WHITE PAPER

Eliminating the Pain of Migrating Your Unstructured Data

This white paper documents a user-transparent pull methodology that minimizes unstructured data movement while maximizing its value.

Introduction

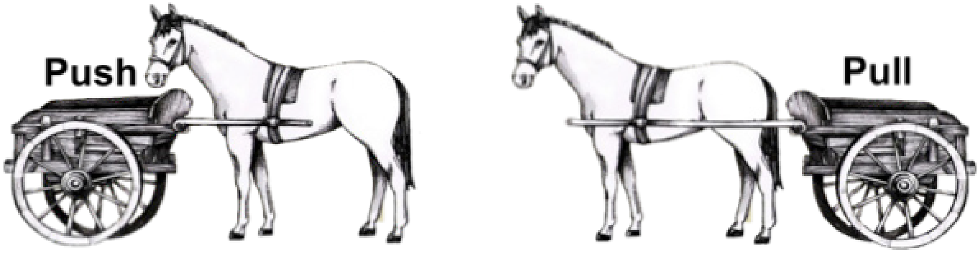
Unstructured data ranges from 60 to 80% of most organizations' stored data. Unstructured data is increasing at more than 3 times the rate of structured data with a compounded annual growth rate of between 60 to 75% per annum as reported by IDC and Gartner. It's big and it's getting bigger. That growth rate means the amount of unstructured data being stored is doubling approximately every 18 months or less. This by itself creates several problems for IT organizations. How can the view of this unstructured data be aggregated in a single or multiple storage systems as it continues to accelerate? How can it be visualized? How can it be searched? How can it be harvested and analyzed to provide useful actionable insight? How can it be protected with minimal risk, cost, and rapid recoveries? How can it be migrated? How can it be archived and still be available in real-time?

These are serious problems that need to be resolved, and resolved sooner than later. There are many storage vendors who claim to have the answer. Some opine their object storage, scale-out file storage, software-defined storage hyper-converged storage, or public cloud storage will solve most all of the problems. (NOTE: none of them claim to solve all the problems.) They scale to huge amounts of capacity, usually in the dozens to hundreds of petabytes, with quite a few scaling as high as thousands of petabytes. Several will provide a global namespace with visibility to all the unstructured data in their storage system. Many can distribute data across geographic regions. Most can tier across different performance and cost profiles. Numerous have extensive unstructured data interfaces including NFS (Network File System), SMB (Microsoft Server Message Block previously called Common Internet File System or CIFS), and S3 (object storage de facto RESTful API standard based on AWS Simple Storage System), with some even providing HDFS (Hadoop Distributed File System). They have data protection software and tools.

Sounds great. Problems solved. There is just one caveat; the only thing that needs to be done to start reaping the benefits of these storage systems is move all of that unstructured data into their system, and to keep that data coming. Full stop. That's a non-trivial problem and the reason these highly scalable unstructured data storage systems have not dominated storage sales as many had predicted.

Moving any amount of unstructured data, let alone vast amounts (terabytes to petabytes) is difficult at best and exceedingly onerous at worst. It is generally today manually labor-intensive. The way it's most commonly done is by first pushing the data from where it is to where it's required, and then continuing to do so as an ongoing process. The key complaint from IT professionals is that it requires much too much time and patience and is analogous to pushing a string.

This paper takes a deeper look at how unstructured data is actually moved around to solve the aforementioned problems. Then it documents a new, far better, more automated, application and user-transparent pull methodology that minimizes unstructured data movement while maximizing its value.

