WHITEPAPER

Universal Unstructured Data Orchestration

www.datadynamicsinc.com

Business Challenge

Most enterprise data remains 'at rest'. Is that bad thing? Well that depends on what that data is. Most companies only know their data at a high level. An example of this is they have finance database or the marketing file share. We know what should be in there. Can we take advantage of what's in there? Databases tend to be highly managed and well organized, but your NAS (network attached storage) is probably not in that same shape. In fact, because 90% of your employees are not "IT" related, they have no constant rhyme or reason for how their files and data are stored.

CDO's and the business units they work with have a constant challenge of evaluation and taking action on that evaluation. As business changes and new needs are brought in, the evaluation and landscape of data zones will need to be reconsidered. The ability to work with Data Stewards and have that translated to the Data Specialists is key. The abilities of an organization to manage through the constant change are what set them apart.

- Gartner recently stated that without a plan for the NAS infrastructure, your Data management is likely to fail.
- With NAS data growth in 50% range, the cost needs to be managed. The recent evaluation of $1100 a TB of storage cost, means a considerable impact can be made by taking advantage of the data and turning it into and business asset.
- Many companies have lifecycle management in place, but that's for moving old data. The usual idea is to move the data, not eliminate it. The only evaluation criteria is access time.
- Evaluating risk is key. Sometimes the exposure of data can cost more than the cost of the storage. Just recently…
  - Marriot London sued, could cost 1.7B euros
  - Capital One fined $80M.

Taking steps to orchestrate your data can not only save you money but by putting data where is can be most useful can help give you a business advantage.

Storagex by Data Dynamics is designed to scale to the largest Enterprise and provide value Data Orchestration across the entire NAS environment. With its history of data movement, combined with its visionary abilities to analyze at both the file and file content level, the StorageX family of products can enable a successful Data Orchestration lifecycle.

The Basics of Orchestration

The Process

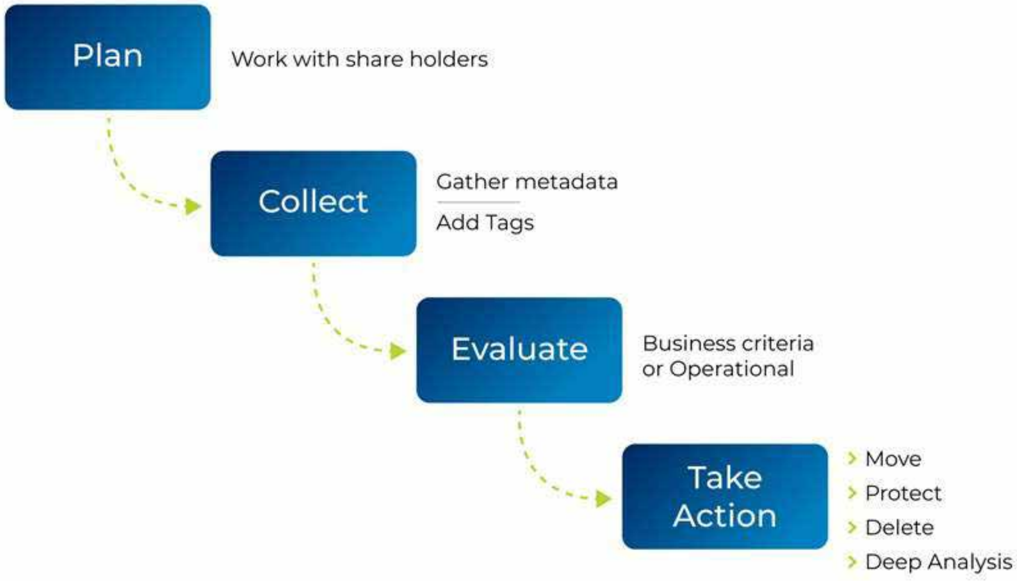


Figure 1: Data Orchestration Process

Planning

Getting clear direction from the Steering Committee is important. Working with the Data Stewards will help translate Policy into standards for search and analysis.

Working with the business units and stakeholders you identify target data to do analysis scans on.

