Library Hi Tech

Cloud storage for digital preservation: optimal uses of Amazon S3 and Glacier

Yan Han

Cloud storage for digital preservation: optimal uses of Amazon S3 and Glacier

Yan Han
University Libraries, The University of Arizona, Tucson, Arizona, USA

Received 19 December 2014
Revised 19 December 2014
Accepted 11 February 2015

Abstract

Purpose – The purpose of this paper is to use cloud storage in digital preservation by analyzing the pricing and data retrieval models. The author recommends strategies to minimize the costs and believes cloud storage is worthy of serious consideration.
Design/methodology/approach – Few articles have been published to show the uses of cloud storage in libraries. The cost is the main concern. An overview of cloud storage pricing shows a price drop once every one or one-and-a-half years. The author emphasize the data transfer-out costs and demonstrate a case study. Comparisons and analysis of S3 and Glacier have been conducted to show the differences in retrieval and costs.
Findings – Cloud storage solutions like Glacier can be very attractive for long-term digital preservation if data can be operated within the provider's same data zone and data transfer-out can be minimized.
Practical implications – Institutions can benefit from cloud storage by understanding the cost models and data retrieval models. Multiple strategies are suggested to minimize the costs.
Originality/value – The paper is intended to bridge the gap of uses of cloud storage. Cloud storage pricing especially data transfer-out pricing charts are presented to show the price drops over the past eight years. Costs and analysis of storing and retrieving data in Amazon S3 and Glacier are discussed in details. Comparisons of S3 and Glacier show that Glacier has uniqueness and advantages over other cloud storage solutions. Finally strategies are suggested to minimize the costs of using cloud storage. The analysis shows that cloud storage can be very useful in digital preservation.

Keywords Cloud computing, Cost analysis, Amazon S3, Cloud storage, Glacier
Paper type Research paper

1. Introduction

Over the past 20 years, millions of manuscripts, serials, audio and videos resources have been digitized globally for access and preservation. Almost every academic library, archive, and museum have some digital initiatives to make their unique materials available on the internet for access. Best practices and digitization standards have been published and adopted by libraries, museums, and archives to produce high-quality uncompressed digital surrogates. In addition, the new trend of preserving born-digital big data from research requires a way to store and save these critical data for the future. In the past, the data (i.e. master copies) have been typically stored and backed up in traditional local storage (e.g. hard disks and tapes), while derivatives were loaded to repositories or web sites for daily access. Access to the master copies is very infrequent and usually consists of two types: first, data integrity and verification: the primary access to them is to verify the checksum to ensure data integrity in a predefined schedule per preservation policy (e.g. once every six month); and second, data review and update: typically only a small portion of data is required to be accessed. Sometimes a master copy is untouched for a long time of period (e.g. 5+ years).

Library Hi Tech
Vol. 33 No. 2, 2015
pp. 261-271

0737-8831
DOI 10.1108/LHT-12-2014-0118

While "Lots of Copies Keep Stuff Safe" (LOCKSS) is certainly the right concept, not all organizations are able to have all their data in LOCKSS boxes. As a result, in order to assess the possibility of losing data, it is critical to have scientific data and modeling to understand the risks and to justify digital preservation efforts with reasonable costs. The goal is to have enough copies to minimize data loss while at the same time understanding the costs using either traditional storage or/both cloud storage. A report from OCLC and CRL (2007), "Trustworthy repositories audit & certification: criteria and checklist" suggests that "Repository has effective mechanisms to detect bit corruption or loss," and gives an example: "If a repository's policy limits loss to no more than 0.001% of the collection per year, with a goal of course of losing 0% […]." Some academic libraries have been participated in the evaluation for the certification. Several articles have been published to model data failure rate and evaluate possibility of losing data using case studies. A paper presents a model of evaluating data loss and suggests that three copies of data have an unreliability rate of 0.033 percent (CDP, 2005). A case study of four copies of data have a unreliability rate of 0.001693 percent using sample disk and tape failure rates (Han and Chan, 2008). It is recommended that each institution have a digital preservation policy to understand the risk while keeping in mind on related resources and costs.

