BACKUP? ARCHIVE? OR BOTH?: WHY YOUR DATA BACKUP STRATEGY IS NOT A DATA ARCHIVE STRATEGY

This paper is written to assist the C-level executive and the IT manager. It distinguishes data backup from data archiving, and it outlines strategic data management practices for using both data protection methods.

INTRODUCTION

It's widely proposed that 90 percent of the world's data has been created in the last two years. Considering that the information onslaught is in no way slowing down, many corporate systems are approaching overload. This overload tightens the budget, but the service expectations keep rising, and the compliance mandates keep getting stricter.

Most companies rely on data management strategies to ease the pain. Many strategies focus on data backup as the mechanism to preserve digital data assets. However, for corporations that have mandated data retention rules, backup alone won't be sufficient. Regulatory compliance may require data archiving in addition to data backup. Data archiving provides specific features that ensure full digital data preservation.

This paper will review data backup and data archiving, highlight why they are different, and discuss where each is necessary. A complementary backup and archiving strategy helps ensure regulatory compliance while at the same time optimizing data backup and storage budgets.

DATA BACKUP

Let's start out with a definition for data backup.

WHAT IT IS

> Data backup, or the process of backing up, refers to the copying of computer data so that it may be used to restore the original after a data-loss event. Backups have two distinct purposes:[1]

1. To recover the most recent data after its loss, be it by data deletion, corruption or destruction.
2. To recover data from an earlier time, if current data is not valid.

Data backup focuses on preserving company data for recovery from loss.

WHY DO WE DO IT

Data backup is driven by internal company needs. Every company, regardless of size, needs a reliable backup solution, and this solution is typically part of a broader disaster recovery plan. The value provided by good disaster recovery planning becomes apparent when disaster strikes. In this situation, quality disaster preparedness directly affects the ongoing viability of a company. Backups are a critical component.

WHAT IT LOOKS LIKE

A data backup system is a storage infrastructure that is completely independent of any primary storage you already have. It may take the form of spinning disk (SAN/NAS/DAS) or it may take the form of robotic tape libraries. Smaller companies have additional options. Data backup is best achieved by a combination of on- and off-site storage. The off-site piece is commonly provided through a private or public cloud solution.

In general, a quality backup solution provides you with:

- An integrated system tailored to your specific business infrastructure and operational requirements.
- Protection for data stored on both physical and virtual environments.
- An off-site component that keeps your valuable data secure:
    - o On a separate server in a separate location over 100 miles away.
    - o In the cloud.
- Multiple layers of security, including strong encryption.
- Manipulation of the data being backed up to improve backup /restore speeds, data security, media usage and network bandwidth requirements.
- A detailed revision history that identifies where key data resides and from which platform it was generated.
- A timed, automated process that ensures backups are made regularly and to reliable standards.
- Third-party monitoring and reporting to ensure the integrity of the process and the data.
- A comfortable balance of capacity, accessibility, security and cost.
- A practical business continuity plan that ensures your ability to operate in the event of a data disruption.

DATA ARCHIVING

WHAT IT IS

Data archiving (digital data preservation) is a formal endeavor to ensure that digital information of continuing value remains accessible, searchable and usable.

It combines policies, strategies and actions to ensure access to content, regardless of the challenges of media failure and technological change. [2]

The goal of data archiving is the **accurate** rendering of **authenticated** content over time. [3,4]

→ Data archiving focuses on preserving non-changing content that might not be essential to the daily operation of the business, but may be required for historical, compliance, legal or other reasons.

WHY DO WE DO IT

Data archiving is often driven by external factors (e.g. legal compliance). In most cases, archiving does not affect the day-to-day operations of a company. Instead, it is a required overhead that is performed by mandate or regulatory requirement. However, storage-heavy companies often benefit from lower costs of overall storage when they adopt a good archiving system (e.g. medical PACS data).

While it may have little impact on day-to-day operations, a good archive's value comes if and when there is an event that requires the company to reproduce data in the archive. Correctly satisfying a legal discovery or compliance audit can save a company time and money.

ARCHIVING IS DIFFERENT THAN BACKUP

We have defined data backup, why we back up, and what a good data backup system looks like. We have defined data archiving and why we archive. Now, before we examine what a good archiving system looks like, let's briefly discuss why archive storage is different than backup storage.

