SNIA
Advancing storage & information technology

Storage Networking Industry Association
Technical White Paper

Data Protection Best Practices

Version 1.0

October 23, 2017

SNIA DPCO | DATA PROTECTION & CAPACITY OPTIMIZATION

Abstract: ***This white paper covers the Storage Networking Industry Association's (SNIA's) position related to data protection best practices using common data protection technologies, as identified and recommended by the SNIA's Data Protection & Capacity Optimization (DPCO) Committee.***

USAGE

The SNIA hereby grants permission for individuals to use this document for personal use only, and for corporations and other business entities to use this document for internal use only (including internal copying, distribution, and display) provided that:

1. Any text, diagram, chart, table or definition reproduced shall be reproduced in its entirety with no alteration, and,

2. Any document, printed or electronic, in which material from this document (or any portion hereof) is reproduced shall acknowledge the SNIA copyright on that material, and shall credit the SNIA for granting permission for its reuse.

Other than as explicitly provided above, you may not make any commercial use of this document, sell any or this entire document, or distribute this document to third parties. All rights not explicitly granted are expressly reserved to SNIA. Permission to use this document for purposes other than those enumerated above may be requested by e-mailing tcmd@snia.org. Please include the identity of the requesting individual and/or company and a brief description of the purpose, nature, and scope of the requested use.

All code fragments, scripts, data tables, and sample code in this SNIA document are made available under the following license:

DISCLAIMER

The information contained in this publication is subject to change without notice. The SNIA makes no warranty of any kind with regard to this specification, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The SNIA shall not be liable for errors contained herein or for incidental or consequential damages in connection with the furnishing, performance, or use of this specification.

Suggestions for revisions should be directed to http://www.snia.org/feedback/.

Revision History

Revision	Date	Originators:	Comments
V0.10	*November 16, 2015*	Thomas Rivera, Gene Nagle, Mike Dutch	Initial Draft
V0.53	*October 19, 2017*	Thomas Rivera, Gene Nagle, Mike Dutch	Final DRAFT
V1.0	*October 23, 2017*	Thomas Rivera, Gene Nagle, Mike Dutch	Final

Suggestions for change or modifications to this document should be submitted at: http://www.snia.org/feedback/.

Forward

This whitepaper was prepared by the SNIA Data Protection & Capacity Optimization (DPCO) Committee to provide IT professionals with best practices and guidance on data protection from a storage perspective.

Table of Contents

Executive Summary

Data protection is often viewed as consisting of the execution of backup operations that are assured of providing data recovery if a loss of the original data (production data) occurs. In fact, data protection encompasses much more than backups and recovery techniques, such as dealing with issues related to data corruption and data loss, data accessibility and availability, as well as compliance. This whitepaper covers the aspects of data protection that relate most directly to storage systems.

1 Introduction

This white paper covers the Storage Networking Industry Association's (SNIA's) position related to data protection best practices using common data protection technologies, as identified and recommended by the SNIA's Data Protection & Capacity Optimization (DPCO) Committee. SNIA has developed this position in order to influence future engagements with Standards Development Organizations (SDOs), as well as to educate the end-user storage networking community.

1.1 Data Protection Overview

Data protection[1] is defined as the assurance that data is not corrupt, is accessible for authorized purposes only, and is in compliance with applicable legal and regulatory requirements.

The above definition of data protection goes beyond the notion of "data availability", which is defined as the amount of time that data is accessible by applications and users during those time periods when it is expected to be ready for use. Unacceptable performance can lower productivity levels such that access to applications and related data is effectively unavailable. Note that data security[2] and compliance issues are also intimately involved in data availability, as the ultimate goal of data protection is to mitigate vulnerabilities, costs, and downtime.

This concept of "data availability" can be further defined in terms of accessibility, integrity, and timeliness. Accessibility involves assuring that the data is accessible at the right place, for the right uses and in a timely manner. Integrity refers to the data being correct in all aspects, the "same" as it was stored, with no alteration or corruption to the information, whether intentional or accidental.

[1] SNIA Dictionary (https://www.snia.org/education/dictionary)
[2] SNIA Storage Security White Paper (http://www.snia.org/sites/default/files/Storage-Security-Intro-2.0.090909.pdf)

The concept of data "usability" is defined as the protected data should be "usable" for its intended purpose. Usability may require that steps be taken to provide data integrity, application consistency, versioning, availability, and acceptable performance.

The concept of "resiliency" is defined in the SNIA Dictionary as the ability of a storage element to preserve data integrity and availability of access despite the unavailability of one or more of its storage devices.[3]

"Journaling" will be mentioned later in this paper. In the context of data protection, journaling refers to using a mechanism to "track" all the changes of a volume (or file system), thereby simplifying recovery in the event of a volume or file system failure. Journaling is also often used to support recovery between separate physical sites that are being replicated or mirrored. Multi-site replication will be covered in Section 2.2.1 of this paper.

Finally, another aspect of Data Protection is Business Continuity Management (BCM), which is defined as the processes and procedures for ensuring continued business operations[4]. BCM involves the recovery of data, access to data and associated processing through a comprehensive process of setting up redundant sites (equipment and work space) with recovery of operational data to continue business operations after a loss of use of all or part of an infrastructure. This involves not only the essential set(s) of data but also the essential set(s) of all the hardware and software to continue processing of that data and business. Any disaster recovery may involve some amount of down time, depending on the specific BCM implementation. BCM will be covered in section 2.2.2 of this paper.

[3] SNIA Dictionary: (http://www.snia.org/education/dictionary)
[4] ISO/IEC 27000

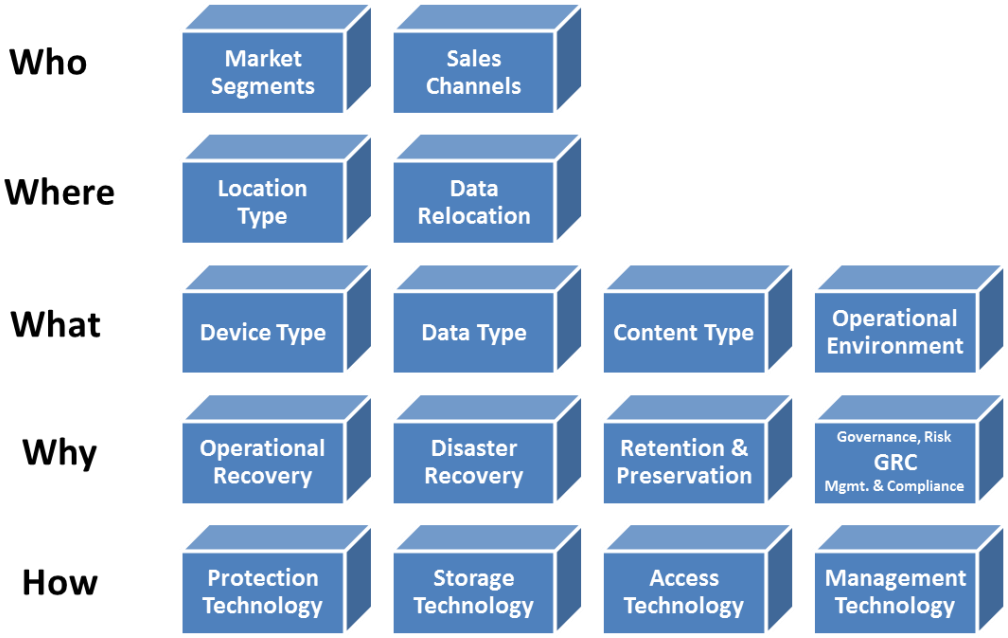

Figure 1. Data Protection Taxonomy (SNIA)

Figure 1 is a high level overview of the SNIA data protection taxonomy. This taxonomy uses boxes to represent distinct "lenses" through which to view a data protection solution.

Each lens is independent, but there are many relationships between theses lenses, and the data protection taxonomy encourages examination of these relationships. Each row of boxes addresses a particular question, namely, "who", "where", "what", "why", and "how". Please refer to the SNIA Data Protection Taxonomy White Paper[5] for the specifics of this taxonomy layout.

