Storage in the Cloud: What You Need to Know and Why

David A. Pease

IBM Distinguished Engineer
Manager, Exploratory Storage Systems
IBM Almaden Research Center
San Jose, California, 95120, USA

ABSTRACT

In recent years there has been huge growth in the amount of data stored on public cloud services. Much of the data stored on these services is placed there through the use of mobile apps (such as Facebook or DropBox), with little consideration given by the owners to the potential consequences of storing their data on a public service. While these services typically spell out the conditions under which user data is stored in their Terms of Service, most users don't take the time to read these and they are seldom written in a way that encourages real understanding of the terms. However, there are many issues that should be considered when choosing to store personal data in a public cloud service, both from the perspective of the user and the cloud service provider. This paper begins by explaining some key concepts of cloud storage. It then provides a list of items that should be considered when choosing a cloud service provider, explains why the issues are potentially important, and gives suggestions for actions the user might make in order to achieve their desired goals.

Keywords: Cloud, Storage

INTRODUCTION

The term "cloud" is the technical buzzword of the decade; it is used to describe a multitude of related concepts and implementations. As a result, the term itself is almost meaningless except as a high-level concept. The term "cloud storage", though still open to various interpretations, is somewhat more well-defined. In general, cloud storage refers to storing data in a remote system to which the user or application is attached by a network. While cloud computing may suggest many varied attributes, cloud storage is usually understood to include two main concepts: network access and elastic storage capacity. The term cloud storage may also suggest other, operational attributes such as object storage models and RESTful interfaces, but these are orthogonal to our discussion.

The idea of remote data storage is hardly a new one; protocols and services for storing and retrieving data remotely have existed for decades. As a technology remote storage has been available in one form or another since before the term “cloud” was coined, and before the development of the world wide web and the widespread availability of consumer internet (Sandberg *et al.*, 1985). However, several fairly recent developments have led to a profusion of low cost, easily accessible remote data storage services:

- the rapid growth in disk storage capacities combined with the plummeting cost of such storage

- the nearly ubiquitous availability of medium- to high-speed network connectivity in many parts of the world
- the explosion in the use of mobile devices

These factors have combined to fuel the growth of varied models of data storage services, from social sharing networks such as Facebook (Johnson, 2008), to file sharing sites such as Dropbox (Drago *et al.*, 2012), to data backup services such as Carbonite (Carbonite, 2014). We refer to these services generically as cloud storage services. Users may put files such as photos on social sharing sites without even thinking of them as cloud storage services; however, the considerations for storing their data "in the cloud" are no less applicable.

This paper presents some of the issues and considerations that users should be aware of when using any cloud storage service. It focuses on individuals' use of public cloud services, though many of the issues presented are equally applicable to business or other organizations, and to other remote storage models.

CONSIDERATIONS FOR USING CLOUD STORAGE SERVICES

The goal of this paper is to make users aware of some of the questions they should consider when choosing to store data in the cloud. There may be situations where a minimal set of storage service guarantees is sufficient for a user's needs, but many times it is not. Not all services are the same, and users should be aware of what they are agreeing to and what they are receiving when they use a cloud service to store their data.

Different cloud storage providers (CSPs) have different end-user cost models, but in the end the level of service a CSP can provide will depend on the cost of providing the service and the income they receive in return.. A CSP must have some way of recouping the not insignificant costs associated with providing their service (not to mention making a profit) in order to provide dependable storage services. CSP costs include capital expenditures such as processors, storage, networking equipment, and power and cooling distribution equipment; they also include operational costs such as data center premises, power, and human maintenance and administration. The quality of service (QoS) guaranteed in the CSP's terms of service will directly affect the costs associated with the service; for example, keeping more than a single copy of data increases the cost of storing that data. Users "pay for" cloud storage in various ways, whether it is by overtly paying a fee, exposing themselves to (often targeted) advertising, providing an unintended source of data that can be "monetized" by the provider, or some combination of the above. Additionally, in general users will get what they pay for. That is to say, users should not expect to receive first-class service guarantees for ostensibly free storage services.

What follows are some of the issues to consider when choosing to store data in a public cloud.

Confidentiality

Confidentiality refers to the idea that the contents of a user's data files are not visible to others. A user storing income tax returns, for instance, probably has an expectation that his or her data will not be visible to others, either other users of the service or employees of the service. Most users simply assume that the CSP takes appropriate steps to protect the confidentiality of their data. However, even if that is true, confidentiality cannot always be guaranteed in the case of a rogue employee or of a hacker's break-in to the service. There have also been egregious cases of a CSP creating confidentiality exposures, such as when Dropbox turned off user authentication for two days due to authentication server technical problems.