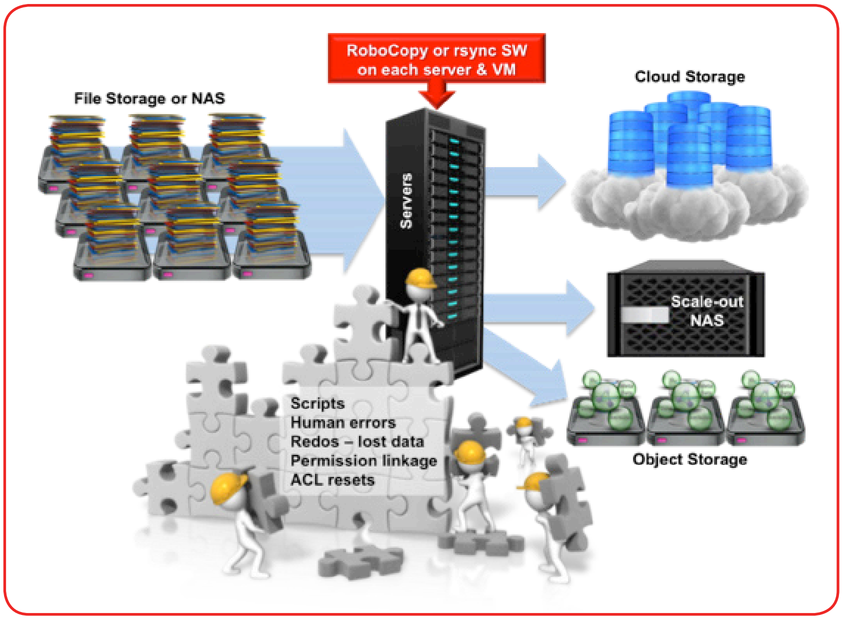


Pushing unstructured data

The typical pushing methodologies for moving unstructured data from where it's stored to an object storage system, scale-out file system storage system (NAS), software defined storage, cloud storage system, or simply for an unstructured data storage tech refresh, all tipically require pushing the unstructured data (files or objects). These include:

1. Data Migration Projects
2. Archiving Software
3. NAS Cloud Gateway + Data Migration Projects Or Archiving Software
4. Intersystem NAS or Object Storage Tiering (Including Cloud Storage)
5. Copy Data Management

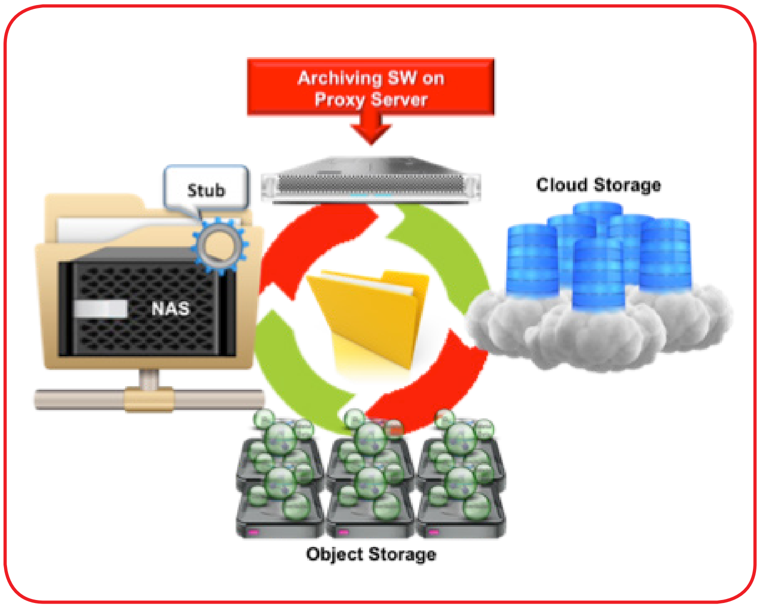
1. Data Migration Projects



These tend to be point events for NAS or filer tech refresh or the one-time movement of large amounts of data. They are not meant to be ongoing. The reason for that is how difficult they are to accomplish. It is not just pushing the unstructured data from one storage system to another, it is making sure the security access and permissions have moved with the data. And of course there is the application and user disruption that takes place as they are repointed to the data's new location. That disruption invariably requires IT to schedule the cutover to take place on weekends, late nights, or holidays. Limited time windows make the entire process highly stressful and mistake prone. Data migration projects are consistently painstaking tasks dependent on open source tools such as rsync or RoboCopy, and scripts. The whole process is manual, labor-intensive, fraught with human errors, a ton of steps, do-overs, and lost data. It is so frustrating that it's considered the worst job in the data center. Bloor Research has reported that 84% of data migration projects are over time, over budget, or both. That budget is considerable. The average third party professional services charge for a data migration project is 30% of the cost of the new target storage system. And consider that the median average data migration project is approximately nine months where both the old and new storage are running, paying maintenance, and consuming operating expenses. There are third party software that can reduce the errors, the time, and the cost, but fundamentally it is still a push-the-string operation. And in the end, all of the unstructured data must reside in the new target storage to have a global namespace and view. Anything outside that storage target is invisible.