[5] SNIA – DPCO "Data Protection Taxonomy" White Paper (http://www.snia.org/sites/default/files/A_Data_Protection_Taxonomy_V51.pdf)

1.2 Data Protection and Data Management

1.2.1 The Two Main Facets of Data Protection (Backup and Restore)

Data protection is an important component of any Information Technology (IT) system, and the methods used for data protection and how they are configured have important inter-relationships with other aspects of the data center. By its nature, data protection has two sides, the backup or replication side and the restore or recovery side.

The backup side of data protection is the process or processes performed on a regular, or even a continuing basis to create one or more copies of an organization's production data at a particular point in time. Backup processes may well differ from one type or subset of data to another, and they must be chosen with care to minimize the impact on the availability of production data to all applications and users that need it. The backup must also provide for recovery of data in the way prescribed by the organization's service level objectives (as defined by the Service Level Agreements) with regard to each set of data. Thus traditional daily backups (copies of data to a different media) may be used for some subsets of data, while a real-time mirroring process may need to be used for other, highly critical data sets, in order to assure continued access to the data.

Successful recovery operations are the result of having put appropriate backup processes in place, and recovery of lost or corrupted data is vital to an organization's health. A recovery operation may be required just to replace a file that a user accidentally deleted or a corrupted set of data (operational recovery), or to replace a major portion of an infrastructure, in case of a disaster such as a multiple device failure, a virus (e.g., ransomware), a denial of service attack, or destruction by a fire or flood (disaster recovery). There are two important considerations or objectives for data recovery that in turn determine how it needs to be backed up; they are the Recovery Point Objective (RPO) and the Recovery Time Objective (RTO). RTO and RPO are important factors in deciding what backup or replication strategy the business needs to use, and they need to be a part of any organization's SLAs with regard to data protection.

The Recovery Point Objective (RPO) value is the amount of data loss the business can afford for each respective business process. In other words, the RPO tells the business how old their backup data can be. For example, if the last backup took place at 2AM, and the storage system disaster occurred at 5AM, then all the data that was created and modified within those 3 hours will be lost.

The Recovery Time Objective (RTO) value is what the tolerable (to the business) duration of the process disruption is. In other words, how soon the respective processes have to resume in order to avoid significant negative consequences. Like the RPO, the appropriate RTO is usually determined in cooperation with the respective business process owners. See the RTO/RPO diagram below.

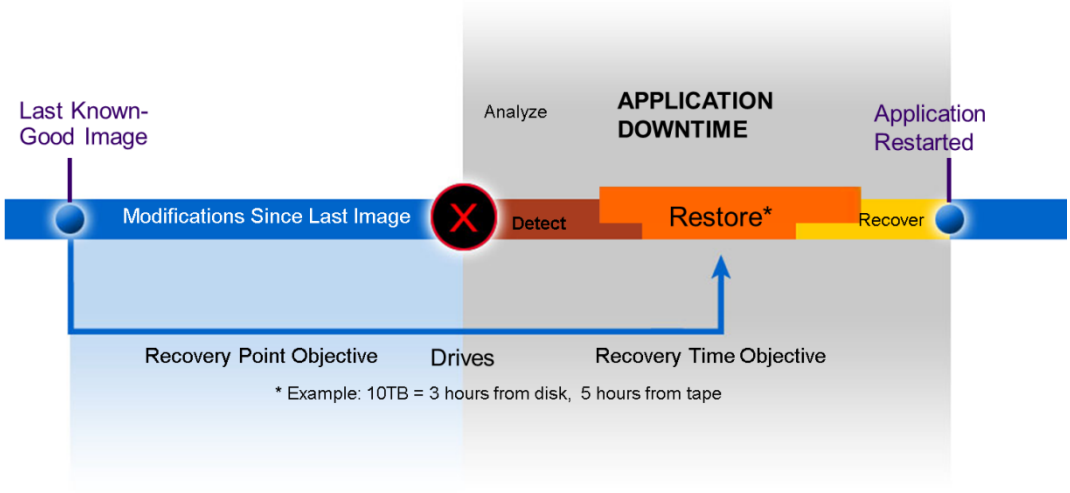

Figure 2: Recovery Point Objective (RPO) & Recovery Time Objective (RTO)

1.2.2 Data Protection and Digital Archives

Although a digital archive represents another set of copies of production data like those intended for backup or disaster recovery (DR), an archive is more immutable in nature, with changes either not allowed, or strictly controlled by a journaling process. Also, archives themselves require data protection, but they are not intended to be used for data protection.

Archives may be divided into two types based on their intended longevity, with those intended to last more than ten years being considered a long-term archive[6]. Long-term archives typically require different methods for storage, security and management. Data protection of medium and long-term archives is discussed in section 2.1.6.

1.2.3 Other Important Characteristics of Data Protection Devices and Procedures

In order to provide secure and reliable protection of primary storage, the devices used as destinations for data protection must themselves be: “resilient”, “secure” and “compliant”.

As noted in section 1.1, resilient (or resiliency) is defined as the ability to preserve the integrity and availability of stored data despite component failures. “Secure” refers to making full use of safeguards such as access and distribution controls via both physical means (facilities) and logical means, (e.g., encryption). The act of being “compliant” involves full conformity with corporate and legal requirements.

[6] SNIA-100 Year Archive Requirements Study (http://www.snia.org/sites/default/orig/100YrATF_Archive-Requirements-Survey_20070619.pdf)

2 Drivers of Data Protection

The SNIA DPCO has categorized data protection technologies into three sections or "Drivers":

1. Data Corruption / Data Loss
2. Accessibility / Availability
3. Compliance

This paper will specify the data protection technologies within each "driver", and then will provide appropriate references (e.g., existing standards), and finally recommend the SNIA Best Practices for each data protection technology.

2.1 Data Corruption / Data Loss

This section describes the appropriate application of multiple technologies used to avoid data corruption and data loss, such as: data protection algorithms (e.g. RAID, etc.), snapshots, backup and Continuous Data Protection (CDP), etc.

2.1.1 Data Protection Algorithms (RAID, Erasure Coding, etc.)

This section will describe data protection algorithms, which are used to "protect" the data on the storage media itself, from data corruption due to bit errors or media failures. Data protection algorithms such as RAID (Redundant Array of Independent Disks) and erasure coding are ways of protecting data.

RAID (Redundant Array of Independent Disks)

RAID[7] is an enabling technology that leverages multiple disk drives (spinning disk, solid state or virtual drives) as part of a disk drive set that provides data protection against disk drive failures as well as media failures (e.g., unreadable sector). RAID can also serve to improve storage system performance since input/output (I/O) requests may be serviced by multiple disk drives independently.

Although RAID may be technically considered a form of erasure coding, generally, erasure coding refers to a very different approach to data protection, and is described in the next section.

The most common RAID types in use are: RAID 0, RAID 1, RAID 5, RAID 6, and RAID 10. There are other variants of RAID used in the industry, but these are the most widely used. Below are descriptions, as well as the common pros and cons of each of each common RAID type.

[7] https://www.snia.org/tech_activities/standards/curr_standards/ddf

RAID Level	Description
0	RAID 0 describes a method of writing data across two or more disk drives, typically to achieve higher throughput. Data is stored without any parity data or other form of backup of the stored data. ***Pro:*** More performance (compared to writing to a single disk), since the data is being written ("striped") across multiple disk drives, and this involves more disk heads, which enables parallel access to more data records. ***Con:*** No protection against data corruption or loss.
1	RAID 1 involves creating exact copies of data on two or more disk drives, often referred to as "mirroring". ***Pro:*** Offers data redundancy in case of a disk drive failure. Performance gains are achieved since reads can be retrieved from either member of the "mirror". ***Con:*** True storage capacity is only half of the actual capacity, since the data is written twice. This also doubles the cost of the storage.
5	RAID 5 consists of a minimum of three disk drives, with each drive storing both data and parity. If a disk failure occurs, parity data from the remaining disk drives is used to recreate the missing information that was on the failed disk drive. ***Pro:*** Balance of data protection, capacity overhead and cost. It ensures fast read performance (due to reading from multiple disk drives), and also ensures quick data recovery if a single disk failure occurs. ***Con:*** Writes may be slower since the system has to calculate parity before writing data. This may not be an issue if the RAID operations are performed in hardware and/or utilize a large cache. (Also applies to RAID 6).
6	RAID 6 is similar to RAID 5, except a minimum of four drives are required, and there is an additional parity block, so that two disk drives can fail, but still have access to the data. ***Pro:*** Two disk drives can fail, but still have access to the data. ***Con:*** Disk drive rebuild may impact data access performance. Because of this, rebuilds may be scheduled for a later period in time so as to not impact specific application performance requirements. (Also applies to RAID 5).
10	RAID 10 involves both striping data and mirroring data. This can be implemented in two different ways: striping sets of mirrored disk drives, often called "RAID 1+0", or using two or more mirrored sets of striped drives, often called "RAID 0+1". ***Pro:*** RAID 10 provides fast read and write speeds, as well as maintaining data protection based on the redundancy (mirroring) of the data. ***Con:*** Cost, since the data is written twice.

