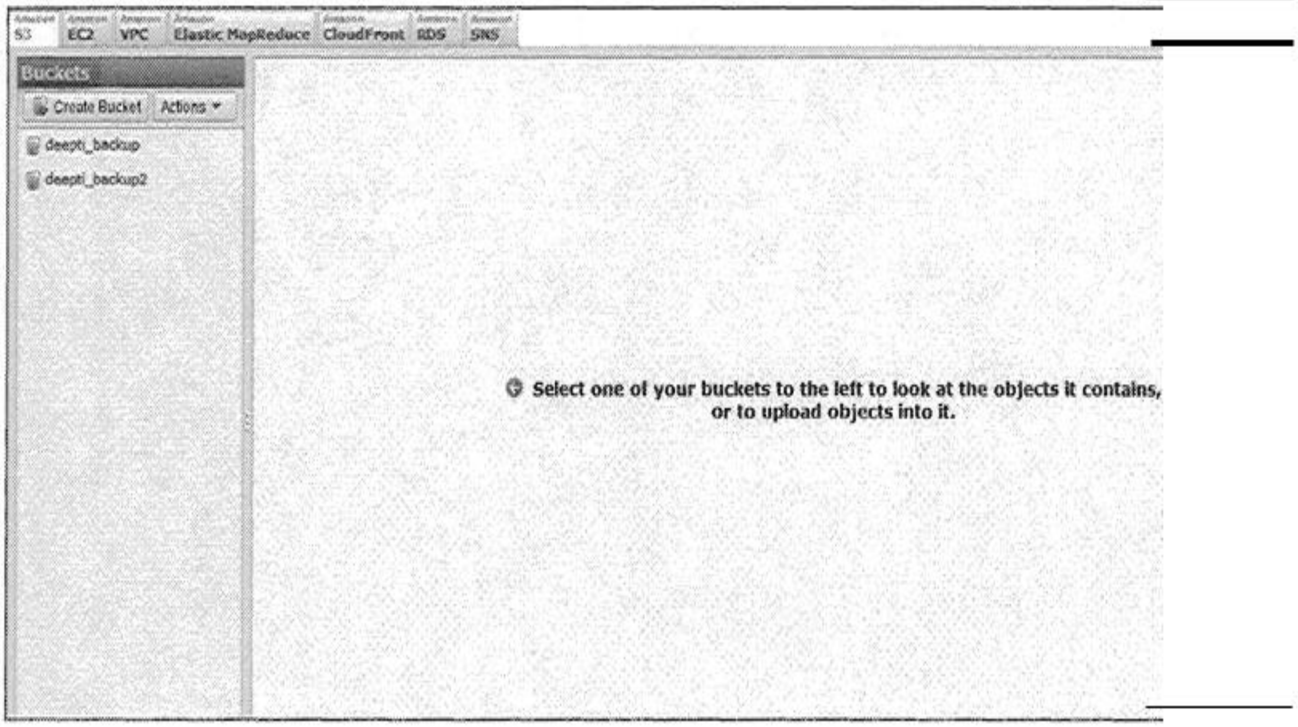
Internet services are growing rapidly, with most of the computation now taking place virtually on Cloud systems. This implies that the data involved in computation is stored in data nodes located on remote servers. Although this sort of storage has led to a huge decrement in physical storage costs and maintenance charges, an in-depth study is required to be performed in areas such as "Data redundancy" in order to further minimize storage costs on the Cloud. Consider small and medium scale organizations. These companies aim at avoiding capital costs by shifting towards Cloud computing. Hence data records maintained by these companies are shifted to the Cloud, available as virtual data resources. Virtualization involves maintenance of replicated information at various data centers, available as data servers on Cloud systems.

Data replication of critical data is highly essential for enterprises as well as businesses. There are several benefits of redundant/ replicated data. Firstly, it provides for easy availability since several clients as well as servers are required to access the same information at a given point in time. Also replicated data helps in the recovery of the data system in case the data server renders unavailable. More benefits of data replication has been elaborated in the "Data redundancy" chapter of this paper.Now, consider a situation where a company decides to shift its entire database and computation to a Cloud. After gaining a brief understanding of data storage on the Cloud, we can conclude that several data servers hold redundant information of this company's data. The old versions of certain data are simply present for the organization's records. Such data is also being replicated as it's the Cloud's property to provide replicated files.In such a case, there would be minimum need of replicating such data every time a new version is created. Storage costs for saving non critical data can be minimized by deciding where to store this sort of non-critical data.

H. Existing solution, Advantages and Disadvantages

Amazon's S3 in particular is studied, among the several data centers available on different Clouds involved in storing data on virtual data Clouds. Simple Storage Service(S3) contains storage buckets that store and replicate data. There are storage buckets that hold critical data, as well as replicated data. This implies that the oldest versions of certain files are also stored in these storage buckets.