2. Overview of cloud storage and its uses in libraries

When Amazon released Glacier in August 2012, it was called "cold storage" and was marketed for long-term storage for medical records and archival materials. The "cold" term was used to describe materials that do not need to be accessed often and quickly. Before Glacier, as early as 2006, cloud computing providers such as Amazon, Google, and Microsoft marketed their cloud computing services, such as virtual servers, databases, and storage, toward regular end users and business. Quite a few articles have been published to describe the advantages of cloud computing and the uses in libraries since 2009 (Kroski, 2009; Tonjes, 2010; Mitchell, 2010; Han, 2010). However, few have explored the uses of cloud storage and related costs.

Cloud storage solutions have certain advantages over traditional ones, such as availability, scalability, off-site storage, on-demand, and multi-tenancy. For example, Amazon S3 claims that it aims for 99.99999999 percent durability and 99.99 percent availability in a year. Popular enterprise cloud storage solutions include Amazon S3, Google Cloud Storage, Microsoft Azure. In business world a lot of popular applications and web sites are utilizing the big three's infrastructure while building their own applications. For example, Dropbox, Tumblr, and Pinterest use Amazon S3. Amazon reported that S3 has two trillion objects in 2013[1].

2.1 Use of Amazon S3 and Glacier in libraries

Central Connecticut State University Libraries used Amazon S3 and compared the costs between OCLC Digital Archive and Amazon S3 for digital preservation (Iglesias and Meesangnil, 2010). The University of Arizona Libraries compared the costs of using S3 to local storage and concluded that local storage was still an attractive solution over cloud storage (Han, 2011). DuraCloud stated that it used Amazon S3, Rackspace Cloud Files, Microsoft Azure, and San Diego Supercomputer Center cloud storage (Branan, 2011). There was a big pitfall of cloud storage: its costs were still higher at that time. "LOCKSS boxes in the cloud" (Rosenthal and Vargas, 2012) has details on running LOCKSS on Amazon E2, EBS, S3, and Glacier. The authors also compared the economics of running on the cloud to local storage. They concluded that "until there is a radical change in one or other of these cost curves it [is] clear that cloud storage is not even close to cost-competitive with local disk storage for long-term preservation." Regarding Glacier, they also concluded that it would be more competitive if customers can tolerate delays in accessing data and/or infrequent integrity checks (Rosenthal and Vargas, 2012). Mitchell discussed Glacier to address two major concerns: the cost and durability of cloud storage (Mitchell, 2012).

Starting in 2013, DuraCloud started to offer digital preservation plans which use extensively Amazon S3 and Glacier. The plans come with features such as synchronization, healthy check, and file recovery. Technical details, plans, and costs of using these cloud storage were described in a 2013 document (DuraSpace, 2013). Currently DuraCloud offers both open source software and subscription plans. The open source software is available for download for all major operating systems. In addition, it offers four levels of subscription plans for institutions who do not want to run the software themselves (DuraSpace, 2014):

(1) Preservation Plan (using Amazon S3, one-copy): standard features; $1,800 for the first 1 TB, $700 for additional TBs;

(2) Enterprise (using Amazon S3, one-copy): standard features + access control + more; $5,500 for the first 1 TB, $500 for additional TBs;

(3) Preservation Plus (using Amazon S3 (one-copy) and Glacier (one-copy)): standard features + synchronization + file recovery; $1,925 for the first 1 TB and $825 for additional TBs; and

(4) Enterprise Plus (using Amazon S3 (one-copy) and Glacier (one-copy)): standard features + synchronization + file recovery + access control + more; $5,625 for the first 1 TB and $625 for additional TB.