The use of RAID levels includes aligning the use of the appropriate RAID level that will give the right balance between cost, performance and protection. For example, when wanting to achieve very high data availability, while maintaining the best possible performance, then a RAID 10 configuration may be ideal. The downside of course, is that RAID 10 doubles the cost of storing data, versus a RAID 0 (striping) configuration.

Erasure Coding

Erasure Coding (EC) can be used for data protection instead of RAID, and is becoming quite common as larger and larger multi-terabyte disk drives are being manufactured. For example, many of the "Object Storage" systems on the market today use some form of EC rather than RAID. Also, EC systems are typically software-based, as compared to traditional RAID 5 and 6 systems which have often used specialized hardware to perform necessary I/O processing.

EC is a forward error correction technology that's used to provide data resiliency and long-term data integrity. Erasure codes are often used instead of traditional RAID because of their ability to provide a more granular correction process, thereby reducing the time and overhead required to reconstruct data (drive rebuilds).

EC parses incoming data into multiple component blocks, then, somewhat like a parity calculation, expands each block with some additional information, creating a slightly redundant but more resilient superset of data. With a mathematical algorithm, the system can use these expanded blocks to recreate the original data set, even with missing or corrupted blocks. This allows the storage system to still deliver data, even after multiple drive or node failures. There is little overhead for "reads" when using EC, except when there are drive failures, since the calculations happen during "writes". Most EC schemes allow the user to configure the level of resiliency, essentially by increasing the amount of parity data generated for each block. There are also different levels at which EC can be applied: at the array level, at the node level (for scale-out architectures) or at the system level – which can affect how much processing overhead it consumes.

EC can be combined with data distribution or dispersion to improve resiliency and eliminate the need to make dedicated copies for off-site storage. This process essentially spreads data blocks across multiple nodes or systems, usually in different physical locations. However, using a distributed architecture where data blocks are spread between different physical locations can create a latency problem, since network bandwidth quickly becomes the limiting factor when blocks are pulled across the WAN. Some object storage systems combine EC and replication, using ECs at the local system level and copying data between geographic locations to alleviate latency.

Here are some of the pros to the use of EC:

- Depending on how EC is set up, more disk drives can fail as compared with RAID without losing access to the data, achieving greater fault-tolerance[8].
- Better space efficiency may be achieved versus RAID, since EC does not need the extra space for parity calculations that RAID requires.
- Better space utilization may be achieved versus data replication since replication often uses more "landing" space as compared to some of EC's algorithms for recovering data.

Here are some of the cons to the use of EC:

- As compared to replication over a Wide Area Network (WAN), there is a performance impact with the use of EC, since the EC parity calculations must also take place which is CPU intensive. This may not be an issue if EC is implemented in hardware and could therefore operate at line speed, in which case, the operations become transparent.
- Write-intensive applications may not be a good fit, since the performance penalty for EC is mostly during "write" activity. This may not be an issue if an efficient, hardware-based system is used.
- The overhead that EC represents is dependent on where erasure codes are applied (at the array, at the node or at the system), and the level of resiliency chosen.

Best practices for EC includes using it when considerations for recovery time and latency make it a good fit for particular applications, such as archive data. For using EC with remote (off-site) data, the main considerations will be performance requirements and overall capacity efficiency savings, versus replication. In other words, the capacity savings of EC must out-weigh the extra complexity that comes with it, versus replication. The reason that archive data is a good fit for EC (other than the fact that more simultaneous failures can occur without losing data access), is that most of the activity for an archive is "reads" rather than writes, and EC does not incur much of a performance penalty for reads. The probability of recovery of data from multiple disk failures will vary, depending on the vendor implementation of EC. Some vendors will sacrifice failure recovery probability (e.g., recover from 99.9% of disk failures), as a tradeoff for faster EC.

Another consideration is that drives are getting larger (multi-terabyte), and rebuild times using standard RAID algorithms are getting longer (multiple hours). Because of this, the need for EC is

[8] SNIA Dictionary definition of fault tolerance is: "the ability of a system to continue to perform its function (possibly at a reduced performance level) when one or more of its components has failed."

becoming much more important to sustain multiple drive failures and not have to suffer the prolonged period of "read/recovery" time per drive, as seen with many RAID-protected storage systems.

2.1.2 Snapshots

A snapshot is a point-in-time copy of a defined collection of data. A "delta snapshot" is a point in time copy that preserves the state of the data at an instant in time, by storing only those blocks that are different from an already existing full copy of the data.

Snapshots are a way to create distinct "point-in-time" views of a data set, when performing data protection actions such as backups and/or replication, since the "view" of the data set is "frozen" and in a known state, and usually alleviates issues such as open files. The exact implementation of snapshot execution will vary by vendor.

Snapshots are usually taken regularly, as part of a backup strategy (see "Backups" below). The interval of the snapshots is usually based on the granularity requirements for restoring from a specific point in time. For example, taking snapshots every minute will provide greater granularity in restore point capability, versus taking snapshots every hour. The criticality of the data will help in deciding how often snapshots should be executed. So, if the business requires data to be restored to a point in time with a granularity of one-minute, then snapshots should be executed every minute.

Here are some of the pros to the use of snapshots:

- Allows for the recovery of files from a specific point in time (based on snapshot schedule).
- Backup applications can use the snapshot as a "quiescent" view of the data set to be backed up, so that there will be no issues with open files, ongoing modifications, etc.
- The snapshot-based backup can be performed transparently to ongoing processing.

Here are some of the cons to the use of snapshots:

- Space is consumed for each respective snapshot taken.
- There could be performance degradation during the execution of the snapshot, and also afterwards while the snapshot is maintained.

Best practices for snapshots include using them for backups, so that the source of the backup is the snapshot, such that the backup as well as the restore can be executed from a "quiescent" view of the

data set. This assures that there are no "open" files[9] in the snapshot from the data set that was backed up. Also, the use of the snapshots for restores allows for a finer granularity versus regular daily backups, for the restore of a given data set to a certain point in time, also known as Recovery Point Objective (RPO). See RPO in Section 1.2.1.

Another best practice for snapshots includes executing the snapshot interval in line with the Recovery Point Objectives (RPO) requirements of the data sets that are being protected. So, for example, if the business requires that a specific data set needs to be recovered to within a one-minute point in time, then the snapshots should be taken once per minute.

2.1.3 Backups

A "backup" or "backup copy" is defined by the SNIA as a collection of data, often stored on non-volatile storage media for purposes of recovery, in case the original copy of data is lost or becomes inaccessible[10].

In addition, there is a difference between a "file" backup versus an "image" backup.

A 'file" backup is a file/folder-based backup, in which the smallest unit that could be restored is a file or folder. The typical use for such a system is to restore a file or folder that has been lost on an otherwise healthy system. This is a "selective" backup, where the business chooses what data should be backed up, and only those files and folders are backed up. The total backup is much smaller in size, with less capacity needed and a lower overall cost. The downside is that should a disaster occur, the data restoration can take much longer than anticipated since it must start from scratch to restore your system to working order – installing the operating system, all the software applications, the software used to back up the files and folders, and then finally the files and folders themselves. Only the files and folders selected for the "file" backup are restored, and any other files lost in the data disaster cannot be retrieved.