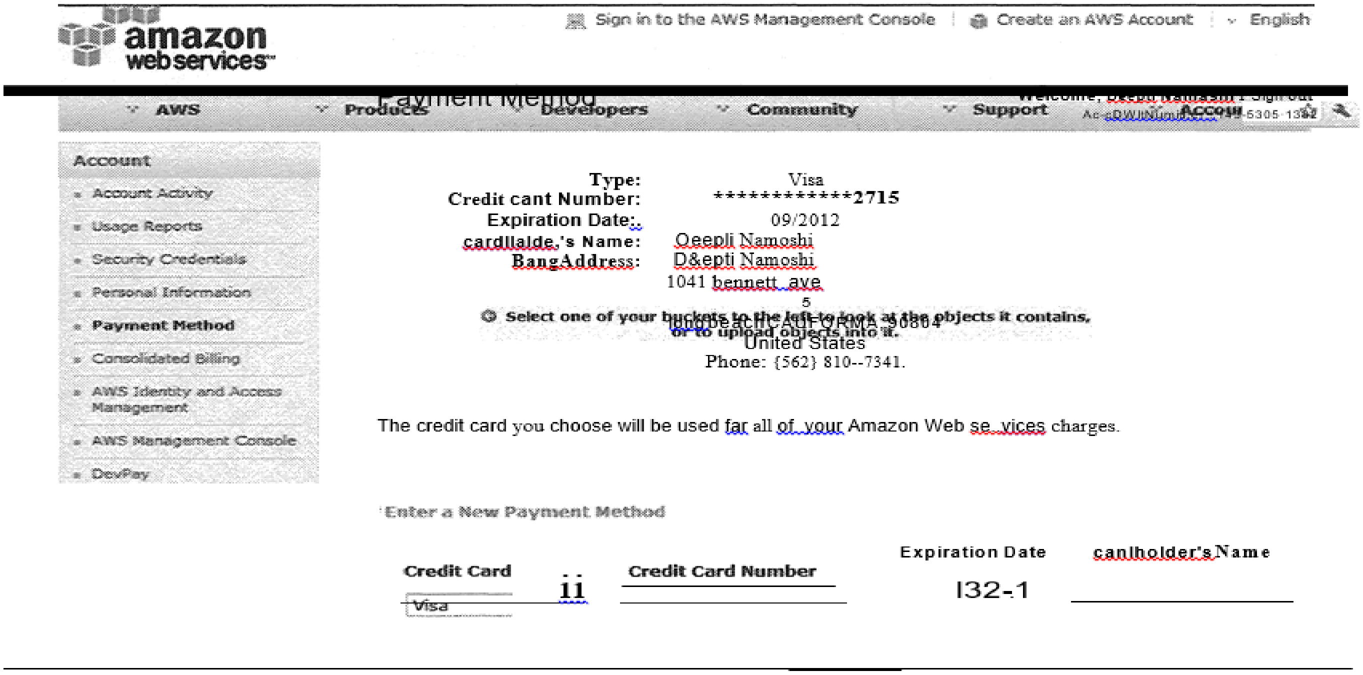
above is the figure of a storage bucket of S3. The creation of this storage bucket is done by first creating an account in the Amazon's AWS console. A registration process requires the user to provide details such as Name, Billing information, location, etc. Figure 2 shows a screenshot of the registration procedure as well as the billing information.



Select one of your buckets to the left to look at the objects it contains, or to upload objects into it.

Fig.2: S3 Data Storage Bucket.

The account information is required for the registration process and also the credit card details are required by the Cloud provider in order to maintain logs of usage and subsequent payment for the same.



Type: Visa
Credit cant Number: *************2715
Expiration Date:. 09/2012
cardllalde,'s Name: Oeepli Namoshi
BangAddress: D&epti Namoshi
1041 bennett ave
5
lungbeachCAUFORMA 90804
United States
Phone: {562} 810--7341.

The credit card you choose will be used far all of your Amazon Web se..vices charges.

Enter a New Payment Method

Credit Card		Credit Card Number	Expiration Date	canlholder's Name
Visa	ii		I32-1	

Fig.3: Account information and payment for account in S3.

