Data Protection: Where It Is Broken, Why It Is Getting Worse And How To Fix It

A Comprehensive Data Protection Solutions Methodology

By Hitachi Data Systems

May 2014
Contents

Executive Summary .................................................................................................................. 2

Introduction ............................................................................................................................ 3

Amount of Data That Needs to be Protected ......................................................................... 3
Protection of Large Objects .................................................................................................... 3
Protection of a Large Number of Small Objects ..................................................................... 3
Long Retention .......................................................................................................................... 3
Multiple Copies ......................................................................................................................... 3

A Bad Situation Is Getting Worse ......................................................................................... 6

Data Growth ............................................................................................................................. 6
Server Virtualization ............................................................................................................... 6

A Comprehensive Approach to Data Protection ..................................................................... 7

1. Adopt Service-Level Protection ......................................................................................... 8
2. Reduce the Amount of Data to Be Protected ..................................................................... 9
3. Enable Application Intelligence and Consistent Recovery .............................................. 9

Comprehensive Solution in Practice ...................................................................................... 11

The Data Protection Journey ................................................................................................ 12

Summary .................................................................................................................................. 13

Resources ................................................................................................................................ 14

Storage-Based Data Protection Solutions ............................................................................. 14
Host-Based Data Protection Solutions ................................................................................... 14
Executive Summary

The objective of this white paper is to discuss the challenges storage administrators face when trying to protect data in their environments. It considers inadequacies of existing processes and industry trends that are likely to exacerbate these issues, and then profiles the comprehensive Hitachi Data Systems solution to affectively address these challenges. The intended audience for this document is storage and data protection practitioners.

For the most part, data protection is presently synonymous with backup and recovery, and few organizations, if any, are happy with their existing environments. This white paper advocates taking a holistic view by leveraging other technologies, such as snapshots, replication, archiving and object storage.

Traditional backup and recovery is still a viable option for large portions of the data management spectrum, where reasonable amounts of downtime and data loss are tolerable. However, this tolerance is not the case for the more important and critical applications and data that the organization relies on to complete its mission. For those applications, downtime and data loss can be measured in significant costs, even to the extent of threatening the continued existence of the organization.

It is true that the compounding rate of data growth, at 40% to 60% or more per year, is straining scalability and performance limits of traditional backup solutions. It is also true that tight service level objectives of critical applications are creeping down to the more standard data sets, further challenging legacy backup solutions. However, the largest threat to the incumbent solutions has been the rapidly increasing complexity and distribution of information technology. Just in the past 10 years, we've seen the following changes that require new approaches to data protection:

- Server virtualization has moved, in a big way, into the production environment.
- New applications and compute categories have come to market, such as SAP HANA, big data and analytics.
- Application and storage users, and their end point devices (smartphones, tablets), have become mobile.
- Globalization efforts, through the development of remote and branch offices, have been met with many country-specific data security and privacy regulations.

Each of these changes has led data protection process owners to find new approaches, often in the form of individual point solutions, to meet the new challenges. But each additional point solution only compounds the complexity of the environment and adds incremental hardware, software and management costs. The “fix” really isn’t any more sustainable than the previous solution.

MEASURING DATA PROTECTION EFFECTIVENESS

Service level objectives for the protection and recovery of applications and data can be measured in 4 ways:

- Backup Window: the amount of time available to perform data protection processes.
- Recovery Point Objective (RPO): the amount of time between backups; equals the amount of new data at risk.
- Recovery Time Objective (RTO): the amount of time to restore operations following any outage or data loss event.
- Total Cost of Ownership (TCO): the costs of all aspects of data protection infrastructure, software, services and processes.
Introduction

In the situations where backup isn’t working, administrators generally are unable to meet the shrinking protection (backup) window. They can’t meet recovery time and recovery point objectives (RTO and RPO) and/or operate the environment within the available budget. These challenges primarily stem from:

- Amount of data to be protected.
- Protection of large objects.
- Protection of a large number of small objects.
- Long retention times.
- Multiple copies.

It is becoming increasingly clear that data protection approaches of the past, based on backup as a separate, brute force, one-size-fits-all approach, no longer suffice. They are unable to meet all current, much less future, recovery needs.