One solution to the problem of confidentiality is not to store data in the cloud in a form that is accessible to others; for example, for the user to encrypt the data locally then transmit only the encrypted data to the CSP. (This provides protection for the data both in transit and at rest.) While a good practice, at least for sensitive data, this leads to other challenges such as encryption key management. While software such as 1Passwd exists to assist with this task, the databases associated with such software are also a prime target for hackers.

Confidentiality may be the most difficult problem for the CSP to provide a technical solution for. A CSP may, for instance, attempt to mitigate the potential for data confidentiality breaches by storing the data in an encrypted form. However, the system must still be able to present the data back to the user in the form in which it was originally stored, so the knowledge of how to decrypt the data is part of the system, and thus the potential for others to access the data still exists.

Privacy

Privacy is related to Confidentiality, but we use it to mean the privacy of the users of the service. This includes confidentiality not only of file data, but also metadata (that is, data about data). Obvious user metadata could include file and directory names associated with an account; iCloud apps, for instance, have visibility to all of a user's iCloud file and directory names, and could use that information to search for specific types of files or even well-known file names. Other metadata could include Google Maps navigation history or photo location metadata; this type of data becomes of concern when we realize that the NSA requires only 6 location data points to uniquely identify an individual.

Giving files and directories random or meaningless names is not likely to be a good alternative for most users, and actions such as stripping metadata from files such as photos is not always desirable. Beyond these types of actions, there is little a user can do to protect meta-information beyond being aware of the terms of use of their potential CSPs and choosing one whose terms meet the user's expectations or requirements.

Reliability

Other terms for the concept of Reliability could include Durability or Resiliency. This refers to the likelihood of being able to retrieve data at some point in the future. To some users' surprise, free cloud storage services have no liability for data loss (though this may only become apparent after a data loss event).

Durability of data in the face of system failures is a problem that can be effectively addressed by storage technology. Different approaches may be chosen depending on the storage hardware infrastructure being used and the level of reliability required. Some modern cloud systems use large numbers of low-cost disk drives, and can achieve data reliability through multiple copies of files. Others may use more sophisticated hardware and/or software, and techniques like RAID for reliability. Remote replication of data (copies of data in multiple, isolated data centers) can provide even better reliability.

However, each of these solutions requires an investment in hardware, software, and potentially administration on the part of the CSP. Multiple copies of data doubles or triples the raw storage cost of data; RAID solutions typically require more expensive hardware or software, and also increase storage utilization (though by a smaller factor than multiple full copies); remote replication not only multiplies the raw storage cost, but also requires additional data centers and a powerful networking infrastructure.

It is unlikely that a user can find a free service that will provide one of these types of reliability technologies, at least for more than a very small amount of data; the cost is simply too high. So a user must decide what their reliability requirements are and how best to meet them. One choice is to pay for a service that provides reliability guarantees. An alternative choice is to manage the reliability themselves. In some cases, keeping a local copy as well as a copy in the cloud may be sufficient; another choice could be to store copies of files in multiple CSP environments, counting on the unlikelihood that they would all lose the data. With either of these approaches, a user may want to access data periodically just to insure that their data hasn't become inaccessible for one (or more) of the copies.

Integrity

Integrity refers to the integrity of the data when it is retrieved from the cloud storage service. Spontaneous undetected data corruption is probably not a major concern in modern storage systems (though it does occasionally happen). The bigger concern is whether the CSP processes the user's data so that it is no longer what he or she originally stored. For example, both Flickr and YouTube may compress photos or videos in order to save storage space, and no photo storage system today stores “raw” photo files in their uncompressed form.

If, for example, photos and movies are uploaded to a cloud service strictly for sharing, and a lower resolution or higher level of compression is acceptable, then this is of no concern. However, if the only copy of a valued photo could be compressed or reduced unexpectedly, it may be more a more important consideration.

Availability

Availability refers to the ability to retrieve the user's data when desired. Considerations here include both agreed-to QoS constraints, such as with Amazon Glacier which has an advertised potential 4-hour retrieval lag for any data requested, as well as unforeseen events. Unforeseen events can be accidental like power and network outages, or through deliberate intervention, such as government censorship or curtailment of access to Internet services (as when Egypt experienced a shutdown of its cellular phone networks in 2011, or the more recent moves by the Turkish government to ban twitter and other social media services).