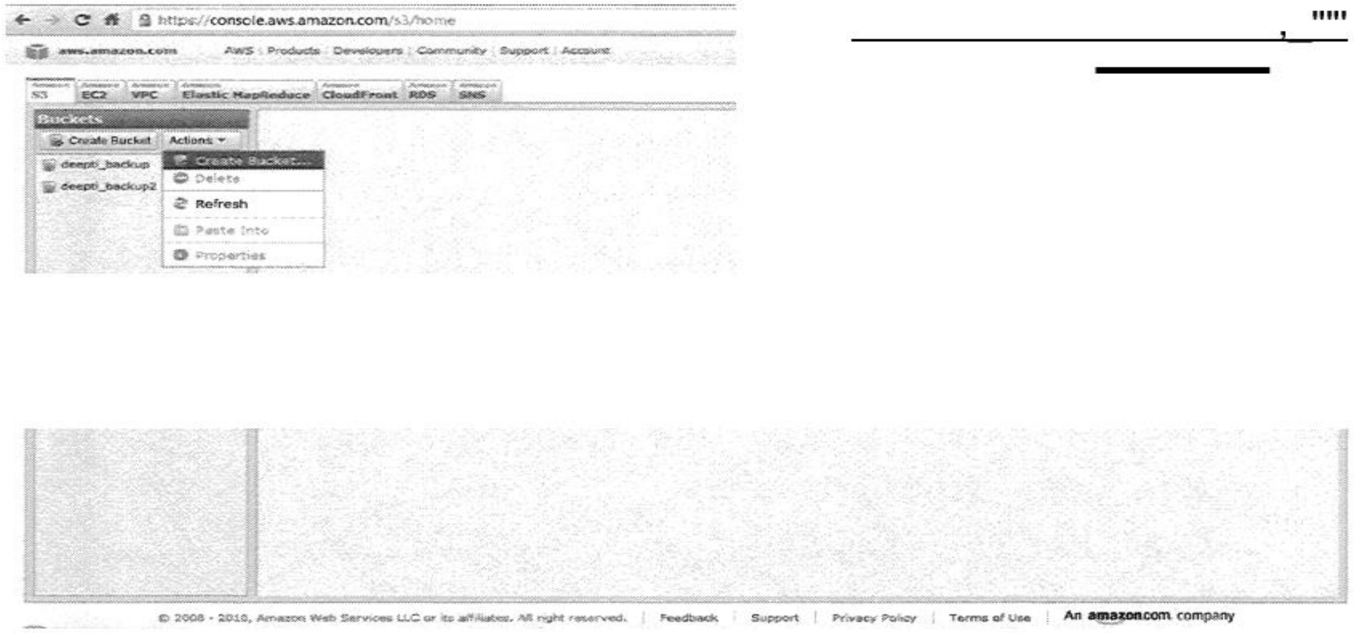
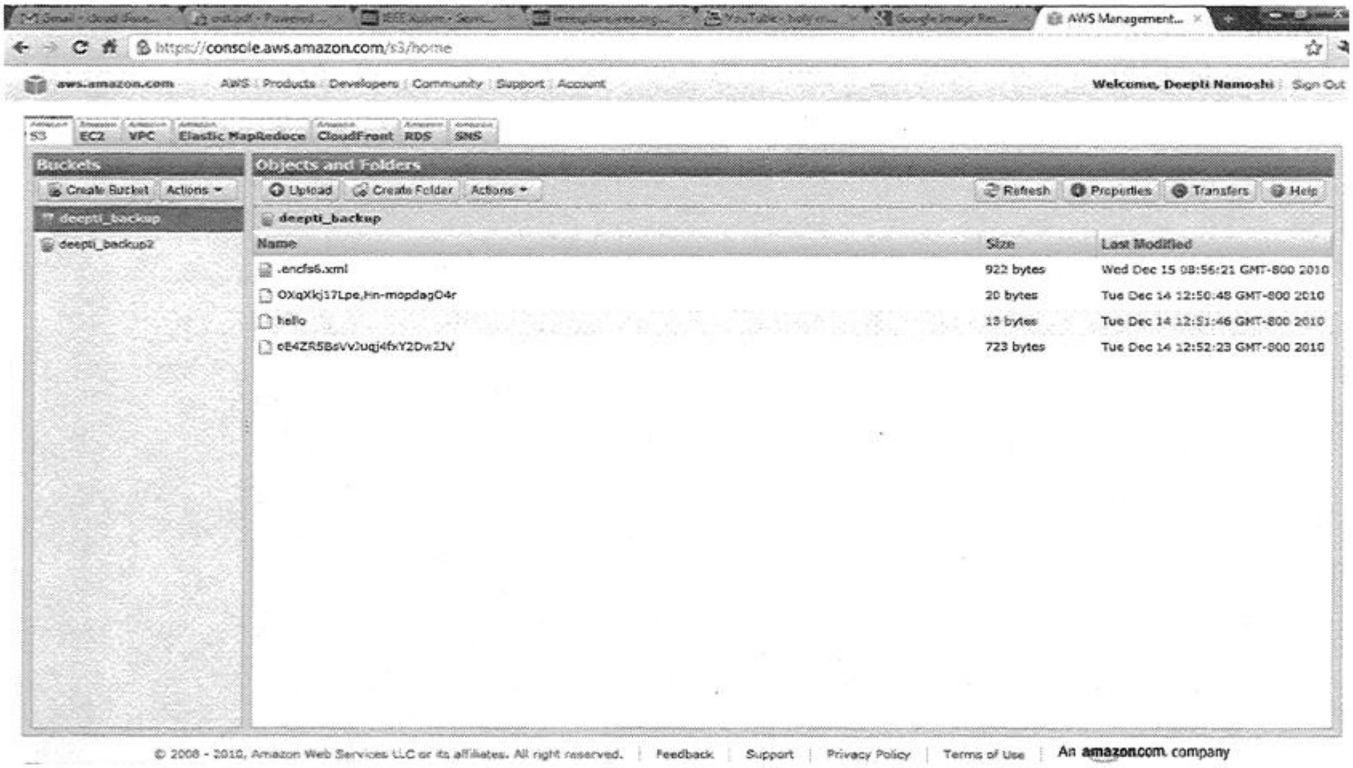


Fig.4: Create bucket Procedure.

