

This article first appeared in the March 2008 issue of *Novell Connection* magazine.

www.novell.com/connectionmagazine

Backup and Recovery

Trend Talk

by Amin Marts

Novell.

Introduction:

Novell Open Enterprise Server, now in its second generation, has a deeper, richer and more capable ecosystem of support encompassing it. A major facet of that ecosystem is backup and recovery. As with any platform that serves as a foundational component of enterprise computing and collaboration, having compatible tools to recover lost files is not a luxury—it's a must have.

The world of backup and recovery is chock full of technologies that offer much greater capabilities today than they did a few short years ago. No longer is backing up data incrementally or differentially a standard best practice. Nor is it the only way to approach the task. Technologies such as synthetic backup, capacity-optimized storage/data deduplication and continuous data protection are a few of the newer capabilities vendors are offering in this space.

These technologies are especially compelling when paired with Open Enterprise Server. Unlike many technologies focused directly on the knowledge worker, implementing Open Enterprise Server lays the foundation for a comprehensive backup and data recovery plan.

This article provides an overview of the latest backup and recovery technologies available for Open Enterprise Server 2, with particular focus on the new and often misunderstood offerings in the space.

Synthetic Backup:

One of the newest and most important technologies in the backup and recovery space is synthetic backup; however, in the strictest sense, the term 'newest' is misleading. Synthetic backup has been on the scene for close to three years. It was introduced by a number of niche data management startups and has most recently been added to the portfolio of tier-one vendors such as Symantec and CommVault.

A major differentiator between this technology and traditional backup methodologies is the ability to create a full backup without accessing the original, online data. Enterprise backups are generally responsible for the safekeeping of terabytes of data. Depending on the available bandwidth between the data being copied and the backup device, a full backup can take upwards of an entire weekend. Many times while this is happening, endusers and applications are changing and manipulating the data. Synthetic backups mitigate this challenge by reducing the amount of time needed to complete the full backup.

Borrowing a typical scenario from the real-world, incremental backups take place Monday through Thursday with the full backup starting on Friday. During these incremental backups, data is streamed from the primary data store to a backup medium. Whether it's tape or disk, the same rules apply. Regardless of the destination, bandwidth and processing overhead, resources are consumed by both the media server and the targets. The aggregate amount of data can be relatively small per target, but that changes drastically when a full backup takes place.

During a full backup, all of the target's data is streamed to the backup device. The data sent to the backup device is quite different from that of an incremental backup, because it includes everything. Everything—meaning data that has been altered as well as data that hasn't been touched in weeks, months or years. As you might guess, this is a high-touch, resource-intensive process.

Synthetic backup transforms this resource-intensive scenario at the full backup stage. This is accomplished by leveraging the backup file meta data. Instead of streaming data from the backup target to create the full backup, data is messaged from the incremental dataset. The heavy lifting in this case is done by the media

server, which orchestrates the entire process. The metadata, which is comprised of backup dates, times, data locations, and the like, is then used to create a full backup without requiring data to traverse the network.

Organizations adopting this technology have seen vast improvements in backup times. An example is the University of Montreal. They went from having to shut down production servers to conduct backups for 12 hours each weekend to performing a standard full backup only once a year. Synthetic backups pave the way for improved resource management and flexible backup strategies. This winning combination also provides for substantial cost savings in media and power consumption.

Capacity-optimized Storage (COS):

Generating a great deal of buzz within the storage industry is capacity-optimized storage, commonly referred to as data deduplication and aligned with data reduction practices. Capacity-optimized storage was first introduced to the data center in support of existing backup solutions. More recently it has migrated into a primary storage role. Although this Darwinian evolution is intriguing, the focus of this article will remain on the role of capacity-optimized storage as a complementary backup and recovery technology.

Due in part to the “data tsunami” many organizations are currently experiencing, meeting Recovery Time Objectives (RTOs) is an ongoing challenge. Organizations that are forced to replicate recovery data offsite find it especially challenging. Simply put, the overriding objective of capacity-optimized storage is to reduce the total amount of data housed on a storage medium.

Data deduplication or data reduction hinges on the illumination of patterns to identify redundant data. Redundant data, or data that remains untouched or in its original state, can consume more than 60 percent of an organization's storage capacity. Deduplication technologies simply distinguish data that has not changed from data that has—and then save only the latter. This technology eliminates the redundancy of backing up information that is unchanged.

Two techniques are used when analyzing data for recurring patterns: byte and block level. At the byte level the file itself is in play. The deltas, or changes, are recorded as file versions. Think of this as the version control WebDAV associates to collaboration. Typically they are stored in 100MB chunks, as opposed the 8Kb chunks in which block data is stored. The type of data, median file size, and amount of data changed daily, weekly or monthly are typical factors to take into account when determining the appropriate solution. At a high level, fewer 'data chunks' means less file system defragmentation. As capacity-optimized storage is a disk-based solution, preservation of file system performance is critical to the overall health of the backup solution.

An analogous battle-tested solution within the Novell Workgroup stack is iFolder. Designed to provide mobile users with access to their data anytime and anywhere, it has also seen action in augmenting backup and recovery solutions. By design, it transmits only the deltas from the directory of origin (such as a desktop) to the target, back-end server. Replication and transmission of the deltas from the desktop to the data center preserve bandwidth while simultaneously reducing data upload times. In this way, iFolder represents the same conceptual idea as capacity-optimized storage—but applied to a different use case.