A good starting place can be running the "Open Shares Report". This identifies shares that are open to everyone and targets risk immediately.

Planning the evaluation criteria should be discussed during this phase. Understanding the business goals and data policies is key. What data is important, are there metrics such as data that's 10 years old should be deleted. There could be several simultaneous criteria to collect and evaluate.

- Example 1: Files that haven't been accessed in 3 years AND modified more than 5 years ago
- Example 2: Files that have "invoice" in their name and were created more than 18 months ago
- Example 3: Any file in an open share

As part of your planning, develop a set of tags that will help you identify data by means other than just file name. Tags can be created to make business, IT, or project-based classifications during the evaluation phase.

These tagging examples are ideas of how you can use tagging.

Tag	Value	Alternate1	Alternate2
Organization	Finance		
Sub-Org	AP	Payroll	AR
DataCenter	NYC	College Rd	
Project	Purge	11-20-21	CTO-45-2020
Location	4-114	Row 32 cab 2	Room 5

In StorageX you assign one tag per value. The concept is two-fold. First, your analysis can be helped by grouping and sub-grouping your searches and dataset creation. For instance, if Finance is spread across multiple datacenters (NYC, WDC, and Dallas). While you have Finance at the top, AP, Payroll, and AR all included as Sub-Org's. An example of tags might be:

DataCenter	Share(s)	Tags

NYC	\\xntap32\finance
	\\xntap12\ap_share
		DataCenter=NYC
		Org=Finance
		Sub-Org=AP

WDC	\\Isilon02\finusers
	\\Isilon02\ap_dump
		DataCenter=WDC
		Org=Finance
		Sub-Org=AP

WDC	\\Isilon02\ar_dump	DataCenter=WDC
		Org=Finance
		Sub-Org=AR

Dallas	\\server4\Payroll	DataCenter=Dallas
		Org=Finance
		Sub-Org=Payroll

You can setup datasets like so:

Search	Result
IF Organization = Finance	Gives you all of the data
IF Sub-Org = AP AND DataCenter = NYC OR DataCenter = WDC	AP data located in NYC or WDC

The second part of the value is when you archive. These tags show up as HTTP tags if you archive the files to an S3 target. This gives you yet another data point on where the files came from.

Collect

StorageX can generate reports that will give you lists of shares or exports. These can be used project manage the entire process.

- Create policies with appropriate tags for targeted shares.
- This can be done by the StorageX Administrator Browser or via Rest API
- You can choose multiple shares or exports
- Schedule or execute Policies

StorageX generates a Discovery policy per share or export. This allows simultaneous discovery and increases over all velocity of the discovery.

Each discovery will bring in a full complement of the metadata on each file. This is much more than you can see in a directory. An example of this is the SID string of the system. To most users this is invisible.

Each file is associated with the Tags you placed in the Discovery policy

StorageX has been shown to gather 10M files metadata per hour in a production deployment.

Evaluate

StorageX provides a plethora of ways to evaluate the data you've gathered.

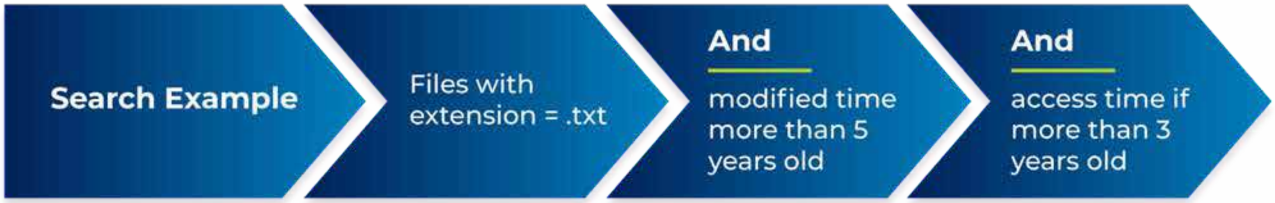
Methods for evaluation

Method	Granularity	Data Orchestration
Meta Data Analysis	File	Individual files can be archived and removed. Analysis set can be used in the Privacy Scan to identify PII
Meta Data Analysis	Share	Entire shares can be identified via Metadata and migrated
Migration Archive	Folder	Automatically creates a migration policy targeted the folders identified
Privacy Scan	File	Remediation includes archive to object, in place encryption with an immutable chain of custody
Open Share Report	Share	Identifies shares open to everyone. Orchestration happens via migration and/or archive
Duplicate File Report	File	Works across volumes within the same dataset