Figure 3, shown above highlights how the creation of a new bucket is done on the S3 console. The create bucket icon is clicked. After doing so, a suitable name is provided to the bucket. Once the bucket is created, local files are uploaded onto the Amazon's S3 bucket with the help of the FUSE API procedure, discussed in the earlier chapters of this article. This is used in order to mount local files onto the Cloud. Also, the Boto interface is used for Amazon S3.



Name	Size	Last Modified
.encfs6.xml	922 bytes	Wed Dec 15 08:56:21 GMT-800 2010
OXqXkj17Lpe,Hn-mopdagO4r	20 bytes	Tue Dec 14 12:50:48 GMT-800 2010
hello	13 bytes	Tue Dec 14 12:51:46 GMT-800 2010
cE4ZR5BsVvJuqj4fxY2DwZJV	723 bytes	Tue Dec 14 12:52:23 GMT-800 2010

© 2008 - 2010, Amazon Web Services LLC or its affiliates. All right reserved.
Feedback
Support
Privacy Policy
Terms of Use
An amazon.com. company

Fig.5: Created bucket with files, size and usage.

Figure 5 shows the details of the data residing within the bucket. This information is available as a log that contains fields such as "Name" of the files stored, its storage "Size", along with the "last modified" date field. The files maintained in the buckets on S3 are encrypted using the GPG software for encryption. This ensures authenticity of the data accessed by the users. The basic functionality of the Amazon S3 is as below:

Performs Read, write and deletion of objects that have data ranging from 1 byte to 5 terabytes each. Storage of objects is unlimited. The objects stored in bucket are stored and retrieved through a unique secret key for security.S3 bucket is available for storage in any of the regions, the regions can be chosen to reduce the costs involved, to have optimal latency as well as to manage address spaces.Objects are specific to the region they are stored in. They can only be transferred explicitly by the user to another region.Usage and data manipulation rights are given to users by providing authentication. This ensures security of data. The data objects can be set as private or public.

The users of Amazon S3 are required to read and adhere to the norm of the Amazon S3 service level agreement [19].The above solution provided by S3 explains the technique of data storage in S3. Several end users imply this method for computation using the Cloud. This could ease the organization by reducing the expenses involved in allocation of physical memory to store file systems, as well as maintenance costs, etc. The organization is required to only pay for the amount of storage space they use on the Cloud. This cost is significantly cheaper than physical memory costs. Also, data encryption included in S3 ensures data security.

But a disadvantage with this system is encountered when versioned files are required to be stored. Replication is essential to ensure data availability and for recovery of data during disastrous conditions. Nevertheless, storage of old versions of files leads to a down curve in the performance of the system due to the possession of unnecessary redundant data files. The cost involved in storing old versions of files cannot be avoided.

In the further part of this chapter, a brief solution is provided that aims at cutting costs involved in storing non-critical information on the S3 Cloud.The below figure gives the current costs for data stored per GB in S3. The first column gives costs for S3 storage, whereas, the second column gives the cost for another storage bucket, called the Reduced Redundancy Storage (RRS) bucket.

I. *Proposed Solution*

Amazon's Simple Storage Service (S3) has recently launched Reduced Redundancy Storage as a new feature. Reduced

Redundancy Storage (RRS) is a recent storeroom option that allows customers to mitigate costs by storing less important, reproducible files or data objects. The redundancy level here is lower than the one provided by S3. This is the main reason for costs to drop. The solution is good for distributing content that is durable on any storage space and for easily reproduced data. The object and data replication done here is not as much as seen in S3. "RRS provides durability and availability of 99.9% throughout a year." Hence the data loss is about 0.01% annually.

Making use of this new feature, we intend to resolve the issue of having to pay a heavy price to store replicated data on the Cloud. A simple scenario is explained as below, to derive a flow chart of our solution. The results, comparison with previous solution and probable implementation in real world would be included in further parts of this chapter.

J. Scenario :.

Consider a small/medium scale organization that has recently decided to move all its business onto a Cloud after understanding the benefits of the new Cloud computing methodologies, in order to scale better in business and manage costs efficiently while doing so. Now, this requires the organization to move its data objects onto the Amazon S3 for storage. The cost calculations can be made by referring to the chart in Figure 12. Clearly, the RRS costs are much less when compared to actual S3 bucket. The necessity of storing important/critical data on S3 is no doubt a basic requirement by the organization, since S3 provides for data recovery and security up to 99.99999%.

However, the organization might contain several redundant data that is never used on a regular basis, maybe even for more than two years. But since the company has shifted its entire database onto the Cloud, it is required to store the unused non critical data as well.

To resolve this issue, we have devised a small approach called the "CES method." This investigates the data in question and concisely decides on what type of data is required to be stored it in the RRS bucket. This would minimize redundancy of non critical data objects. Also, since the costs involved in storing data in RRS is cheaper than S3, the overall expenses for storage would be reduced.

We also consider the fact that RRS does not take too much of redundant data, hence the CES method we propose would do the job of removing redundant/replicated information from the old file versions and store only the non replicated data.

