The Modern Approach to NAS Data Protection

THE CHALLENGES OF NAS DATA PROTECTION

Network-attached storage (NAS) data comprises the largest unstructured dataset in most organizations today. Due to the sheer amount of data, the significant differences between how block and file data must be handled, and the resources and time required to make regular copies, backing up NAS data can be a daunting, if not insurmountable, task.

In 1996, when the average NAS dataset was in the TB range, Network Data Management Protocol (NDMP) was the answer, but today it is not uncommon for a large NAS data footprint to exceed 1PB; most average at least 150TB. At the same time, there has been a demand to reduce backup windows in an effort to reduce impact on production processes. This has led to data protection trade-offs and compromises that would not be acceptable for any other production data.

DobiSync® addresses this issue by filling the void that exists in today's NAS data protection strategy in a flexible, scalable, and reliable way.

A BRIEF HISTORY OF NDMP

Built in 1996 as an open standard protocol for backing up NAS devices, NDMP was last revised in 2003. Primary contributors to the project were NetApp and Legato (which later became the Dell EMC® NetWorker product). According to the Storage Networking Industry Association (SNIA), there is no further development planned; therefore, NDMP is effectively dead. It was built in the age of 16GB hard drives and is rapidly becoming irrelevant in an era of PB+ single filesystems.

Furthermore, most of the backup media or devices on the market today are smaller than most of the source NAS systems and rely heavily on deduplication and compression to shrink the backed-up datasets to fit. Unfortunately, NDMP backup streams are not easily compressed.

In recent years, storage vendors have introduced proprietary extensions to NDMP to extend functionality as a short-term solution. We can call this NDMP+. Isilon® Changelist API and multistream NDMP and NetApp® SnapDiff are good examples of this, but neither provide a fundamental fix to the underlying problems. Any of these extensions also require that the backup application vendor add support for these specific enhancements as well.

KEY ISSUES WITH NDMP DATA BACKUPS

- NDMP backups are almost universally filesystem specific, which means that data backed up from a NetApp WAFL filesystem cannot be restored to an array from Dell EMC using another filesystem type (UXFS, UFS64, OneFS), or vice versa.

- No two vendors store multiprotocol permissions in the same way; therefore, backing up these datasets is inherently complex.

- With the broader move from 32-bit to 64-bit filesystems (or aggregates), larger datasets are less likely to be able to finish in a given backup window.

- Proprietary NDMP extensions do not provide a permanent fix because they must be implemented in the same way on the backup application.

- Once NDMP is used for backup, that dataset cannot be restored with any other backup application, nor to any other NAS device. Users are tied to NDMP for the life of their retention requirements, and to that specific source platform. As a consequence, companies are often forced to keep a single old system in the data center into perpetuity on third-party maintenance for no other reason than the unlikely potential need for data restoration.

- Different vendors implement parts of the NDMP "standard protocol " specifications , but not all; e.g., restartable mode is supported on NetApp but not VNX.

SNAP AND REPLICATE: THE NDMP WORKAROUND

In the past several years, many organizations have given up on doing traditional backups once they exceed 150TB of NAS data. Instead, they have accepted the compromise of relying on snap and replicate as a "good-enough" solution. This involves using storage-level snapshots, which were designed for short-term recoverability (one to two weeks), and replication to a disaster recovery (DR) array. While this is good-to-have, it is simply not enough.

Circumstances and the lack of better technical solutions have forced the hand of backup and storage administrators.

Snapshots

- Snapshots are intended for short-term recoverability only. If malware or ransomware infects, corrupts, or encrypts your primary data, and you don't catch it quickly, the snapshots are affected as well.

- The more snapshots you have and the longer you keep them, the greater the impact will be on disk and CPU resource overhead on the production storage.

Replication

- Replication is designed for business continuity and failover or protection against ransomware, not versioning.

- To meet RPO and RTO goals, data must be replicated frequently, which results in corruption or deletions (accidental or malicious) being replicated quickly as well.

DOBISYNC: RELIABLE, RISK-FREE NAS DATA BACKUP AND RESTORE

DobiSync is backup software uniquely designed to handle the complexity of today's large NAS data backup tasks and to fill the void in a data protection strategy that ignores the importance of proper data backups. DobiSync is specifically architected to enable the fast, accurate, hassle-free backup and restoration of any amount of data, from any NAS platform to an on object storage system (public or private cloud).