Methods for evaluation

Meta Data Analysis by File - StorageX gathers not only the meta data (see appendix for complete list) for each file, but also information at the file server level. Much of the data is invisible to the average user. This includes things like the Security Identifier for the server itself. The advanced search features in StorageX gives the Data Specialist capabilities to search for any combination of data. It also gives Data Stewards the ability to create specific search and evaluation criteria. The combination of the Meta Data combined with tagging gives the Data Specialist the power to create search criteria. A standard search can be defined and reused. The example below could be given a common name and then called up when needed. This supports a data categorization standard being developed

Data Governance	✓
Data Orchestration	✓
Data Analytics	✓
Optimization	✓
Data Security	✓



Search Example

Files with extension = .txt

And

modified time more than 5 years old

And

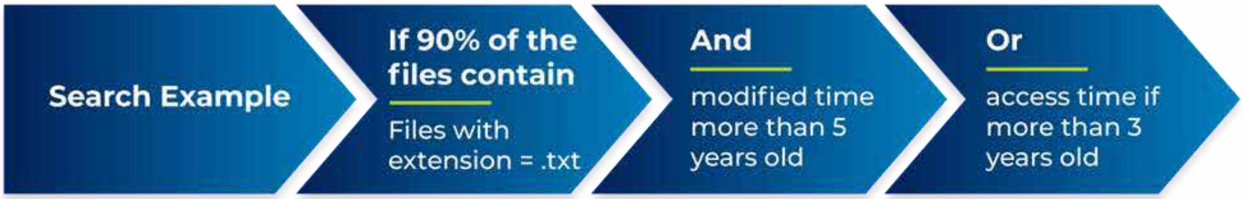
access time if more than 3 years old

Why it matters			
CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Gives empirical data on where you stand in terms of compliance	Once defined, the search criteria is easy to pass and have executed.	Easy to implement, once created. Repeatable and can be changed without penalty	Optimization of resources and understanding what's important

Methods for evaluation

Evaluation by entire Share or Export uses the same meta data as criteria, but the entire share or export is evaluated.   This evaluation allows the user to pick the percentage or number of files that meet the meta data chosen. You can also choose by number of files.   This feature allows you to assess the information before you take action to migrate the data to a different tier of data.

Data Governance	✓
Data Orchestration	✓
Data Analytics	✓
Optimization	✓
Data Security	