K. CES Method

A simple formula to decide on what data needs to be extracted from S3 and stored into RRS in order to minimize costs is conceived. This can be called as the CES (Compare, Extract, and Store) Method to move non-critical data into RRS from S3 storage.

Firstly, before we start the execution of this method for non critical data, it is essential to classify and select non critical data from a pool of file versions in the S3 bucket. This is done by a simple comparison made in the "last modified" field of the log maintained by the bucket. In case the file version has never been modified since several months, with newer versions of this file already present in the S3 bucket, we target those specific file versions. The flowchart seen below gives a clear picture of this procedure.Figure 13 shows a procedure to first qualify what data needs to be stored in RRS. Hence, to do so, the latest entry of file version is compared to the already existing versions of the same file. Since version numbers are sequential, the existing version number is subtracted from the latest one. If the version is four versions old, then it is still regarded to be critical information and necessary to be replicated. But in case it is older than four versions, then it is removed from the S3 bucket to carry out the CES method, thereby preparing the data to be suitable for storage in the RRS bucket.

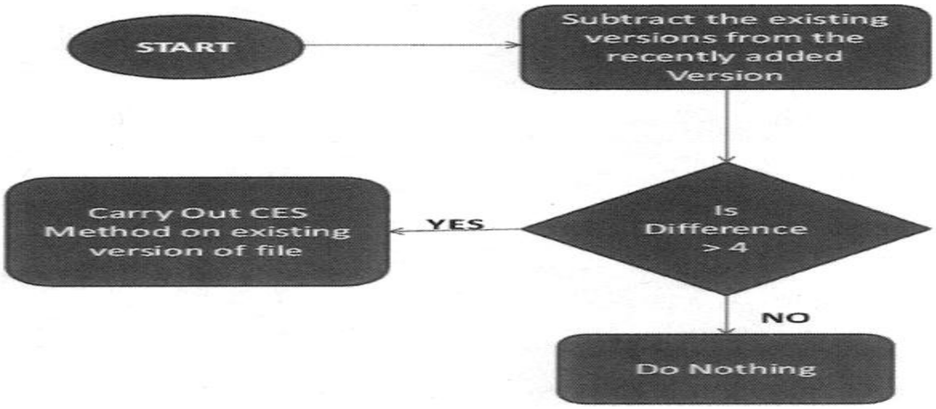

START
Subtract the existing versions from the recently added Version
Is Difference > 4
YES
Carry Out CES Method on existing version of file
NO
Do Nothing

Fig.6: Decision making flowchart to carry out CES on files.

In the above figure, one might wonder as to why the comparison is being made with the number "4". This value is taken as an example in our article. We have selected this value presuming that these many versions are crucial and could still be required to be saved in the main storage bucket. Consider only two versions being saved in the main bucket. This could be questionable, because must there be any data loss in the current version and if the previous version is required for reference, the previous version might require the version earlier than that for reference. Therefore, we have been safe and decided in storing four file versions in the main storage bucket. But, the 4 can be replaced with "n" where the company or user could decide on how many versions they would like to save in the main storage bucket, based on its usage and last used date, as well as the space availability and the amount of money they are willing to spend for such storage.

L. CES- Compare Extract Store

The foremost thing that is required here is that the oldest version of the file is stored in the RRS bucket implicitly. After this happens, the CES method is applied to the following file versions.

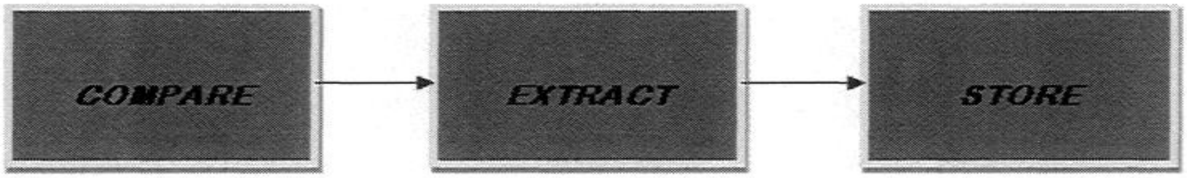


Fig.7: CES COMPARE

Here, several existing tools are used to compare two version of a file. The file that qualifies for the CES method is taken and compared with the pre existing oldest version in the RRS bucket. Tools such as "Difference viewer", etc. are used to compare the actual objects in the files. The comparison result highlights objects that are not replicated. The replicated objects are hence not highlighted.

In this procedure, simple Amazon data base queries are applied to extract the non- replicating objects from the data tables that are highlighted. Hence, the queries extract only that data which is new and non redundant after the comparison between oldest version and next version. Once this data is extracted, the final step of this method is performed.Once the non redundant data has been extracted from the file version that is immediately newer than the oldest version, this data is stored as the next version of RRS. The version numbers are sequential here, to ease identification.This way, the entire process is repeated every time new file versions enter into S3. The "COMPARE" method compares the qualified file version with the previously added version present in the RRS bucket. This leads to saving information that is non- redundant, and also saving so much storage space in S3.