Texas Digital Library (TDL) is running DuraCloud's open source software using Amazon S3 for storage price of $360 per TB/year and data out price of $120 per TB/year. Amazon Glacier for storage price of $120 per TB/year, and data out price of $120 per TB/year (Texas Digital Library, 2014). DuraCloud's implementation shows that Glacier is used to complement S3 as an alternative solution to hold an additional copy of data. The Preserving Objects with Restricted Resources (POWRR) is an IMLS funded project to investigate, evaluate, and recommend scalable, sustainable digital preservation solutions for libraries with small amount data and limited resources. The project published a white paper titled *From Theory to Action: "Good Enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions*. The project surveyed and investigated multiple digital preservation tools including Internet Archive, DuraCloud, MetaArchive, Preservica, and Archivematica. It mentioned that "Preservica is a vendor solution which also uses Amazon S3 and Glacier using OAIS-compliant workflows" (Schumacher *et al.*, 2014).

3. Overview of history of cloud storage pricing

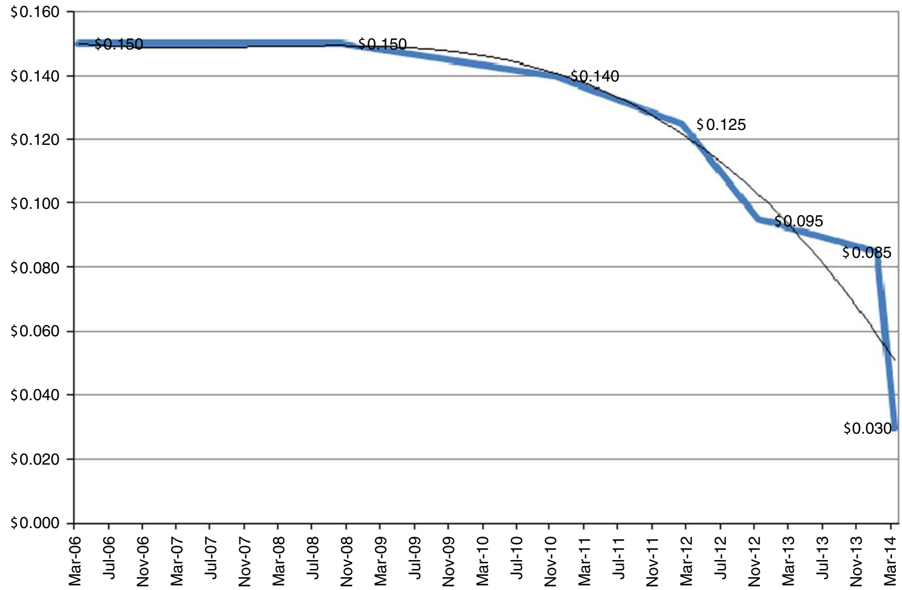
IT service competitions and price reductions are always ongoing in business. This is also true in cloud computing, as the big three providers: Amazon, Google, and Microsoft have been dropping their pricing lower and lower over the years. In the past eight years, the cloud storage pricing generally gets lower once every one to one-and-a-half year. Amazon S3, Google Cloud Storage, and Microsoft AZure pricing has been dropping from USD $0.15/ GB for Amazon S3 in 2006, $0.17/ GB for Google in 2010, to $0.10/ GB for Amazon S3 and Google in 2012, then to $0.03/ GB for Amazon, Google, and Microsoft in 2014. As Amazon is the most used and leading cloud computing and cloud storage providers, the author lists its S3 rate history below (Figure 1).

Few publications emphasize data transfer costs. However, the data transfer rate is a critical to determine the total costs of cloud storage. It is not difficult, but tricky sometimes, to calculate the real cost, as it varies on the usage, location, and sometimes the speed of retrieval. When the cloud storage service offered eight years ago, the cloud storage providers charged fees for both data transfer-in and transfer-out. Later on, these cloud storage providers removed the data transfer-in fee, but still charge data transfer-out fee. Figure 2 shows the data transfer-out fee history in the past seven years.