QoS constraints are presumably known beforehand, and should come as no surprise to the user. Yet it is not uncommon for a typical consumer to be unaware of the implications of such constraints for the usability of the service. Therefore, the implications of such constraints should be carefully considered before entering into an agreement.

Obviously, no one can predict the unforeseen event. However, when the only copies of data are stored in one or more cloud services, the user must always be prepared to accept the possibility of such events and the unavoidable delays they imply. If immediate access to data is of real importance, a copy that can be used without access to the CSP should be maintained. For some users, this consideration needs to be weighed against the benefit they are obtaining by treating the cloud service as a convenient and consistent data repository. This is because keeping a local copy of data stored in a cloud service requires effort to identify and maintain the latest version. The effort rapidly grows when “local” copies of the data are actually being used on multiple devices in the user’s possession. This is increasingly common with the wider range of devices we use today, both to communicate and for digital media consumption.

Lifetime

Lifetime refers to how long data is expected to exist, and how it is eventually disposed of. Most users probably expect to eventually delete data when it is no longer needed, though when storage services are free of cost this may not always happen. When storage services are not free users must consider how they will manage their data if and when they decide to discontinue paying for the service, or when they decide to move to another service.

A real concern regarding lifetime is what happens when data outlives its owner. Some CSPs have policies or procedures for handling data belonging to deceased customers, while others do not. Google, for instance, gives the user the ability to set up a “dead man switch” that can define actions to be taken when the user no longer accesses the system. If users are concerned about what will happen to certain kinds of data (e.g., family photos, legal documents) after their death they should make certain that someone they trust has the credentials to access the storage as well as directions for handling the data after they are deceased. They should also be aware of the provisions, if any, regarding such an occurrence in their service agreement.

Cost, Capacity, and Provider Viability

Cost and capacity are interrelated issues. As noted earlier, users typically get what they pay for. A CSP may provide a comparatively small amount of storage for little or no cost, but may have higher rates for large amounts of data. Users should be aware of what their data storage needs are (or will be), what the CPS's storage policies and constraints are, and what the consequences of exceeding those constraints could be.

Provider or CSP viability should at least be considered when storing data on a cloud service. Providers have been known to fail or be taken offline and thereby lose massive amounts of user data, as happened with the Megaupload service (Cohn & Samuels, 2012). In a more unusual scenario, Major League Baseball (MLB) sold digital rights management (DRM) protected copies of sport videos that could only be played using software that required contact with a remote server. When the service was discontinued, the remote servers were eventually shut down; this resulted in many upset fans who felt that they had purchased videos that had unexpectedly become unplayable due to some external change outside their control for media they felt they owned. In this case, the users of this service had not considered (or possibly never realized) their dependence on a CSP.

When choosing a CSP, users should weigh the possibility of failure, service disruption, or provider continuity and viability, against the cost of the service and the importance of the data they are storing.

CONCLUSIONS

There are many issues and considerations when storing data in a public cloud. The best advice is to know what issues are important and to understand the Terms of Service that a potential CSP offers. In other words, read the service agreement, or at least ask the important questions ahead of time before choosing to store data in a public cloud.

This paper has presented some of the issues to consider when deciding to store data in a public cloud, and has attempted to give some background and guidelines to help with the decision.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge the assistance of Ahmed Amer, who not only inspired the topic of this paper but also provided input and editorial assistance.

REFERENCES

Carbonite (2014). “Carbonite Web Site” http://www.carbonite.com/ last viewed on April 10th, 2014.

Cohn, C., & Samuels, J. (2012). Megaupload and the Government's Attack on Cloud Computing. Electronic Frontier Foundation.

Drago, I., Mellia, M., Maurizio, M., Munafo, M., Sperotto, A., Sadre, R., and Pras, A. (2012) “Inside dropbox: understanding personal cloud storage services”, Proceedings of the 2012 ACM conference on Internet measurement (IMC ’12). p.481-494. NY, USA: ACM.

Johnson, A.N. (2008) “Looking at, looking up or keeping up with people?: motives and use of Facebook”, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. p.1027-1036. NY, USA: ACM.

Sandberg, R. Goldberg, D. Kleiman, S. Walsh, D. and Lyon, B. (1985). “Design and Implementation of the Sun Network Filesystem”, USENIX Technical Conference.