TANEJA GROUP
TECHNOLOGY ANALYSTS

INDUSTRY ARTICLE

The New Case For Open Source Data Protection

June 2008

Open source tools, utilities, and products have been available for many years. While these alternatives tend to offer low acquisition costs, companies have been hesitant to adopt them for several reasons. These reasons include spotty technical support, poor or inconsistent documentation, unreliable release schedules, and lack of a driving commercial focus to address issues and provide sustained development directions.

The open source data protection market has been no different in the past, but recent developments should make small and medium-sized enterprises (SMEs) take notice. Evolution in maturity, product functionality, and commercial backing warrant a re-evaluation of open source data protection alternatives.

This article reviews data protection requirements for SMEs, and evaluates how today's open source data protection alternatives are able to meet them.

SME Data Protection Requirements

SMEs generally have limited IT resources. Surveys confirm that in the data protection arena, these companies look for ease of use, low cost, and then functionality—in that order.

In the ease-of-use area, SMEs need simple yet powerful solutions that can back up and restore data across heterogeneous clients, including Windows, Linux, Unix, and MacOS. Ease-of-use features include centralized management consoles and common tool sets across heterogeneous platforms, as well as "business" functionality, such as simple licensing schemes and responsive technical support. Data protection is a required administrative task, but because it does not really contribute to a company's competitive advantage, IT administrators naturally seek to minimize the amount of effort required to ensure recoverability.

Low cost applies not only to the initial purchase price, but also more importantly to the ongoing maintenance and administrative costs. Simpler, easier-to-use products generally exhibit lower ongoing management costs, so there is good synergy between the "ease-of-use" and "low-cost" requirements. Also, costs associated with maintaining ongoing access to archived data should be taken into account.

In terms of functionality, there are specific requirements that most SMEs look for. Backup-and-restore scheduling and management must cover heterogeneous clients and support

multiple storage architectures, including DAS, SAN, and network-attached storage (NAS). Alternative client restores should be a supported option. Support for off-host backups that leverage snapshot technologies such as Windows VSS and others are also becoming requirements. Backup media support should include a variety of both disk and tape devices and provide media management capabilities with features such as media labeling and retention, overwrite protection, and tape duplication. Finally, scalability should be considered as well. Although an environment may start small, SMEs may grow to hundreds of systems that need to be backed up over time.

Open Source Options

Administrators in SME environments often write their own custom scripts around open source backup utilities. Many of the open source utilities have matured enough to provide excellent building blocks for custom-made backup-and-restore systems. Shipping along with most operating systems, these utilities impose no license fees, but they tend to have platform-specific quirks, and the backup systems built around them require time and effort to develop and maintain. Because they are really just a loose assembly of generic tools designed to locate and move data from one location to another, they lack the data protection focus that commercial backup-and-restore products have. Restores can be challenging with these utilities, and the backup systems built upon them tend to lack scalability as well as a reliable support model.

In addition, the actual developer of these solutions is usually the only person familiar enough with them to maintain them, and generally this person has a number of more pressing responsibilities. If that person leaves the company, there is no documentation that will enable another administrator to quickly come up to speed on the code and no standardization that can be relied on to shorten the learning curve.

Common open source backup/restore utilities

All non-commercial backup systems are built around a set of basic backup utilities that basically copy data from one place to another. They all lack built-in scheduling abilities as well as any sort of searchable backup catalog, making restores more difficult. Common utilities around which homegrown backup systems are built include *rsync* and its close cousin *rdiff-backup*, *rsnapshot*, *dump*, *cpio*, *dd*, and *tar* on Unix systems, and *ntbackup* and System Restore on Windows.

Rsync and *rdiff-backup* are software applications for Linux and Unix systems that synchronize files and directories from one location to another while minimizing data-transfer requirements. *Rsync* differs from *rdiff-backup* in how it stores older backups and file metadata, with *rsync* storing older backups as complete files while *rdiff-backup* stores only the compressed differences between current files and their older versions.

Rsnapshot is a file system backup utility that uses *rsync* and hard links to make and keep multiple full backups instantly available on disk while consuming minimal disk capacity. *Dump*, *cpio*, and *dd* are utilities used to make copies of files, with each accessing file systems somewhat differently.