4. Overview and history of Amazon Glacier

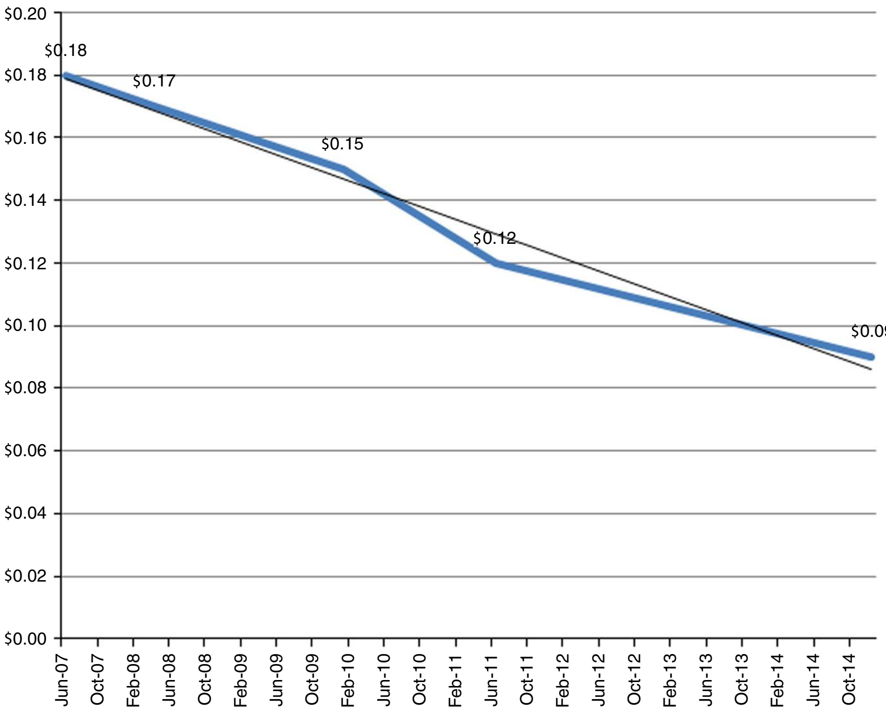
The release of Amazon Glacier in 2012 was a game changer in cloud storage, as its rate is so much lower than others and its pricing model is complicated as well. Even today, it is still one of a kind in cloud storage. The technology behind Glacier must be very unique, as Amazon still keeps the technology secret. For people who are interested in, Cooper wrote an article to speculate it (www.itproportal.com/2013/11/09/one-of-techs-most-elusive-mysteries-the-secret-of-amazon-glacier/).

The data model in Amazon Glacier is very simple with two concepts "Vault" and "Archive." A vault is a container which can store many "archives," while an "archive" is any data such as a document, audio, video or photo. An "archive" is the basic unit, which must has a unique ID within a vault and an optional description. In addition,



Source: Amazon Web Services Blog: http://aws.amazon.com/blogs/aws/what-can-i-say-another-amazon-s3-price-reduction/; http://aws.amazon.com/s3/pricing/

Figure 1.
Amazon S3 storage pricing history



Source: Amazon Web Service Blogs: http://aws.amazon.com/blogs/aws/more-bits-for-y/; http://aws.amazon.com/blogs/aws/aws-lowers-its-pricing-again-free-inbound-data-transfer-and-lower-outbound-data-transfer-for-all-ser/

Figure 2. Amazon S3 data transfer-out pricing history

a "job" and "notification" are information/actions to access and notify the users with "vault" and "archive." There are three ways to store data in Glacier via: Amazon web interface, applications, or software development kits (SDK). For regular users, the web interface is very straightforward to use. For developers, they can use SDK with two APIs: low-level and high-level API.

Retrieval of stored data in Glacier takes a two-step process and also requires longer time compared to getting it from other cloud storage like S3 or local storage. The first step is to request data to a staging area and wait for notification. To do so, a request "job" must be made, and Glacier will move the requested data into a staging area. Once the data is ready for download, a notification will be sent to the user. Then the second step is to actually download the data. Users have a 24-hour period to download the requested data. Data can be retrieved by a range or portion of an archive. This is useful if an archive is very large or users cannot completely download all the data within the 24-hour period.