To show the above process in a better way, we present a simple example. This would help us get a clear picture of the CBS method.

Steps:

- Consider a file with version number 1.1, this is the oldest version of a file which has been edited and replicated a number of times in the S3 bucket. The current file version of the same file is 1.10.
- We check to find out that versions 1.1 to 1.10 of a same file are present in S3, hence we decide to move versions 1.1 to 1.6 into the RRS bucket to avoid storage costs in S3.
- Version 1.1 is implicitly moved in the RRS bucket. Next, the immediate version after 1.1 is 1.2, so the CBS method is applied on this first.
- Consider version l.l=data A, Version 1.2= data AB (Compare Process)
- Extract Process= AB-A= B
- Store Process. In this, version 1.2 of the RRS bucket contains only the non redundant data between 1.1 and 1.2 , therefore, Version 1.2=B
- After doing this, the process is repeated for the next immediate version in S3, which is 1.3 and is compared with 1.2 of RRS.

A cost comparison is made after the old file versions are removed from S3 and stored into RRS. Clearly, there is a significant different in the cost curve. Thereby reducing costs involved in storing non-critical data on a Cloud.

The above solution holds when the comparison made with the immediate version results in common (redundant) data. Now, consider an alternative scenario, where the data in Version 1.1 = A, version 1.2= AB, then we store the data B in the Version 1.2 as per the CES method. But the following version 1.3 contains data = AC. Here, we see that when we compare versions 1.3 with its immediate previous version 1.2, there is no common data. But version 1.3 has redundancy, since it contains AC, where "A" is present in version 1.1. In order to resolve this issue, the previous versions are first merged and later compared to the latest entry. This would mean that it solves the problem of storage and hence reduces costs, but the comparison and merging of all previous versions have a time complexity. However, the issue here strictly concentrates on expense and space.

Another alternative solution is presented in the below Figure 8, which clearly shows how the redundancy in the RRS bucket can be reduced. The function and example of the below flowchart is also provided for a better understanding.

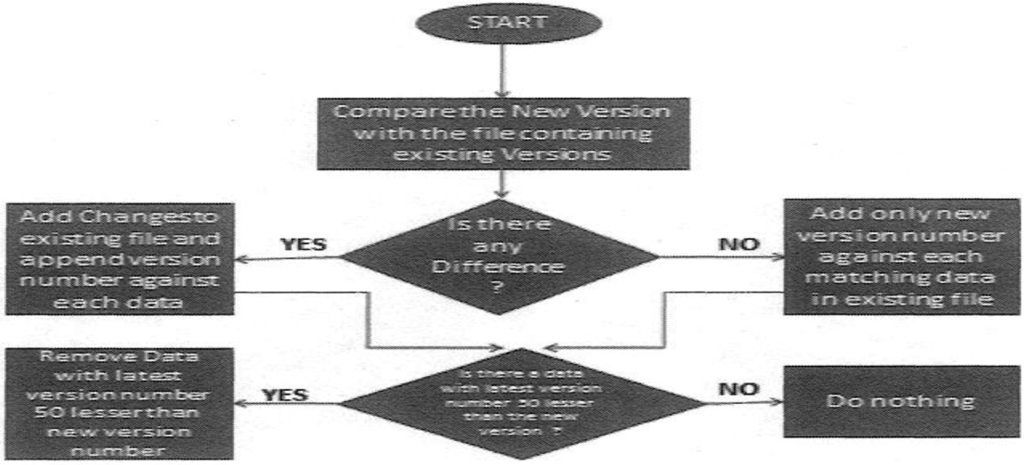

START
Compare the New Version with the file containing existing Versions
Is there any Difference ?
YES
NO
Add Changes to existing file and append version number against each data
Add only new version number against each matching data in existing file
Is there a data with latest version number 50 lesser than the new version ?
YES
NO
Remove Data with latest version number 50 lesser than new version number
Do nothing

Fig.8: Alternative CBS.

The above Figure shows a flow chart to select the type of data required to be stored in the RRS bucket. Considering the scenario mentioned in the previous paragraph, the above flowchart gives a solution to store the versions of files, without having to save redundant information. The explanation to the flowchart can be given in the following steps.

In the first step, we compare a new file version that goes into the RRS bucket, holding an already existing file version. Consider, version 1.1 is already present in RRS, with data "ABCDE". The next version that enters the bucket would be 1.2, with data "ABCDEFGHU". Therefore, we can do a comparison between these two versions and save a file with the version number, against the data it already holds.

When the comparison has a change or addition in data, for example, 1.1 = A. 1.2 = AB, after comparison, we have data "A" in both 1.1 and 1.2. Hence, the version numbers 1.1 and 1.2 are written against this data "A". Similarly, data "B" has version number 1.2 against it. This process is repeated such that data is stored with the version number associated with it. Consider a next version, 1.3 with data "AC". This means, the data "A" has version numbers 1.1, 1.2, 1.3 against it. Data "B" has 1.1, 1.2. Data "C" has 1.3 against it. Therefore, all the versions are stored as data, with version number logs present. Upon requirement, they can be reproduced by seeing what data belongs to the respective version number.

