# Storage Lifecycle Management: A New Approach to Reducing Data Storage Cost and Complexity

**An Exploration of Storage Lifecycle Management and How Spectra's StorCycle® Storage Lifecycle Management Software Makes it Possible**

800-833-1132 • 303-449-6400 • 6285 Lookout Road • Boulder, CO 80301 USA SPECTRALOGIC.COM

# Table of Contents

# Part I: What is Storage Lifecycle Management?

Storage Lifecycle Management (SLM) is a new concept that allows organizations to have total control of where data resides and for how long. With the proper tool, administrators and users can easily identify and move files and objects to appropriate storage tiers based on cost, performance, digital preservation and security needs. These files and objects may easily be accessed and recalled by the user for future use.


To sum up this ability as simply "tiered storage" or "primary storage offload" would be a grave disservice to the power of SLM. While tiering and offload can be powerful functions, it often overlooks SLM's ability to make other aspects of storage function more efficiently – allowing archives to be active; making tier-1 storage upgrades easier; decreasing backup windows and cost; integrating high-latency mediums such as archive cloud or tape, into active access devices; incorporating machine/sensor/edge data to management policies; and categorizing storage for charge-back or departmental tracking. The list of benefits with SLM is extensive and all aimed at increasing efficiencies, decreasing complexities and delivering substantial cost savings – typically 70% or more.

Spectra's *StorCycle Storage Lifecycle Management* software was designed to create a new approach to working with existing storage as well as planned or future storage – enabling the best of what organizations currently have as well as what new storage technologies will deliver in the future.

# Part II: Applications for Storage Lifecycle Management and StorCycle Storage Lifecycle Management Software

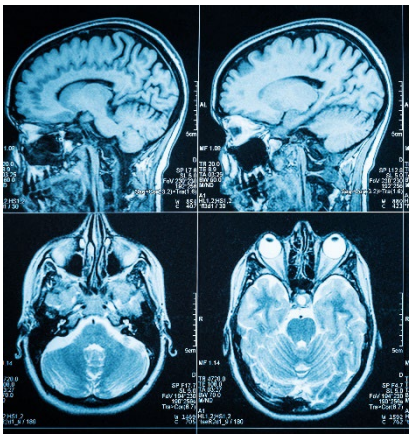
## Project Archive

Project Archive is a key differentiator in StorCycle's approach to Storage Lifecycle Management. Project archive is used by StorCycle customers in a myriad of creative ways across virtually every industry that works with digital storage. While this could be used in any vertical market, it takes on slightly different connotations and even terms across different industries. In this sense, the term "project" may refer to any data set that has some commonality across its contents. It could be the output of an experiment; a collection of yearly financials; content created for a specific commercial, movie or sporting event; and even a group of files created by an individual employee.

Project archive allows users to identify any files or directories associated with a "project" and archive them as a group. In High-Performance Computing (HPC), this can be done immediately after a large project or experiment is completed. Archived data sets can be tagged with additional information to identify anything of importance to the project, be it grants associated with the project, researchers involved, project names, etc. This metadata can easily be searched at any point in the future. Before StorCycle's project archive feature, there were very few other approaches that made sense in this realm.

Often the data sets were too big to be archived via a backup solution – the cost would be enormous, and most of the granular features of backup aren't needed since the data in a project to be archived is usually not changing. It's probably safe to say that every HPC site in the world could use project archive capability.

Machine data is another category of data which benefits greatly from project archive. Wikipedia classifies *machine data* as "information automatically generated by a computer process, application, or other mechanism without the active intervention of a human."



In healthcare, think of incredibly powerful medical instruments like X-ray or CT scan imagining. In Oil & Gas, think of millions of sensors on oil pipelines monitoring oil flow. In merchandising, think of edge computing at hundreds of stores across the country sending back cash register totals at end of day. In engineering, think of modeling or wind tunnel testing creating huge amounts of data. In the entire world, think of The Internet of Things (IoT) where everything from appliances and doorbells to medical monitoring devices and self-driving cars are creating and sending data. According to IDC and other storage analysts, machine-generated data is the largest area of data growth in the world.

