On the Feasibility of Data Loss Insurance for Personal Cloud Storage

Xiaosong Ma
Qatar Computing Research Institute, xma@qf.org.qa

Abstract

Personal data are important assets that people nowadays entrust with cloud storage services for the convenience of easy, ubiquitous access. To attract/retain customers, cloud storage companies aggressively replicate and geo-replicate data. Such replication may be over-cautious for the majority of data objects and contributes to the relatively high price of cloud storage. Yet cloud storage companies are reluctant to provide costumers with any guarantee against permanent data loss.

In this paper, we discuss the viability for cloud storage service to provide optional data insurance. We examine major risks associated with cloud storage data loss and derive a crude model for premium calculation. The estimated premium level (per unit declared value) in most scenarios is found significantly smaller than that accepted in mature businesses like shipping. Therefore, optional insurance can potentially provide cloud storage services with more flexibility and cost-effectiveness in resource management, and customers with both peace of mind and lowered cost.

1 Introduction

People are generating, as a new asset class [11], more and more personal data, such as photos, videos, legal and financial documents, and digital receipts. These data are increasingly generated and accessed from non-traditional computing platforms, e.g., mobile devices. A recent Cisco report [4] states that global mobile data traffic grew 81% in 2013 to 1.5 ExaBytes/month. Also the number of mobile devices is projected to surpass the world's population by 2014 and monthly mobile tablet traffic alone will exceed 2.5 ExaBytes by 2018.

With such trend, plus the growing adoption of public cloud platforms like Amazon EC2 [2], it becomes common practice for people to store personal data in one or more cloud-based facilities. Popular storage services iCloud and Dropbox reported having over 300 and 200 million users by late 2013, respectively.

These wide-spread services provide users with multi-fold advantages compared with traditional personal data storage (on PCs, laptops, and/or household external storage devices such as USB hard disks). First, users can easily and transparently access/share their data across multiple devices. Second, remote cloud storage provides quite reliable file backup to recover data unavailable on local platforms. Third, cloud storage is elastic: users get charged for the actual usage while the capacity seamlessly grows with demand. Fourth, cloud storage is (almost) maintenance free and works as a consolidation repository for the always upgrading collection of active devices in a family. Finally, cloud storage comes with many value-added features such as integration with mobile applications, data sharing via social network, and direct connection with online media/software stores.

However, it is hard for current cloud users to trust service providers with *all* their data or to use one cloud service as a sole household data repository. A 2012 Gartner report predicted that consumers would store more than a third of their digital content in the cloud by 2016 [5]. Very commonly, people adopt cloud storage, but supplement it with traditional storage, self-administered backup, and/or third-party backup services. Sometimes people use more than one cloud storage services simultaneously. When doing so, users forfeit part of the benefits brought by cloud storage, in particular ease of maintenance and integrated storage. For example, if a user suddenly needs to find a set of Power-Point slides last modified fifteen years ago and could not remember where the file was saved, he/she may need to search through cloud storage, active and retired laptops, a dozen of old USB thumb drives, and backup CDs/DVDs. Users are responsible for scheduling backups and checking the functioning of storage devices. It might be challenging to find devices that read old media types. Accessing replicated files (intentionally or not), without using a central repository, easily generates content divergence and may require tiresome manual reconciliation.

The reason for consumers' partial adoption is likely associated with both cost and control [9]. While cloud storage providers commonly provide a few GBs of free space, upgrades typically cost significantly more than commodity hard disks of similar capacity. For example, Dropbox charges $99.99 per year for 100GB space, while users can buy a 2TB external drive at the same price. The cost is non-trivial for a modern family to store its many TBs of data in cloud, with the majority of which being cold media files that do not demand the distributed access convenience of cloud storage anyway. On the other hand, people hesitate to put the only digital copy of their great grandmother's early pictures in the cloud (without local or secondary remote backup), for

fear of data loss or corruption.