An "image" backup consists of the block-by-block contents of a data set, virtual machine (VM), or disk drive. All of this data is backed up as a single file, called an "image". In the event of a data disaster, a business' entire data set is preserved, sometimes allowing for a move to new hardware and a quick restore of all the associated information required to get back up and running. Many modern environments use this for VM management, for the creation of a single image ("golden image") that allows rapid deployment of fully patched and configured operating systems and associated applications.

[9] Open files are an issue since they will usually be skipped when the backup occurs.
[10] SNIA Dictionary: (http://www.snia.org/education/dictionary)

File-based backups are well-suited for certain data sets in when wanting to recover individual files, although it will take longer periods of time (e.g., hours to days) to restore an entire system to operating status. Image-based backups are often best suited for business-critical data that needs to be recovered quickly (e.g., minutes to hours).

Image or file backups can be performed in many different ways, very often in conjunction with snapshots, and often backing up onto various types of storage media. The best methods to use will vary based on specific requirements of the organization. Also, at the data set level, different data protection levels may be deployed – based on the criticality of the data. So, for example, a data set that includes financial trading logs for a financial trading firm may have a more stringent data protection policy than the data set that includes the trading firm's daily lunch menu.

In most environments, individual backup data sets are retained for a limited time period, usually between 30 to 90 days, since backups are not typically considered a good method for creating archives. (Archives will be discussed below, in Section 2.1.6). If an organization elects to do full backups of all data on a daily basis, the most recent day's successful backup is the one that will be used for any necessary restores. This makes all previous days' backups obsolete except as future references for particular points in time, such as the state of financial files at the end of a month or quarter before accounting procedures are run. In some jurisdictions, backup sets may be subject to electronic discovery[11], therefore consider that the backups only be kept for a minimal amount of time, and in compliance with the organizations retention policies for backups.

Performing lengthy full backups on a daily basis, however, is seldom considered now, given large data volumes and the need to run most organizations on a 24x7 basis. Although snapshots can be used as the source for a backup rather than the actual files, backing up from snapshots that are stored on the original devices may cause significant performance degradation, depending upon the specific vendor implementation.

One alternative to daily full backups is to use incremental backups, whereby a full backup is done periodically, such as once per week, followed by backing up just the changed files during each succeeding day. One variation to the above is "differential" backup, whereby the files that are backed up are all files that have changed since the last "full" backup. The other main difference between an incremental and a differential backup is the time needed to restore data in the event of an emergency. When using a combination of full and differential backups, only two backups need to be restored – the most recent full and the most recent differential. Otherwise, if using a combination of full and incremental backups, then a combination of the most recent full and all of the incrementals since the last full backup need to be restored in the event of an emergency.

[11] Discovery that includes the identification, preservation, collection, processing, review, analysis, or production of Electronically Stored Information. [ISO/IEC 27050-1]

Another variation of this method is called “incremental forever”, or “synthetic fulls” which allows the backup administrator to do one full backup and then only incremental backups after that. To accomplish this, the backup software must maintain a secure index of all changed files that allows a restore to access the latest version (or some other specified version) of every file that needs to be restored.

Another consideration for backups is to consider the data sensitivity of the data set(s). Based on the sensitivity and/or criticality of the data set(s), the backups may need to be handled differently. For example, if there is a concern of backup media being stolen, the data should be encrypted. Another example is if the backups of specific data set(s) need to be restored within specific time frames (see “RTO” discussion above), then the media selected should be appropriate to allow for “fast” recovery of those data set(s) after a disaster/disruption has taken place, e.g. disk drives versus tapes.

Here are some pros and cons to using each of the backup methodologies:

Daily Full Backups:

Pros:

- Any files that need to be restored are a part of a single backup, and therefore the restore is usually faster.
- No need for multiple “mountings” for a restore.

Cons:

- The backup time is usually much longer and may interfere with other production operations.
- The space required to store the full backup is greater than the other backup types.

Daily Incremental Backups:

Pros:

- The backup time is shorter for incrementals versus daily full or differential backups.

Cons:

- The restore will need to involve the most recent full backup, as well as all of the incrementals, up to the most recent daily incremental that is available, which will be longer than restoring from a combination of full and differential backups.

Differential Backups:

Pros:

- The restore will be shorter than using incremental backups, since the restore will only require a maximum of two "mountings", the full and the differential backup in either order.
- Differential backups save backup time versus a daily "full", since the daily differential backup will only include the data that has been modified since the last "full" backup.

Cons:

- Differential backups take longer than incremental backups.
- The amount of daily backup storage will be more with differential backups than performing daily incremental backups, since the daily backup will include the data that has been modified since the last "full" backup.

"Synthetic Fulls" or "Incremental Forever" Backups:

Pros:

- Faster backup times, since only need to perform incremental backups after the initial "full" backup.
- "Synthetic fulls" or "incremental forever" backups save backup space versus a daily "full", since the daily backup will only include the data that has been modified since the last backup.

Cons:

- Relying on one initial "full" backup, and would need to restore from potentially many incremental backups in order to restore a single file or data set.
- The restores would need to be restored in the appropriate order, which could be time consuming.

The storage of backup media is also important. For critical data, the SNIA's recommendation is to maintain at least 3 copies (one primary and two secondary) of each data set across at least two geographically disparate locations, with at least one copy write-protected, preferably isolated and on different media for business continuity purposes. The overall goal is to ensure that the appropriate data set(s) can be restored in the case of any type of disaster, including an entire data center

becoming unavailable, within the specified restoration goals of your organization for each respective data set. This level of data protection for backups may not be cost effective for non-critical data.

Another important best practice regarding backups is to ensure that the backups that are executed can be used for successful data recovery. This can be accomplished with a regular routine of data recovery testing, to ensure data recoverability. The worst time to find out that the backups are un-recoverable is when a restore needs to be executed just after a disaster took place. Keep in mind also that the restoration time for a given data set will highly depend on the backup type (e.g., incremental versus full), the backup media to be restored from (e.g., tape versus disk), network topology (e.g., LAN versus WAN), and the overall size of the data set.

There are a few additional considerations for backing up virtualized environments. For virtualized environments, it is best to use a backup application designed to work with virtual machines, so that the backup will take place directly at the hypervisor layer (without involving the Guest OS layer or the VM host), and then all the appropriate resources are made available for the backup session workload. If there is external data (outside of the VM image), then other methods will be required to make sure that all external data is backed up appropriately. Also make sure that the VM APIs are used whenever possible for backups, so that there will be direct access to the VM disk files, and the "Changed block" feature is used, which allows for quicker incremental backups. Always remember to back up the individual host servers, along with the appropriate VM server(s) configuration files.

Additional backup requirements may be in order based on specific business requirements, including the criticality of each data set, regulations, contractual commitments, audit logging, etc.

2.1.4 Continuous Data Protection (CDP)

Continuous Data Protection (CDP), also known as "Real-Time Backup", is a class of mechanisms that continuously capture or track data modifications, enabling recovery to previous points in time. CDP traces every single change of data (usually at the block level), and allows a restore of the data set as it was at any point in time. The ability to "roll-back" to any given point in time using CDP allows the user to recover the state of a data set from any range of points of time within the data set.

Although the primary purpose of CDP is to provide low RTO/RPO protection against logical problems, the fact that there is also an additional physical copy of the data means that CDP also delivers physical data protection as a by-product. This can prove to be very useful if there is a need to failover to the CDP copy to act as the production copy on a temporary basis.

Some drawbacks on CDP such as additional CPU and capacity overhead can be mitigated through the use of "near CDP", by lengthening the sampling interval of the changes to the data set, if supported by the CDP vendor.

Here are some pros and cons to using CDP:

Pros:

- CDP, with its granularity at the byte level, is very space-efficient, compared to regular backups, with its granularity at the entire file level.
- CDP can provide the ability to restore to any previous point in time, since the backups are taking place near-instantaneously; therefore, the potential for data loss is very small.

Cons:

- CDP may heavily impact LAN or WAN performance, if writing over a network to another storage device.
- For data sets that are being frequently changed, the overhead involved with tracking the CDP changes may adversely affect the storage and/or server system(s) involved.