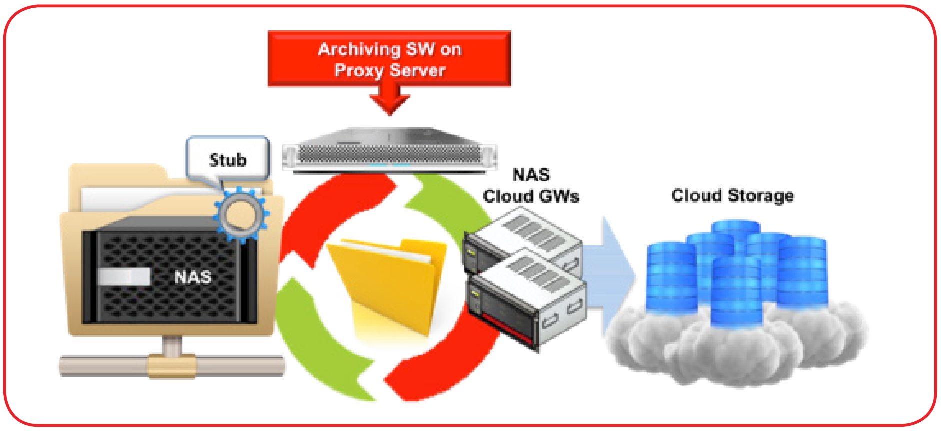
2. Archiving Software

Archival software has been around for quite some time. It works quite similarly to hierarchical storage management (HSM) in that the software copies and pushes the unstructured data from the storage where it resides to the new targeted storage. The data is moved based on policies such as time since last accessed, when it was created, compliance policies, and more. Then the original data is deleted and a stub is left in its place. When the user or application attempts to read or alter a file or object that has been moved, the stub copies and moves the data back



to the original storage location. As time passes and the policies are activated once again, that altered file or object is copied again and moved a third time back to the target storage system. The copy on the original storage is again deleted with a stub left in its place. This means every time that file or object is read or altered it will have to be moved a minimum of two more times. And each time it is altered it creates a new file or object that consumes more storage. This process is quite onerous when moving large files such as videos, films, MRIs, CAD/CAM, seismic files, etc. It takes much too much time, makes the user or application wait for the file, hogs bandwidth, makes multiple copies of files or objects, thus consuming excessive amounts of storage. It explains why the push technologies of HSM and HSM-based archiving software never took off in the market.

3. NAS Cloud Gateway



NAS cloud gateways are file storage targets that push files to cloud storage. They deduplicate and compress the files, copy it to public cloud storage, delete the original on the gateway while leaving a stub. They act as a cache for frequently accessed files. But similar to all HSM-like products, it ends up having to move files multiple times to be read or altered if it falls outside the cache. In reality, it gets a bit worse. The NAS cloud gateway is not the original storage for the vast majority of unstructured data. That means the original files reside on a different storage system. They must be moved from the original storage system to the NAS cloud gateway. So, either it becomes a data migration project or archiving software must be used.