Amount of Data That Needs to be Protected

Increasingly, more and newer types of data are being added to the mix that needs to be protected. Until recently, laptops, desktops, remote offices and testing and development environments were not generally included in enterprise data protection processes. Due to a number of government regulations and litigation, business-critical information, which is often stored in a widespread manner, now needs to be protected adequately.

Protection of Large Objects

Protecting a single large object becomes difficult, as there is no easy way to break the object into smaller pieces before copying it (backing it up). For example, copying an 84TB dataset over a single 10GB connection takes 24 hours, which makes protecting a larger dataset on a daily basis virtually impossible. If your business cannot afford to lose more than 1 hour of this data (the RPO), you can see that you have a big problem.

Protection of a Large Number of Small Objects

Sequentially opening, reading, copying and closing a large number of objects (files) takes an inordinate amount of time. In file systems with a large number of files (tens of millions of files), this process could take several hours and exceed the backup window. For example, assuming 100 files can be backed up every second, over a 24-hour period, only 8.6 million files can be backed up in a single stream.

Long Retention

Given existing technologies, it is difficult to ensure that data can be reliably recovered after long retention periods, such as 20 or more years. Every electronic storage medium, whether it be disk, tape, optical or flash, has a limited lifecycle. Plans and processes must be in place to migrate long-term data to newer technologies every few years. The media itself may last for longer periods, but the read/write mechanisms, operating systems, applications and other software may not.

Multiple Copies

Figure 1 assumes full backups (100% of the data) on weekends and incremental backups (5% change rate) on weekdays. It also depicts an industry-standard 90-day retention. In this scenario, administrators need 40TB to protect 1TB of data. Assuming a 10x deduplication rate, this requirement can be brought down to 4TB. Having multiple copies (onsite, offsite, and so forth) or longer retention magnifies the amount of storage needed, which increases complexity, floor space and costs.
Figure 1. Full Backups Generate Massive Amounts of Duplicate Data (With a 10x deduplication rate, the number of terabytes required to protect storage can be reduced significantly.)

<table>
<thead>
<tr>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weekend Full 100%</td>
<td>Mon Incr</td>
<td>Tue Incr</td>
</tr>
<tr>
<td>1TB</td>
<td>50GB</td>
<td>50GB</td>
</tr>
<tr>
<td>1TB</td>
<td>50GB</td>
<td>50GB</td>
</tr>
<tr>
<td>1TB</td>
<td>50GB</td>
<td>50GB</td>
</tr>
<tr>
<td>Total 90 days</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De-dup 10:1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Backup to Disk Storage Consumption: 1.62TB
With Replication: 3.21TB
Raw Storage (20% RAID overhead): 4.1TB

Incr = incremental, De-dup = deduplication

A more efficient model for capturing backup data is one that employs an incremental-forever technology. In this model, a full backup is taken on the first instance, but then only changes and new data are captured on subsequent backup jobs. This method can reduce the amount of raw backup disk consumption by 90% or more after just a few months.

The key is to find an incremental-forever solution that can perform a single-pass restore. Most legacy backup solutions need to restore the last full backup set, and then each subsequent incremental backup set, in order, up to the desired recovery point.

Hitachi Data Systems offers several incremental-forever data protection solutions:

- Snapshots: Hitachi Thin Image supplemented by Hitachi Data Protection Suite with IntelliSnap application-consistent snapshot management.
- Continuous Data Protection (CDP): Hitachi Data Instance Manager captures changes at a block level (Microsoft® Windows® only).
- Synthetic Full Backup from Incrementals: Symantec NetBackup with NetBackup Accelerator performs file-level incremental backup, and creates a synthetic full backup in the background.
The benefits of utilizing an incremental-forever model, both in saving backup storage space and in simplifying the recovery process, are shown in Figure 2.

**Figure 2. Incremental-Forever Backup Reduces Backup Storage and Speeds Recovery**

This example is a bit extreme, with a 10% per day change rate, but it makes it easy to see the difference between backup approaches. The legacy backup application takes an initial full backup (100TB), then incremental backups each day of the week, then another full backup the next weekend. It needs to take this approach to make the restore possible.