5. Costs and analysis of storing and retrieving data in Amazon S3 and Glacier

Rosenthal and Vargas (2013) reviewed the pricing history of raw disk with Kryder's Law, described the pricing history of major storage services, and projected the future cloud storage pricing. Their conclusion is that the 30-year history of raw disk costs showed a drop of at least 30 percent per year, while the six-year history of cloud storage dropped at most 3 percent per year. The recent large price drops is definitely out of the projection, as the big three's pricing is currently set at USD 0.03/ GB, a 60 percent of drop from 0.085/ GB in January 2014. S3's data transfer-in is now totally free, and data transfer-out has also dropped from 0.12/ GB (2011) to 0.09/ GB (2014). Amazon Glacier is still priced the same as it was in 2012. From the above figures, we can conclude that the storage fee will keep dropping, but it is difficult to project the future fee. In addition, Rosenthal and Vargas article primarily focussed on evaluating costs of running LOCKSS, which requires running a server instance at all times. Certain preservation operations do not require a server instance to run all the times. For instances, a server instance can just verify data checksum and then shut itself down once completing the task.

The author believes that the combination of big price drops and free data transfer within the same data zone makes the cloud storage a very attractive solution for long-term digital preservation, especially using Glacier. After the huge price drop in S3, S3 or other cloud storage services, they can also be considered to use in digital preservation in libraries.

There are a few posts and/or articles discussing use of different cloud storage providers. Only a few mentioned Glacier and very few compared the costs. Font wrote an article to describe the process of using Glacier including software to upload, uploading, accessing, and deleting data. The fee structure was mentioned (Font, 2013). Readers should be aware that the above literature review mostly only mentioned storage pricing, but few emphasized the data transfer-cost, which can be a big portion of the total bill. In comparison to the traditional storage model, the cloud storage service providers charge both storage and data transfer fees. The transfer-out fee in Amazon S3 or Microsoft Azure is very easy and straightforward to calculate, while that fee in Glacier can be very tricky and shall be watched very carefully because its pricing model is somewhat confusing to understand. Finley specifically gave an extreme example of a 3 TB archive which cannot be split into smaller chunks that could lead to a retrieval fee of $22,000 (Finley, 2012).

At the libraries, a program was written to use API to create several vaults, upload and download a few archives in November 2014. To reduce the size of data, we used different popular compression algorithms such as gzip, bzip2, and 7ZIP to reduce size of digitized monographs from 1.4 GB to 275 MB. In other words, that was a 80 percent reduction in size. The data were uploaded it to both S3 and Glacier, and after two weeks retrieved all the data from Glacier as quickly as possible. See Figures 3 and 4 for a cost comparison.

The retrieval cost is called peak retrieval rate, which is $0.48 calculated as (275 MB/ 1,000 GB /4 hours × 0.01 GB/retrieval × 720 hours = 0.49) per the Glacier's pricing model. As you can see from Figure 3, the retrieval fee accounts for 98 percent of the bill, while the storage fee is only 1 percent. This is due to the peak retrieval rate set up in this pricing model. Amazon Glacier pricing model gives free 5 percent of monthly data spread out evenly across the number of hours. In other words, assuming a month has 30 days, then each hour the free retrieval data is about 0.006944 percent of the total amount of data (5 percent/30 days/24 hours = 0.006944 percent). Exceeding that will result in a peak retrieval charge. It is interesting that Amazon does not have its own cost estimator for Glacier. A cost calculator is available at http://liangzan.net/aws-glacier-calculator/.

▾ Glacier		$0.50
US East (Northern Virginia) Region	Usage	
Amazon Glacier Peak-Retrieval-Bytes-Delta		
$0.010 per GB - Retrieval Fee	48.235 GB	$0.48
Total:		$0.48
Amazon Glacier Requests-Tier1		
$0.050 per 1,000 Requests	18 Requests	$0.01
Total:		$0.01
Amazon Glacier TimedStorage-ByteHrs		
$0.010 per GB / month - Storage	0.143 GB-Mo	$0.01
Total:		$0.01
Region Total:		$0.50

Note: Amazon Glacier

Figure 3. Sample monthly cost of storing and retrieving 275 MB data with peak retrieval rate

Services | **Estimate of your Monthly Bill ($ 299.57)**

Estimate of Your Monthly Bill

☑ Show First Month's Bill (include all one-time fees, if any)

Below you will see an estimate of your monthly bill. Expand each line item to see cost breakout of each service. To save this bill and input values, click on 'Save and Share' button. To remove the service from the estimate, jump back to the service and clear the specific service's form.