Search Example

If 90% of the files contain

Files with extension = .txt

And

modified time more than 5 years old

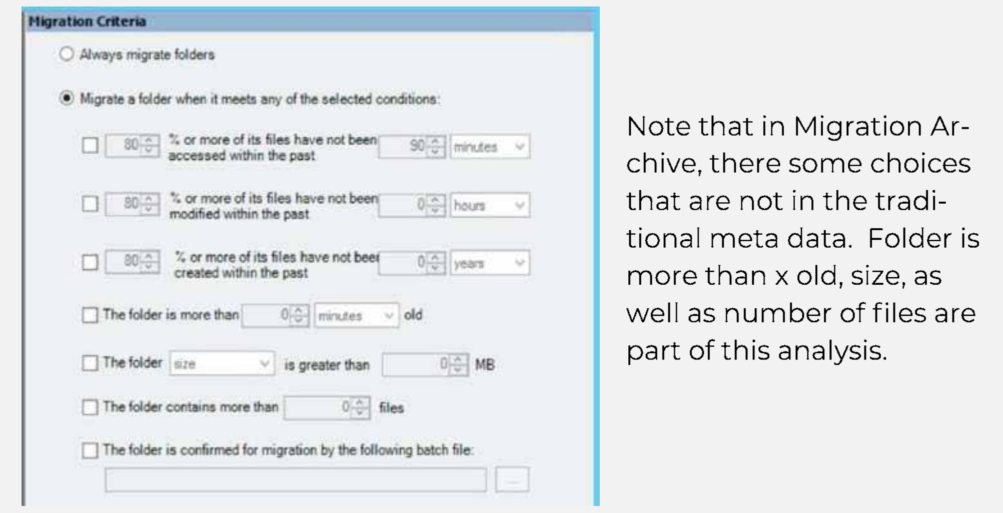
Or

access time if more than 3 years old

Why it matters			
CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Gives empirical data on where you stand in terms of compliance	Once defined, the search criteria is easy to pass and have executed.	Easy to implement, once created. Repeatable and can be changed without penalty	Optimization of resources and understanding what's important

Migration Archive

This feature allows you to evaluation down to the folder level. These are based on a more restricted list shown below.



Migration Criteria

○ Always migrate folders

● Migrate a folder when it meets any of the selected conditions:

- □ 80 % or more of its files have not been accessed within the past 90 minutes
- □ 80 % or more of its files have not been modified within the past 0 hours
- □ 80 % or more of its files have not beer created within the past 0 years
- □ The folder is more than 0 minutes old
- □ The folder size is greater than 0 MB
- □ The folder contains more than 0 files
- □ The folder is confirmed for migration by the following batch file:

Note that in Migration Archive, there some choices that are not in the traditional meta data. Folder is more than x old, size, as well as number of files are part of this analysis.

Data Governance	✓
Data Orchestration	✓
Data Analytics	✓
Optimization	✓
Data Security	

This feature allows you to assess the folders and either present results or assess and create migration policies to actually migrate the data. The migration policy is also capable of running batch files that can perform specific operations that can help with the data movement or even leave a file with the results of the movement policy.

Why it matters

CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Gives empirical data on where you stand in terms of compliance	Once defined, the search criteria is easy to pass and have executed.	Easy to implement, once created. Repeatable and can be changed without penalty	Optimization of resources and understanding what's important

Privacy Risk Scan

This scan works by opening files and looking for regulated information, such as PII, GDPR, HIPPA, or CCPA related. The Privacy risk scan gets its dataset from the StorageX File Meta Data scan. That way, you can efficiently identify only the files that are important to your business. Examples of this might be files with the word "Invoice" in their name or just looking through JPG's (via OCI) to find drivers licenses. There are 49 entities that can be identified.

Data Governance	✓
Data Orchestration	✓
Data Analytics	✓
Optimization	✓
Data Security	✓

- Full Names
- Vehicle Registration
- IP Address
- Credit Card Number
- Date of Birth
- Email Address
- Driver License Number
- Social Security Number
- Telephone Number
- Genetic Information
- Physical Factors
- Banking Information
- Birthplace
- Country
- Occupation
- Fingerprints
- Digital Identities
- Gender
- Blood Type
- Address
- Emergency Contact
- Facial Recognition
- Health Related Data
- Zip Code
- Economic Factors
- Cultural Factors
- Tools
- IOT Identifiers
- Cookie Identifiers
- Geo Location
- Applications
- Social Identities
- RFID Tags
- Salary
- Device
- Health Insurance Data
- Medical History
- Race
- Immunization Dates
- Vital Signs
- Diagnoses
- Allergies
- Ethnicity
- Lab Tests Resus
- Major Diagnoses
- Principal Language
- Radiology Images
- Medications
- Progress Notes

This is both an 'Execute' and 'Collect', because you've identified in a previous discovery via Meta Data File assessment and Collect,' as it gathers and categorize even more information about the targeted file systems. There is further action that you can take with remediation of the data.

Why it matters			
CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Gives empirical data on where you stand in terms of compliance and risk of exposure	Pre-defined regulations and configurable reporting criteria can be fitted to a variety of business needs	Reporting is easy to understand and in the remediation phase they have support via approvals	Huge Risk reduction on exposed data not known before.

Open Share Report

The open share report examines the protections on CIFS shares to provide administrators with possible security holes. You can use this report to "dial-in" to places to use Meta Data File discovery or the Privacy Scan to immediately assess risk and exposure.