When archiving is mandated by external or regulatory groups it is useful to understand what those groups may require from the solution. For example, a legal or regulatory event may trigger a data discovery process that requires the production of expired information in its native state. A compliance mandate may require retention of records in a format that is unalterable. Your team must be able to find, reproduce and provably verify data content over some defined length of time. Depending on the nature of the query, you may also be expected to be able to preserve data beyond its intended expiration date (such as for legal hold). For some liability issues, you may be required to provide records of who accessed specific data and when. Finally, because archived data may need to be accessed very far in the future, there are issues that must be addressed that are not typically part of the backup landscape.

Archive systems keep track of a lot of information that describes the data that they store. This information is commonly known as meta-information (information about the information). Specifically:

Serialization

A good archiving system will provide serialization of data. This is the notation that every file gets a unique serial number assigned to it. Serial numbers guarantee that when you are examining a file that may have multiple counterparts, you are always looking at the same file. It also ensures the accuracy of the records and that they are readily accessible by indicating the order in which records are stored. Specific records are easier to locate, and the storage process is authenticated.

Fingerprinting

Fingerprinting is the notation of creating a unique checksum for each file. The checksum is created based on the content (and possibly other information) of the file. If changes are made to the file, re-calculating the checksum will result in a fingerprint that does not match the original. Fingerprints are one way to help guarantee the integrity of data in files by providing verification that a record has been accurately stored during the initial write and that it maintains its integrity over time.

Secure Timestamps

In many regulatory scenarios, the time and date that a file was generated or accessed can lead to a company being in or out of compliance. If the timestamps on a system are not correct, then archiving reports can be inaccurate. A secure and reliable timestamp mechanism is critical to maintain the accuracy of records.

Indexing and Searching

Indexing is designed to ensure that the records are accessible. It enables the search for specific records among the many stored. Indexing can be manual or automated. With manual indexing a person enters relevant contract numbers, customer names or other search terms into an enterprise content management system (CMS). Examples of automated indexing are the full-text index of the contents of a file and OCR of scanned images.

Legal Hold

A Legal Hold is a mechanism that marks data that must be preserved in all forms, independent of any policy that may allow for deletion. Typically the information is relevant to ongoing litigation.

Audit of Access

Authentication and Access Control are of paramount importance in determining who should have access to data. Audit logs for data access should be a part of all data archive systems. In a compliance review or under litigation, having the ability to validate who accessed specific data and when that data was accessed can make a difference in the outcome of the proceedings.

Policy to Keep Data

Known as a data retention policy, this policy dictates how long data ingested into the archive system must be preserved. Once the policy expires, data may be deleted according to the data deletion policy. If there is a Legal Hold, then data must be preserved.

Policy to Delete Data

Once the data retention policy has expired, the data deletion policy dictates what to do with the data. The policy may take different forms:

- Keep the data as long as space permits.
- Delete the data as soon as possible.
- Delete the data by using a randomized data overwrite of the files' original storage locations (if possible).

In some instances, such as with optical media, the media must be destroyed. A Legal Hold can override this policy.

WHAT IT LOOKS LIKE

Like a data backup system, a data archiving system is a storage infrastructure that is completely independent of any primary storage you already have. Also, like data backup, data archiving is best achieved by a combination of on- and off-site storage.

By conscious choice or by policy, data is placed into the archiving system. Once there, it may be safely removed from primary storage (optionally leaving a reference pointer in its place). Data placed into the archiving system will undergo a process whereby all the meta information about the file is generated. Then, the meta information and the file are written to the archiving system and kept according to the data retention policy.

Like a backup system, the storage for an archive system can take the form of disk or tape. However, the storage has additional features:

Immutable Storage/WORM

WORM (Write Once Read Many) is also known as "immutable storage." The concept is that once data is written to a storage location, it can never be changed (even by an administrator). Many of us are familiar with the concept from using CDROMs/DVDs. Once a write session is closed to the disc, data written on these optical media can never be changed or erased. (But the media can be physically destroyed.)

Some compliance organizations (notably SEC 17-A) state that certain digital information must be written in a non-rewriteable and non-erasable format.

