TIP

Backup vs. archive

Pierre Dorion

This was first published in October 2008

When discussing the difference between backups and archives, we should probably start with introducing some generally accepted definitions for the terms "backup" and "archive." I use the "generally accepted" qualifier because in IT, terminology is rarely absolute. Furthermore, the world of IT is full of smart people and smart people like to argue a lot.

Backup: A backup refers to a copy of data that may be used to restore the original in the event the latter is lost or damaged beyond repair. It is a safeguard for data that is being used.

Archive: An archive refers to a single or a collection of historical records specifically selected for long-term retention and future reference. It is usually data that is no longer actively used.

One of the differences worth noting in the above descriptions is that a backup is always a copy while an archive should be the original that was removed from its initial location and sent elsewhere for long-term retention.

Frequently, backup software is configured in an attempt to fulfill both roles. Data backup schedules often include daily incremental backups kept for seven days, a weekly full kept for a month, a monthly full kept for a year and finally, a yearly full kept for seven years. The yearly full backups are often referred to as seven-year archives. The problem with that scheme is that it becomes very difficult to single out specific files for long-term retention out of an entire backup job, so a user of this method often ends up having to lump everything into on large "archive" package, which is a full backup. Some email programs allow you to create so-called archives from a user's mailbox, but they all end up in one large file (i.e., a PST file for Exchange). So, retrieving an archived record can become a daunting task.

The lifecycle of backup copies

When we no longer need a backup copy of a file, usually because we have a sufficient number of point-in-time copies, we simply delete the oldest copy or backup job of which the file is part. This step is typically automated for us. Likewise, when we restore a file from a backup copy, we know what we are looking for and we may opt to select a point-in-time copy or version based on a date criteria.

If long-term backups are used in lieu of archives, things can get a lot more complicated. Other than to satisfy a legal or regulatory compliance requirement, archives are used to free up primary (production) disk storage space from data that is no longer actively used but must be retained. Keeping an entire backup job for seven years as per our earlier example is not a very cost-effective way to use storage. Because backups only copy data, the original file is left in place, which frees up no space at all. In addition, the backup creates yet another copy of the data, which means we are now using twice as much storage space as before unless files are deleted manually after the backup. This only adds more data management overhead.

Furthermore, because a lot of backups are organized in a sequence (i.e., full and incremental), a full backup has to potentially be taken out of sequence to capture the desired point-in-time full copy of a file. Day 3 of an incremental backup sequence would be useless without the previous full and incremental backups.

Archives

One way that archiving software is different from backup software is the cataloging and search capabilities. Metadata about archived objects is stored in a database and can be searched based on user-supplied criteria. Some backup software products also offer basic archiving capabilities where files can be archived, individually or in groups, and retained independently from the backups. Archived objects can be named or labeled based on the type of data, date, ownership or any other searchable criteria deemed appropriate to ease the search process for future reference.

However, for organizations generating and handling large volumes of archives such as email, imaging data, etc., the ability to create a simple searchable label is no longer sufficient. More sophisticated archive solutions providing search capabilities based on content, ownership, etc. are required. Just picture trying to find a 5-year-old patient record or court case transcript using your backup software's search wizard!

Another challenge with archives that backup software simply can't handle is the static nature of aging archived data vs. the very dynamic technology used to access that same data. As much as software vendors try to maintain backward compatibility with some applications, data eventually ages to a point where it can no longer be easily or usefully accessed with the latest release of an application. Some sophisticated archive solutions have integrated conversion capabilities to allow future access to certain data using a more universal format. An example would be the conversion of reports to PDF format, which allows future access to those reports without requiring the application that created them. Archived files are typically maintained "as is" and are not alterable.

So, an archive really isn't a backup copy kept for a long time. We can also add that using backup software to produce large amounts of archives that may eventually need to be searched (i.e., legal discovery) is a bad idea.

About the author: Pierre Dorion is the Data Center Practice Director and a Senior Consultant with Long View Systems Inc. in Phoenix, AZ, specializing in the areas of business continuity and disaster recovery planning services, and corporate data protection.