Data Governance	✓
Data Orchestration	
Data Analytics	
Optimization	✓
Data Security	✓

Why it matters			
CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Gives empirical data on where you stand in terms of compliance and risk of exposure	Better understanding of risk and the ability to re-check as needed	Ability to run and report as needed.	Huge Risk reduction on exposed data not known before.

Duplicate File Report

Data Orchestration	
Data Analytics	
Optimization	✓
Data Security	

After running a discovery scan with md5 hash enabled, StorageX compares the hash and the file name to find duplicate files. This can happen across several shares, exports or even volumes. Duplicate data affects companies because it not only represents a waste of space, but data duplicated could be used for purposes outside of its initial business use.

Why it matters			
CDO's/Steering	Data Stewards	Data Specialists	IT Teams
Better understanding of data efficiency can give future direction on management, cleansing, and reduction.	The empirical data can show either the need for better management or a better understanding of data flow.	Give's empirical data to be able to create management plans with Data Stewards	Optimization and rationalizing of NAS footprint and budget.

Take Action

Execution on the strategy defined in planning is not the end of the process, but only the next leg until you go back to Collect and Evaluate. This is will be a cycle that is constantly examined and tweaked as you learn to understand the data you encounter. Taking action happens in variety of ways.

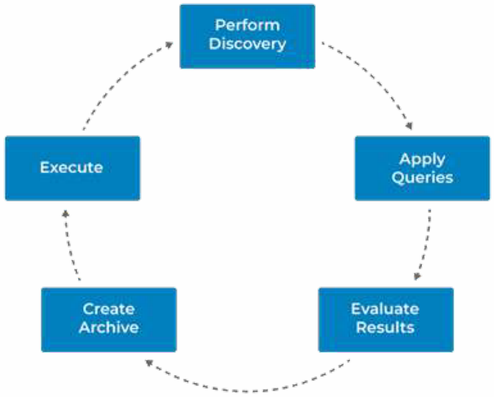
Action to Assessment Matchup

Action	Eval Source	Data Orchestration
Archive	Meta Data Analysis - File	Individual files can be archived to object storage and removed
Privacy Scan	Meta Data Analysis – File Or Open Share report	Data sets are generated from Meta Data Analysis.
ControlX Remediation	Privacy Scan	Privacy scan generates lists of files at risk to remediate via movement to object storage or in place encryption
Migration Policy	Migration Archive Policy	Migration Archive Policy generates the appropriate Source/Target for a migration
Migration Policy	Business Need	Includes new storage allocation, introduction of Cloud storage. These policies can be generated from the console, imported via CSV, and Rest API.
Replication Policy (NAS to NAS)	Business Need	D/R, backup, Data Cleansing, Distribution, or other Business-oriented need. Generated via the Console
Replication Policy (NAS to Object) F2O	Business Need	D/R, backup, alternate processing. Governance needs such as Data Isolation could also be considered.
Workload Tiering	Meta Data Analysis – Share/ export	Generates Migration Policies to move the share or export to the appropriate storage tier

With all the options, the workflows that result should become part of the operations manual and integrated into the Orchestration specific documentation.

The workflows for each of these allows users to dictate the outcome they'd like. For instance, if you're evaluating based on file activity, individual files can be moved. If you're evaluation is on folders, then entire folders will be moved.

Archive Workflow



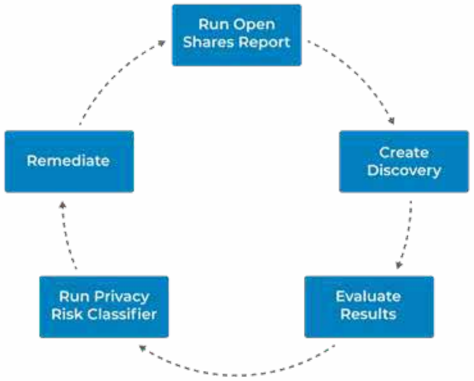
Note: Queries are directly applicable to archive policies. Choosing file modified date older than 5 years

Figure 1 - Archive Workflow

As each cycle finishes in the Archive workflow, the files that identified, are moved to the target object storage, and removed from the source system. This allows IT to reclaim space, lessen backup time, and reduce costs. Each archived file is written with a unique ID. This guarantees there will no two objects with the same name. Additionally, StorageX writes a companion object. This is specially created at a JSON formatted mini database of every piece of metadata. At the very least, this companion object can tell, without a doubt, exactly where and when the primary object was created, as well as who did it.