Self-Healing

Self-Healing is the process of automatically validating the integrity of a file by periodically recalculating the file's fingerprint. If the fingerprint fails to match, then the data has become corrupted. Since redundancy is part of archiving, the file can be properly repaired by using a redundant copy of the file where the checksum matches properly. Over time, self-healing can help protect against data degradation caused by bit-rot, cosmic rays or other data decay.

Technology Refresh

Archived data may be around much longer than the lifecycle of a piece of storage hardware. Archive data has meta information associated with it (serialization information, indexing information, timestamps, checksums and data policies). Therefore the method of migrating data off an old storage platform to a new storage platform must be done in such a way as to protect not only the data, but the meta data.

Since data written to an archive system may be removed from primary storage, you can immediately improve your primary storage space. The process (known as storage offloading or static-data offloading), also immediately improves your backup processes by reducing the amount of data to be protected. Data in the archiving system should be redundant and not require backup.

BACKUPS OF YOUR ARCHIVE

By its nature, WORM/immutable storage is impervious to some of the issues that require us to back up primary storage. Specifically, accidental file deletion, file data corruption and modification by malware are not possible on a WORM system. It is not possible to modify or delete the data. This is true even if you are the administrator.

However, hardware failures, site failures, and data decay can cause data loss from your archive. If your archiving system supports writing multiple copies of data, offsite replication, and autonomic healing, you don't need to maintain a separate backup of your archive. By its nature, it already provides you with the best practices.

If you do not have multiple geographically separated copies of your archive, then you will need to protect your archive via a backup that has an offsite component. Having to back up your archive will mitigate some (but not all) advantages that archiving provides to your backup. In either case, you still maintain the advantages gained from storage offloading.

AT-A-GLANCE: COMPARE BACKUP TO ARCHIVING

	DATA BACKUP	DATA ARCHIVING
What does it do?	Active or inactive data copied to storage for the purpose of internal recovery.	Inactive data and meta data stored long-term for internal recovery or external discovery.
Why do it?	To protect critical operational internal data and computing processes.	To protect critical operational, historical, legal or customer data in its native form for retrieval on request.
When does it matter?	Weather disaster, outages, employee error, criminal mischief, hacking, virus, malware, lost or outdated devices.	Interruption of business to comply with lawsuit, regulatory audit or open records requests.
Who benefits?	Internal users, stakeholders, partners, customers.	Public entities, regulators, auditors, courts, clients, partners.
Recovery Profile	Files, folders, databases, system images saved at periodic intervals (hourly, daily, weekly, as required).	Files, folders and complete data determined by the corporate retention policy or by regulatory mandate.
Data Management Benefits	Version control, disaster recovery, business continuity, peace-of-mind.	Compliance, legal protection, operational assurance, additional storage resources.
What's at risk without these?	Loss of trust, reduced employee productivity, potential demise of the company.	

BACKUP WITH ARCHIVING: BETTER TOGETHER

When effectively combined, data backup and data archiving together create a value-adding synergy to your data management plan. Each solution delivers a separate set of functions, features and benefits. With the right policies, processes and technology, you can combine the two to achieve greater levels of economy, assurance and scale.

As your data assets progress through their data lifecycle, they may first be candidates for primary storage. There, they will also be candidates for backup. As those data assets age, they may then become candidates for data archiving. They can be written to an archiving system and removed from primary storage, creating a tiered solution. If you have data assets that are candidates for direct archiving (i.e., scanned images, PACS images, reports, etc.), you can immediately place those into an archiving system through automated policies. This provides the best protection against modification and loss.

With archiving, your backup stores can be smaller. The reduced load on network bandwidth gives faster processing and better system I/O. By reducing the load on your backup system, you can run backups more efficiently. At the same time, you improve the overall integrity of your organization’s data assets. Archiving gives you greater storage performance, smaller backup sets and faster recovery times.

ENSURING BOTH RECOVERY AND ACCOUNTABILITY

In practice, a good data management strategy depends on where you are as a company regarding data assets. What policies are in place to guide your data protection and management activities? Are you faced with any of these issues?