This data has to land somewhere, and it is typically on some form of primary or actively-used storage. This data is often considered a candidate for immediate archive until it is needed. So its age or access record isn't helpful in identifying it for archive or migration. You might hear the term *watch folder* – a directory or folder which is set up to move everything that enters directly to longer term storage (BlackPearl NAS, other NAS, cloud or tape).

General IT can also benefit greatly from project archive. We have all had the experience of needing a piece of content or data that an ex-employee created. What happens to the data employees have on their laptops when they leave? That data is an excellent candidate for project archive. IT professionals could also apply this method to HR records, end of quarter/end of year financials, and more.

In Media and Entertainment, raw content for every movie, sportscast, series, documentary and even commercial has tremendous amounts of data that might not make it to the final production, but has sufficient value to be archived in association with the final production. Again, a great use of StorCycle's project archive feature. And of course the final production content itself is also a great candidate for project archive.

In addition to allowing users to tag projects with additional metadata for future search and retrieval, StorCycle produces a manifest for each project archive which can be accessed as a file. The built-in manifest shows exactly what was moved, where the data originated, where it was moved to, and when it was moved. It can be displayed by clicking on the finished project archive and stored with other files in the project. It does not require a query into the database, and can be worked into existing workflows.

## Delete Data as well as Migrate Data

"Total control" of data includes not just what data to move, but also how long it should reside on any given storage medium and how long it should be kept over all. StorCycle offers administrators the ability to manually delete or schedule the deletion of data. This allows administrators to determine how long data lives in the storage environment – hence, "lifecycle." Scheduled delete offers a proactive, "set it and forget" approach while manual delete allows administrators to be responsive as organizational mandates, compliance requirements or new IT projects change.

In addition to controlling the lifecycle of on-premise storage, StorCycle is able to control the lifecycle of cloud storage, allowing for deletion in the cloud as well. Otherwise cloud storage is not functioning as part of the "lifecycle."

## Specialized Backup

StorCycle is not a backup application and is not positioned in that manner. Yet it can be used for protection of infrequently changing data where traditional backup solutions simply are not feasible or are cost prohibitive. One growing area in this category is organizations that have unrealistic amounts of rarely changing data being backed up by traditional means.

One StorCycle user example is a Spectra U.S. government customer who oversees and administers several thousand SQL database servers. Applying a backup license to each of these servers could literally run into tens of millions of dollars. The storage industry tends to be fixated on "unstructured" data. That is understandable since it is the greatest area of growth, but what about "structured" data such as these large databases? They tend not to be as concerned with incremental backup or directory/subdirectory backup, etc.

This particular user found that they could increase performance and decrease costs by having the SQL servers dump their database backups to disk files, to, in essence, create a point in time "snapshot." StorCycle then automatically moves these backups off to disk and tape in the background, and eventually automatically deletes them after a period of time when no longer needed.

StorCycle is also used to back up unique types of data such as primary object storage. Traditional file system backup solutions often do not make sense in these environments.

## Reduce Backup Load, Backup Windows and Backup Cost

Where traditional backup is required, our Spectra's software partners do an excellent job. StorCycle can be of assistance here as well. Unchanging, rarely accessed data is often left on primary storage because administrators have no way to identify it, move it and assure users can get it back. This cold data will be backed up in every full backup that occurs – sometimes for years.

Using StorCycle's scanning, migration and reporting abilities, cold data can be moved off of primary storage, significantly reducing the need for additional licenses (freeing up future budget) and often lowering ongoing support costs by decreasing existing licensing (also freeing up current budget).

Other benefits include shorter backup and restore windows and newly freed up space on tier-1 storage which also leads to greater performance and less need for adding additional primary storage at significant costs.

