Is It Safe to Preserve Data in The Cloud

by W. Curtis Preston, Senior Analyst

We are living in the information age, and modern companies rely on their data more than any other time in history. Many companies are leveraging the cloud for preserving this vast ocean of data, but some companies are concerned about real or perceived risks of data integrity. Corruption can happen during the transit of data to cloud, to data at rest once it reaches the cloud, as well as to data at the staging area prior to the move to the cloud. Data is an asset that is at risk to corruption, deletion and even media degradation. Organizations must takes steps to mitigate these risks to maintain data integrity.

Damage Due to Human Error or Malice

The most common cause of data loss or corruption is human error and it can take many forms. The most common type of human error is where someone accidentally deletes the data. Perhaps they "fat fingered" something and accidentally deleted a file or over wrote a file with a version they did not intend to save. Even worse, a system administrator can accidentally delete an entire user, file folder, or storage bucket. The damage caused by such an action goes way beyond inconveniencing an individual user.

Consider, for example, a well-meaning system administrator who was told to manage the cloud storage data and accounts. He logs into cloudvendor.com and sees various storage entities, but in unfamiliar formats called blobs or containers etc. It is difficult to see if the storage accounts are tied to on-premises or off-premises assets, and its difficult to ascertain who owns them. The syntax and vernacular is different for each cloud. While checking on the data in one of the storage accounts, he accidentally deletes a blob, or worse the entire storage account or subscription.

Another risk to your data that falls under human error is software bugs. Software is tested against all manner of conditions, but there are always edge conditions that are not conceived or tested for. A perfect example of this is the Amazon Web Services (AWS) outage of September 2015. Too many customers were using a new service, which caused that service to be temporarily taken offline. Unfortunately, other safeguards kicked in that created even more I/O traffic, resulting in a perfect storm that caused a cascading failure of AWS for a period of time. It did not ultimately result in data loss, but it is an example of an edge condition that was not tested for.

The next two potential risks to your data have to do with what the industry calls black hats, which are people seeking to either steal or damage your company's data. (The term comes from old Westerns where bad guys always wore black hats.) The latest example of black hats seeking to damage company data is ransomware, which encrypts your data and is followed by a ransom demand to decrypt it.

Ransom-based attacks can also be like what happened to cloudspaces.com, where a black hat gained control over their entire AWS account. Subsequently they were told to pay a large ransom or have their company deleted. They chose not to pay the ransom and to try to "freeze out" the attacker. The attacker deleted their entire AWS account and as a result the company went out of business.

Black hats are also known to steal your data. Perhaps they are a competitor who is trying to steal company secrets to compete against you, or an electronic vigilante. Consider, for example, what happened to Sony Pictures when all of their confidential correspondence was made public.

Silent Data Corruption

Every time you transmit data and every moment that data is stored it risks silent data corruption. A one becomes a zero and suddenly your file is no longer useful for its intended purpose. Silent corruption can occur in transit and as data is written to the media. Every transmission and storage medium has an undetectable bit error rate, which is the odds that the data that you stored is not the data you thought you stored. The odds of any individual write being corrupted are quite low (e.g. 1:1015-21 depending on the medium), but the odds of it happening to you and your data go up every time you transmit a new file or object.

Silent data corruption can also occur over time. The only question is how long before your data is corrupted and how much of it will be corrupted. There is a formula (KuV/kt) that can help determine how well a particular medium will store your data over time, but the data is clear – all data stored on magnetic media will degrade over time. (Flash media will also degrade but for different reasons.)

Intrusion Detection and Prevention

No matter where your data is stored, you need an intrusion detection and prevention system, as well as a corruption detection and prevention system. These are related but very different data protection techniques, both of which are possible in the public cloud.

Access to your data should be protected at all costs. Inappropriate access can be prevented using enhanced access control and all access can and should be monitored. Role-based access and two-factor authentication are two of the best ways to protect against unauthorized access and to limit exposure if unauthorized access were to occur. All cloud providers provide access logs that should be extracted and saved for future use in the case of a data forensics situation. They can also be analyzed for trends, such as a significant increase in access of a particular user, or a particular user accessing their data from a different location than usual.

Corruption Detection and Prevention

Eventually, it seems, that even the most secure organization can be infiltrated. As a result, organizations should also have a corruption detection and prevention system. Of course data stored in the cloud should be stored or sent to the cloud in such a way that a company can recover from any corruption or accidental deletion, and make it harder to accidentally delete cloud data, and the data must also be proactively monitored for corruption or attacks.

Most cloud vendors do have the ability to have multiple copies of data in different geographies. However this is more of a HA rather than a preservation function as typically snapshot integration is not native nor simple in the cloud across tiers and across clouds. Traditional cloud vendors also do not yet support WORM functionality that could prevent accidental or malicious deletion or corruption of data, nor do they support automated integrity checks of the data.

But this doesn't mean that a third party platform can't provide such functionality. If a platform acted as a gateway to a traditional cloud vendor, it could add a number of integrity data functions, such as creating a WORM or golden copy of data that cannot be overwritten via that interface. Customers should choose a platform where you can specify retention locks, access and WORM copies before data goes to the cloud. This control is essential, so the organization knows the right data is being preserved, that it can't be accidentally deleted and can't be made available after the expiration date has passed.

While a third party platform can prevent accidental or malicious deletion or corruption via its interface, it is still possible that a malicious person could go around their interface and attempt to damage data by directly interfacing with the cloud storage vendor the platform stores its data on. In addition, it is also possible that a copy in the cloud could suffer from bit rot (i.e. silent data corruption) over time. For these reasons, the platform should automatically check the data integrity of the golden copy on a regular basis.

Such a service should in a separate environment, or else errors & attacks could wipe both the data and the verification system together. Verifying data by bringing it back from the cloud or creating an infrastructure in the cloud for such a purpose could be expensive, so such systems should be built in leveraging cost effective compute components such as LAMDA from AWS or Micro functions from Azure. The system should be able to deliver verification reports on demand and be able to catch the "cloud backdoor" that was described earlier. It can also automatically process data access audit logs to perform root cause analysis should the unfortunate occur. These logs should also capture administrative overrides on WORM locks/retention periods should they occur. It is important to have a consolidated end to end audit of logs of all data from when it leaves your company, during its transit and finally at its rest in the cloud. This would mean a consolidated view of both the cloud vendor logs as well as the platform logs in a cohesive manner.

StorageSwiss Take

There is increasing pressure to consider the cloud for data preservation due to its cost, agility and other reasons. Before companies embark on this journey, they are advised to consider mechanisms to provide data assurance at the on-prem staging area, during transmission and when at rest in the cloud. We encourage companies of all sizes to follow these best practices before they consider a large scale move to the cloud for data preservation.

Sponsored by CloudLanes