Save and Share

⊟ Amazon S3 Service (US-East)		$ 295.54
Storage:	$ 295.52	
Put/List Requests:	$ 0.01	
Other Requests:	$ 0.01	
⊞ AWS Data Transfer In		$ 0.00
⊟ AWS Data Transfer Out		$ 5.88
US-East / US Standard (Virginia) Region:	$ 5.88	
⊞ AWS Support (Basic)		$ 0.00
Free Tier Discount:	$ -1.85	
Total Monthly Payment:	$ 299.57	

Note: Amazon S3 and Glacier

Figure 4. Monthly costs for storing 10 TB and transferring out 5 percent of data

In order to illustrate the costs associated with cloud storage, the author made the following assumptions:

(1) using 10 TB space for digital preservation;

(2) less than 5 percent of data transfer-out per month evenly in any given hour; and

(3) using the most current pricing and data transfer models (2014) with understanding that:
   - data in: \$0.00; and
   - data out (the same region): \$0.00.

The 10 TB data costs \$1,200/year in Glacier, compared to \$3,600/year in S3.

6. Discussions

Neither DuraCloud nor TDL emphasized the limitations on the speed and the costs of getting data out of Glacier, probably due to its software design of using both S3 and Glacier and/or Glacier's confusing pricing and retrieval models. The combination of using Amazon S3 and Glacier in DuraCloud software makes the speed and data transfer costs irrelevant. However, if users would like to use Glacier alone, it is highly recommended to fully understand the speed and costs.

6.1 Comparisons of cloud storage S3 and Glacier

There is no doubt that cloud storage offers some significant advantages over traditional local storage. Cloud storage can handle any amount of data (even PBs), and its on-demand feature allows users to use and only pay the needed amount. Multi-tenancy allows different applications or different users to access the same resources to fit their needs. A specific advantage for digital preservation is that data in cloud storage can be easily transferred and duplicated globally to minimize data loss due to natural disaster.

Readers should not directly compare the above costs (Figure 4) to these of DuraSpace's plans, as the above have not considered additional data recovery services such as synchronization and recovery provided by DuraSpace. However, this basic storage service may be suitable for certain customers who use different digital preservation strategies. For example, one may organize data into logical trunks, compress them, create checksum, and use Glacier only to take advantage of its pricing and data transfer model.

In comparison to other cloud storage solutions such as Amazon S3 and Microsoft Azure, Glacier is unique and has certain advantages:

(1) Lower storage costs: the above example shows that by working with Glacier's pricing and data retrieval model, the cost of Glacier is about one-third of that of S3 and transferring no more than 5 percent of data evenly in a month of time period. It is critical to understand that users must break up retrieval data over the longest period of time to minimize the data transfer fees. When using Glacier, also consider how to use the data transfer free feature within Amazon's same data zone. The combination of both storage cost and data out cost makes Glacier unique in cloud storage.

(2) Lower data transfer costs if operating under Glacier retrieval model: Glacier offers up to 5 percent of free data retrieval. If planned carefully, data out can be 100 percent free of charge. Amazon just released a retrieval limit policy via web interface or an API. With this feature, users can monitor and reduce the retrieval cost. In comparison, currently Amazon S3, Google, and Microsoft Azure charge fees for data transfer-out (note: only offer very small amount such as 1-5 GB free data out per month).

(3) Isolation: traditional backup typically uses offline storage such as tapes. The advantage of offline data is that it is not accessible except during the read/write operations. As a result, tapes are immune to failure of online backup models, which provide better data protection in certain cases such as intentional or unintentional operations by human beings and natural disaster. Giving that Glacier has longer retrieval times and a two-step model, it is good for long-term preservation materials that do not require immediate access.