Also, another option of deleting old versions of files in the RRS bucket can take place by checking the log to see when they were last modified. If the data has never been accessed or changed, it means it has not been used. This data can be removed or deleted from the RRS bucket in order to make more space for other files, and hence reduce costs.The above scenario and solution is proposed just in case a file comparison is required for files with data that is not common between sequential versions. Also, since there are latest versions of files already existing in the critical storage bucket, the RRS bucket would be apt to store only few versions of the non critical data rather than making an investment in storing several versions of the same file.

M. Justification of the Solution

A graphical comparison is made between the costs involved in storing data in S3 and the new storage in RRS after the CES Method has been carried out. The graph is drawn between the Data space usage on the X-axis and Cost on the Y-axis. This comparison can be shown in the below Figure 9.

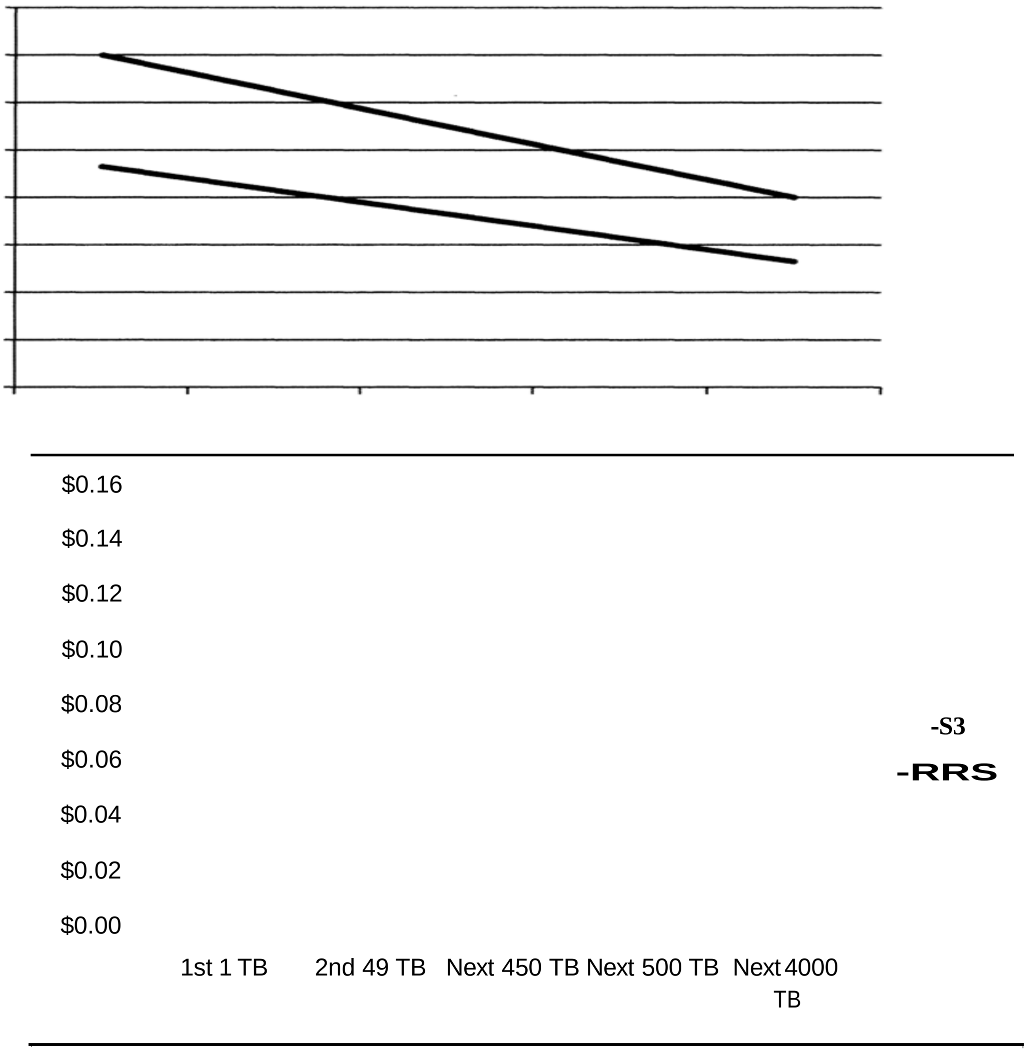


Fig.9: Graph Showing the difference in costs.

Clearly, in the above graph in Figure 17, the costs for RRS is much lower than S3 storage. Upon removing more redundant data from S3, extracting the non- redundant data from it and saving this data in RRS, the cost is further reduced. This is so due to the removal of unused redundant data from S3's bucket and moving only non redundant information into RRS.

RRS also reproduces data, therefore, in future, must a need occur to reproduce the older versions of files, the latest stored version can be compared with its previous version in order to merge data for reproduction of the required files. The above theory can be applied to run small scale and medium scale businesses over the Cloud. A huge difference in storage expenses will be noticed with the help of this theory. CES method can also help in reducing the storage space in S3 bucket by removing redundant information that has never been used in a long time. S3 bucket holds the latest replicated file versions, as these data files are necessary in the restoration of data at the time of a disaster.

