Three Simple Complexities of Data Protection

DENNIS WENK, Symantec

Data is the foundation of information; information leads to knowledge and knowledge is power. There is little disagreement that data has value. In fact, digital data seems to be the new world currency. So protecting valuable data assets is a central concern for business continuity management.

Data loss, data unavailability and data corruption all have adverse economic impacts on the organization. Not only do we need to ensure that data is usable and available, we also need to ensure that sensitive data is protected from unauthorized use. Protecting digital data doesn't sound particularly challenging, as it typically begins with a simple task: make an extra copy. Making and managing these extra copies, however, remains one of the most common pain points for any organization.

There are three fundamental aspects of data protection that contribute to its complexity:

1. **There are lots of data-copy options.** There are a number of data-copy techniques and most techniques have multiple variations. These options vary in both cost and functionality; there is no one-size-fits-all solution. To be optimal, the data-copy solution used should directly correspond to the value of the data.
2. **There are lots and lots of digital data to copy and protect.** Digital

data changes fast and often, and there simply isn't enough budget to make a backup copy of every piece of data every time it changes. While making backup copies is expensive we can't afford to lose the important stuff.

3. **Digital data lacks governance information.** Data can't be protected if we don't know who owns it, or how the data is being used. Lack of this fundamental item, that is, the data ownership, increases both operating expense and operational risk. Operating costs increase because irrelevant data can be copied unnecessarily or data is copied far more frequently than required. While copying too much data or copying the data more frequently than necessary increases costs, if mission critical data is not identified and that data is not copied then it is vulnerable to be lost permanently. The result of losing mission critical data permanently could be catastrophic.

Protecting digital data means that we have to make the right choice about the right solution for the right type of data.

Range Of Data Copy Methods

Backup refers to copies that are created in response to a specific point in time or a specific event. Backup ensures that there is a secondary copy of critical data; providing a consistent point from which to recover the data. Backup typically describes a process of re-writing all the data from the original source to a secondary source.

Tape

Backup is well established as the de facto standard of data protection. Tape has been the media of choice for backup and the historic cornerstone of data protection. Tape media has two notable qualities: it is relatively inexpensive and it is highly portable. That is, tape is cheap to move and store. Tapes can be easily loaded into a truck and moved to an off-site storage location.

In addition, if the backups were needed then the tapes could be moved from the off-site storage to a recovery site. Time is the most valuable resource in a recovery and tapes become a hindrance when recovery time is reduced. Tape is slow — it takes time to locate, move and restore tape.

Versioning

Backup copies can also be versioned. A version can be created at some specific point in time, say every 15 minutes or in response to some specific policy, such as a number-of-changes event. An event policy could be something like the number of updates performed to the file or the number of transactions processed. So a backup would occur regardless of the timeframe — the backup happens every time there are more than say 50 updates to that file.

Versioning also refers to managing multiple point-in-time copies of data sets throughout their lifecycle. It is used to minimize recovery time by increasing the number of intermediate checkpoints from which the data can be recovered. File

versioning products can be thought of as providing an “undo” function at a file level.

Replication

Replicating data is rapidly replacing back up as the data copy method of choice when recovery time is critical. The goal of replicating data is not simply to keep a second identical image of critical data, but to make it possible to quickly reconstruct critical data for recovery purposes. Real-time copy technologies are solutions to maintain current copies of data where every write to the primary data is immediately sent to a replica.

A recent count of disk-array vendors revealed 27 different replication products. Add to this server-based replication and other infrastructure products and the choice in replication is dizzying. Replication falls into one of two categories: synchronous and asynchronous.

Synchronous replication intercepts write-IO requests (write-IOs change the data through add, update, change or delete functions; read-IOs do not change the data) and only allows them to complete when both the primary and the secondary replica have been written. In other words, the system waits until all the write-IOs have been completed, therefore the replicas appear identical, are interchangeable, and most importantly, are entirely transparent to the application.

A primary consideration with synchronous replication, however, is that write-IOs take longer to complete. The adverse impact of this additional time can be significant, as distance creates propagation delay or connectivity limitations interrupt acknowledgements from the secondary replica indefinitely.