Continuous Data Protection (CDP):

According to the Storage Networking Industry Association, Continuous Data Protection (CDP) is “*a methodology that continuously captures or tracks data modification and stores changes independently of the primary data, enabling recovery points from any point in the past. CDP systems may be block-, file- or application-based and*

can provide fine granularities of restorable objects to infinitely variable recovery points.” (For more information, see snia.org/forums/dmf/programs/data_protect_init/cdp/.)

In essence, the technology provides a framework from which aggressive RTOs can be met with minimal effort. The technology can be best thought of as a robust file system journal capturing data writes. Because of its architecture, CDP technology is targeted at the application. Its primary role is to get the application and its associated data back online quickly in case of failure.

A typical use case scenario for this technology is not complete disaster recovery, but repairing database corruption or accidental file deletions. Typically, the technology provides the greatest benefit in environments that meet the following criteria:

- Frequent data change
- Traditional backup is not an option
- Datasets are large transactional systems

Introducing newer technologies like continuous data protection to an existing disaster recovery strategy provides for additional layers of data survivability. Many organizations understand this, but don't know exactly where or how to implement it.

When it comes to backup and recovery best practices, it's best to standardize on a solution where possible. Successfully mixing and matching best-of-breed software components from multiple vendors poses significant interoperability challenges. Additionally, it's a disaster recovery nightmare waiting to happen. The reasons for this are simple: More so than with hardware, the backup and recovery market is not built around interoperability between vendors, an important caveat as it relates to the media server. CDP is the brains of the operation and, as such, must be compatible with other applications in the backup and recovery system. Its primary responsibility is to manage backup operations—mainly scheduling; however, it also serves as the location of the backup catalog. The catalog is a record of all the data that has been backed up. To have a catalog that spans years and terabytes of data is common.

Case in point: HIPPA regulations dictate a seven-year capture of information. Think about the number of digital records encapsulated within a single patient's file in that time frame. Then multiply that by an entire hospital's files. Real-world examples such as this demonstrate why data backup and recovery have become so critical to enterprise operations.

Although special agents, such as open file agents and database agents, are installed directly on the systems that require them, all file version tracking is bound to the media server. The media server can also serve as the conduit between the backup targets and the offline storage repository in both disk-to-disk and disk-to-disk-to-tape environments. In enterprises with robust data stores, positioning the media server in-line with the tape autoloader can be a bottleneck, especially if there are numerous Network-Attached Storage devices specified as targets. Backup vendors savvy to this issue enable their enterprise solutions with Network Data Management Protocol (NDMP) capabilities.

The beauty of NDMP is that it allows a media server to control/manage a backup device without being physically attached to it. Removing the physical attachment allows data from backup targets to flow directly to the device. Providing a direct path to the device creates an environment where multiple (or large) backup jobs can take place simultaneously. The result is a great reduction in backup times.

Because of the lack of interoperability among backup and recovery software vendors, mixing and matching components would necessitate multiple media servers. Because it's best to avoid this, you'll likely want to stay in

the same vendor family as you take advantage of new technologies; however, adopting backup and recovery technologies is far from an all-or-nothing proposition. For the most part, vendors have built their solutions to allow you to adopt the appropriate technology at the appropriate time. The policy of choice is paramount above all else. In a similar way, Open Enterprise Server provides you with ability to introduce not only Linux, but a number of back-end technologies, at your own pace. The point is that the ability to *choose* is critical and a growing number of software vendors are responding to that. A community of Novell representatives, partners and other customers are here to provide real-world knowledge to help you make your choice.

Introducing a CDP solution simultaneously with a data deduplication technology is possible, but it can be a challenging undertaking. Identifying the need is the first step in determining the appropriate technology. Simply put, what issue are you attempting to address?

1. Are backups running through the weekend into Monday morning?
2. Are backups failing because your media server is low on available resources?
3. Are you spending too much money on tape media?
4. Does your current backup solution support the file system where your data lives?
5. Do you need to improve Recovery Time Objectives (RTO) and Recovery Point Objectives (RPO) in databases?
6. Has your storage architecture changed to accommodate additional storage or a more reliable solution?

Traditionally, the adoption is staged, with metrics being taken along the way to evaluate progress. Beginning with a deduplication appliance will significantly reduce the amount of tape media required for “offsite-ing.” Additionally, because deduplication employs a disk-to-disk-to-tape methodology, you’ll likely see a positive change in your RTOs. The migration of disk to a primary backup medium can also open the door to more aggressive and flexible backup policies. This flexibility is augmented by technologies like CDP. When implemented correctly, CDP can significantly enhance the protection of data.

As you plan your own backup and data recovery strategy, remember that no one technology or feature set can do it all. Carefully assessing your organization’s business objectives, regulatory compliance needs and technology infrastructure will help you hone in on the technologies that will bring the greatest return on investment. It’s also important to understand, a backup solution is a living system that must evolve alongside an organization’s growth plan. It is not something that can be tacked on at the last minute and then expected to provide a high level of insurance when the need arises.

Whatever backup and data recovery solutions you choose, rest assured that market-leading vendors and their latest technologies are fully supported on Open Enterprise Server 2. This ecosystem is critical when introducing Open Enterprise Server 2 to your data center. The ability to recover lost data is not optional in a highly competitive, global market. Disaster recovery, collaboration, business continuity and anti-virus solutions take center-stage in the agile 24x7 business model. **N**