Possible use case for the companion object include applying Machine Learning, AI, and custom queries. With these options finding long term patterns for any piece of meta data is possible. This can lead to predictive analysis and proactive remediation.

Privacy Risk Workflow



Note: Open shares represent a security risk. Evaluating the contents of the files helps understand the risk

Figure 2- Privacy Risk Classifier Workflow

In the case where files are found to contain PII (or any regulated information) identification is the first step. Options for remediation can include isolating the data into an Object storage, encrypting the file in place or working with IT to fully secure the share.

Control-X Remediation

This is one of the critical orchestration events. At this point, the Insight AnalytiX, Privacy Risk scan has identified files with PII or Regulated information in them. Your corporate process defined by the Data Stewards will dictate the output of the remediation. Approvals depend on process defined by this same set of standards. Insight AnalytiX software allows you to facilitate this workflow. Note the actions will continually be updated by Data Dynamics.

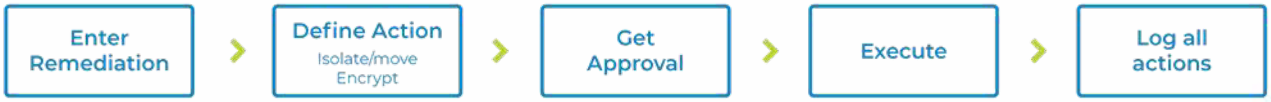


Figure 3 - Remediation Workflow

Migration Policy, Archive

The archive template in Migration, has the ability to generate policies that rationalize folders against a given set of meta data as explained in Evaluate.   While you can choose to automatically

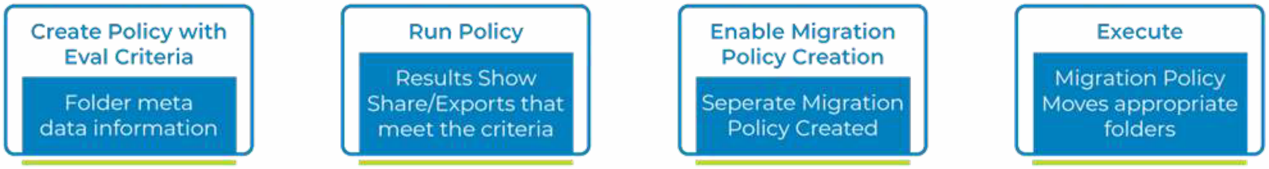

Create Policy with Eval Criteria
Folder meta data information
Run Policy
Results Show Share/Exports that meet the criteria
Enable Migration Policy Creation
Seperate Migration Policy Created
Execute
Migration Policy Moves appropriate folders

Figure 4 - Migration Policy, Archive Workflow

generate the migration policy, the workflow below breaks it out to assessment and the action of performing the migration.

Migration Policy, Business Need

The business need behind migrating data can be as simple as a new hardware platform. Using StorageX not only mitigates risk, but also reduces cost through the efficiency automation and having a centralize platform. StorageX's migration is capable of facilitating mergers, acquisitions and even divestitures with its advanced Security Identifier management capabilities. The ability to do one to many or many to one type of migration gives the business units the capability to re-org their data as they move to a new structure.

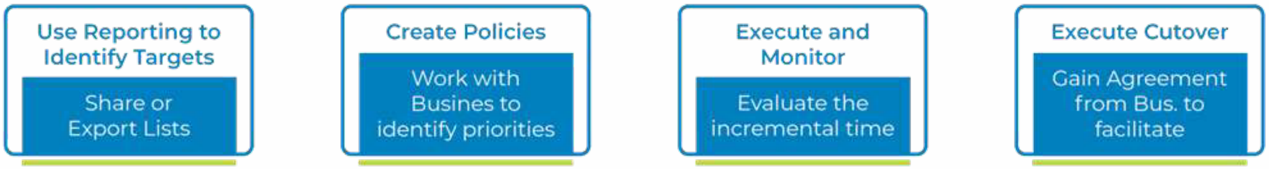

Use Reporting to Identify Targets
Share or Export Lists
Create Policies
Work with Busines to identify priorities
Execute and Monitor
Evaluate the incremental time
Execute Cutover
Gain Agreement from Bus. to facilitate

