How to Manage Data Without Putting it in Jail

by George Crump

Data growth requires a response. For some organizations, that response is upgrading to a more scalable higher performing NAS or an on-premises object storage. For others it is moving some data sets to public cloud storage.

Historically, these migrations are a manual copy or move of data from the origin to the new destination. The problem is these manual movements of data are time consuming and have a high chance of error. There is a good chance that not all the data is moved, the incorrect data is moved or when the data is moved, links back to applications needing that data are not updated.

Recently, data management solutions have emerged to automate this data movement process, making both the movement and access to this data more seamless. Data management solutions have three primary objectives: To identify and organize data, to move data to the most appropriate storage system and to provide access back to that data.

Providing access back to the archive data is what causes the biggest challenges. The vendor solutions to create a transparent response to the recall request make the data management solution appear more like data jail. Once an organization commits to one of these solutions they are locked into that solution for the foreseeable future. Also, these solutions tend to take a lowest common denominator approach sacrificing each storage platform's unique features to homogeneity.

The Battle of Transparent Recall

In most cases, to facilitate transparent recall, these solutions leave stub files in the original location that links to where the data has been moved. Alternatively, vendors create a centralized meta-data controller that routes users to the location of the file. While this solution eliminates the need for stub files, it does create the potential for a centralized bottleneck which impacts performance and scalability. These solutions also typically create a unique file system on the storage system that replaces the capabilities of the underlying storage system.

Both situations create a total dependence on the data management solution; neither stub files or metadata is universal across vendor solutions. While the vendor may suggest the user could manually path to the file, it certainly removes the need for the solution in the first place.

Great angst is spent trying to create a transparent recall environment. By implementing a data management with transparent recall solution, the organization is potentially putting its data in jail. IT needs to decide if their organizations really needs that level of convenience. It is important to remember most data movements are to either archive a dormant set of data to a more cost effective storage area or to migrate active data to newer storage system and in some cases both.

IT Should Still Be in Control of Data Management

Clearly, because of the rapid growth of data IT can't afford to manually move data nor can it take the risks associated with that movement. But that does not mean it needs to give up total control.

The reality is, in both the archive and the upgrade use case, data movement is a one time occurrence. And in both cases IT is well aware of the activity. IT needs a solution that will help identify files based on a given criteria, move those files and create a way to find those files if a user requests them.

"Data management solutions have three primary objectives: To identify and organize data, to move data to the most appropriate storage system and to provide access back to that data."

Archiving with Freedom

The main attraction to archiving data is to save the organization money. The goal is to move in-active data from primary storage to a secondary archive that frees up capacity on the primary store. The result should be less frequent upgrades of primary storage, at least for upgrades for reasons of capacity.

One of the realities archive vendors tend to miss when discussing ROI, is the primary storage is already bought and paid for. Archiving all of it to have an empty primary storage system doesn't make any sense. Instead, IT really only needs to free up enough capacity on the primary store so they don't need to buy more storage.

Instead of archiving the entire 90% of data that hasn't been accessed in the last 90 days, just archive enough to meet the current capacity demand. This means the organization can archive the oldest 10% of data, which probably hasn't been accessed in years. Archiving data this old also means the chances of that data ever being recalled decreases considerably, which minimizes the need for a stub file or metadata management structure.

The value of archiving data without the need for a stub file or metadata controller is data is now stored in the native format of the target file system, free from any data management construct. It also means it can fully exploit the capabilities of the archive target, which if that target is an object store means the organization can take advantage of advanced metadata tagging. If the target is the public cloud, it can leverage cloud compute to run processes against that data.

"The main attraction to archiving data is to save the organization money."

To accomplish this type of archive, IT needs to be able to analyze the current data set, have that data set identified by various parameters (like oldest 10%) and then be able to give a command to move that data based on those parameters. The solution should also help the IT find and give access to data if it is requested for by a user.

Upgrading with Freedom

The other data management use case, upgrading to a new storage system, also benefits when IT is in more control of the data management process. The upgrade use case is also another excellent opportunity to identify data and to only move certain data types. For example, if the organization decides to upgrade to an all-flash array, it doesn't make sense for the organization to move data that hasn't been accessed in years to that upgraded array. Idle data performs the same on a hard disk-based storage system as it does on a flash array.