6.2 Strategies to minimize costs

Simply put, cloud storage solutions charge fees for data transfer-out besides storage cost. To fully utilize Glacier for its low cost, it is critical to understand fully Glacier's pricing model and its two-step data retrieval model. Getting all data out as soon as possible will result in a large bill which completely defeats the purpose of using Glacier. The data transfer-out cost can be large in other cloud storage services as well. Therefore, the strategies of minimizing costs consist of two parts: data storage cost and data transfer-out cost. For S3 and Glacier, users should consider how to operate data within the same data zone and minimize data transfer-out. For Glacier, users should focus on operations how to utilize the 5 percent monthly retrieval cap of total storage evenly in hour:

(1) Reduce data size: use local computers to organize archival data into logical parts (e.g. each folder per file), then zip and compress them. In our example of compressing TIFFs, the compress ratio is around 80 percent.

(2) Minimize data transfer-out by operating and synchronizing data in the same data zone:

- Upload the compressed files to Amazon EC2 with attached storage S3, and calculate checksum.
- Synchronize the storage with Glacier.
- Turn off Amazon EC2 instance to minimize cost (or even terminate the data in S3 to minimize the cost).
- Verify data and fix them if errors are found per digital preservation policy. This can be operated like: start a new Amazon EC2 (either manually or automatically) with S3; transfer data from Glacier (note: transfer data in the same data zone is free); run checksum checking, verify data and fix them if error found; shut down the EC2 instance and S3 after the verification.

In December 2014, Amazon released data retrieval polices which allow users to integrate the free retrieval policy and max retrieval policy to limit the cost. It is possible to operate almost all of IT functions given the availability of a full set of cloud computing services: virtualization such as EC2, database such as RDS, cloud storage such as S3 and Glacier, and event-driven services such as SNS and Lambda. Fully utilizing related services and retrieval policies will significantly reduce the total costs in IT operations. The low storage cost in Glacier and the huge price drops in S3 also suggest that institutions can focus on other related costs such as IT staffing and process improvements.

7. Summary

Many articles have been published to discuss uses of cloud computing in libraries. However, cloud storage was not popular due to its cost. Two articles published before 2013 provided examples that cloud storage might not be better than local storage due to costs. Since then, cloud storage pricing dropped significantly 70 percent (from 0.10/ GB to 0.03/ GB) and data transfer-out rate also dropped from 0.12/ GB to 0.09/ GB. In addition, data transfer within the same data center is now completely free. The paper is intended to bridge the gap by re-examining use of cloud storage for digital preservation. An overview of cloud storage and the use of cloud storage in libraries are provided. The author presents S3's storage and data transfer-out pricing history to show the drops of cloud storage pricing over the years.

The paper specifically emphasizes the cost of data transfer by demonstrating data transfer-out fees can be very large. An analysis of the optimal use of Glacier demonstrates the need to understand and operate data retrieval to fully utilize the benefits of Glacier. By adopting strategies to operating data within the provider's data zone and minimize data transfer-out, cloud storage like Glacier can be very attractive in terms of costs and isolated backups. Along with all these advantages of cloud storage such as scalability, reliability, and multi-tenancy, cloud storage solutions such as S3 and Glacier are worthy of serious consideration for institutions.

Note

1. https://aws.amazon.com/blogs/aws/amazon-s3-two-trillion-objects-11-million-requests-second/

References

Branan, B. (2011), "DuraCloud: a technical overview", available at: www.slideshare.net/DuraSpace/dura-cloud-technicaloverview201111finis1 (accessed December 12, 2014).

DuraSpace (2013), "Using the DuraCloud service to archive content in Glacier presentation slides", available at: www.slideshare.net/DuraSpace/51613-using-the-duracloud-service-to-archive-content-in-glacier-presentation-slides-21330261 (accessed December 10, 2014).

DuraSpace (2014), "DuraCloud subscription plans", available at: http://duracloud.org/pricing (accessed December 15, 2014).

Finley, K. (2012), "Is there a landmine hidden in Amazon's glacier?", The Wire, available at: www.wired.com/2012/08/glacier/ (accessed December 15, 2014).