CDP may be expensive to implement, along with the drawbacks listed above. Some backup software vendors support their own version of CDP, and can be implemented as part of an existing backup infrastructure. The specific implementation of CDP will vary by vendor.

Best practices for CDP include CDP being used as a complimentary approach to an overall data protection model for a given data set, where appropriate. CDP does not always negate the need for executing separate backups and/or creating archives.

2.1.5 Replication and Mirroring

Although “replication” and “mirroring” (or RAID-1) are often used interchangeably, they will each be defined separately in this white paper.

Replication is a class of mechanisms for copying a collection of data at a point in time, typically asynchronously. The replication can be executed either locally or within the same Data Center via a local area network (LAN), or a storage area network (SAN), or remotely, via a wide area network (WAN). Although archives are often created via replication, they are meant to store data for a specified time frame; archives are typically not suited for day-to-day restores.

Mirroring on the other hand, is used for “continuous” (not a point in time copy) writing of data to two or more targets, and having each target member of the mirror capable of being used for read operations. Each member of the mirror could be “local” (within the same storage array), or a different storage array in the same data center, or it could be “remote” (in another building on campus or across town, or in a different geographic region, or even in another country). Note that

mirroring is not typically used as the only backup, since any change or corruption would be reflected on the other member(s) of the mirror.

There are two types of remote mirroring: “synchronous remote mirroring” and “asynchronous remote mirroring”. Synchronous remote mirroring and asynchronous remote mirroring are both continuous processes, which means that they are not dated and transmit (or cache) each I/O. The difference is that in synchronous remote mirroring, no further I/Os are allowed until the remote array acknowledges that it has successfully written the current I/O. However, that is not true with asynchronous remote mirroring where there could be a nonzero RPO in case of a disaster. The distance between local and remote sites may be so far that the latency for acknowledgement may be unacceptable so that synchronous remote mirroring cannot be used. For really critical time sensitive applications, a three site model where both synchronous and asynchronous mirroring are deployed may provide the best solution.

Remote replication is often used in conjunction with standard backups, however data replication does not require a “restore” of the data (as backup does), before the data can be used/accessed. In other words, assuming the appropriate resources are in place at the replication target location, the replicated data can be used instantly or nearly instantly (depending on the specific vendor implementation).

Whether replication is used as part of the data loss/availability procedures greatly depends upon the importance (criticality) of the data to the business, along with its availability requirements. Requirements for availability of the data set(s), are often depicted as Recovery Point Objective (RPO) and Recovery Time Objective (RTO) requirements. For example, for a banking customer’s transaction logging data set, is likely to have a “zero” RPO, and an RTO of a few seconds (or less). In this case, even a minute of disruption can pose an unacceptable threat and can lead to devastating consequences for the business. Data “backup” is not powerful enough to prevent such data loss/availability at this level. So, as the criticality of the data rises, more sophisticated and costly geographically dispersed replication solutions are often deployed.

End-users typically deploy 3 primary categories of replication and mirroring:

(1) Complete replication and mirroring of data between 2 or more sites

(2) Caching of frequently used data at remote sites with complete data sitting at a home site

(3) Hybrid complete replication and mirroring with caching

Complete replication and mirroring is often used to provide continued access to data in the case of a disaster and the "copies" of the data are often in geographically dispersed locations. In addition, complete replication and mirroring is utilized in standard production environments to provide faster

access to the data by localizing the data access for the user or application that is requesting the data. Replication and mirroring (in particular) can even be used within the same data center or in another data center within close proximity simply to maintain more than one copy of the data and/or distribute the access load. One example of complete replication would be for a financial institution such as a credit card company where complete data copies must reside in differing geographical locations to account for natural disasters, terrorism, etc.

Caching of frequently used data at remote sites allows many of the same benefits of complete replication and mirroring, without the heavy investment in hardware at all sites. End-users can access and alter smaller subsets of data at geographically dispersed sites while the aggregate data set of all cache sites can reside in one or more central locations. This is often referred to as a “hub and spoke” configuration because the hub would contain all data used by all users and each spoke would be a cache site utilized by a subset of users that only need access to a subset of the data. The trade-off of requiring less storage at a cache site is that in some cases data may need to be transferred from the home site to the cache site (or vice versa) if it is not already at the cache site. An example of a “hub and spoke” replication configuration could be a hospital's digital image repository. A central repository would contain all medical images and cache sites would reside within each department and contain only those images specific to that type of medicine.

The third type of topology used for replication and mirroring is a hybrid approach of the previous two described topologies; a combination of complete replication and mirroring between multiple, geographically dispersed sites with smaller, cache sites fanning out from each of the complete sites. This can be thought of as a "multi-hub and spoke" topology that is often used in very large organizations where requirements necessitate multiple copies of data (often for regulatory reasons) and fast access to data in many sites. Complete copies of data reside in each of the hubs and cached data can be accessed quickly from one of many spoke cache sites. In reality, any of the complete replication and mirroring examples or any of the caching replication and mirroring examples can utilize a hybrid approach. For example, in the case of the hospital, there could be a network of hospitals where each hospital contains a complete copy of all medical images and each of the departments within that given hospital can act as a cache site.

The appropriate use of replication and mirroring will depend on the defined business requirements of the respective data sets – based on the importance of the data to the business. For example, for financial data that is critical to the organization, the type of replication used may be complete replication and mirroring of the critical data sets to another geographically dispersed data center, to protect against disasters, terrorists, etc.

Here are some pros and cons to using replication and mirroring:

Pros:

- Replication can assist in complying with Business Continuity requirements, which often include ensuring that the backups are stored in a remote location, at a sufficient distance to escape any damage from a disaster at the main site.[12]

Cons:

- Both replication and mirroring involves significant resources and adds more complexity to the environment. This is important because some data is not necessarily "critical", and therefore that non-critical data should not be on the same enterprise storage device.

- Replication to a different type of architecture may be cost-prohibitive from a restore perspective, although such replication may be inexpensive from an (initial) storage perspective.

- If using a wide area network (WAN), time may be an issue for restoring, due to the available bandwidth. Latency over a WAN connection is also a common concern.

Best practices for replication and mirroring include using the appropriate implementations of each, for meeting the specified business requirements, specifically in the areas of Business Continuity Management (BCM) planning.

2.1.6 Archive

An archive is a collection of data objects that represent an official working copy of the data, but is managed separately from more active production data, for such purposes as long-term preservation and better cost economics.

Archives are often used for storing data sets that need to meet specific regulations and/or legal/contractual obligations. Archives also play an important role in the data protection technologies that are used in an organization.

Archives are normally used for auditing or analysis rather than for application recovery. After files are archived, online copies of them are typically deleted and must be restored by explicit action. Archives are typically stored in lower cost storage, and this reduces the amount of data that is being backed up from the primary storage (storage for active data sets), which in turn, reduces the time for data backups of the active data sets on the primary storage.

[12] ISO/IEC 27002:2013 Information technology -- Security techniques -- Information security controls

Some common approaches to the implementation of archives include: online, nearline, or offline solutions:

- Online archive: A storage device that is directly connected to a host that makes the data immediately accessible.
- Nearline archive: A storage device that is connected to a host, but the storage device must be "mounted" or "loaded" in order to access the data.
- Offline archive: A storage device that is not ready for use, and some type of manual intervention is required to "connect", "mount", or "load" the storage device before data can be accessed.

To explore the data archive a bit deeper, it is important to understand that an archive is a collection of selected data for the purposes of long-term preservation, but a data archive plays a transformational role in data protection not only from a risk perspective, but also from a governance and compliance perspective.

Archived data must be protected with some form of embedded integrity assurances. However, archived data does not have to undergo a regular process of backup. Backup is a recurring cyclical process; if full backups are run weekly, a file that had not changed in a year would be backed up 52 times even though only a limited number of copies of the file are available at any one time (since only a certain number of copies are retained at any point in time).

A best practice for archiving data is to use some form of replication, whereby all the qualified data can be moved into the archive at the appropriate time.