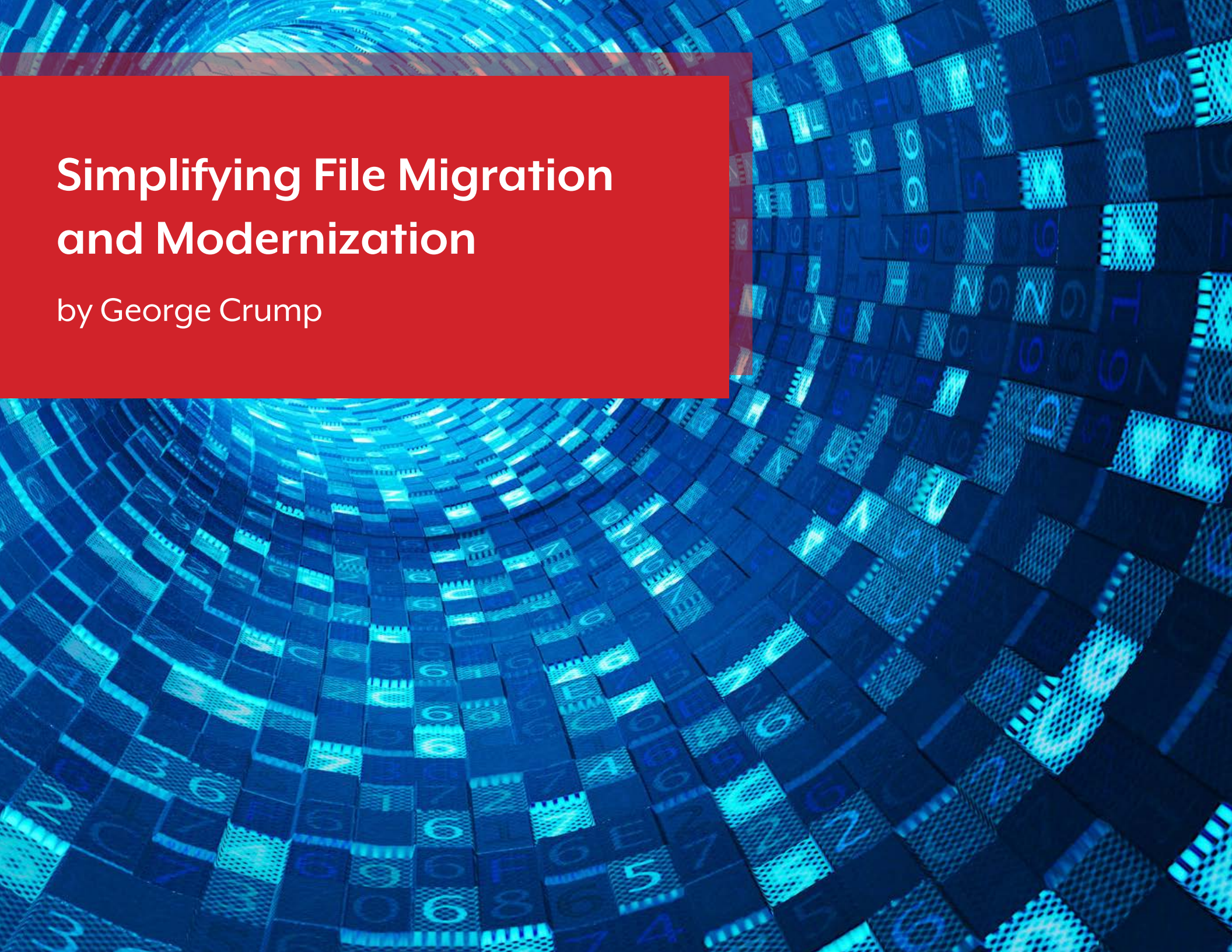
Instead, IT should identify the active and near-active data sets on the current storage system and then move just that data to the new system. The old data could stay in-place on the current NAS if the organization intends on keeping it or it could be archived to less expensive storage either in the data center or in the cloud. Again, these two data sets are moved once. The need to move them again is rare until the new primary storage system fills up. Once that situation occurs, IT either begins the archive process above or starts another round of upgrade migrations.