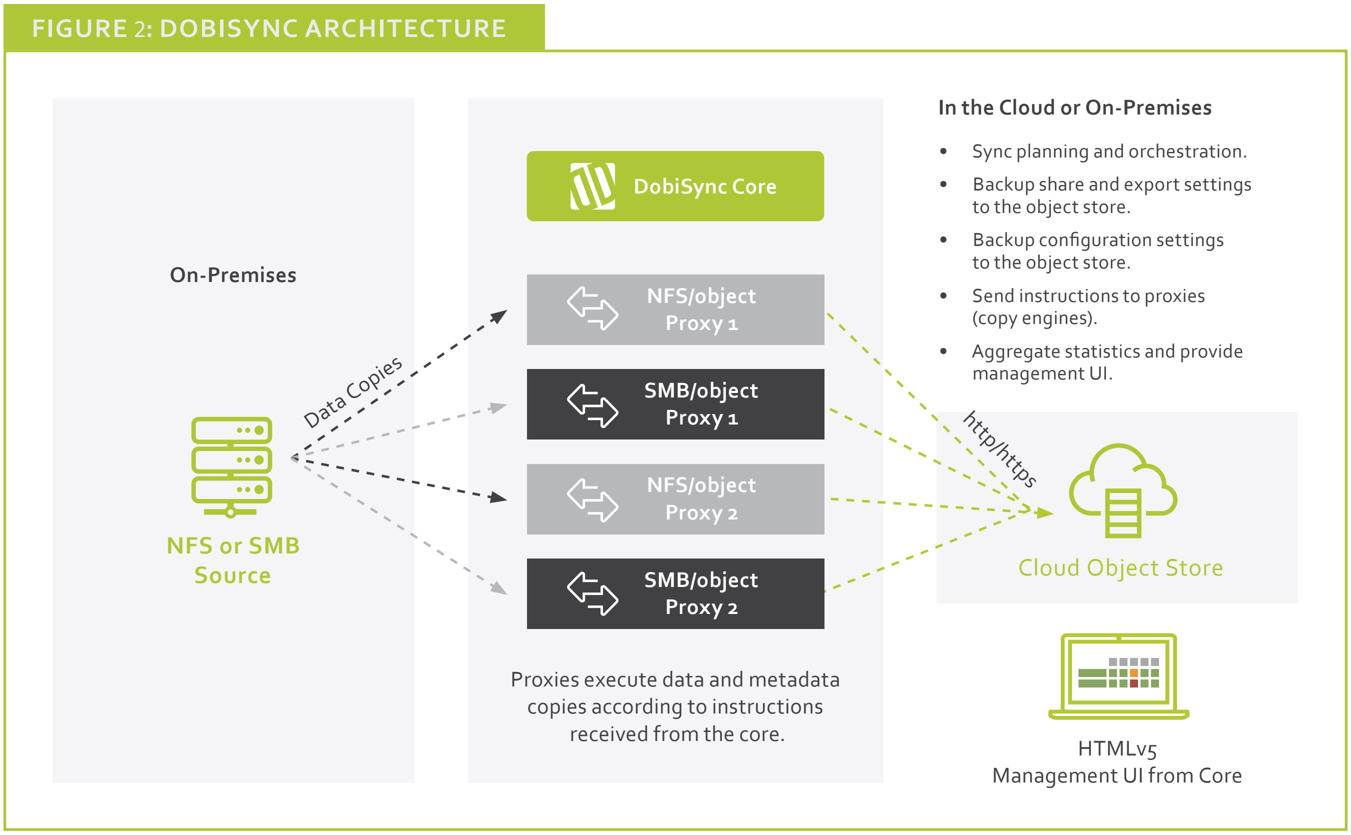
FIGURE 1: THE MODERN APPROACH TO NAS BACKUP

Key characteristics of successful NAS data backup software	Best practice	How DobiSync delivers
Modern	The same protocols are used to back up the data that users and applications use to access the data (SMB up to v3, and NFS up to version 4.1).	DobiSync enables backup of any file data and metadata (SMB or NFS) to your choice of object store. Also protect every version of share and export definitions with each backup set.
Immutability and integrity	Data integrity must be checked on every file, every time it's backed up. The target must be a reliable media. There is no sense backing up data if you 're not completely confident you can restore it.	DobiSync calculates an MD5 checksum on every file on the source, gathers the checksum from the target object, and ensures that they match. Using an object store can give you geo-replication and data consistency out of the box, plus encryption in flight and at rest.
Scale	The backup target must be able to grow larger than the source filesystem and accommodate as many copies of the data as the business requires.	DobiSync lets you back up petabytes from one user interface (UI) to an object store that scales however you want. It allows you to keep as many versions as you want, for as long as you want.
Granular and flexible restoration	The target must allow recovery of any dataset or portion of a dataset that originated on any source, to any target that speaks the same file access protocols.	Because the data is read over NFS and SMB, DobiSync can restore any portion of the data, down to the file level, from any version, to any target that supports SMB or NFS.
Lifecycle	The backup application must be able to migrate the backup set as trends and targets change over time without losing any of the aspects mentioned above.	Datadobi was founded to enable large object data migrations and has developed solutions that make unstructured data migration and protection painless and hassle free, whether it is in the cloud, on-prem, off-prem, or any combination.