Tar is an archiving program designed to store and extract files, with support for both disk-based and tape-based archives.

Ntbackup has been the Windows native backup utility since Windows NT, but System Restore, a utility that backs up the Windows registry and other critical files to create a bootable image, is another commonly used utility in Windows environments.

OpenSSH allows administrators to open a secure shell on remote systems to execute a variety of tasks and is often used in conjunction with backup utilities.

Traditionally, many SMEs faced three choices: Buy expensive proprietary backup software, write their own scripts around a set of operating system-specific utilities, or use an open source backup product.

Open source backup products such as BackupPC and Amanda have typified some of the pros and cons of open source software: They require no license fees and offer more of a "product" orientation than homegrown scripts, but lack technical support, properly maintained documentation, and an orderly release schedule. Legacy open source software has relied on a development community that does not have the focus provided by a commercial vendor to update and maintain a product in a predictable, reliable manner. But in the past year, a commercial offering has emerged in the open source space built around the Amanda open source data protection product. These two open source products—BackupPC and Amanda—were designed with slightly different objectives in mind. First, let's take a quick look at each of them, noting their design tenets and unique functionality.

BackupPC

BackupPC is an entirely disk-based backup-and-recovery system that was designed for heterogeneous environments of desktops, servers, and laptops. Because it is entirely disk-based, BackupPC lacks media (tape) management capabilities. To complete backups, it leverages standard tools and utilities (non- proprietary media formats and standard device drivers) that come with each OS so that it can support a wide variety of clients, but it adds to that user control of and access to backups through a Web browser-based interface. BackupPC also provides good support for DHCP and disconnected clients.

While somewhat weak in the areas of media management and security, BackupPC excels in its ease of use, and it offers a unique feature in the world of open source data protection: integrated data de-duplication. All backups go to a set of one or more disks called a "disk pool." While BackupPC stores a directory tree per client backup, it checks to see whether any file has been stored before from any other client running the same OS. If so, then BackupPC uses a hard link to point to the existing file in the common disk pool, saving a lot of hard disk space. Because BackupPC uses hard links to store identical files, the entire backup repository must be on a single file system, limiting the scalability of BackupPC configurations.

Use of proprietary media formats and device drivers

Most commercial backup software products use proprietary media formats and device drivers. By implementing their own proprietary media formats, vendors can offer a common set of backup tools and utilities that work across platforms. Native tools and utilities tend to have platform-specific quirks, with inconsistent functionality across platforms. Unfortunately, any data backed up using a proprietary media format can only be restored using that same format, effectively imposing a vendor-specific lock-in that can become troublesome when multiple backup solutions are present (as they almost always are) or when restoring old, archived data (where the original backup software may no longer be available in the environment).

By providing a layer of management abstraction around native tools and utilities, the goal of a common set of cross-platform management tools can be achieved based on open media formats. Open media formats support restores using either the backup product or the native tools and utilities, avoiding vendor lock-in.

Use of proprietary device drivers allows backup software vendors to enforce use of the same block sizes for data transfers, thereby optimizing transfers during backup and enabling other advanced features, such as tape drive sharing on a SAN. Although operating systems all ship with standard device drivers, different default block sizes for data transfers are used on each platform. Cross-platform solutions that leverage the standard device drivers may not transfer data as optimally as those that use proprietary device drivers, but they are less risky. They are not dependent on the development of a special device driver and will support any device supported by the operating system. They also remove the risk that upgrading the backup software, which is generally done at least once a year (if not more) due to maintenance releases, will break support for a device that is integral to backup activities.

Amanda

Amanda was first developed at the University of Maryland in 1991 with the goal of protecting files on a large number of heterogeneous servers and workstations with a single backup server. Amanda allows users to set up a single master backup server to back up multiple Linux, Unix, Mac, and Windows clients to a very large selection of tape, disk, storage grids (e.g., Amazon's

Simple Storage Service, or S3) and optical devices, including tape libraries, optical jukeboxes, RAID arrays, NAS devices, and many others.