Each weekend, most of the data in the full backup is the same as the previous week. The more backup sets that you retain, the more duplicate data you'll have to store. This old model forces the periodic full backup because of the way it performs restores: If a server crashes on Friday, it needs to restore the last full backup and then the incremental data sets from each weekday.

Since active files are changing frequently, different versions of the same file can exist in multiple backup data sets. Each version will have to be recovered and overwritten during the multistep recovery process. This approach adds time and complexity, and therefore also adds risk.

In the incremental-forever model, an initial full snapshot is also taken, but that's the last one. Every subsequent snapshot captures only new or changed data (files or blocks). Therefore, there is no duplicate data in the backup set, unless copies of the same file are captured from multiple sources. Deduplication can be used in these cases.

The secret to the power of an incremental-forever protection solution is in the index database: It knows which blocks in the backup repository are needed to restore data to any recovery point, and can pull that data together in a single pass.
A Bad Situation Is Getting Worse

The challenges noted in the previous section are further exacerbated by explosive growth of data and the widespread adoption of server virtualization into production environments.

Data Growth

According to IDC (The Digital Universe of Opportunities, April 2014), overall data volumes will continue to grow at 40% per year into the next decade. It will grow from 4.4 zettabytes (4.4 trillion gigabytes) in 2013 to 44ZB in 2020. That’s a 10x increase in only 7 years.

However, also according to IDC, 90% of the digital universe is made up of “dark matter,” data that is looked at only once, if at all. Properly categorizing data can help avoid overspending on data protection and other data management tasks.

Server Virtualization

This technology is becoming mainstream, even in production application environments, due to several tangible benefits. It supports better asset utilization, which leads to lower infrastructure capital costs, floor space and HVAC (heating, ventilation and air conditioning) consumption. Unfortunately, one of the significant challenges organizations face when operating a virtual environment is the difficulty associated with protecting them effectively. This challenge is primarily due to the lack of available physical resources in the servers to handle multiple, concurrent protection and recovery processes.

Traditional data centers revolved around physical servers. The amount of data per server was limited and the servers had plenty of unused processing cycles available for other tasks, such as backup. Traditional data protection solutions deploy resource-intensive backup agents in the physical server, which copy and move data from production storage to a back-end disk or tape target. This approach worked well for limited storage capacity and on servers with plenty of processing cycles available. Unfortunately, a virtual environment eliminates most of the excess physical resources, such as CPU cycles and network bandwidth.

Deduplication, essentially source side, can be an effective way to move data in a shorter window; however, its value in a VMware environment is limited as it needs even more CPU cycles for the deduplication process.

Figure 3 provides an example. Assume that 5 physical servers are virtualized. Before virtualization, cumulatively, 5 sets of separate network connections and CPUs were available to process the backups. Unfortunately, after virtualization there is less than 1 (as the utilization of even the 1 set of CPU and network becomes higher), but the cumulative data volume has not decreased. In effect, what is being asked is to back up 5x the amount of data with less than 1x the resource.
A Comprehensive Approach to Data Protection

To solve these problems, Hitachi Data Systems has developed a 3-pronged approach to meeting data protection objectives. The prongs include shrinking the backup window, driving RPO and RTO toward zero, and reducing costs. This approach is shown in Figure 4.

Figure 4. The HDS 3-Pronged Approach to Effective Data Protection and Recovery

1. **Adopt Service Level Protection**
   - Establish tiers of protection service levels to meet objectives
   - Leverage the appropriate technologies in each tier
   - Tune objectives for operational, disaster and long-term recovery

2. **Reduce the Volume of Data to be Protected**
   - Move inactive data to a content store with built-in protection
   - Reduce the amount of storage needed for protection
   - Less data requires less time for protection and recovery

3. **Enable Application Intelligence and Consistent Recovery**
   - Places application in backup-ready, quiesced state using standard APIs
   - Leverages hardware-based snapshot and cloning technologies
   - Fast, non-disruptive, incremental-forever operational recovery
1. Adopt Service-Level Protection

Most organizations follow a policy of full backups on the weekends and incremental backup on weekdays to protect all data. This “one-size-fits-all” approach is increasingly becoming inadequate, as all data is not equal in importance.