- Does the time it takes to execute a system backup or recovery keep growing?
- Do the larger data sets keep increasing the procurement of primary storage?
- Do legal and regulatory requirements weigh into the mix, further complicating data strategy?

All companies deal with changing data contexts. As software systems evolve over time, backed up data becomes obsolete. If this presents a threat to your organization’s health, let archiving reinforce your backup plan.

When you are responsible for preserving digital data integrity and long-term availability, data archiving is essential. In particular, archiving is required for securities brokers, publicly traded companies, hospitals or CPA and law firms. To meet regulatory mandates such as Sarbanes-Oxley, SEC-17, HIPAA, PCI DSS, GLBA, and legal hold, you need a legitimate (and strategic) archiving solution.

IN SUMMARY

To manage risk and maintain historical records, be sure your solution delivers:

- Integrity Features
    - o Immutable storage/WORM
    - o Fingerprint of all files
    - o Serial numbers
    - o Secure timestamps
    - o Constant data verification
    - o Self-healing

- Security and Privacy
  - o Separation of data stores
  - o AES-256 encryption
  - o Built-in key management
  - o Authentication of users
  - o Access control for users
  - o Active directory integration
- Policy Driven At All Levels
  - o Data ingest
  - o Data indexing
  - o Data retention
  - o Data deletion (with secure delete)
- Redundant, scalable and highly available
  - o Multiple copies of data
  - o Cloud integrated with remote replication (does not need backup)
  - o Failover/failback
  - o Self-healing
  - o Scale performance and capacity independently
- Meets regulatory mandates
  - o Auditable
  - o Searchable
  - o Legal Hold-enabled
  - o Data retention policies
  - o Data deletion policies
  - o Reporting
- Technology
  - o Technology upgrade path
  - o Energy-efficient
  - o Cost-effective
  - o Fully private cloud or multi-tenant cloud
- Manageable

RESEARCHING SOLUTIONS FOR BACKUP AND ARCHIVE

Safe data requires a safe management partner. When researching service providers, look for these criteria:

1. *Do they offer backup services, archive services or both?* For best value, you want both; archiving and backup work best together.
2. *Do they provide cost-effective, long-term, hybrid and cloud-based data storage and services that meet verifiable security and compliance standards?* Should you face an audit, you won't want anything less.
3. *Is their program designed specifically for long-term data storage, resting on a compliant foundation that can scale globally?* The last thing you need is to hit a scalability wall with your provider. No company needs the data growing pains of having to migrate everything to a new solution.
4. *Are their solutions validated to meet or exceed security levels offered by mainstream cloud services?* High security equals maximum peace-of-mind.

5. *Is their infrastructure geo-distributed and engineered to the highest levels of durability?* A local disaster may wipe out a single data center. If your data rests in another location, you'll still be in business.
6. *Do they have a history in the data protection industry with verifiable references and case stories to support their sales pitch?* Recovery requires a provider who can walk the talk.

NETMASS SETS THE PACE IN DATA PROTECTION

Since 1998, NetMass has provided industrial strength, worldwide data backup and data archiving services. Our long-term commitment to data protection is the reason so many clients trust us with their data.

THE NETMASS SERVERBACKUP.COM SOLUTION

Our backup solutions, offered under the operational arm ServerBackup.com, span the gamut of deployment options:

- Software only
- Hybrid, multi-tenant cloud
- Fully private hybrid cloud
- Fully private cloud
- Fully private customer premise

We support companies of any size and using any number of specialty applications and remote locations.

THE NETMASS DATAARCHIVING.COM SOLUTION

Our archiving solutions, offered under the operational arm DataArchiving.com, provide long term data archival using:

- Software only
- Hybrid multi-tenant cloud
- Fully private cloud deployments

We support the archiving, security and compliance needs of any organization. For more information, contact us at 1-800-731-2737 or email info@netmass.com.

[1] Derived from the Wikipedia definition of information technology backup.
[2] Digital Preservation Coalition. "Introduction: Definitions and Concepts". Digital Preservation Handbook. York, UK. Retrieved 24 February 2012.
[3] Evans, Mark; Carter, Laura. (December 2008). The Challenges of Digital Preservation. Presentation at the Library of Parliament, Ottawa.
[4] Derived from Wikipedia definition of digital preservation