VI. CONCLUSION

We conclude the article by providing an overview of Cloud computing and the storage issues faced by Cloud users. Users opt to consider Cloud computing as a solution to minimize capital costs involved in their business strategies, its ease of use as well as hassle free maintenance. We also have managed to propose a solution to reduce redundancy of data involved with old versions of files, called non-critical data. The proposed CES method concisely gives way for cheaper storage of non-critical data on a Cloud environment, specifically, Amazon's S3. Also, the transfer of files from S3 to RRS allocates free space for critical data in the main storage bucket. The removal of unimportant data from S3 helps in better performance of S3 as it is involved in the computation of several applications that deal with critical information. A cost comparison of S3 and RRS led to the derivation of the CES method, which minimizes capital costs involved in data storage. We have also been able to prove the difference between costs with the help of a graph curve between the S3 and RRS storage.

The CES method is restricted to non-critical data and old file versions for small scale organizations. An effort can be made to reduce redundancy of critical information as well. This would increase performance and efficiency of Cloud services. Also, even though the costs have been reduced, we have not been able to test whether the data can be reproduced from the RRS bucket. Hence this can be an area of research to reproduce old file versions when necessary. This form of storage need not be restricted only to Amazon's S3 and RRS. We could also aim at generalizing the solution for all Cloud environments by creating two types of storage options, one for critical information, and another cheaper storage for non-critical information. This would cut down non- critical redundant data and save the storage space available in Clouds.

VII. REFERENCES

[1]. Q. Wang et al. "Enabling public verifiability and data dynamics for storage security in Cloud computing," in 14th European Symp. Research in Computer Security (ESORICS 2009), Saint Malo, France, 2009.

[2]. A.G. Briscoe et al. "Community Cloud Computing," in Proc. Cloud Computing: 1s t Int'l. Conf (CloudCom 2009), Beijing, China, 2009.

[3]. L. Wei et al. "SecCloud Bridging Secure Storage and Computation in Cloud," Proc. IEEE, Int'l. Conf Distributed Computing Systems(ICDCSW 2010), CS Digital Library, 2010, pp. 52-61.

[4]. C. Wang et al. "Ensuring data storage security in Cloud computing," in 17th IEEE Int'l. Workshop on Quality of Service (IWQoS'09), Charleston, South Carolina, USA, 2009.

[5]. G.Reese, "Using the Cloud for Disaster recovery .O'Reilly Community biogs." Apr. 2009; Available: http://broadcast.oreilly.com/2009/04/using-the-Cloud-for-disaster-recovery.html

[6]. J.R.Row, "Cloud computing and disaster recovery plans." Bright Hub, 30 Dec. 2010; Available: www.brighthub.com/environment/green-computing/articles/71273.aspx

[7]. C.Wang et al. "Toward Publicly Auditable Secure Cloud Data Storage Service." In IEEE Trans. on Parallel and Distributed Systems, Aug 2010, pp. 19-24.

[8]. M.R Tribhuvan et al. "Ensuring Data Storage Security in Cloud Computing through Two-way Handshake based on Token Management." in 2n d Int'l. Conj. Advances in Recent Technologies in Communication and Computing (ARTCom), 2010, pp.386-389.

[9]. N.V.Nguyen, "Cloud Computing Security." Slideshare, 2009, Available: http://www.slideshare.net/xoai/cloud-computing-security-2153773

[10]. A.Mohamed, "A history of Cloud Computing." Computer weekly, Service oriented architecture and Web Services. Mar 2009; Available: http://www.computerweekly. com/Articles/2009/06/10/235429/A-history-of-Cloud-computing.htm

[11]. M. Lamoureux, "Cloud computing properties are not benefits," Feb 201O; Available: http://mikelamoureux.net/Cloud-computing-properties-are-not-benefits/

[12]. N.Suganth, "Cloud Computing, an Overview." Authorstream, Feb 2010; Available: http://www.authorstream.com/Presentation/aSGuest37188-316419- introtocloudcomputing-entertainment-ppt-powerpoint/

[13]. TechTarget.com, "Infrastructure as a Service." SearchCloudsecurity, 2009; Available: http://searchcloudcomputing.techtarget.com/definition/lnfrastructure- as-a-Service-IaaS

[14]. M.Miller, "Are You ready for Computing in the Cloud." lnformit, Sep 2008; Available: http://www.informit.com/articles/article.aspx?p=1234970.

[15]. K.Hwang and Y. Hu, "Cloud Security with Virtualized Defense and Reputation- based Trust Management," in Proc. IEEE 8th Int'l Conf Dependable, Autonomic and Secure Computing ,2009, pp 717-722."Fuse API." Sourceforge, 2009; Available:http://fuse.sourceforge.net"Amazon Web Services, Simple Storage Services." Amazon,2011; Available: http://aws.amazon.com/s3/"MSSQL Difference Viewer." Wintestgear, Aug 2010; Available: http://www. wintestgear.com/products/MSSQLSchemaDiff/MSSQLSchemaDiff.html.