Hitachi Data Systems recommends a tiered protection approach that is based on service-level requirements of the data and is focused on recovery objectives. Administrators protect data to recover for 3 broad reasons. Each reason requires different technologies that are optimized for that specific recovery type.

**Operational Resilience** enables operations to continue, either automatically or with minimal manual intervention, following planned or unplanned events that impact local operations. This approach is often the most expensive path to recovery as it entails the need for standby hardware, software and data in a separate location.

**Operational Recovery** includes recovery from operational issues, such as inadvertent deletion, localized hardware or software failure, and malicious actions such as virus attacks. It is the most common form of recovery in data protection operations.

**Disaster Recovery** includes recovery from catastrophic site disasters such as fires, floods, earthquakes and tsunamis. Fortunately, this is infrequently required, though it is highly difficult and expensive, and usually involves moving operations to an alternate data center. In addition to systems, software and data, you may also need to consider staffing requirements.

**Long-Term Recovery** includes recovery of data that is retained for very long periods of time, such as 20 or more years. The reasons for retaining data long-term are many. Organizations must meet regulatory and corporate governance mandates, to enable data mining and big data analytics, to support product lifecycles and warranties, for historical preservation purposes, and more.

In addition, for each of these recovery types, it is recommended that organizations protect the more valuable data more aggressively than less valuable information. In this way, they will help reduce the risk for the higher value data and reduce the cost of protection for data with lesser value to the organization. Table 1 breaks this categorization of importance into 3 tiers, and lists the suggested best practice approach for each. This list should be customized for each individual environment and its service level objectives.

<table>
<thead>
<tr>
<th>Protection Objectives</th>
<th>Tier 1 Critical Data</th>
<th>Tier 2 Less-Critical Data</th>
<th>Tier 3 Noncritical Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational Resilience</td>
<td>Active-active storage clusters</td>
<td>Active-passive clusters</td>
<td>Multiple access points</td>
</tr>
<tr>
<td>Operational Recovery</td>
<td>Frequent copies on arrays</td>
<td>Copies on-site, on disk</td>
<td>Copies on-site, on disk or tape</td>
</tr>
<tr>
<td>Disaster Recovery</td>
<td>Real-time off-site copies</td>
<td>Near-real-time off-site copies</td>
<td>Periodic off-site copies</td>
</tr>
<tr>
<td>Long-Term</td>
<td>Replicated content store</td>
<td>Content store with local protection</td>
<td>Content store</td>
</tr>
</tbody>
</table>
2. Reduce the Amount of Data to Be Protected

Typically, no more than 40% of data in a production application server is active; the remaining 60% may still be needed, but hasn't been accessed or modified for a long time. Why not move it off of expensive Tier 1 storage and into a content store like Hitachi Content Platform that has built-in data protection and disaster recovery capabilities? This change will reduce the load on the production servers and slow the need to acquire more Tier 1 storage capacity, reduce the amount of time to perform backups and restores, and reduce the costs of the backup repositories and disaster recovery services. All these improvements can be accomplished transparently to applications and application users.

As depicted in Figure 5, by moving inactive data to a content storage platform, the amount of data that needs to be protected is reduced. As a result, less time is needed to perform data protection processes (the backup window). It also reduces the data protection infrastructure needed. However, the critical step in this solution is choosing a content platform that has built-in data protection. Without such a platform, the problem is only moved and not solved, as the archive target needs to be further protected.

Figure 5. Reduce protection costs by moving inactive data to a self-protected content store.

3. Enable Application Intelligence and Consistent Recovery

Taking a one-size-fits-all approach to blindly back up everything in the environment may seem like a simple method, but it will usually result in disaster when something goes wrong. Each different application and operating system, as well as virtualized systems, must be handled individually, using the tools, interfaces or processes prescribed by the vendor of those products. That's why enterprise data protection solutions have so many options.