Asynchronous replication on the other hand disconnects the primary write-IOs from the secondary replica write-IOs. Asynchronous replication does not wait for all the write-IOs to complete at the replica. While this reduces wait time, it also means that replica will not be identical to the primary. Replica updates will be some point in time behind the primary update. Network interruption will cause replica-updates to accumulate in a queue until the interruption is resolved. Additional care therefore must also be taken to ensure that all the write-IOs at the secondary replica occur in exactly the same sequence as the primary.

There are literally dozens of techniques for data copy solutions: server based, infrastructure based, or disk-storage based — all are available with both synchronous and asynchronous modes. Synchronous replication keeps secondary copies in lockstep with the primary but results in longer write-IO service times. Asynchronous replication has minimal impact on write-IO services times and can therefore be used to span any distance but will “lag” behind the primary. There is no perfect solution — requirements vary, and so do the solutions.

Big, Big Data

There is certainly lots and lots of digital data today. This ever-increasing avalanche

of digital data increases complexity and can overwhelm an organization's ability to both manage and protect data. In the book *Abundance*, Peter H. Diamandis and Steven Kotler put the enormous size of digital data into perspective: "If we digitized every word that was written and every image that was made since the beginning of civilization to the year 2003, the total would come to five exabytes. By 2013, we will be producing five exabytes of data every 10 minutes." This enormous volume of data creates issues for business continuity, particularly as it relates to traditional backup.

The growth of disk-to-disk, disk-to-disk-to-tape, virtual tape, incremental backups and data de-duplication techniques have all provided some alleviation to the pressure that large data volumes create on the shrinking backup window problem. That is, how to complete a point-in-time recovery point for the data while an application is online 24 hours a day, seven days a week. Most of the improvements to traditional backup solutions focus on addressing the front-end issue of the backup window, that is, how to complete a backup.

Some of these front-end improvements actually prolong the restoration process. Restoring from incremental backups can take an order of magnitude longer to restore than it took to complete the front-end backup process. Why so time intensive? Because it takes time to locate the various incremental backup versions, more time still to organize and synchronize the various copies just to get a consistent point-in-time view of the data.

The primary purpose for maintaining a copy of data is to restore and restart business operations as quickly as possible. Taking too much time to get a consistent point-in-time view of the data could be just as consequential as not having a copy. All that matters to the organization is that the data can't be accessed. The objective is to restore access to the data as quickly as possible, not just maintain a copy of the data.

While copying the data is the prime component of the solution, restoring business operations promptly is the primary objective. Missing a backup might create a potential exposure; missing a recovery, however, would be a disaster.

Point-in-time backups are no longer a practical solution because there is simply too much data and it takes too darn long to restore full copies of the data. For large volumes of mission critical data replication is the solution of choice. There is no need to restore the data with replication solutions; copies are either identical or nearly identical and they are transparent to the application. The obstacles for replication are all the cost trade-offs due to network bandwidth, additional workload and additional storage capacity.

Whose Data Is It Anyway?

It is estimated that 80 percent of all potentially usable business information originates in an unstructured form. Unstructured data simply means data that's not in a structured data model like a Relational database. This includes Microsoft Office documents, photos, music, video files, log files, etc. Companies have thousands

upon thousands of digital photos and videos, Word documents, Excel spreadsheets and PowerPoint files floating around — and usually in multiple copies. For most organizations the challenge in protecting unstructured data is its lack of identity.

IT is often tasked with data governance objectives to protect the data, reduce risk, reduce costs, improve efficiency and achieve compliance. The main challenge in realizing these objectives is that data ownership characteristics and usage patterns are non-existent. Seemingly simple choices to protect data, such as permissions to view a confidential file or identifying files critical for recovery, are nearly impossible without some basic information. To make informed choices about data protection it is necessary to know the data's heuristics, the owners, and how the data is being used.

Don't Gamble & Don't Overspend

Protecting data has never been more important. Data is one of three irreplaceable corporate resources, along with life and time. That said, not all data is created equal and not every piece of data is irreplaceable. Data, like risk, comes in gradations. And like risk, the cost of copying data needs to be in balance with the benefits that it provides. After all, the only rational reason to spend money is to achieve a benefit. To protect all your data in a cost-effective manner you must select the right copy solution for the right data.

For more information, please visit www.symantec.com *[1].*

Source URL (retrieved on *03/11/2014 - 10:52pm*):
http://www.chem.info/articles/2012/10/three-simple-complexities-data-protection?qt-most_popular=1

Links:
[1] http://www.symantec.com