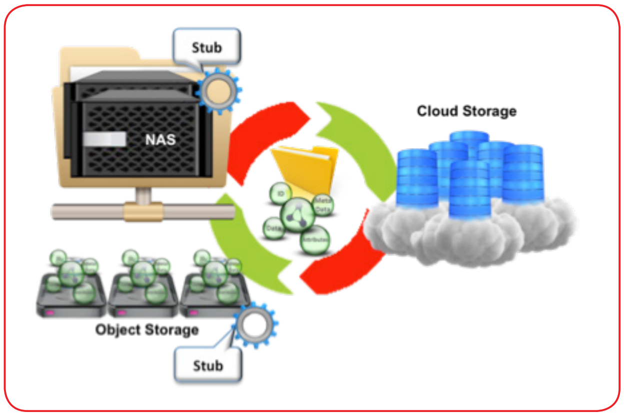
If the choice is a data migration project, the files no longer become accessible on the original storage. The NAS cloud gateway must be mounted for the user or application and files relinked. This assumes the NAS cloud gateway becomes the primary storage over time. Not usually the choice based on the complexity, difficulty, and cost of most file migration projects as well as the limited performance and functionality of most NAS gateways.

If archiving software is the choice, it makes NAS cloud gateways potentially a dual movement dual stub process just to move the data to the cloud. And remember when the user or application wants to read or alter the files they must be recalled first to the NAS cloud gateway and then back to the original storage. Unless those files are not going to be accessed, this results in too much data movement. And then there is all the extra cost of the archiving software, cloud storage transit and retrieval fees.

There is one other disturbing aspect about NAS cloud gateways. If it goes down or goes away, there is no way to access, read, or alter the data. That means there must be a minimum of two gateways and typically more.

It is easy to see that either push choice is less than an ideal solution.

4. Intersystem NAS or Object Storage Tiering (Including Cloud Storage)



Intersystem NAS or object storage tiering functions quite similarly to the NAS cloud gateway except it eliminates the intermediary storage target (NAS cloud gateway) from the path. It still pushes the unstructured data from the original storage system to a different storage target. But, there are only two instead of three storage systems in the file or object path. Otherwise it is still an HSM-like process utilizing stubs and multiple unstructured data movements when read or altered. Many storage vendors additionally demand a license surcharge to move data to public cloud storage or another storage system they do not sell.

On the downside, just like the NAS cloud gateway, intersystem NAS or object storage tiering system is the only path to the data. Thus, it too requires high availability or multiple systems to provide access, which adds significant cost. Accessing or altering files and objects put in a public cloud is comparable to the NAS cloud gateway in that it adds unnecessary transit and retrieval fees.

5. Copy Data Management

Copy data management is a very limited solution. It is primarily file and image-based backup software integrated with a scale-out NAS. The NAS snapshots the backups using pointer-based snapshots creating virtual copies. These virtual copies can be used for dev-ops, test-dev, and search. The backup software is primarily aimed at hypervisor APIs from VMware and Microsoft. The hypervisors push the data to the integrated NAS and backup. Some copy data management utilizes agents (client software that runs on a server) for non-hypervisor hosts to push the data.

There are several issues with this type of solution.