A few common examples of backup-related tools and interfaces include:

- VMware: vSphere APIs for Data Protection (VADP) enables protection of virtual machines (VMs) without the need to install an agent on each VM.
SAP: BackInt is the interface that communicates between the backup application and the SAP BR*Tools utilities for backup, archive and restore of databases, control files and redo logs.

Microsoft® applications: Microsoft Volume Shadow Copy Services (VSS) quiesce Microsoft SQL Server®, Microsoft Exchange and other applications for snapshot and replication operations.

Storage-system-based snapshots and replication have become the modern methods for enhancing application availability, as measured by RPO, RTO and backup windows. These snapshot and replication capabilities offer a range of advantages over traditional backup methodologies.

Hitachi Thin Image (HTI) software, for example, provides industry-leading levels of backup and recovery performance, scalability, reliability and functionality. It is available with the Hitachi Virtual Storage Platform family and Hitachi Unified Storage VM (HUS VM).

However, native array snapshot tools have varying degrees of automation and application awareness. For example, a database application may not write all parts of a new transaction to disk at the same time. Creating a snapshot without the full set of data can result in an inconsistent copy and cause database corruption at the time of a restore.

The IntelliSnap feature of Hitachi Data Protection Suite (HDPS), powered by CommVault, brings application awareness to snapshots. CommVault IntelliSnap is deeply integrated with Hitachi Thin Image. Together, they consolidate, automate and manage snapshot-based protection and rapid recovery of applications, systems, VMs and data.

- Recover applications faster to increase business uptime.
- Maximize investments in Hitachi virtual storage with centralized snapshot management.
- Eliminate snapshot-related scripting and management complexity, align to more stringent service level agreements (SLAs), and accelerate test and development operations.
- Reduce risk by eliminating the backup window and increasing the frequency of backup.
- Reduce IT costs by eliminating multiple point products and management silos with a single software platform.
Comprehensive Solution in Practice

Consider how this strategy can be applied in an organization’s storage environment. Assume that an organization has a very large Microsoft SharePoint environment and is struggling with current protection, which includes full backups on weekends and incremental backups on weekdays. The application stakeholders have determined that the RTO and RPO for this application are 8 hours. They also determined that the data needs to be retained for 10 years. For this application, a possible solution could be:

1. Use Hitachi TrueCopy to replicate the SharePoint data in Hitachi Virtual Storage Platform to a disaster recovery site to meet the disaster recovery requirements.

2. Use Hitachi Storage Optimization for Microsoft® SharePoint® to archive data not accessed in the last 30 days to Hitachi Content Platform (HCP). Storage Optimization for Microsoft SharePoint can also migrate binary large objects (BLOBs) out of the SharePoint databases, improving their utilization and cutting down on the number of database instances needed. This efficiency will ensure that only active data is being actively protected. With built-in data protection through constant self-management, self-correction and seamless technology refresh ability, HCP can ensure that data remains accessible. HCP can keep it in an unaltered form for much longer than 10 years, meeting long-term recovery objectives.

3. Leverage the intelligent application-consistent snapshot management of Hitachi Data Protection Suite (HDPS) IntelliSnap with the Hitachi Virtual Storage Platform family. Take a snapshot of the SharePoint data every 8 hours and then move it to a disk storage system to meet operational recovery requirements. HDPS application integration allows granular recovery of data such as documents, lists, templates and more.

Another organization has a broad mix of Microsoft and Oracle applications, including Exchange, SQL Server and SAP on Oracle. The associated data sets have varying recovery point, recovery time and retention objectives, based on data importance. The more critical and important applications, including email and customer resource management (CRM) systems need to be available continuously, eliminating the possibility of a traditional backup window. Other applications can be stopped during the evening. They also have a number of remote offices and mobile employees that generate important data.

A possible solution for this organization could be:

1. Use Hitachi Data Protection Suite, powered by CommVault, as the core data protection platform. Agents can be deployed to perform automated, scheduled backup of standard applications, remote offices, as well as employee desktop and laptop systems.