## Be Better Prepared for Ransomware Attack

It has become clear that an "air-gapped" copy of data is the only guarantee against losing data in a ransomware attack. This air-gap copy could also protects from other forms of data loss such as natural disaster or accidental deletion as well.

By migrating less active data to virus-proof secondary storage (such as in a tape library), the smaller primary data sets can be restored from traditional backup and replication applications more quickly. The smaller amount of data required to restore, the faster business operations can resume. Depending on the size of the organization, this could take ransomware recovery from months down to weeks, or from weeks down to days.

## Eliminate Storage Quotas

A constant challenge for virtually all IT professionals is managing the amounts of data users want to add to the organization's storage systems. There has been no better way to manage this than applying storage quotas. True Storage Lifecycle Management offers a different approach. It focuses on service levels (SLAs) rather that storage amounts.

Rather than tell employees there's a limit to how much they can store, or their data is being archived with no way for them to individually view it or retrieve it, we can now offer them infinite, accessible storage with varying levels of retrieval time – directly retrievable by the user. This is enabled by StorCycle's unique ability to use HTML links for retrieving data. HTML links were developed by Spectra specifically for high-latency storage mediums such as tape or "cold" cloud.

Users will still see any archived file via the file system. The icon will display slightly differently to indicate a migrated file. When the file is selected, an HTML file will display showing where the file resides (that could be cloud, disk or tape). A simple "click" will bring the file back.

Tape in this sense can be used as a repository of data which is inactive but not necessarily stagnant – i.e., more than backup or DR copies. In Windows, think of setting up an "X" drive for Marketing to store videos, campaigns, graphics, etc. with no storage quotas. Storage Lifecycle Management should be able to make a tape library act as that X drive. Just as a marketing employee might move data daily from the "L" or "M" drives (which do have a space limit), they could just as easily drop data on the X drive with virtually no storage quota. In the Mac and Linux world, the same could be applied via mounted shares. StorCycle's unique HTML links make this possible.

If StorCycle has been used to migrate data to lower cost disk, a "symbolic" link is left in place of the original file, and the data read will simply be redirected to that storage location. This offers complete transparent access as no restoration is required.

If information is being used to reach organizational goals, limiting the amount of information available is like the proverbial, "tail wagging the dog." With some storage types now costing fractions of a cent per GB, storage availability can now be based on organization objectives rather than storage costs.

## Eliminate Storage Silos

Stored data is useless if we can't find it. That is why it's so often left on primary storage. There has been no easier way to look for something than via the file system where the file was originally placed. Most migration software offers links that allow users to locate data on disk and cloud, but almost all of them struggle with the most economical levels of storage like tape or "cold" cloud, such as AWS Glacier. In addition to the HTML links listed above under "eliminating storage quotas," StorCycle allows administrators to eliminate storage "silos." This is accomplished by offering two methods of linking to migrated data. Symbolic links are used to direct the read request to lower cost NAS, the HTML links are used to retrieve the data from cloud, disk or tape.

By having a single migration software that can do both, there's a "single pane of glass" to view all data that's been moved, migrated or archived via StorCycle, regardless of the medium it is been moved to. Data is no longer "lost" to dark silos where users can't find or access it.

## Build a Massive (multi-petabyte) Indexed Storage System

Archives in virtually every industry in the world are becoming massive. More data creation, combined with longer, sometimes infinite, retention periods are creating storage systems of unprecedented size. The International War Museum in the UK recently converted old film footage to digital frames allowing Peter Jackson to create an award-winning documentary on WWI, "*They Shall Not Grow Old*." New technology such as AI has been applied to archived data at NASA, recently discovering new planets, possibly habitable by humans. New technology has been applied to seismic data in the oil industry allowing the U.S. to become a major oil exporter vs. importer. Medical research archives across the world have aided in the most rapid development of vaccines for a life-threatening virus that the world has ever seen. There has never been a greater demand for digital storage.