DOBISYNC ARCHITECTURE

DobiSync uses a core-and-proxy model with the core acting as the orchestration layer – so it can manage scheduling and system activities, as well as provide the UI and the proxies acting as data movers with an ability to scale out according to your requirements. Need more speed? Increase the number of proxies. Need to throttle on and off hours? No problem, it's all built into a single easy-to-understand user interface. And because backing up your data from a file-based system to an object-based system requires a protocol translation, all the proxies speak both object and file protocols.

FIGURE 2: DOBISYNC ARCHITECTURE



Proxies execute data and metadata copies according to instructions received from the core.

In the Cloud or On-Premises

- Sync planning and orchestration.
- Backup share and export settings to the object store.
- Backup configuration settings to the object store.
- Send instructions to proxies (copy engines).
- Aggregate statistics and provide management UI.

WHAT DIFFERENTIATES DOBISYNC

1. A NEW APPROACH TO STORING FILE DATA

One of the greatest challenges to storing file data in an object store is how to organize and save all the associated metadata. DobiSync keeps each version of the file as an object, only copying a new one if the content changes. It also keeps each version of the metadata as an object alongside the file (only copying it again if the metadata changes).

By doing this, DobiSync can effectively store an unlimited amount of metadata about each file, and an unlimited number of versions. All data can be stored in one bucket or split it up across multiple buckets each with its own set of keys, physical characteristics on the storage system, geo-protection, encryption at rest, performance, and more.

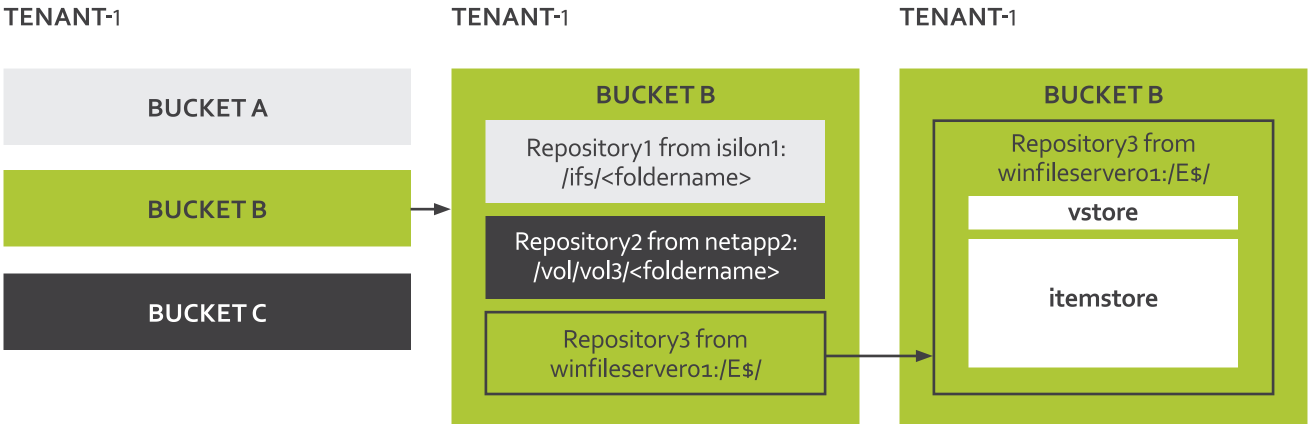
2. THE BACKUP CATALOG IS IN THE VERSION STORE

A backup catalog can be thought of as an index keeping track of where each bit of backed up data is stored. Traditionally, this catalog is stored in a SQL database on a backup server, which must itself be backed up. In DobiSync, it is stored in the object store itself. The following is a diagram of how this works.

A. Inside a tenant or namespace of an object storage system are multiple buckets.

B. Inside of each bucket, DobiSync creates a logical construct called a repository for each unique source path.

C. Inside each repository, two further logical constructs are created, just for that specific source path: the version store and the item store.


TENANT-1
BUCKET A
BUCKET B
BUCKET C
TENANT-1
BUCKET B
Repository1 from isilon1: /ifs/<foldername>
Repository2 from netapp2: /vol/vol3/<foldername>
Repository3 from winfileservero1:/E$/
TENANT-1
BUCKET B
Repository3 from winfileservero1:/E$/
vstore
itemstore

The version store contains:

- Versioned share and export settings
- Job settings and statistics (start time/stop time/completion status of each pass)
- Locations of all versions of all objects in the item store (data and metadata)
- Source directory structures
- Source fileserver configuration for the configured jobs (no passwords or other sensitive information)

The item store contains:

- Each unique version of each file or folder as an object
- Each unique version of the metadata for each file or folder
- An index of the data and metadata for each file or folder

The advantage of this approach? The version store is the backup catalog, but is stored alongside the data itself; therefore, it doesn't need yet another backup solution.