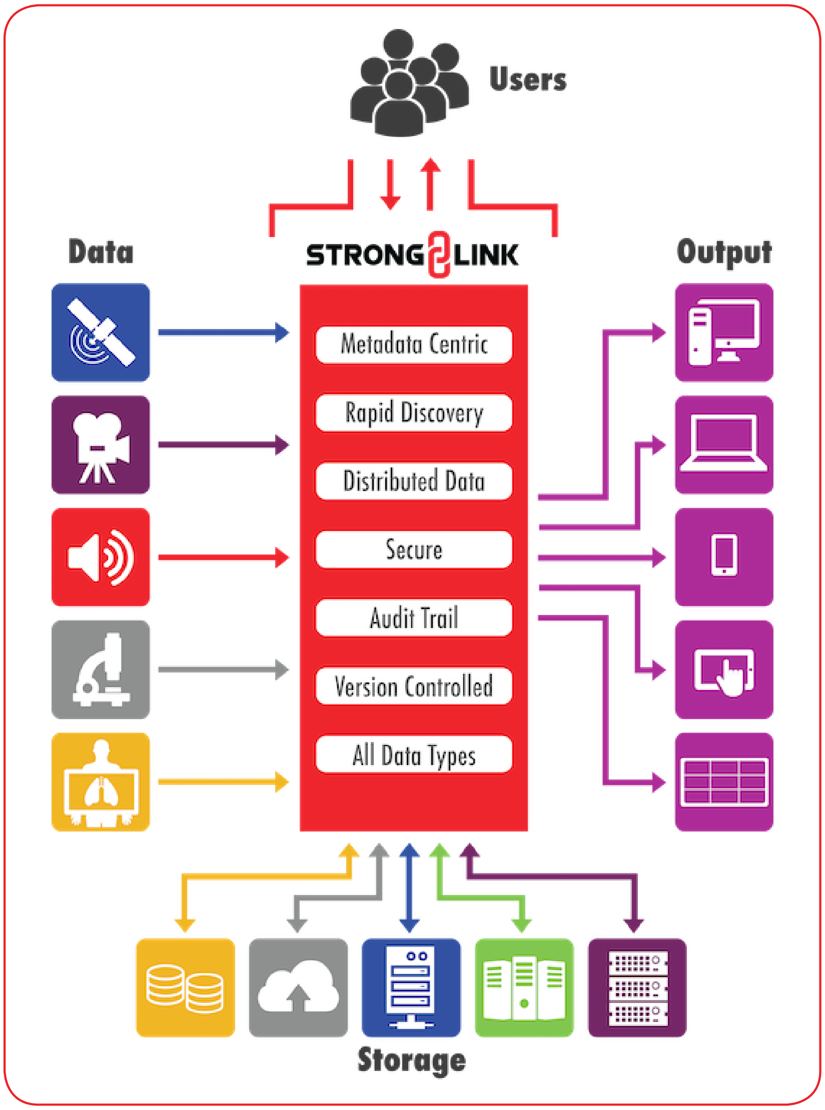
- First and foremost, it is primarily a data protection system with limited repurposing of the data. And an expensive one.
- The data copy is typically 24 hours behind.
- There is no harvesting of metadata or global view. Each backup looks and feels like a completely different data silo.
- Applications, VMs, and the data are not specifically mountable for day-to-day operations. They have tomust be recovered (moved) back to the original systems for that purpose.

It's a narrow fit that does not solve the IT unstructured data movement problem.

That is why StrongBox Data Solutions has come up with a better way to completely solve the unstructured data movement problem based "Pulling" vs "Pushing" the unstructured data. It's called StrongLink.

What Is StrongLink and Why Is Pull Better?

StrongLink® is a synergistic blend of unstructured data management, metadata harvesting, and unstructured storage creating a new storage market category called "Cognitive Data Management". To applications and users, it looks like file and object storage. It's not. It's actually an abstraction layer of the file and object storage that sits behind StrongLink. It is conceptually a virtual file system. To the NAS and object storage behind or in front of StrongLink, it looks like the application or user. That enables StrongLink to mount the NAS or object storage, read all of the unstructured data (assuming it has been given the right permissions), pull the data by copying and moving it to any of multiple unstructured storage systems, including: NAS, filers, object storage, public cloud storage, and even LTFS tape utilizing the StrongBox® LTFS NAS appliance – which makes LTFS tape and tape libraries look feel and act as NAS. StrongLink can also delete the original files or objects from the originating storage after they have been copied and verified. From that point onward, the applications and users will mount and access their data via StrongLink. StrongLink will move the applications and users to the data instead of moving the data to them. This reduces data movement, storage consumption, and costs by as much as 80%.