New storage technologies and approaches are also growing to meet this demand. StorCycle offers state-of-the-art technology that supports management of file systems, object storage, block storage, NAS, cloud and tape. StorCycle allows data centers of all sizes and industries to streamline these efforts at a fraction of the cost, in many cases completely replacing outdated or no longer needed software. All migrations are done in non-proprietary formats assuring future accessibility regardless of the solutions in place at that future date. Terabyte-scale archives can be stored on cloud or disk, and archives scaling into tens or hundreds of petabytes may be easily stored in automated tape libraries.

StorCycle can create multiple copies of any given file, object or migration set. Likewise, StorCycle can place and track files and objects on multiple tiers for various time periods – perhaps keeping new content for 90 days on NAS disk, keeping the full set on a tape library for the next year and always keeping a DR copy offsite, either on ejected tapes or in the cloud. All copies will be tracked, and upon request, StorCycle will restore a copy from the "fastest" location.

## Enable Tier-1 Storage Upgrades

As the cost of flash/NVMe/solid state storage has rapidly decreased, more and more organizations are considering a move to all-flash or hybrid storage solutions for tier-1 active data. This is problematic when tier-1 storage has been growing unchecked for years or even decades, clogged up with months or years of inactive data that must still be retained. StorCycle Storage Lifecycle Management software allows administrators to prune tier-1 storage of cold data, allowing only the most active data to be placed on a state-of-the-art primary storage system while still maintaining the cold data on a more appropriate storage level. This allows for a much more cost effective tier-1 upgrade.

As data is migrated down the tiers of storage, administrators may find that much of their expensive yet obsolete storage, such as RAID arrays, can actually be decommissioned – saving thousands in ongoing support costs. By approaching primary storage refresh as not only adding new technology, but also decommissioning old technology, budgetary planning becomes much more compelling for moving to faster, more reliable technologies.

## Control Ongoing Growth and Cost of Tier-1 Storage

Offloading tier-1 storage allows many of the benefits discussed in this paper. It is also important to think in terms of *continuing* to monitor and manage tier-1 storage utilization. StorCycle allows administrators to automate the process of continuously identifying and archiving cold data. Scan and migration operations can be rerun at any point in time to assure that the unnecessary buildup of cold data on tier-1 storage doesn't reoccur.

## Implement Archive Storage Charge-Back

StorCycle allows administrators to associate a department or even an individual with a storage scan and migration move. If you have a current charge-back program in place, departments or employees can be incentivized to move data to lower-cost tiers of storage. Even if there isn't a current charge-back policy in place for tier-1 storage, storage quotas can be augmented with low-cost archiving which can be much more cost effective than individual departments acquiring external drives or even thumb drives for additional storage once they meet or exceed a storage quota. In addition to cost savings, this would also allow for more reliable, secure storage of organizational information.

## Create Custom-Built Interfaces

Starting with release 3.3 (December 2020), StorCycle will "open" its internal API to allow users to programmatically scan, migrate, and restore data using their own custom-built applications and scripts for their own specific workflows. This allows user to integrate the power of StorCycle into virtually any desired workflow.

## Summary

These needs all fall into the category of Storage Lifecycle Management. StorCycle can assist in every one of the above needs, and the above list is not exhaustive. StorCycle isn't just a product in and-of-itself – It is a solution that makes what you have *better* and makes what you want to have *possible*. StorCycle *is* Storage Lifecycle Management.

# Part III: Tenets of Storage Lifecycle Management

For a deeper understanding of Storage Lifecycle Management, Spectra has created a list of tenets required to successfully implement a Storage Lifecycle Management solution.

A "tenet" is a principle or belief. The main principles of a philosophy are described as the tenets of that philosophy. This can be an especially useful tool to understand, describe and differentiate product solutions as well as philosophies, especially when solutions are created to carry out a philosophy – such as Spectra's StorCycle and Storage Lifecycle Management.