The amount of recalls that come from an archive is largely dependent on how aggressively the organization moves data to the archive. If data is moved after a few months of inactivity then the chance for a recall is quite high, and these organizations might consider a transparent option. But most organizations have data that is years old on their storage systems. If they start by migrating the oldest 10% of data, they can forestall new hardware upgrade but not have to deal with the potential complexities and vendor lock in of transparent recall.

With this type of data management solution in place, upgrades are really a form of archive. Active and near active data are moved to the new system and old data is either left in place (the old system becomes the archive) or old data is moved specifically to the a cost effective storage tier.

Simplifying File Migration and Modernization

by George Crump



There are plenty of solutions that promise to help an organization migrate, or transform, data in the cloud era.

The problem is most of these solutions are just too complicated and try to do too much. Most organizations simply want to move data from Point A to Point B. The problem is the available solutions are either too simplistic or too complicated for the task.

Unstructured, or file data, is increasingly critical to the enterprise. The problem is when this data is created it is often isolated to the system it was created on. The custodians of the data are the infrastructure team, they do not have context of the data. Data does not tend to flow through the enterprise. Even though most unstructured data after its creation quickly becomes dormant and is an ideal candidate for movement to an alternate location. The challenge is identifying and then moving this data requires careful analysis and planning. A simple drag and drop will not do the job.

Also unstructured data is no longer stored to meet some regulatory requirement, it is stored to be mined. Now or in the future. Using analytics processes organizations are interested in gaining insights or discovering new trends. But in order for these secondary analytics to occur data must be migrated to a native, future accessible store or file system.

The Date Movement Challenge

There are several data management and movement solutions on the market. The problem is these solution often require an additional component that either bottlenecks performance or locks the customer into the migration vendor's solution.

For example, gateways are commonplace. These system translate between one file system type and another (SMB to Object, for example). The problem is the gateway becomes a bottleneck and is a roadblock to maximum performance. File virtualization, as another example, locks the customer into a particular vendor's meta-data management engine. If the provider of the file virtualization solution goes out of business it is very difficult for the customer to move to another provider. Then there are solutions that leave stub files in the original files location. These stub files are vulnerable to all sort of problem and actually increase file count.

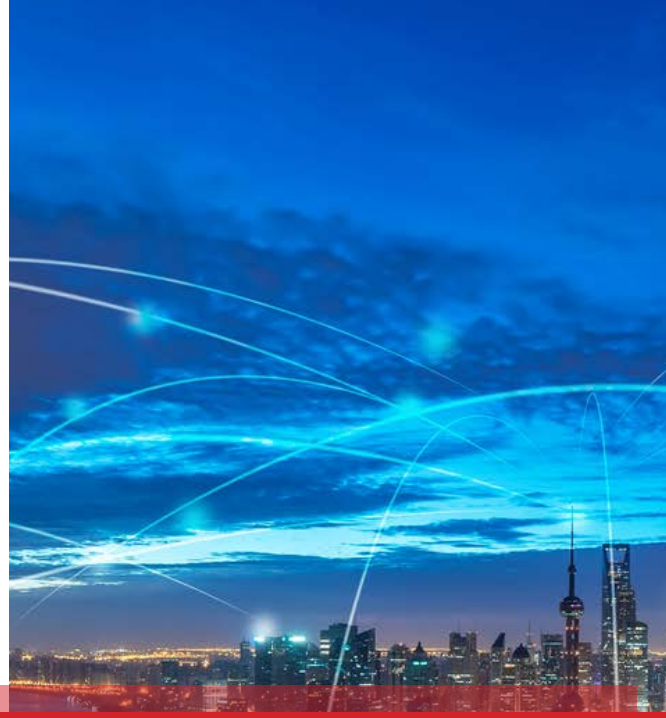
What Data Centers Need

Data centers need to analyze, move, manage and modernize their architectures that they use to support unstructured data. These tasks cannot be accomplished manually, there is simply too much data and too many discrete files across too many systems for humans to be able to manage them all. Instead, IT needs a platform that can scan and analyze data across traditional network mounts like NFS and SMB as well as object storage mounts and cloud storage.

