GDPR IS FOR EVERYONE - DESIGNING A DATA PRIVACY INFRASTRUCTURE

by George Crump

Storage Switzerland, LLC

CHAPTER 1: WHY ALL BUSINESSES NEED A DATA PRIVACY POLICY

Many IT Planners look at the European Union's (EU) General Data Protection Regulation (GDPR) as a distinctly European problem. The reality is the GDPR and regulations like it, impact businesses around the world. GDPR impacts any business that does business in the EU, and many countries are adopting legislation that is very similar to this regulation. In the US for example, California recently passed the Consumer Privacy Act, which has much in common with GDPR.

From a data perspective, GDPR and similar legislation make two processes critical within an organization; data protection and data retention. Unique to these regulations is the focus on the protection and retention of data belonging to users and customers even after those customers have signed a usage agreement.

The data protection components of these regulations are straightforward. They require that the organization must protect all customer and user data and that it must all be recoverable. A challenge in these regulations is their use of language like "in a timely manner", but they don't define what "timely" means.

UNDERSTANDING DATA RETENTION AND PRIVACY

A data retention policy defines how long an organization will keep data. Data privacy defines how an organization will ensure who has access to the organization's or a users' data. Regulations like GDPR present a new challenge to data retention and privacy policies. These regulations allow the user or customer to demand that an organization remove all data that it may be storing about the user or customer.

Often called "the right to be forgotten", these policies are essentially retention policies in reverse. The organization, instead of making sure it always stores a copy of data, now has to make sure that it removes and never stores another copy of that user's data again. It has to redesign its backup architecture to respond to requests like "delete John Smith from every backup and archive."

THE PROBLEM WITH BACKUP AND DATA PRIVACY

Executing a data privacy request requires an organization to have granular detail of the data it stores. The problem is that most backup solutions, for years, have protected data in bulk. In most modern backup applications organizations use today backup data is stored as an image of a volume or a server. Image backups are more efficient at data transfer than most other forms of backup, especially when protecting hundreds of thousands files. While most data protection solutions can extract individual files from an image backup, searching across multiple image backups for every instance of a particular group of files is almost impossible.

> The answer for organizations is to look for solutions that integrate data protection and data management into a single solution.

THE PROBLEM WITH DATA MANAGEMENT

The granular, discrete access to specific sets of files and making sure those files are retained, or not, is typically the responsibility of a data management or archiving solution. The first problem with counting on a data management solution to enable an organization to comply with GDPR like data privacy demands is most organizations don't have a data management solution. These organizations count on backups for their data retention needs, which won't meet the GDPR requirement. The second problem is that data management solutions don't provide data protection so the organization has to implement, manage and monitor at least two separate solutions.

DATA MANAGEMENT WITH INTEGRATED PROTECTION

The answer for organizations is to look for solutions that integrate data protection and data management into a single solution. The solution will still need to provide very rapid backup performance while having the granularity to search across those backups to find specific data patterns that need to be either retained for a long period of time or deleted.

In our next chapter, Solving "The Right To Be Forgotten" Problem", we'll discuss how organizations can address the this problem and how data management and data protection software needs to change to better meet these demands.

CHAPTER 2: SOLVING THE RIGHT TO BE FORGOTTEN PROBLEM

An aspect of the European Union's (EU) General Data Protection Regulation (GDPR) and similar regulations like California's Consumer Privacy Act, is the "right to be forgotten." Simply stated this means that a user or customer of an organization's resources has the right to ask that organization to no longer store their data. While removing data from primary storage is relatively easy, this aspect of these regulations causes a particular problem when it comes to secondary storage formats.

RIGHT TO BE FORGOTTEN AND BACKUPS

The right to be forgotten in relation to backups is particularly troublesome. Most removal requests will involve unstructured data like documents and images. The total capacity of unstructured data sets as well as the large number of files that they contain, leads many backup software developers to backup these data stores as images instead of individual files. The problem with image backups is the software loses individual granularity across backup jobs, meaning searching all backups for "John Smith's" data is almost impossible.

Even if this data is backed up file by file, most backup applications are still job based. Removing data from within a job is a very rare capability. In most cases, retention policies have to be set at the job level. A right to be forgotten will require that the entire job be deleted which invalidates the backup and potentially breaks other retention requirements.

> Data retention is either the responsibility of production storage or a separate archive process.

WORKAROUNDS FOR RIGHT TO BE FORGOTTEN AND BACKUP

There are several proposals for working around the requirements of right to be forgotten. They all hope that backup data is somehow excluded from the requirement because it is not in production or in a usable format until it is restored. This hope is unproven thus far.