Storage Lifecycle Management (SLM) requires moving data across multiple storage mediums. What if a bit is flipped in the movement process? What if the data or information needs to reside in multiple places for varying lengths of time? What if the application which created the data is not the application that's moving it - how will it be retrieved? What about formats? Will they change? What if the "data mover" used to accomplish SLM is decommissioned? How will data be retrieved? These are a mere subset of the myriad of questions and challenges storage professionals must contend with.

## The ten tenets of Storage Lifecycle Management are the founding principles upon which successful Storage Lifecycle Management can be implemented.

### 1. Allow for automated or manual direction and control of data

SLM must allow for proactive (automated) workflows as well as responsive (manual) actions. Moving data based on age and access patterns falls into the "proactive" category – this is only a portion of true SLM. Unexpected or "one-off" events happen all the time – an employee leaves and data must be archived, an experiment is completed and newly-captured data must be moved, etc. These are events which require a more responsive or manual approach.

Examples may include "delete on demand" vs. "scheduled delete," "scan for data to move" versus "drag and drop," "age/access-based move" versus "project archive."

"Total control" includes all mediums including cloud and tape.

### 2. Move data without changing the format

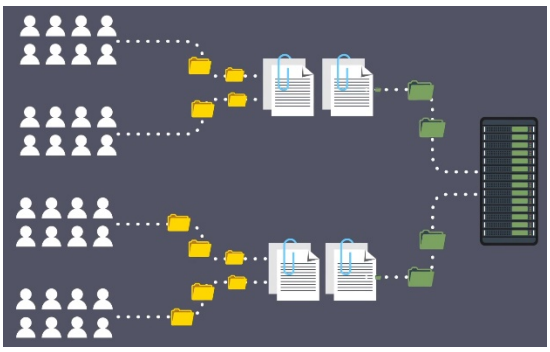
Traditional data movers are notorious for changing the format of data they move. This allows the data mover more flexibility, but it comes at the cost of vendor lock-in to that data mover solution.

Small files are often "packed" to make data movement more efficient. SLM must be able to perform such functions, but allow individual files, folders or directories to be restored individually if required. In some instances (commonly in M&E), even partial file recoveries must be supported.

Deduplication works well as a part of data backup. It creates problems in SLM. If the lifecycle of the data requires moves across multiple mediums, individual file deletes, or access by multiple applications, deduplication cannot qualify as a viable means of data management throughout the entire life of the data/information/storage.

Non-proprietary formats are mandated in SLM. Often, retention dates span years, decades, or even lifetimes. Not all storage solutions will last as long. For this reason alone, proprietary formats are unacceptable in SLM.

If data is stored on tape or disk, there must be a process for migrating data to newer generations of storage. SLM should be capable of performing this in the background as part of SLM, not requiring an additional application, or manual recall and archiving all over again.



### 3. Scalable to hundreds of billions of objects or files

All too often, "scalability" refers only to the amount of storage growth an individual storage solution can support. While that is an important factor (Spectra manufactures the only uncompressed Exabyte scale tape library in the world), the database which tracks petabyte-sized data repositories must also be able to scale. This becomes even more important as the database will oversee the storage of multiple storage systems.

The SLM database must take a centralized approach for ease of access and visibility meaning it must be able to scale to billions of objects or files. In order to manage a repository of this many objects/files, the database must also allow for the addition of metadata.

### 4. Calculate checksum on reads as well as writes

SLM requires that data be verified after all moves. This requires checksums to be calculated on data migrations as well as data recalls. SLM should calculate the checksum of each entire file as it is written to its target storage. A checksum acts as a thumbprint of the file that can be reproduced to ensure that the file has not changed. Each file's checksum must be stored in the SLM database, and this value must be accessible and exported as part of a migration job's manifest.

When a moved file is recalled, the database must be able to validate the checksum to ensure that the file has not changed. If a change in checksum value is detected, the SLM database should report an error.

Part of the role of SLM is Digital Preservation. If one copy of a file suffers from corruption (which would be identified when reading that file), the SLM system should enable the user to restore a different copy of that file, perhaps stored on a different media.