Figure 5- Migration Policy, Business Need

Replication Policy (NAS to NAS), Business Need

The StorageX replication policy can be configured to run on the scheduled needed down to minutes. Furthermore, it can be configured one to many for data distribution. There are any number of use cases, including D/R in a heterogeneous environment. This could be also be used for a secondary copy, either local or remote. Replication in CIFS can be pointed at sub directories to protect data in specific use cases. This can be used for data testing, data cleansing, or out of band deep inspection. The workflow for this is similar to the Migration Policy, but there no cutover. If used for a D/R situation, recovery should e included in the planning.

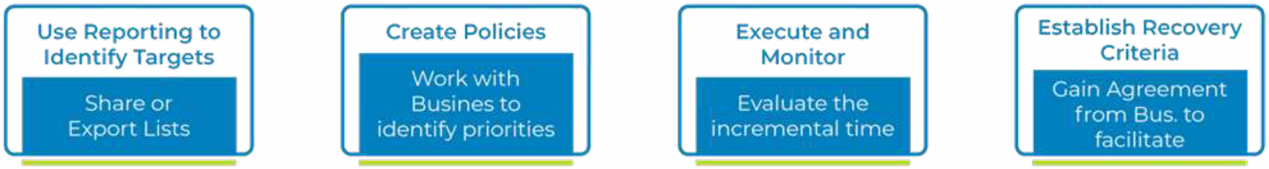

Use Reporting to Identify Targets
Share or Export Lists
Create Policies
Work with Busines to identify priorities
Execute and Monitor
Evaluate the incremental time
Establish Recovery Criteria
Gain Agreement from Bus. to facilitate

Figure 6 - Replication, Business Need

Replication Policy (NAS to Object – F2O), Business Need

AS to object replication, simply replicates file data and its structure to an S3/object store. By using 'path as key', the file folder structure can be visualized using an S3 browser.   While it is possible to evaluate the NAS source via discovery and meta data analytics, most use cases would not need those steps. The business decision to place entire information stores is usually because there is a need for S3 protocol access from numerous machines. The ability replicate via schedule and create multiple, simultaneous replications expands to the use cases to a variety of applications including genomics, security analysis, and many others.

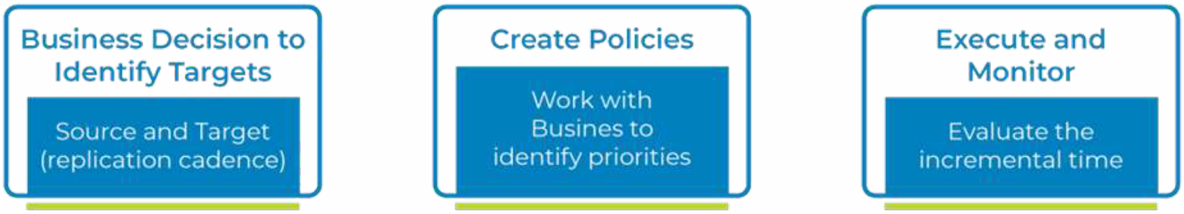

Business Decision to Identify Targets
Source and Target (replication cadence)
Create Policies
Work with Busines to identify priorities
Execute and Monitor
Evaluate the incremental time

Figure 7 - Replication, File to Object

Workload Tiering Shares/Exports

The concept behind workload tiering is simply putting files on the cheapest storage where possible. Of course, it can work the other way and find highly used files and putting them on the best server for that purpose. Bringing the Business and Data Stewards in to define the evaluation is necessary to make sure that the business has the optimal response to the files and data it needs to complete its tasks.

this workflow, entire shares or exports are evaluated on the targeted meta data. The decision is based on either the percentage of the total files or by the number of files that meet the criteria. Once the targets are identified, the user is given the opportunity to create the migration policy to move the data to a new tier.