If backup data is somehow excluded from consideration, then backup software vendors still have work to do. Most are promising to deliver a "delete on restore" capability. Delete on restore will require an organization to keep a list of people requesting to be forgotten. It is also unclear if keeping a list of people requesting to be forgotten is in compliance. The backup software will then, during restore, see if the data it is restoring belongs to a user on the list, if it is then it will restore it to a "null" device, essentially making sure that the user's data never comes back into production. It is unclear what impact the constant checking of every file being restored will have on restore performance but it is reasonable to assume it will have a significant impact. It is important to note that at this moment, no vendor provides this capability and adding it won't be an easy development effort.

Another alternative is to restore all data to a quarantined area, then remove all data belonging to users requesting to be forgotten prior to moving data back into production. This method is more readily available today but is full of concerns. First, it assumes it is acceptable to have a list of users requesting to be forgotten. Second, it assumes that is acceptable to restore all data to a quarantined area. Neither assumption is proven acceptable at this point.

Once the restore to the quarantined area is complete it also assumes that the organization has the tools to scan the data to find data that should be removed. It again also assumes that it is acceptable to keep a list of forgotten users. Finally, this method means that every restore becomes a two step process. First, IT restores the data to the quarantined area and then has to restore it again to production. This method doubles the time to restore even without factoring in the time to scan data, which could easily triple restore times.

Another alternative is to only maintain backups for a very short period of time, five days as an example. Data retention is either the responsibility of production storage or a separate archive process. Retaining data via production storage means never allowing the deletion of production data and possibly maintaining infinite version tracking capabilities, driving production storage capacities (and spending) to record levels.

The alternative, archive everything, means that the organization needs to implement an archiving solution. Storage Switzerland finds that most organizations do not have a formal archiving process in place today. Most organizations use backup for their archive, which won't work for reasons described above. It also requires the archiving of all data, not just old data. This requirement means that the archive solution will need to scan the environment almost as frequently as the backup solution, which impacts overall performance.

SOLVING THE RIGHT TO BE FORGOTTEN PROBLEM

The solution to the right to be forgotten problem is multi-faceted. First, backups of unstructured data need to be done file by file, not by images. Vendors need to develop technology that enables file by file backup without greatly impacting the time it takes to protect data. Second, backup and archive need to integrate into data management. In this model backup becomes the method by which data is transferred but archive is the manner is which it is managed. The data management software then provides the ability to search and remove data directly from the archive/backup copy.

CHAPTER 3: CREATING A DATA PRIVACY PROTECTION ARCHITECTURE

The General Data Protection Regulation (GDPR) forces organizations to evolve from a data protection mindset to a data management mindset. IT can no longer let backups store data on secondary storage as giant blobs of ones and zeros. The process that transfers data must also possess an intimate understanding of the data it is storing.

THE BACKUP PROBLEM

Most organizations count on their backup process not only as a means to recover from data corruption or hardware failure but also for data retention. The problem is because of the growth of unstructured data, especially as it relates to the number of files, most backup solutions now backup and store that data as images. For legacy backup products, an image based backup of a large file-system with hundreds of thousands of files is actually faster than a file-by-file backup. While individual file recoveries from a particular backup job are possible, searches across backup jobs is difficult.

Even if the files are backed up file-by-file, most backup solutions have relatively rudimentary metadata databases. They typically only provide a file name and media location. Legacy backup products also store data by backup job, with all data backed up during that job's execution stored together regardless of data type. While searching across a file-by-file backup is possible, removing files from within a job and having that job still remain viable, is not.

THE ARCHIVE PROBLEM

An alternative is to not use backup for retention and only store backup jobs for a few days. Retention is then done by an archive. Most archives store each file discretely so finding and removing files is more straightforward. The challenge with this alternative approach is that it requires two time consuming passes across the file-system. Data is also stored twice, once by each process, and in most cases in two separate storage systems. The second pass, performed by the archive software, is not optimized for performance like the backup pass is. The result is the archive pass takes even longer to complete.

Additionally, unless data is aggressively removed from primary storage, which many organizations are not willing to do, the archive approach is more expensive and more time consuming than traditional backups.

SOLVING THE GDPR BY INTEGRATING BACKUP AND ARCHIVE

A more logical approach, since data needs to both be protected and retained, is to integrate the two processes into a data management solution. The transport component of the solution performs a file by file backup of the environment, but uses a journaling like approach so that after the first backup job is complete, subsequent data transfers of new or changed data complete quickly.