Leveraging a client/server architecture with a single backup master, Amanda stages all backups directly to one or more "holding disks," allowing later migration to other media types. If a tape drive target is not available, Amanda keeps the backup images on the holding disk until the tape drive becomes available, at which point it migrates them to tape. Because disk is a random-access medium, use of the holding disk concept allows multiple clients to be backed up at the same time. And because of the high data-transfer rates supported by disk, dumps to tape can be done while keeping tape drives streaming for optimal performance.

Amanda supports both client-side and server-side compression and encryption options. Interestingly, Amanda has been certified for use by the US Department of Homeland Security—the only open source data protection product to achieve such a distinction. Optimized for backup to disk and tape, and enabling simultaneous backups to dissimilar backup targets, Amanda does not use proprietary device drivers. Instead, it uses standard utilities available on all operating systems such as *dump* and *tar*, and offers a unique backup scheduler that arguably provides the best load balancing in the industry.

The biggest advantage of Amanda over most other backup software is that it offers a scalable solution that does not use proprietary data formats or special device drivers, effectively freeing users from vendor lock-in. Because it uses standard utilities, data can be recovered even without Amanda, obviating concerns about recovery and archiving present with products that use proprietary data formats.

Amanda's Intelligent Scheduler

Most backup products provide a scheduler that allows administrators to set up any backup schedule they desire, but they all rely on the administrator to tell them exactly what to do and when to do it. A common concern in setting up a backup schedule is load-balancing to smooth out resource (backup server CPU, network, I/O, target devices, etc.) requirements so that peak load requirements are not much different from average load requirements. The inability to manage this effectively can result in purchasing of a lot of additional capacity that remains underutilized except for peak load conditions.

Amanda's approach allows an administrator to specify a set of parameters within which the software will calculate the backup schedule to optimally smooth resource requirements across the days in each week. For example, instead of giving Amanda the exact instruction, "Do a full backup every Sunday for clients A, B, and C; full backups on Wednesday for clients D, E, and F; and incrementals at all other times," the administrator sets a few parameters that define how Amanda calculates the backup schedule: "For every client, do at least one full backup within each seven-day period, and do incrementals all other days with a maximum time

between full backups of seven days." If this appears simpler with only six active clients, imagine how much simpler it is when an environment has hundreds of clients that need to be scheduled.

Amanda's Intelligent Scheduler also provides a great solution for disconnected clients. If a client is disconnected on a particular day, the scheduler takes that into account, allows the backup to complete while skipping that client, and then makes backup scheduling adjustments (such as promoting the backup level for that client) to ensure backups of that particular client will still meet the parameters established by the administrator.

To Open Source, or not to Open Source?

Certain organizational profiles have historically favored open source adoption. First, an organization must have technically astute IT personnel who are willing to take on the added burden of tool maintenance and, in some cases, development. Second, these organizations' environments have tended to be less complex, and can get by without some of the advanced functionality that is available with commercial software. And third, the possibility of unreliable technical support must not cause risk to business operations.

Because technical support is basically provided by the open source community—a self-proclaimed special interest group who does not get paid for its efforts—it is inconsistent. At times it can be responsive and of high quality, at other times less so. While the same can also be said of commercial support offerings, they do at least provide an escalation path lacking in open source that can focus a vendor on resolving a problem in a timely manner. For these reasons, open source data protection products have generally been deployed in smaller, less-complex environments for non-mission-critical applications.

In exchange for taking on these risks, an organization has access to tools, utilities, and products that impose no fees or media format lock-ins. For organizations with available technical resources, an open source product can be tailored for their environment without paying any source or binary license fees, and in some cases patched faster than comparable commercial products. Historically, when the problems to solve have been relatively simple, open source alternatives have provided more cost-effective, less-complex solutions. Ongoing maintenance costs, particularly for archiving, have tended to be lower with open source options. Unlike commercial backup products that use proprietary media formats, most open source data protection utilities use readily available, industry-standard tools such as *tar*, *dump*, and others for backup. This precludes the need to maintain expensive proprietary software to ensure access over time to data archives if backup software is ever replaced. It is not uncommon on the discussion threads for various open source backup products such as BackupPC or Amanda to see posts from users that have tried commercial software offerings but found them too complex and costly, or just plain overkill, for their environments.