Although individual files or records do not change, an archive is not static in its contents. An archive has both inflows and outflows. Inflows are simply additive – a new piece of data has been added to the archive. Outflows are subtractive; data is removed from the archive. If the outflow process results in data being migrated to another piece of storage media, the data is preserved on the target piece of storage media (and deleted from the original archive). If no data migration is involved in the outflow process, the archive copy of the data is simply deleted[13]. For the data to be truly destroyed, all other copies of the same data would also have to be destroyed as well. Data sanitization in covered in Section 2.3.4.

Authorized users can access information in the archive for proper uses. The archive can use data protection technologies for both physical and logical protection, and one or more data protection copies of the archive need to be made as well.

[13] Although not necessarily destroyed (sanitization)

Here are some pros and cons to archives:

Pros:

- Archive can be used to address future legal and regulatory obligations.

Cons:

- Recovery/recall of data from an archive can be long (hours to days).
- If specific data potentially exists, the organization may be forced to look for it, e.g., for eDiscovery requests from ongoing litigation.

Best practices for storing data sets in an archive include understanding the statutory, regulatory, and legal requirements for the retention of the specific (relevant) data sets, as well as the associated costs.

2.1.7 Data Protection in the Cloud

Cloud backup refers to backing up data to a remote, storage-as-a-service cloud provider (public, private or hybrid), rather than storing the data locally on a physical medium such as a hard drive, solid state drive or tape.

A cloud backup service is not a pre-defined, fixed solution and must be considered in the overall context of a business data protection or disaster recovery strategy. Cloud-based backup appeals to many businesses because it offers a low-cost way to protect business data off-site but there are multiple considerations to be aware of when planning such an implementation. Considerations for using Data protection in the cloud include:

- Network design (bandwidth, redundancy, latency, etc.)
- Appropriate Service Level Agreements (SLAs)[14], including a termination clause
- Risk Management policies and technologies
- Legal requirements (management, security, location, availability and access management)
- Appropriate mix of technologies (disk, tape, replication to multiple sites, etc.)
- Use of compression, data deduplication, and/or encryption (if a combination of compression and/or deduplication is being used with encryption, then encryption must be done last before data is transmitted)
- Data Security (ensure that the cloud backup provider maintains compliance with both the appropriate data protection laws and the individual organization's security policies).

[14] ISO/IEC 19086-1 Cloud SLA Framework is an excellent reference for Cloud SLA checklist items

Here are some pros and cons to public cloud-based backup as part of the data protection strategy:

Pros:

- Making the data backup expense an operational expense (OPEX) rather than a capital expense (CAPEX), whereby the backups are transmitted to a cloud backup provider, instead of the business purchasing backup equipment, and maintaining an off-site data center.
- The cloud backup scheme can make backup implementation and maintenance very easy to use without burdening the workload for the IT staff. This could also broaden the use of backup strategies by adding things like laptop backup, which can be provided as a user self-service solution, especially for employees in the field.

Cons:

- The restore of data from a cloud-based backup could cause severe issues with required Service Level Agreements (SLAs), for the ability to restore a given data set within a specified Recovery Time Objective (RTO).
- The location of the off-site data may be a problem if your business has to comply with legislation that requires data to remain in country, or a certain physical minimum distance from your primary data center.
- There could be some amount of "control" requirements that may not be able to be met with a certain cloud service provider. For example, there may be a need for specific monitoring and management controls over the backup data that the cloud service provider may not be able to be provide.
- There may be issues of data migration to another cloud service provider (CSP) or back to your own data center, in the case of a termination of your agreement with the cloud service provider or if the CSP goes out of business.

2.2 Accessibility / Availability

In the context of Data protection, "Accessibility" refers to the ability for applications to have appropriate access to data, and the technologies that assist with data's accessibility include: replication, disaster recovery and basic infrastructure redundancies, e.g., dual power supplies, backup generators, multiple LAN/WAN links, etc. The concepts of "Accessibility" and "Availability" are covered in depth, in section 1.1 of this paper.

2.2.1 Replication (Multi-Site)

Some businesses elect to have their data replicated to more than one site, hence "multi-site" replication. This is to ensure even higher availability in case of disasters that may affect an entire geographic area. It also allows for data to be potentially accessed from a physically "closer" data center (e.g., data caching), to reduce latency of data access. The decision to perform multi-site replication will depend on the value of your data versus the cost of implementing this safeguard in order to mitigate the risk of data loss and/or access.

Basic replication concepts are covered earlier, in section 2.1.5.

2.2.2 Business Continuity Management

Business Continuity Management (BCM) involves the processes and procedures for ensuring continued business operations. As such, BCM includes Disaster Recovery (DR) which involves the coordinated process of restoring systems, data, and the infrastructure required to support ongoing business operations after a disaster occurs. But, DR is only one aspect of a business's BCM plan. A BCM plan includes technology, people, and business processes for recovery, after different disaster scenarios occur.

BCM identifies critical business processes and established policies and procedures that should prevent or minimize interruption of these key processes, in the event of unexpected incidents that could negatively impact a business. The goal is to reduce the negative impact of incidents on the business and to mitigate the risk involved in appropriate areas of the business.

The implementation of a BCM plan is often governed by regulations and standards. In some countries, it is a legal requirement to have a plan for handling business-critical operations. The standards provide a methodology that helps to create a functional BC plan.

There are multiple BCM standards, which can be used as BCM guidelines, including the ISO/IEC 27001:2013 Information Security Management Systems, and the North American Business Continuity Standard (NFPA1600) standard, as well as the British Standard for Business Continuity Management (BS25999).

The proper BCM plan will vary from business to business, depending on the regulations, contractual requirements, etc., and the appropriate BCM requirements will be determined by a thorough analysis of the different disaster scenarios. These requirements are often described as values, such as: recovery time objective (RTO), and recovery point objective (RPO). These values will tell the business what the level of redundancy and availability that is needed, as well as how costly it will be to meet these requirements.

Even with having a good BCM plan, this means nothing if the systems that contain the data cannot be reconstructed in an appropriate time frame, so that the organization can continue to function. This is where the concept of resiliency becomes vital. Basic infrastructure resiliency is further discussed in section 2.2.3.

2.2.3 Basic Infrastructure Resiliency

Achieving high availability of data, leads to a requirement of architecting full resiliency[15] into the entire business infrastructure. This means fully redundant hardware, such as dual-controller storage architecture, redundant SAN components (switches, HBAs, NICs), and also redundant servers. In addition, it may be necessary for architecting remote sites, with replication. The key is that everything is at the very least, redundant on a local level from a basic infrastructure perspective.

Basic infrastructure includes things like power supplies, air conditioning within data centers, Network connections, etc. In the case of a power outage, even though a storage system powers down immediately, batteries are usually in place to protect data in cache. Batteries must be adequate to maintain the data in the cache until the data can be flushed from cache to disk. This avoids the potential loss of the data that was not written to disk before the loss of power.

To overcome short power outages, uninterrupted power supply (UPS) units are usually installed. UPS units are equipped with large powerful batteries, and provide electricity from the batteries for the data center while the power is out. Some businesses use redundant power sources via two separate grid power circuits, and/or some businesses will have power generators, where diesel engines are a common power source that lasts much longer than batteries.

The air conditioning (A/C) in a Data Center is necessary since IT equipment produces a great deal of heat, and if there is a disruption in the functioning of the A/C, then the IT equipment can incur damage due to overheating. The IT infrastructure should have triggers to automate the shutdown of the equipment if the ambient data center temperature rises to a specified level. For this reason, the A/C units should be redundant, along with the proper maintenance schedule for each respective A/C unit.

[15] The SNIA Dictionary defines resiliency as: "The ability of a storage element to preserve data integrity and availability of access despite the unavailability of one or more of its storage devices."

Best practices call for Internet connectivity that is contracted from at least two separate connectivity providers. In addition, if providing private links between buildings, across campus, or to the other end of town, a minimum of two redundant links should be installed, and preferably redundant links configured via different paths to the other data center, to avoid the case of a single link getting severed.