Font, V. (2013), "How to back up your data with Amazon Glacier", available at: www.technologyguide.com/howto/how-to-back-up-your-data-with-amazon-glacier/ (accessed December 15, 2014).

Han, Y. (2010), "On the clouds: a new way of computing", *Information Technology and Libraries*, Vol. 29 No. 2, pp. 87-92.

Han, Y. (2011), "Cloud computing: case studies and total cost of ownership", *Information Technology and Libraries*, Vol. 30 No. 4, pp. 198-206.

Iglesias, E. and Meesangnil, W. (2010), "Amazon S3 in digital preservation in a mid-sized academic library: a case study of CCSU ERIS digital archive system", *The Code4Lib Journal*, No. 12, available at: http://journal.code4lib.org/articles/4468 (accessed December 5, 2014).

Kroski, E. (2009), "Library cloud atlas: a guide to cloud computing and storage", *Library Journal*, available at: http://lj.libraryjournal.com/2009/09/technology/library-cloud-atlas-a-guide-to-cloud-computing-and-storage-stacking-the-tech/ (accessed December 5, 2014).

Mitchell, E. (2010), "Using cloud services for library IT infrastructure", *The Code4lib Journal*, No. 9, available at: http://journal.code4lib.org/articles/2510/ (accessed December 12, 2014).

Mitchell, E.T. (2012), "Data and the shift in systems, services, and literacy", *Journal of Web Librarianship*, Vol. 6 No. 4, pp. 325-328.

OCLC and CRL (2007), "Trustworthy repositories audit & certification: criteria and checklist, available at: www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed December 5, 2014).

Rosenthal, D.S.H. and Vargas, D.L. (2012), "LOCKSS boxes in the cloud". available at: www.lockss.org/locksswp/wp-content/uploads/2012/09/LC-final-2012.pdf (accessed December 5, 2014).

Rosenthal, D.S.H. and Vargas, D.L. (2013), "Distributed digital preservation in the cloud", *International Journal of Digital Curation*, Vol. 8 No. 1, pp. 107-119.

Schumacher, J., Thomas, L.M., Drew, V., Stacey, E., Jeff, H., Aaisha, H., Meg, M., Patrice-Andre, P. and Danielle, S. (2014), "From theory to action: good enough digital preservation for under-resourced cultural heritage institutions", available at: http://commons.lib.niu.edu/handle/10843/13610 (accessed December 15, 2014).

Texas Digital Library (2014), "DuraCloud storage and ingestion options", available at: https://tdl.org/duracloud/duracloud-ingestion-retrieval/ (accessed December 15, 2014).

Tonjes, C. (2010), "Cloud computing at DCPL", LITA Cloud Computing Session in American Library Association 2010 Annual Conference, June 25, available at: www.slideshare.net/ctonjes/chris-tonjes-cloud-computing (accessed December 5, 2014).

Further reading

Constantopoulos, P., Doerr, M. and Petraki, M. (2005), "Reliability modelling for long term digital preservation", 9th DELOS Network of Excellence thematic workshop "Digital Repositories: Interoperability and Common Services", Foundation for Research and Technology-Hellas (FORTH), Heraklion, Crete, Greece, available at: http://delos-wp5.ukoln.ac.uk/forums/dig-repworkshop/constantopoulos-1.pdf (accessed December 5, 2014).

Han, Y. and Chan, C.P. (2008), "The modeling system reliability for digital preservation: model modification and four-copy model study", *iPRES 2008*, pp. 281-286, available at: www.bl.uk/ipres2008/ipres2008-proceedings.pdf (accessed December 5, 2014).

Corresponding author
Dr Yan Han can be contacted at: yhan@email.arizona.edu

This article has been cited by:

1. Aleksei A. Sorokin, Sergey P. Korolev, Andrey N. Polyakov. 2015. Development of Information Technologies for Storage of Data of Instrumental Observation Networks of the Far Eastern Branch of the Russian Academy of Sciences. *Procedia Computer Science* **66**, 584-591. [CrossRef]