While they tend to be more expensive, commercial backup software offerings do come with a promise of reliable technical support, consistently updated documentation, regular software updates and releases, and a commercial development focus. Historically, these have been deployed in more mission-critical environments. In the data protection space, commercial software alternatives have also offered more advanced features, such as more application agents, support for vendor-specific snapshot/backup approaches, dynamic device sharing, and more sophisticated vaulting technology.

In 2007, Zmanda Inc. changed the face of open source data protection. By creating a commercial backup software offering around Amanda, Zmanda promises to address the legacy issues with open source data protection alternatives. Amanda is the only open source data protection distribution that has this commercial backing. Zmanda calls its product Amanda Enterprise and targets it at SMEs. Generally, Amanda Enterprise licensing runs about 25% to 30% of the cost of well-known backup applications.

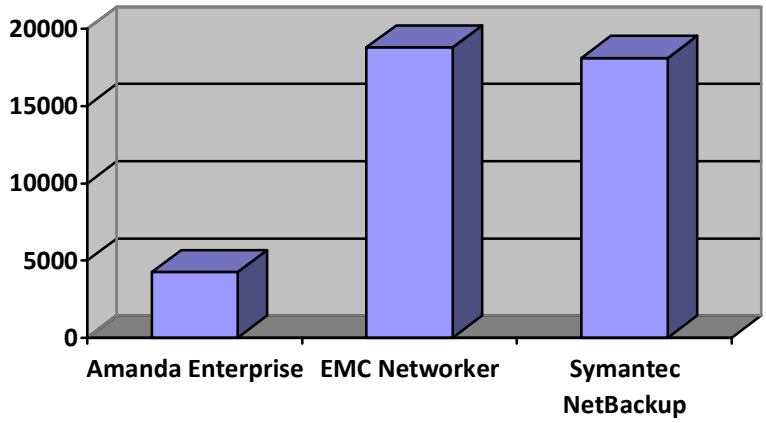

Figure 1 represents a comparison of initial purchase costs across the 3 platforms for the following configuration: 1 Backup Server on Linux, 15 backup clients spread equally across Windows, Linux, and Solaris, support for backup to disk (1TB), and support for a tape library with 2 drives and 40 slots. Data obtained from http://www.sun.com/storagetek/products.jsp.

Users familiar with commercial backup software licensing will also be pleased not only with the acquisition costs of Amanda Enterprise, but also with the simple licensing model. Included with the base subscription price are features such as support for disk-based backup, SANs, Windows VSS, compression, and encryption, as well as alternative client and server-independent restore support. Backup application agents are the only separately priced products in the Zmanda offering, and today they include Solaris, Linux, Windows, Mac, Exchange, SQL Server, and SharePoint, with Oracle and many others on the way.

Conclusion

Open source data protection products have made great steps forward, with options such as BackupPC and Amanda offering feature-rich, cost-effective solutions for SMEs that can be much easier to use than homegrown script-based options and much more cost-effective than commercial alternatives. Both of these solutions are well-suited to heterogeneous environments with up to hundreds of clients that do not require advanced features such as support for vendor-specific snapshot/backup methodologies or dynamic tape drive sharing. But most open source options still suffer from the issues of unreliable technical support, inconsistent documentation, and unpredictable release schedules. And it is this lack of a commercially oriented development and support focus that has historically given enterprises pause when presented with open source options.

With the entry of Zmanda's Amanda Enterprise, SMEs now have a fully supported and documented open source data protection option that has the backing and focus of a commercial company. Amanda Enterprise is a full-function product with technical support, documentation, and predictable release schedules.

The increase in functionality of open source options is putting more pressure on the commercial backup software vendors to differentiate their offerings and justify their higher prices. The emergence of an open source data protection product with full commercial backing – Amanda Enterprise - is turning up that heat even more. For IT managers in SME environments, these developments mean that now is a good time to take another look at open source data protection offerings.

Eric Burgener is a senior analyst and consultant with the Taneja Group research and consulting firm (www.tanejagroup.com).