After the analysis is done IT needs to be able to seamlessly move this data to alternate platforms. This movement is not solely motivated by cost, although cost savings is a big factor. In many cases organizations want to repurpose that data. For example, they may want to move data to the cloud to leverage cloud-based services like indexing for search, video/ audio transcription and other Big Data type of functions.

"Most organizations simply want to move data from Point A to Point B. The problem is the available solutions are either too simplistic or too complicated for the task."

Cost savings comes in at the data management part of the process. Where, thanks to the analytics component, data that is cold can be identified and then moved to a more cost correct tier. That tier may be a high-capacity NAS on-premises, an object storage system on-premises or a cloud-based object store. Again, making sure that data lands in one of these targets in its native form provide flexibility for future access and potential data repurposing .
Finally, many organizations are also looking to modernize their unstructured data infrastructure by moving data from traditional file based storage systems to native S3 Object Storage systems. Once again the key is organizations want to move these data correctly, but also independent of a third party file virtualization or gateway. Doing so allows data to land on the object store in its native format and allows organizations to leverage the advanced capabilities of an object storage like advanced metadata and custom keys.



INTRODUCING STORAGEX 8.0

Data Dynamics StorageX 8.0 is a software application that enables the movement of unstructured data between storage systems in the enterprise. Its primary use was as a migration tool, simplifying the ability to move data from Vendor A's NAS to Vendor B's NAS. The 8.0 release expands the use case, enabling customers to analyze their existing data, make informed decisions, move data from legacy protocols like SMB, NFS to S3 Object Storage, manage their data by archiving to S3 Object Storage and to modernize their environment by moving file data to native S3 Object.

While there are a variety of solutions on the market that promise similar capabilities, they are either gateway solutions prone to bottleneck problems or file virtualization/ namespace solutions that require customers to jump in with both feet and trust the vendor will remain in business for the life of their data. In both cases, the data, when it gets to the new environment, acts like it is still on the old environment, which means it can't take advantage of the features and capabilities of that new environment.

StorageX is at the opposite end of the spectrum. It simply allows customers to move data from point A to point B and when that data gets to "point B" have that data be in the new native format so it can take full advantage of the environment's features. This includes taking advantage of advanced metadata tagging common in S3 and object

storage systems. With tagging in place the ability to perform a analytics on this data is greatly improved. Essentially as part of the migration the data is “modernized”. Finally, after the migration is done, there is no need to keep StorageX running, simply shut it down until the next time IT needs it.

The key new additions to StorageX are its improved analysis capabilities, which include a portal for access as well as the ability to apply custom task and create user defined reports on the data in the enterprise. The other big addition in 8.0 is the inclusion of S3 Object Storage into all aspects of the product including, analysis of S3 data storage, the ability to move data to and from S3 data storage and the ability transform data into a native S3 format.

STORAGESWISS TAKE

Data Dynamics’ StorageX has been proven in the market for a long time. It is a well vetted solution. 8.0’s inclusion of S3, and its enhanced ability to analyze current file assets, are a welcomed addition to the product. Data Dynamic’s approach of not being in the data path, providing the organization direct, native access to its data, and having archived data continue to work for it should have high value. Organizations looking to implement new storage systems, S3 compatible object storage systems or to move data to the cloud, should seriously consider StorageX to help them with that process.

THE ANALYST

George Crump is the founder of Storage Switzerland, the leading storage analyst 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.

THE FIRM

Storage Switzerland is the leading storage analyst firm focused on the emerging storage cataegories 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 appraoches 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.

Storage Switzerland, LLC

DATA DYNAMICS
Manage Your Data®

ABOUT OUR PARTNER

Data Dynamics empowers enterprises to seamlessly analyze, move, manage and modernize their critical data assets—from creation to archival—while maximizing the value of their current and future infrastructure investments, on-premises and in the cloud. Their award-winning StorageX software platform enables businesses to unlock their data, so that they can quickly adapt to changing business requirements. Today, StorageX is trusted by 24 of the top Fortune 100 companies and customer proven with more than 160 PB of storage optimized. The result? StorageX has saved more than 80 years in project time and $80 million in total storage costs.