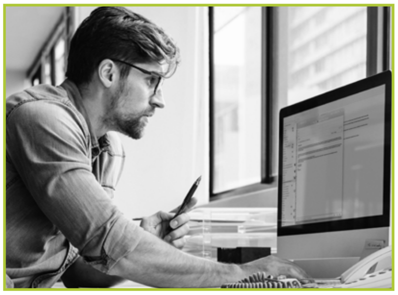
3. RAPID RESTORATION IN THE EVENT OF DISASTER

As long as you have the keys (access/secret) to the object storage buckets, or the ability to generate new keys, you can quickly deploy a new DobiSync environment on-prem, at your DR site, or in the public cloud; give it the keys to the object store, and it will locate all the existing backup repositories in each bucket, and quickly show you everything available for restoration. This DR DobiSync system can even be pre-created and running in the cloud – ready to go to simplify the process should it ever be needed.

4. A RISK-FREE APPROACH TO DATA ENCRYPTION

Many products offer an extra level encryption of data as it is transmitted. The problem with this approach is varied:

- You must now manage an extra set of encryption keys to be able to decrypt the data (and an external or internal keystore).
- Encrypting data that is already encrypted by the protocol unnecessarily burdens performance.
- Additional encryption of the data by the application renders it totally unusable by any outside process.
- The application that wrote the data is the only one that could ever be used to recover it.



DobiSync does not add extra encryption to your data, which means the data is never held hostage. However, the file access protocols themselves can be encrypted. For example, if SMBv3 or NFSv4 is configured on your source NAS, as the data flows from the NAS to the DobiSync proxies, it is encrypted in flight. If you have https set up for your object store, it is encrypted with SSL as it travels to the destination. And finally, if your object store supports encryption at rest, or uses self-encrypting drives, then the whole process is encrypted end to end.

THE TCO STORY

With any backup solution, you have two primary components to the total cost of ownership (TCO), and calculating it can get very complex:

1. The backup application and associated infrastructure
2. The backup storage target or media, which includes:
   a. The backup target
   b. A DR copy of the backup target (at a DR site)
   c. An off-site storage location and trucking service (such as Iron Mountain®)
   d. Possibly a tiered object bucket to store older backup sets

In the past, tape has been used as the backup media of choice simply because it is cheap and is easy to get off-site quickly. The problems with tape are fourfold:

1. It is potentially unreliable (you never know until you try and read one back if it will work or not).
2. Tape libraries and associated SAN infrastructure are very expensive and have an extremely large physical footprint.
3. To allow long-term recoverability, you must move backup sets from older tapes to newer ones as standards change, which is a logistical challenge every few years.
4. Recovery times are often measured in days, not hours, as you generally have to get certain sets of tapes shipped back from a remote site, have a person load them into the library, and then begin a restore process.

Because of the inherent problems with tape, the industry has largely switched to disk-based backup targets in recent years. These have the disadvantage of being much more expensive than tape in terms of raw dollars per TB; however, they offer more rapid recovery abilities (and an easy way to test restore), lower level of manual effort, and built-in resiliency (through RAID or other means), which have made the investment worthwhile.

But now there are more moving parts. Disk-based backup arrays need to be replicated to another copy at a DR site, which means you are now storing at least four copies of every bit of production data:

1. Production storage array
2. DR storage array
3. Production backup array
4. DR backup array

These backup arrays have monumental price tags too, which are almost always viewed as CAPEX, so the money is all needed upfront.

DobiSync is SaaS (software as a service). Organizations simply pay for the volume of source data protected, per TB, per year. When paired with a hosted object storage solution this gives companies the unique benefits of:

1. Paying only for what you need to protect, thereby offering true BaaS (backup as a service).
2. Storing your data off-site, immediately.
3. Nearly instantaneous restore, as there are no tapes to get off a shelf, ship, or load into a library.
4. Substantially lower costs than the huge on-prem, purpose-built backup appliances.

A BETTER SOLUTION

Giving up on NAS backup is not the solution, it's simply where we, as an industry, ended up due to lack of better options. Snapshots and replication offer steps towards data protection, but nothing more. Now, there is a better option. In light of recent ransomware attacks crippling governments and businesses, it is more important than ever to start protecting your NAS data again. DobiSync is the answer to the question of how to do so at scale and affordably.

FOR MORE INFORMATION

To request a demo of DobiSync or to learn more about Datadobi's range of products and services, visit www.datadobi.com.