In this paper, we explore a solution that potentially eases both concerns. We observe that cloud storage providers strive to aggressively protect clients' data with multiple forms of redundancy, yet reluctant to provide concrete guarantee against the rare but apparently possible data loss. Given the abundance of cheap storage options, data centers with tiering capabilities, and mature data protection/restoring mechanisms, cloud providers should be able to explicitly handle the risk of data loss by allowing users to purchase insurance for valuable data contents. We reason that not only the premium is likely to be negligible compared to what customers are used to with similar services (such as shipping or security boxes), such insurance may allow providers to relax redundancy requirement for non-crucial data. As a result, resource utilization can be enhanced and customers can benefit from lower cloud storage cost.

2 Current State of Data Storage Guarantee (or the Lack of)

We briefly surveyed the terms of service given at the official websites of popular cloud storage providers, for clauses regarding data durability or loss. Note that in this paper, we focus on *permanent data loss/corruption*, rather than temporal data unavailability. We consider transient service interrupt, while costly for cloud-based businesses, remains a minor concern for personal data storage, as active data tend to have local cached copies.

Service	Cost/yr	Durability claims
Dropbox	\$0.99/GB	Uses S3 as underlying service; "as-is" with max loss compensation: greater of \$20 or past 3 months' service fee[1]
Box Standard S3 RRS	\$0.9/GB \$0.72/GB	11 nines[2] 4 nines[2]
Box	\$0.6/GB	"As-is"[3]
iCloud	\$2/GB	"As-is"[4]
Google Drive	\$0.6/GB	"As-is"[5]
Baidu Cloud	\$0.6/GB	"As-is", max loss compensation: current storage period's service fee[6]

Table 1: Sample cloud storage services' pricing information and durability claims

Table 2 summarizes these providers' related policies, which turn out to be quite similar. We also list their annual service charge rates, in all cases significantly higher than consumer-grade external storage price per GB. Most companies claim that they provide "best-effort" or "as-is" services, including making a reasonable effort to achieve data durability and avoid data loss. At the same time, the terms of service typically include clauses explicitly stating that the provider does not guarantee/promise its service to be free from loss, corruption, or security intrusion. "To the fullest extent permitted by law", the cloud storage service providers have no liability for the result of such inadvertent events. Among those we surveyed, only Dropbox and Baidu Cloud Storage mention about moderate monetary compensation, on the order of several months of service charges.

Regarding durability claims, only Amazon S3 gives concrete estimated data loss rate. Its standard service provides "11 nines" of durability, i.e., expected 0.000000001% annual object loss rate. It also provides an alternative storage option, *RRS (Reduced Redundancy Storage)*, which is cheaper than the standard S3 but only provides "4 nines" of durability.

In addition to general-purpose storage providers, there are several other commercial offerings targeting data durability. For example, Data Insurance (DI) [1] is an IP licensing company that "licenses the use of its patents, standards and procedures to insurance companies and brokers." DI appears to target businesses as customers, who purchase data insurance policies and are subsequently required to use a DI-approved and DI-audited data management company. This is unlikely to work for industry-leading, well-established cloud storage providers. A 2006 article discussed solutions providing "digital safety boxes" [8]. However, the durability of such providers themselves are in question: the sample service mentioned in the article (xdrive.com) does not seem to exist anymore.

3 Risks in Cloud Storage of Personal Data

Next we characterize the risks that might result in permanent data loss in cloud storage. Also, it is helpful to examine precautions or remedies to these risks.

An incomplete list of hazards that produce data loss risks may include the following:

1. Storage hardware failures that cause data loss/corruption, most commonly (but not limited to) hard disk failures
2. Security attacks resulting in data removal
3. Incorrect handling of data caused by human operation errors or software bugs
4. Environmental accidents such as building corruption and fire
5. Natural disasters such as earthquake, flood, storm, and wild fire

[1] https://www.dropbox.com/terms
[2] http://aws.amazon.com/s3/details
[3] http://box.com/static/html/terms.html
[4] http://www.apple.com/legal/internet-services/icloud/en/terms.html
[5] http://www.google.com/policies/terms/
[6] http://developer.baidu.com/wiki/index.php?title=docs/cplat/bcs/terms

6. Fraudulent claims from customers (data owners)
7. Termination of business due to loss or irrational behavior of management