2.3 Compliance

In the context of Data Protection, there are certain issues that need to be addressed in order to meet specific "Compliance" requirements. This includes the application of specific technologies that allow for the ability to secure data for meeting the appropriate rules and regulations typically related to data retention, authenticity, immutability, confidentiality, accountability and traceability, as well as the more general problem of data breaches. To address the common issues deployed to meet compliance requirements, this section focuses on the technologies and feature sets, including: archives, retention period support, deduplication, storage encryption, data sanitization (electronic data shredding), and monitoring and reporting.

2.3.1 Data Retention and Disposition

There are concepts of archiving data at various lengths of time, e.g., short, medium and long-term. Archives of any length must ensure proper integrity, immutability, authenticity, confidentiality and provenance. This topic of Archive is covered in greater detail above, in Section 2.1.6.

The SNIA best practices for data retention are to keep a minimum of three copies of each data set, on at least two separate types of media, and a minimum of one offsite copy. This need for multiple copies is to maintain data integrity and data availability.

Retention period support for an archive is the ability to set up a period of time that the data set is to be saved (retained), and after that time has lapsed, automatic deletion of that data set takes place.

The appropriate retention period for each respective data set will be dependent upon the regulations that are in place for any particular business.

For "legal holds", the data is to be "frozen", such that no modifications, versions, or deletions can take place on the specified data set.

Media sanitization techniques are often used to make sure that the data is removed at the end of the retention period. Section 2.3.5 below will cover the topic of Data Sanitization.

The appropriate Retention period for each data set will vary from business to business, depending on the specific regulations, contractual requirements, etc., that a given business must abide by. The

retention period requirements are often described within specific regulations, such as (US)–HIPAA[16] (Health Insurance Portability and Accountability Act), (US)–SEC (Securities & Exchange Commission), etc.

Best practices for retention period support includes understanding what the specific regulations and other contractual commitments are, and then setting the retention period for each affected data set.

2.3.2 Data Authenticity and Integrity

There may be certain data sets that will need to be kept in order to meet specific legal and/or regulatory compliance requirements, and therefore some additional requirements will need to be taken into consideration. For example, for data that has been deduplicated, the deduplication process creates hash tables that need to be retained along with the data. So, the hash tables must be protected in addition to the data itself, otherwise, the data cannot be "un-deduplicated" or "rehydrated" for use, when the data is to be subsequently read by a user or by an application.

Also, when deduplicating data prior to replicating to another system locally or to another system off-site, the data will need to be re-hydrated upon restore, so that it can be used, as described above.

Another example includes some forms of metadata[17], which could be critical to the future usability of a data set for a particular purpose, therefore critical metadata may need to be treated the same way as the data itself.

2.3.3 Data Confidentiality

At the storage level there are multiple potential threats to data confidentiality, including tampering with the data, which violates data integrity. Another threat is storage media theft, which violates both data confidentiality and data integrity.

One tool for securing data from unauthorized access is encryption. The data can be encrypted while being transferred to the storage media, often referred to as "encryption in flight", and/or the data can be encrypted on the storage media itself, often referred to as "encryption at rest". As a general rule, data should be encrypted as close to the data origin as possible.[18] For example, an application, or an encryption appliance can be used for encrypting data prior to being written to the storage device, thereby having the ability to encrypt the data in flight. However, many times this is not possible, such as when specific applications do not have the ability to encrypt data.

---

[16] Title II of HIPAA defines policies, procedures and guidelines for maintaining the privacy and security of individually identifiable health information as well as outlining numerous offenses relating to health care and sets civil and criminal penalties for violations.

[17] Metadata is: Data that defines and describes other data. [ISO/IEC 11179-1:2015]

[18] SNIA-Data Encryption & Key Management White Paper (https://www.snia.org/sites/default/files/technical_work/SecurityTWG/SNIA-Encryption-KM-TechWhitepaper.R1.pdf)

Another consideration for encrypted data is the encryption keys. The encryption keys need to be protected, since if the encryption keys are not available, then the data that is encrypted is also not available.

Yet another consideration is if backup data is to be compressed and/or deduplicated as well as encrypted, always make sure that the encryption process takes place last. Otherwise, the backup data will not have any possibility of benefiting from space optimization via compression and/or deduplication.

The SNIA best practices for maintaining data confidentiality include a thorough review of the sensitivity of each data set. To aid in the classification of data sets, a simple data classification scheme can be that, at a minimum addresses the sensitivity of data and whether the data set is production versus non-production data. This helps with determining where each respective data set will ultimately reside, thereby ensuring the appropriate confidentiality.

The table below shows that this would result in four unique classifications:

	Production Data	Non-Production Data
Sensitive	Data Set #1, #3	Data Set #2
Non-sensitive	Data Set #4	Data Set #5…

Production data in any of its forms would take priority over non-production data. Note that in reality, there may be a need to classify the colored boxes with finer granularity, based on the requirements of the organization. This classification also allows for the use of different storage types, which could save money on both the overall storage architecture, and on how the data is protected. This classification also allows, for example, in the case of a data breach, for there to be a legally defensible position. Provided that this classification is a part of the organization's policies and procedures, this goes to proving that due diligence is being exercised.

When using encryption as described above, the encryption keys must also be protected. The best practices for the protection of encryption keys is to back up the keys often and every time there is a change to the media and/or the encryption keys. Manual backups of these encryption keys can become quite cumbersome to manage and is subject to human error, so another best practice is to implement an enterprise key management system, whereby the backups of the encryption keys take place automatically.

In addition, the organization needs to consider the regulations that the business needs to abide by, when deciding which technologies to use for maintaining data confidentiality.

2.3.4 Data Sanitization

It is important to ensure that data is cleared at the end of its lifecycle, which can be referred to as Electronic Data Shredding. Data sanitization involves one or more methods to ensure that data is completely erased from the storage hardware, and cannot be reconstructed.

The purpose of sanitization is to sanitize storage media. "Sanitize" means rendering the data on the media infeasible to recover for a given level of effort. The "level of effort" is an important concept as it will motivate decisions regarding the sanitization process and its implementation. The more effort an adversary might be willing to invest in recovering it, the more careful an organization must be in its choice of a sanitization method.[19]

In addition to electronic data sanitization, there is also physical media destruction, which involves the physical destruction of data storage media, through crushing, pulverizing, melting, etc.

One example of where data sanitization arises is in the European Union (EU). The EU has new data protection legislation that covers some of the privacy concerns of the EU members. One of the EU regulations includes the "right to be forgotten", which gives EU members the ability to require the erasure of their personal data without undue delay by the data controller in certain situations. So, if the EU member withdraws their consent and no other legal ground for storing or processing of their personal data applies, then all of their associated personal data must be "sanitized".

Best practices show that a careful review should be done regarding the decision of whether to keep a particular data set for longer than required by the regulations that the business must abide by. One of the considerations to take into account is the potential value of the data for analytical purposes versus its being a potential liability because of legal risks. For example, some businesses delete all email that is older than one year, even though there are no regulations for their business that require this, because the business understands that emails could be used in future litigation, potentially costing the company losses in revenue, productivity, reputation or even worse: going out of business. The data classification of each respective data set will determine (based on policy) whether the data needs to be deleted (shredded/erased), and at what level of data shredding/erasure that is needed, in order to meet the business and/or regulatory requirements. Proper documentation of the sanitization activities may also needed. (See Section 2.3.5). This classification should be codified in a policy and the organization will be held accountable for the organization following the policy. The failure of an organization not adhering to the policy can get the organization into legal trouble.

[19] SNIA – Storage Security: Sanitization White Paper
(https://www.snia.org/sites/default/files/technical_work/SecurityTWG/SNIA-Sanitization-TechWhitepaper.R2.pdf)

2.3.5 Monitoring, Auditing and Reporting

As it relates to data protection, the monitoring and alerting activities involve gathering information on the elements and services that have access to the business data. Then, reporting on those actions, and finally taking the appropriate steps in case of any type of data breach, data tampering, etc.

Logging records events into various logs, and monitoring reviews these events. Combined, logging and monitoring allow an organization to track, record, and review activity, thereby providing overall accountability.

There is often a need to show proof of sanitization and proof of encryption, which is sometimes referred to as “Proof of Service”.