2. Deploy HDPS IntelliSnap technology, with Hitachi Thin Image hardware-based snapshot technology in HDS virtual storage systems. This combination allows you to perform fast, frequent and nondisruptive application-consistent protection of the critical Microsoft and SAP applications. These snapshot copies can then be replicated off-site by Hitachi TrueCopy or Hitachi Universal Replicator for protection against local disasters.

3. Leverage the integrated archive module of HDPS to move historical, long-term retention information from the backup repository to Hitachi Content Platform. There is no need to recapture this data from the production systems; HDPS manages the scheduling and movement based on business-defined policies.
The Data Protection Journey

Hitachi Data Systems can also help organizations throughout the data protection transformation process with the right levels of technology, operations and skills. HDS offers a broad range of consulting, implementation and management services that help ensure a successful transition and return on investment.

Figure 6 describes the step-by-step journey that HDS helps customers travel to achieve transformation to an efficient data protection environment.

Figure 6. The Journey to Build an Efficient Data Protection Strategy

Phase 1: Virtualized Infrastructure
- Storage / server standards
- Transformation roadmap
- Infrastructure planning, implementation and management
- Data migration

Phase 2: Data Management Standards
- Service management framework standards
- Data protection
- Data replication
- Data security
- Data management
- Application protection

Phase 3: Operational Transition and Readiness
- Organizational change management
- Operational process standards and transition
- Training needs analysis and skills management
- Complete alignment with stakeholders
- Infrastructure cloud-service catalog

Benefits Realization and Measurement along the Journey
Summary

Data protection is currently synonymous with backup, and organizations are finding it difficult to meet their data protection objectives. This problem is getting worse due to:

- Explosive growth in data (especially unstructured).
- Increasingly tight data availability service level objectives (backup window, RPO, RTO and costs).
- Increasing complexity and distribution of the data to be protected (virtualized servers, new applications, remote and mobile users).

Hitachi Data Systems recommends the following 3-pronged data protection strategy that weaves together technologies, such as archive, content store, snapshots, replication, continuous data protection and traditional backups.

- Adopting service-level protection optimizes costs while balancing data protection risks (availability and data loss).
- Archiving inactive data to a content store that has built-in data protection.
- Utilizing application integration to ensure consistent recovery capabilities.
Resources

Organizations rely on Hitachi Data Systems to pull from a wide array of available capabilities to meet their specific data protection and management requirements. More information about some of these capabilities can be found through the following embedded product links. You can also have your questions answered by sending an email to DP-Sales@hds.com.

Storage-Based Data Protection Solutions

- **Hitachi Content Platform**: highly scalable object storage environment for private and hybrid cloud storage services.
- Hitachi global active device: active-active storage clustering makes selected data always available in multiple locations.
- **Hitachi ShadowImage Replication**: high-performance, in-system cloning technology quickly creates full copies of data volumes for a wide variety of use cases.
- **Hitachi Thin Image**: high-performance copy-after-write snapshot solution eliminates the need for backup windows; enables fast and frequent protection (RPO); and supports fast, granular restore (RTO).
- Hitachi TrueCopy: synchronous mirroring and remote replication, across metro distances, between Hitachi storage systems for disaster recovery and business continuity.
- **Hitachi Universal Replicator**: asynchronous replication, across global distances, between Hitachi storage systems for disaster recovery and business continuity.

Host-Based Data Protection Solutions

- **Hitachi Application Protector**: application-consistent snapshot management for Microsoft Exchange, SQL Server and SharePoint.
- **Hitachi Data Instance Manager**: unified copy data management for Microsoft Windows servers, including backup, continuous data processing (CDP), snapshots, replication, archiving and more, with a unique workflow-based user interface.
- **Hitachi Data Protection Suite, powered by CommVault, with IntelliSnap technology**: comprehensive heterogeneous data protection and archiving platform, including application-consistent snapshot management for Hitachi storage systems.
- **Hitachi Storage Optimization for Microsoft SharePoint**: archiving, BLOB migration and e-discovery solution for SharePoint environments.
- **Symantec NetBackup with V-Ray technology**: market-leading enterprise data protection solution with advanced capabilities to protect virtualized servers.
- **Symantec Enterprise Vault**: market-leading archive solution that manages policy-based long-term retention, legal hold and data destruction.