Among these hazards, #1-#5 account for *physical hazards* [10] that may lead to actual data loss. It is remarkable how existing data center and storage design has been preparing for such hazards. The central mechanism is adding data redundancy, in many forms and at many levels/locations. Data are protected with schemes ranging from RAID to geo-replication. This highlights a unique advantage in data risk management: *any correct copy of the original data is as good as the original, while the cost of making such a copy is independent of (and may be significantly lower than) its value*. In contrast, many traditional insurance coverage objects, such as human health (properly functioning body components) and properties (valuable personal items, unique antiques, collectible art) are impossible or illegal to replicate. For those indeed replicable, the cost of such replication typically represents the actual value of the insured item. Besides redundancy, precautionary measures such as disk scrubbing [7] and versioning (e.g., 30 days for free with Dropbox and optional service with S3) are widely adopted to reduce the data loss risks caused by hardware or human operation errors.

#6 is a *moral hazard* [10], similar to those existing in mature insurance business. Fortunately for data such hazard might be much easier to prevent, as to be discussed in the next section. We discuss #7 in Section 5.

4 Potential Data Loss Insurance Solutions

Next, we examine applying insurance, a mature risk management mechanism used for hundreds of years, to the problem of personal data storage in the cloud.

Optional Insurance Coverage for Personal Data We envision a practice of providing optional insurance coverage when users save their personal data using a cloud storage service, just like when users ship items using mail/courier services. Based on the user declared value of the data (e.g., in $/MB), a certain insurance premium is charged per month or per year. If the insured data object is considered lost and not recoverable from the cloud storage provider, the data owner (policyholder in this case) can file a claim. When such loss is confirmed, the cloud service provider (also insurance carrier in this case) will pay indemnity at the insured amount. This way, users explicitly receive risk management against data losses that they often have no control over.

It is hard to imagine that users have to specify the declared value for each individual data objects. Rather, it might be more feasible for them to do so at the directory level, such as having a "precious family pictures" directory, whose content share a declared value of $5000/MB, a "good pictures" directory of $500/MB, and an uninsured "new pictures" directory. A subdirectory may have its own declared value level that overrides that inherited from the parent directory. The total premium charged will depend on the amortized storage volume during the insured period. Note that the value structure of storage content may not align with typical, intuitive object organization. However, given the low premium estimated below, we suspect that fine-granule value specification may not be necessary.

Proof of Loss Unlike in the case of traditional insured property, in data storage *it is relatively straightforward to verify the authenticity of stored data*. Mature technologies such as hashing are widely used in data storage for purposes like endurance, authentication, and deduplication. For each insured data object, the storage provider can calculate a checksum (or an array of per-block checksums), which will be included in a receipt (proof of policy purchase). A data loss is recognized when the provider cannot reproduce a copy of the insured item that carries the correct checksum(s).

Popular hash functions such as MD5 and SHA-2 are not free of collision. However, it is considered computationally impractical to perform a *preimage attack*, where the storage provider forges a data object that carries the given checksum shown in a receipt. By doubling the small overhead of checksum calculation and storage, a service provider can further enhance its protection against insurance fraud by including two sets of checksums calculated with different hash functions.

As a side remark, there is one type of "hazard" not included in the list in Section 3: when a true data loss is detected at the cloud side, a user may still have a local or remote copy of the insured data, but has every financial incentive to keep this fact from the storage provider. He/she can therefore receive full compensation while still possessing the insured data. This may arguably be categorized as a *morale hazard* [10], which traditionally refers to the increased risk caused by the indifference of policyholder due to the existence of insurance coverage.

Premium Estimate Another major difference in data insurance is that unlike health or property insurance, here the insurance carrier has almost total control over the insured item. The total premium collected not only becomes pooled funds from insured entities (exposures) to protect against low-probability risks, it can be directly used to significantly *reduce* those risks. In particular, the incremental cost of higher data redundancy would be much smaller than needed for an express shipping courier to upgrade to safer means of transportation or to reduce human errors in item handling.