Audit logging is used for accountability, traceability and provenance. An example of this is HIPAA, where logging is required, so that all events related to an object will be recorded. Auditing of the events typically takes things further, requiring the inspection of individual events in an environment for compliance.

Best practices include that audit records are used to enable the monitoring, analysis, investigation, and reporting of unlawful, unauthorized, or inappropriate information system activity. This will ensure that the actions of each user can be uniquely traced to that specific user so they can be held accountable for their actions.

3 Summary

Data protection of digital data is a fundamental and mandatory responsibility for all organizations. Therefore, organizations need to understand the basic principles and concepts of data protection. To satisfy that need, this whitepaper has provided an overview of the relevant best current practices for data protection, as defined by the SNIA's Data Protection & Capacity Optimization (DPCO) Committee on behalf of SNIA. As discussed in this paper, there are many factors to consider when it comes to data protection at the storage level. The three main areas that were covered fell into three data protection "drivers":

1. Data Corruption and Data Loss
2. Accessibility and Availability
3. Compliance

Protected data must meet intended uses for all three drivers. Preventing data corruption and data loss ensures that the data is what the organization expects it to be when the data needs to be used. Accessibility and availability relate to the data being made available in a timely manner for intended uses. Compliance ensures that the data usage meets all legal and regulatory requirements.

1. Data Corruption and Data Loss

Data must be protected both logically (such as to prevent data corruption from hacking or other external threats) and physically (such as data loss, such as the irreversible failure of a storage device). Physical prevention of data loss from hardware failure on a random-access storage system can use techniques such as RAID or erasure coding.

Backup and recovery are two of the traditional cornerstones to data protection for both physical and logical reasons. Backup relates to the processes of providing a copy of the data at a point in time and recovery refers to the ability to restore data for intended application use according to the organizational SLAs. One approach on a storage system itself is through the use of snapshots. These snapshots may serve as the basis for the data that is copied to a backup target storage system, but snapshots are not always used. Others approaches include the use of Continuous Data Protection, or to use a public or private cloud as a backup service.

Replication and mirroring are also used to make copies of data. As used in this paper, replication refers to point in time copies whereas mirroring provides for continuous writing of data to two or more targets. Replication may be used for both physical and logical data protection while mirroring is a physical data protection approach.

An archive is an official set of more or less fixed data that is managed separately from more active production data. As such copies have to be made for data protection purposes, but more active measures, such as standard backup or mirroring are not necessary.

2. Accessibility and Availability

For accessibility and availability, Business Continuity Management (BCM) includes the processes and procedures for ensuring ongoing business operations. One key aspect of BCM is Disaster Recovery (DR), which involves the coordinated process of restoring systems, data, and the infrastructure required to support ongoing business operations after a disaster occurs. But a BCM plan also includes technology, people, and business processes for recovery.

As part of accessibility and availability, basic infrastructure redundancies need to be provided, including UPS systems to provide redundancy for power in case of a power outage and extra network and power connections.

3. Compliance

Compliance includes the application of specific technologies that allow for the ability to secure data for meeting the appropriate rules and regulations typically related to data retention, authenticity, immutability, confidentiality, accountability and traceability, as well as the more general problem of data breaches. There are a number of technologies that relate to compliance including:

- Long term retention of archival information is useful for integrity, immutability, authenticity, confidentiality, and provenance purposes.
- Encryption provides support for confidentiality and integrity reasons.
- Data sanitization, i.e., electronic data shredding, provides for the proper deletion of data at the end of its life-cycle.
- Monitoring and reporting gathers access information to determine if data tampering or data breaches have been attempted or have taken place.

SNIA Positions on Data Protection Best Practices

- For critical data, the SNIA's recommendation is to maintain at least 3 copies (one primary and two secondary) of each data set across at least two geographically disparate locations, with at least one write-protected copy, preferably isolated and on different media for business continuity purposes. The overall goal is to ensure that the appropriate data set(s) can be restored in the case of any type of disaster, including an entire data center becoming unavailable, within the specified restoration goals of your organization for each respective data set. The level of data protection described is specified as a recommendation for critical data, as this may not be cost effective for non-critical data.

- For backups, industry tradition has called for a retention/recycle time of one month, but it may be better to consider shorter timeframes. Here are some points to consider in deciding how long to keep backup data:
    a. Backups are typically considered as a short-term recovery mechanism of data in the case of an accidental deletion or incident
    b. If the data backups are kept for longer periods of time, in some jurisdictions, backup sets may be subject to electronic discovery
    c. If the backup data sets are deleted too quickly, then there is exposure to compromises such as ransomware, because the restore points become limited

- For sensitive data that needs to be saved for longer periods of time (longer than the backup retention timeframe), placing that data set in an archive is more suitable. This is because there needs to be specific controls (e.g., encryption, WORM, etc.) on how that data is stored, based on the regulatory and policy requirements that the organization is subject to. There are also considerations for how the data is destroyed (e.g., cryptographic erasure, degaussing, etc.)

- For maintaining data confidentiality best practices include a thorough review of the sensitivity of each data set. To aid in the classification of data sets, a simple data classification scheme has been proposed:

	Production Data	Non-Production Data
Sensitive	Data Set #1, #3	Data Set #2
Non-sensitive	Data Set #4	Data Set #5…

Note that there may be a need to classify the colored boxes with finer granularity, based on the specific requirements of the organization.

4 Acknowledgments

4.1 About the Authors

Thomas Rivera, CISSP has over 30 years of experience in data storage, with specialties in data protection and data privacy. Thomas is currently a data security and privacy consultant and was most recently a Senior Technical Associate in the Emerging Solutions Group at Hitachi Data Systems. Thomas also co-chairs the SNIA's Data Protection and Capacity Optimization (DPCO) Committee, and is an active member of SNIA's Security Technical Working Group, along with serving as the secretary on the SNIA Board of Directors. Thomas also serves as the secretary for the Cybersecurity & Privacy Standards Committee within IEEE.

Gene Nagle has over 30 years of experience in data storage, primarily in product management and applications engineering, and is currently Director of Technical Services at BridgeSTOR, managing the technical aspects of the sales and marketing of their cloud storage products. Gene has been active with the SNIA since it's founding and currently serves as co-chair of the SNIA's Data Protection and Capacity Optimization (DPCO) Committee, and is also a member of SNIA's Long-term Retention Technical Working Group (LTR TWG).

Mike Dutch worked in the computer storage industry since 1980 for IBM, Hitachi Data Systems, Troika Networks, Veritas (Symantec), and EMC (Dell). He was a manager and an individual contributor working as a mainframe developer, field consultant, product manager, and standards contributor (http://www.snia.org/about/profiles/dutch_mike). Mike holds over 50 patents in the data protection space and most recently worked at Dell EMC as a Technical Staff member in software engineering, and is now retired.

4.2 Reviewers and Contributors

The (SNIA) Data Protection and Capacity Optimization (DPCO) Committee would like to thank the following individuals for their contributions to this whitepaper:

Richard Austin, CISSP (retired)	Hewlett Packard Enterprise
Michael Dexter	Gainframe
Eric Hibbard, CISSP	Hitachi Data Systems
David Hill	Mesabi Group
Tim Hudson	Cryptsoft
Glen Jacquette	IBM
John Olson, PhD	IBM
Ronald Pagani	Open Technology Partners
Tom Sas	Hewlett Packard Enterprise
Gideon Senderov	Ciphertex Data Security
Gary Sutphin	
Paul Talbut	SNIA

5 For More Information

Additional information on the SNIA Data Protection activities, including the SNIA DPCO Committee, can be found at http://www.snia.org/dpco.

Additional SNIA materials related to data protection and capacity optimization can be found at: http://www.snia.org/dpco.

Suggestions for revision should be directed to http://www.snia.org/feedback/.

About the SNIA

The Storage Networking Industry Association is a not-for-profit global organization, made up of member companies spanning the global storage market. SNIA's mission is to lead the storage industry worldwide by developing and promoting vendor-neutral architectures, standards and educational services that facilitate the efficient management, movement and security of information. To this end, the SNIA is uniquely committed to delivering standards, education, and services that will propel open storage networking solutions into the broader market. For more information, visit www.snia.org.