Data is then stored not by job but logically, by file. The solution tracks file versions and builds a rich metadata index of all the files it is maintaining. The software could optionally remove files from primary storage if the organization so chooses. Its data structure also makes it easier for the solution to tier data to the cloud so that on-premises secondary storage doesn't exceed data center capacity.

An integrated data management approach means that GDPR's right to be forgotten requests are easily executed. Removing John Smith's data from the secondary data store is as easy as removing it from primary storage. In fact the data management software, since it has a journal of what is on primary storage, can in a single pass remove data, from both primary and secondary storage. The software could eventually log the transaction as proof that John Smith's data is removed.

Secondary data, stored granularly has value beyond GDPR. For example, ransomware malware files often site idle for weeks prior to execution. During the idle time they are backed up. An integrated protection and data management solution could leverage threat lists to scan the secondary storage repositories and remove any malware files that have made their way into them, ending ransomware attack loops before they begin.

CHAPTER 4: HOW APARAVI DELIVERS GDPR COMPLIANT DATA MANAGEMENT, PRIVACY, AND PROTECTION

Data Privacy regulations like GDPR require that data be protected and managed differently than it has been in the past. Organizations need to prove they are protecting data, securing it, retaining it, and they need to, if a user requests it, remove all of a user's data from their storage systems. Organizations are trying to make their backup solution address all of these demands. The problem is the backup process was never designed to address them.

At the heart of the problem for most customers is unstructured data. Because of the number of files that unstructured data represents, most backup solutions resort to an image backup, which is typically a faster way to backup data instead of a file-by-file backup. These image backups further compound the problems organizations face when trying to adhere to privacy regulations. They lose insight into the files which those backups contain.

Aparavi takes a different approach and integrates data protection and data management resolving issues with both and enabling organizations to comply with GDPR-like regulations. First, Aparavi provides a file-by-file backup but without compromising backup performance. After the first backup, ongoing protection of a file system is done intelligently and can finish in almost the same amount of time that a blind image backup will.

As Aparavi is protecting unstructured data, it also creates a rich metadata history about each file it is protecting. It knows the exact location of each file and each version of each file. The rich metadata enables Aparavi to move data to different storage mediums as the data ages, reducing the cost to store information. Aparavi even supports moving older copies of data it is managing to the cloud, so the organization doesn't have to invest in additional on-premises infrastructure.

Aparavi can also service "right to be forgotten" requests because of its rich metadata. When a user requests the removal of their data from an organization's storage system, IT merely searches Aparavi for all occurrences of that data and removes it from all backups. The organization can even use the log to prove to the user that it has taken the appropriate steps.

Aparavi also has value beyond just the right to be forgotten and data privacy. The software provides complete protection of all unstructured data stores. It also enables an organization to reduce its investment in on-premises production storage. Customers can enable Aparavi to remove old data, data that hasn't been accessed for a while, from production storage, slowing the rate of investments in new storage systems.

While removing data from production storage may cause some concern for some IT professionals, Aparavi eases those concerns. First, it not only archives data it protects it. The same process that removes data also has direct access to the process protecting the data. The protection component ensures that the correct number of protected copies exist before removing the data. Second, users can still easily access their data. Aparavi provides a web platform to provide access to archived data allowing it to be retrieved globally from any device.

ABOUT US

Storage Switzerland, LLC

Storage Switzerland is the leading storage analyst firm focused on the emerging storage categories of memory-based storage (Flash), Big Data, virtualization, and cloud computing. The firm is widely recognized for its blogs, white papers and videos on current approaches such as all-flash arrays, deduplication, SSD's, software-defined storage, backup appliances and storage networking. The name "Storage Switzerland" indicates a pledge to provide neutral analysis of the storage marketplace, rather than focusing on a single vendor approach.

APARAVI™

Aparavi, the world's leading SaaS-based Active Archive helps organizations master out of control unstructured data growth. Delivering both on-premises and multi-cloud mobility, Aparavi provides intelligent data management with true storage independence, and together with an open-data format removes vendor lock-in, forever. Aparavi delivers huge savings by slowing secondary storage growth by 75 percent, with guaranteed availability regardless of how long data is retained. A pay-as-you-go model based on usage eliminates up-front expenditures for a better total cost of ownership. For more information visit www.aparavi.com.

George Crump is the founder of Storage Switzerland, the leading storage analyst firm focused on the subjects of big data, solid state storage, virtualization, cloud computing and data protection. He is widely recognized for his articles, white papers, and videos on such current approaches as all-flash arrays, deduplication, SSDs, software-defined storage, backup appliances, and storage networking. He has over 25 years of experience designing storage solutions for data centers across the U.S.