Let us consider the effect of making r simple optional replicas of an insured data object at different data center locations. For simplicity, we target the hard disk failure caused loss here, and make a conservative estimate as-

suming more space-savvy techniques (such as RAID or additional erasure coding) are not used. As actual data center disk failure rate under cloud storage workloads is proprietary information or even trade secret, we use a rough upper bound of failure rates reported in a study based on over 100,000 disk drives in Google production data centers [6]. The study presents annual failure rates (AFRs) of disks by age groups, utilization levels, and average drive temperatures. We use an AFR of 10%, significantly higher than the average value reported. In fact, this rate is higher than that of all but one observed categories (3-month old, high-utilization disks, whose AFR is slightly over 10%). At this AFR, each additional optional replica reduces the overall annual object loss rate by an order of magnitude ($\times 0.1$).

Typically, the total premium P charged by an insurance company to a group of similar policyholders is decided by the formula $P = L + U + E$ [3], where L is the incurred loss, U is the underwriting expense (the cost of risk assessment and policy setting), and E is the insurance profit. In this analysis, we consider U negligible and temporarily ignore E. Our goal then is to find the premium level matching the expected data loss risk.

Under this model, part of the premium can be invested into increasing redundancy using additional replicas. The rest should be pooled to indemnify loss incurred at the enhanced durability level. Given a data object to be insured at declared value of v(\$/MB) and disk price at c_{disk}(\$/MB), the baseline object loss rate of f_{base}, additional replication degree of r for insured items, the premium level p\$/MB matching the expected risk of loss would be

$$p = 0.1^r f_{base} v + c_{disk} r$$

Note that as r increases, the data loss risk (in terms of annual object loss rate) decreases exponentially, while the replication cost grows linearly.

f_{base}	r	$v = 100$	$v = 10000$	$v = 10000$
0.1	1	0.010	0.010	0.010
0.1	2	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}
0.1	3	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
0.01	1	1.0×10^{-3}	1.0×10^{-3}	1.0×10^{-3}
0.01	2	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
0.01	3	1.1×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
0.001	1	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}
0.001	2	1.1×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
0.001	3	2.4×10^{-6}	1.1×10^{-6}	1.0×10^{-6}
0.0001	1	1.0×10^{-5}	1.0×10^{-5}	1.0×10^{-5}
0.0001	2	2.0×10^{-6}	1.1×10^{-6}	1.0×10^{-6}
0.0001	3	1.5×10^{-6}	2.4×10^{-7}	1.1×10^{-7}

Table 2: Premium/value ratios at different baseline durability and additional replication levels

Like in the shipping industry, a convenient way to judge the expense of insurance purchase is to look at the premium/value (p/v) ratio, specifying how much premium one needs to pay to insure unit declared value. Table 2 lists such ratios calculated using the above equation, with a \$100 per 2TB disk price. Compared to recent consumer price listed on a storage hardware pricing history website[1], the price level used is considerably inflated. Together with the over-estimated AFR, such conservative calculation may partially or entirely offset costs not included here, such as networking hardware, data center hosting and energy consumption, plus human resources involved in managing additional replicas.

The baseline annual loss rate of 0.1 corresponds to the durability achieved with storing one single copy of the data, the bare minimum of storage and significantly lower configuration compared to current common practice. Even at this level, a single optional replica for insured items produces a p/v ratio of 0.01 across all value levels. This is equivalent to the amount people accept to pay for shipping: the current minimum p/v ratio is 0.0125 for USPS[2], and 0.009 for both UPS[3] and Fedex[4].

By either enhancing the baseline durability or increasing the optional replication degree, each order of magnitude reduction in object loss rate brings about a similar reduction in the p/v ratio, until when the baseline durability and optional replication degree are both high. In most cases, the p/v ratio is much lower than with the shipping industry. For example, with a 0.01 f_{base} ("2 nines") and two optional replicas, a user would pay 30 cents a year to insure a 3MB video clip at \$1,000 per MB. With three optional replicas, the same premium level covers the same object at \$10,000/MB.

Note that an f_{base} of 0.0001 corresponds to the "4 nines" durability promised by Amazon RRS. This highlights an interesting consequence of optional data insurance: *by explicitly labeling "important data", we also implicitly label "unimportant data"*. Based on people's perceived data value and budget, lowering the default durability level while providing optional insurance may allow cloud storage to be much more affordable and versatile. In addition, providers may be able to profit from insurance, by putting E back to the premium equation.