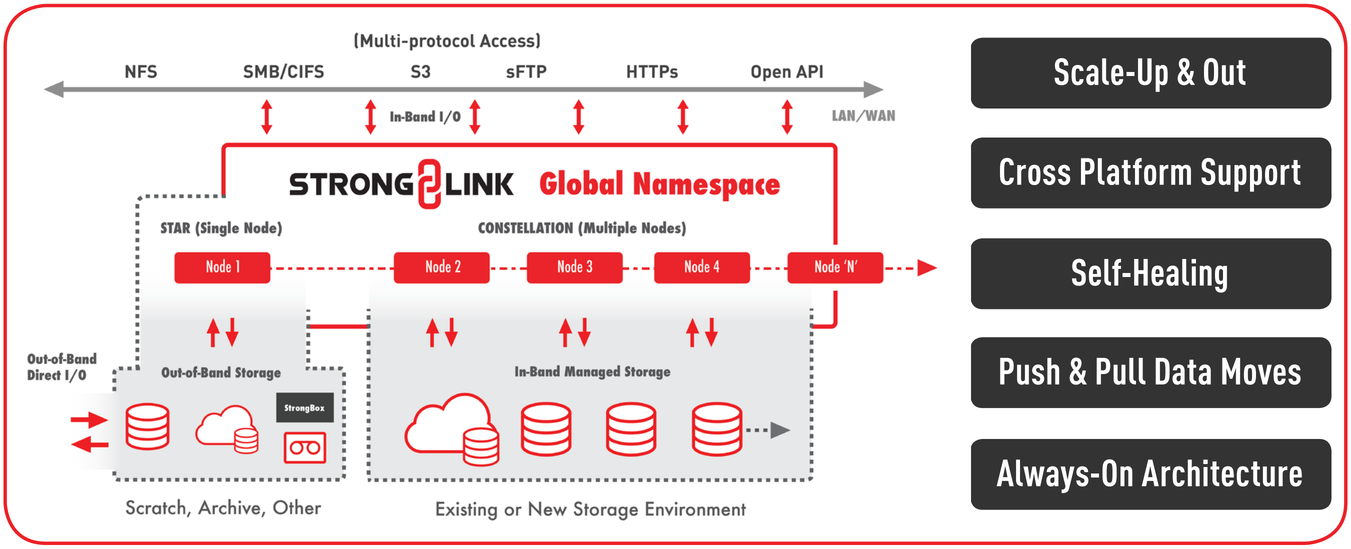


It gets better. Instead of separate data protection, archiving, and compliance software, StrongLink can make and manage as many copies of the unstructured data as required, maintain versions, and keep them on different storage systems. A primary copy may be kept on fast storage. A secondary copy may be kept in a public storage cloud. A tertiary copy can be sent to tape and offsite. Each copy is transparent to the application and user so that in the event of an outage they are automatically connected by StrongLink to the best copy available based on policies such as performance or geographic locality. Copies can be set to be destroyed based on policies. StrongLink's inherent built-in data protection capabilities eliminate the requirement for any further backups or data protection. All versions of files and objects are always available, resilient, and protected. The cost savings in data protection and archiving licensing, hardware to run the software, infrastructure to support the software, and operating expenses are substantial.

And it gets even better. Because all of the unstructured data is pulled through StrongLink, all of the metadata is captured as well. This provides organizations with a global view of their data in a global namespace making searches easy and simple. StrongLink additionally has a machine-learning engine that organizes, parses, and harvests that metadata empowering searches and big data analytics across platforms and vendors making it substantially faster.

StrongLink is highly scalable software with a no-master peer-to-peer architecture. It scales linearly with no known limits at this time. StrongLink operates out-of-band, in-band, or a combination of both. It runs in commodity-off-the-shelf (COTS) white box server hardware and/or as a virtual appliance running in a virtual machine (VM). The virtual appliance empowers StrongLink to provide the same capabilities to clustered hypervisor and hyperconverged infrastructure systems. Doing so overcomes the storage limitations of those systems opening them up to more storage, more variety of storage, and much lower costs.

StrongLink solves the unstructured data movement problem by pulling vs. pushing the data. As a result, StrongLink changes everything. Because in the end it's all about the data not the storage.


(Multi-protocol Access)
NFS
SMB/CIFS
S3
sFTP
HTTPs
Open API
In-Band I/O
LAN/WAN
STRONGLINK
Global Namespace
STAR (Single Node)
CONSTELLATION (Multiple Nodes)
Node 1
Node 2
Node 3
Node 4
Node 'N'
Out-of-Band Direct I/O
Out-of-Band Storage
StrongBox
In-Band Managed Storage
Scratch, Archive, Other
Existing or New Storage Environment
Scale-Up & Out
Cross Platform Support
Self-Healing
Push & Pull Data Moves
Always-On Architecture