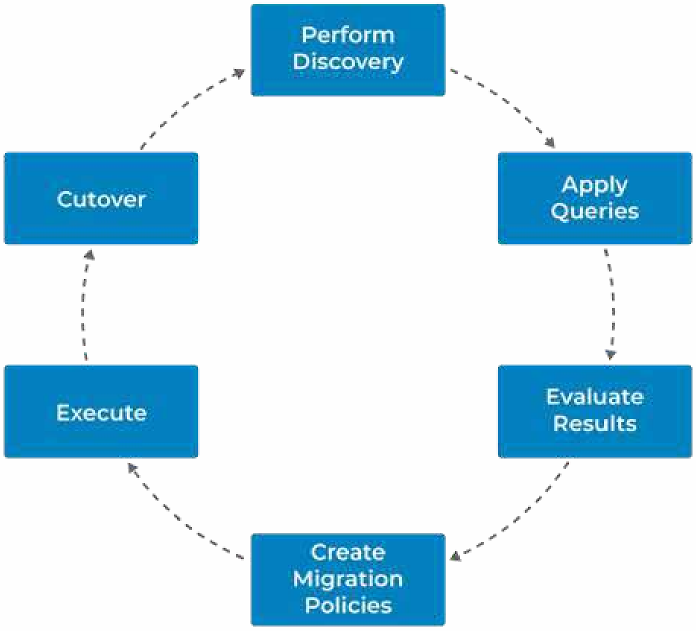


Figure 8 - Workload Tiering

Appendix 1- Metadata

SMB Metadata

File name Path to directory containing the file Machine Name Sharename File Extension File Type Creation/modified/last access/change Time Byte counts – file size, alternate data stream size(s), total	File Attributes Read only Hidden System Archive Normal Temporary Sparse Reparse Point Compressed Offline Encrypted	Security information Owner

NFS Metadata

File name	Size	Group Permission Bit settings
Export	UID	World Permission Bit settings
Path to directory containing the file	GID	Stickybit
	Owner Permission Bit settings	
Modified/Last Access time		

System Metadata

Hosting Resource Name

Benefits of a StorageX Implementation to Data Governance

Benefit	Description
Analytical Driven Process	From the CDO to the Data Stewards to IT, analysis of meta data yields empirical data that allows organizations to decide to move (or not) data or even delete it.
Scalability	A software solution that utilizes a scale out architecture that allows you to run and rerun data discovery. The speed of analysis allows organizations to play "what if" without a time penalty.
Performance	The solution also provides the ability to scale up depending on what performance is required to meet SLA's to support business requirements.
Use Case Coverage	Broad use case coverage makes the StorageX useful across all business units and provides value to CDO organizations by being able to target data and files based on real, current information. The analysis becomes actionable and executes as a synchronous process.
Integration	StorageX provides a robust set of APIs to drive integration with key orchestration and reporting tools that are a must for enterprise deployments. These integrations deliver a true factory functionality by automating many of the tasks involved to drive efficiency and reduce human errors.
Data Mobility	With several methods of data movement, all designed to be efficient and scale to enterprise performance, StorageX allow moving massive amounts of data. StorageX provides the right infrastructure to move that data intelligently either locally or to/from the cloud.
Licensing	StorageX subscription-based capacity license is designed to support as large (or small) an infrastructure you need to get the job done. This provides true cost basis that does not limit your ability to scale up or down

Summary

Evaluator Group Survey, "Trends in Multi-Cloud Data Management", November 2019

Data Dynamics, a global leader in enterprise data management, stands at the forefront of the industry-wide shift towards Digital Trust & Data Democracy. Trusted by 300+ organizations, including 25% of the Fortune 20, the company is recognized for its commitment to creating a transparent, unified, and empowered data ecosystem. Whether addressing data risk, privacy, sovereignty, optimization, sustainability, or facilitating seamless, policy-driven data migration across hybrid and multi-cloud environments, the company is ushering in a new era where data ownership, control, & actionability reside with the data owners.

DATA DYNAMICS®

Contact Sales | Book a Demo