5 Additional Issues

Discontinued Storage Service We have been discussing storage providers as insurance carrier themselves. One obvious reason is that they have the tech-

[1] http://http://www.jcmit.com/diskprice.htm
[2] http://www.endicia.com/price-change-2014
[3] http://www.ups.com/media/en/value-added_pricing_daily.pdf
[4] http://www.fedex.com/us/2014rates/surcharges-and-fees.html

nical capability to accurately assess data loss risks and distribute pooled premium across indemnity, profit, and additional data protection. Meanwhile, unlike in the case of shipping (where the insurance terminates after delivery), data storage customers have to trust the service to be operating properly for an extended period of time. There is always a possibility that a well-established company goes out of business (risk #7 in Section 3).

One possible solution here is for the cloud storage providers to (partially) transfer such risks to third-party insurers using reinsurance [10]. The author is not aware of legislation related to clients' data in case a cloud storage service files bankruptcy. Intuitively, these are assets that should be transferred back to their owners.

Access Pattern Aware Optimizations We are not aware of existing studies on the relationship between perceived data value and use pattern, but intuitively "precious" objects are often read-only: people seldom put items in daily use in a bank security box. Additional "read-only" annotation on insured data could help in further lowering risk and costs (e.g., by choosing low-performance, high-durability media types).

Going Beyond Personal Data If indeed implemented, optional cloud storage insurance can be expanded to business data as well. The risks and liabilities in storing/serving business data are likely much more complicated. E.g., transient data unavailability might be of grave consequences to certain cloud-based businesses. However, we suspect many techniques used in business risk management can be applied or adapted here.

Diverse Ways of Risk Sharing Due to the unique nature of digital data, there might be novel ways for customers to participate. For example, is it possible if customers contribute storage space, rather than monetary premium, to store others' encrypted data? If deduplication and checksums are already deployed for storage efficiency and data authentication/protection, can customers be contacted if it is found that they possess identical data objects that another customer has just lost? From another perspective, covering items with high declared value may introduce new risks to the provider, such as security attacks for insurance frauds.

6 Conclusion

In this paper, we assessed the possibility and implications of having cloud storage services provide optional insurance against permanent data loss risks. Our major observation is that the existing aggressive data replication adopted today may be significantly overkilling for most content, while (psychologically) insufficient for valuable data. Instead, providers may consider lowering the default durability level (along with the baseline storage service charge), but offering a collection of optional data insurance policies. More feasibility study is needed across multiple disciplines: computer systems, customer behavior study, actuarial science, and law.

7 Acknowledgment

The author thanks the reviewers for their constructive comments. She also thanks Lorenzo Alvisi of UT Austin for very helpful discussion and detailed feedback on the paper draft. This work was supported in part by the NSF grant CNS-1318564 and a NetApp Faculty Fellowship, both through North Carolina State University.

References

[1] http://datainsurance.org/.

[2] Amazon ec2. http://aws.amazon.com/ec2, 2014.

[3] R. Brown and L. Gottlieb. *Introduction to Ratemaking and Loss Reserving for Property and Casualty Insurance*. Actex Publications, 2001.

[4] Cisco. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2013-2018. `http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html`, 2014.

[5] Gartner. Forecast: Consumer Digital Storage Needs, 2010-2016. `http://www.gartner.com/newsroom/id/2060215`, 2012.

[6] E. Pinheiro, W.-D. Weber, and L. A. Barroso. Failure Trends in a Large Disk Drive Population. In *Proceedings of the USENIX Conference on File and Storage Technologies (FAST)*, 2007.

[7] T. J. Schwarz, Q. Xin, E. L. Miller, D. D. Long, A. Hospodor, and S. Ng. Disk Scrubbing in Large Archival Storage Systems. In *Proceedings of the 12th IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS)*, 2004.

[8] J. Sturgeon. Beware of holes in bank-box safety net. `http://www.bankrate.com/brm/news/insur/20030117a2.asp`.

[9] TwinStrata. Cloud Storage Adoption Snapshot 2013. `http://www.twinstrata.com/survey-adoption-trends-cloud-storage`, 2013.

[10] E. Vaughan. *Fundamentals of Risk and Insurance, 10th Edition*. Wiley, 2011.

[11] World Economic Forum. Personal Data: The Emergence of a New Asset Class. `http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf`, 2011.