### 5. Create Manifests

Previously we mentioned that each file's checksum should be exported as a part of a migration job's manifest. The manifest shows exactly what was moved, where the data originated, where it was moved to, and when it was moved. It can be digitally displayed by clicking on the finished project archive and

stored with other files in the project. It does not require a query into the database, and can be worked into existing workflows. This manifest may be stored independently and offers another resource for tracking, managing and retrieving data. SLM must be able to create multiple paths to stored data. By producing such a manifest, both manual and automated processes are available.

### 6. User definable access

All of the above tenets in some way contribute to making sure data is available when it needs to be. This involves setting appropriate SLAs, readable formats, scalability without error or loss, assured viability of the data and assured tracking of the data. This is all done so that SLM's most important role can be achieved: Retrieving data.

Older HSM solutions often have no way of informing users that data has been moved. In some instances, this results in workflow conflicts, accidental retrieval of huge amounts of data or simply too much "automation."

Other modern approaches often move data to high latency mediums like cloud or tape and then "time out" when the data can't be retrieved quickly enough – sending an error message rather than a description of where the data is and how it can successfully be retrieved.

SLM for these reasons has met with varying amounts of success and user adoption. None of the effort of using SLM to migrate data is justified if the migrated data can't be accessed in a simple and reasonable manner.

It's imperative that SLM be able to offer access to users and administrators in a straightforward manner regardless of where the data resides. Likewise, system administrators and "data owners" must be able to extend or limit these access methods according to their organization requirements.

SLM must be able to provide the appropriate link for the given medium information resides on. The restore and access process must be appropriate for the expected restore time. The process must clearly communicate to the user when they can expect to have access to their files It should allow for end user restoration of single files/objects, directories, or entire projects. And most importantly, administrators must be able to select or mix and match these approaches as well as open or limit how much end users can see or access.

### 7. Attach Metadata

Files and objects tracked by SLM should not be modified. However, there Is often additional metadata that needs to associated with that file. A viable SLM system must allow for additional metadata in the SLM database, such that searches may be initiated to these "tagged" assets. As a digital asset may in actuality be composed of multiple files or objects, that grouping of objects must be preserved, and be restorable as a group.

### 8. Preserve Access & Security Rights

Files migrated or archived by an SLM system must preserve access restrictions, to prevent unauthorized users from accessing sensitive or protected information.

### 9. Self-Protection

SLM systems must protect themselves by automatically backing up associated databases and other key components. A scalable database is at the core of an SLM system. The system must protect itself by making frequent backup copies of its database which can easily be identified and restored.

### 10. Offer Bandwidth and Insert Performance to Keep up with Near-real-time Applications

SLM systems may be used for identifying, tagging and archiving data that is created at a significant rate of performance. An example could be archiving imagery data associate with incoming video camera cards or being generated by a high speed Cryo-Electron Microscope (common in biopharma research labs, universities and facilities). A successful SLM implementation must be built to migrate and restore up to several hundred terabytes per day. Likewise, it must be capable of adding tracking information into its database for tens of millions of files per day.

***

The above tenets are all required for a successful approach to Storage Lifecycle Management and are based on Spectra's 40+ years of experience in the storage industry. With a full understanding of how Storage Lifecycle Management can be applied to an organization's existing storage as well as future storage, digital storage becomes a true enabler of organizational goals rather than limiting them.

# About Spectra Logic Corporation

Spectra Logic develops data storage and data management solutions that solve the problem of long-term digital preservation for organizations dealing with exponential data growth. Dedicated solely to storage innovation for over 40 years, Spectra Logic's uncompromising product and customer focus is proven by the adoption of its solutions by leaders in multiple industries globally. Spectra enables affordable, multi-decade data storage and access by creating new methods of managing information in all forms of storage — including archive, backup, cold storage, private cloud and public cloud.

To learn more, visit www.spectraLogic.com.