# HOW ARCHIVEIQ DATA DE-DUPLICATION SIMPLIFIES BURA

**By helping eliminate redundant backup data, de-duplication provides a key way to streamline backup, recovery, and archiving (BURA) for organizations of all sizes. Combining Data Storage Group ArchiveIQ™ software with Dell EqualLogic™ PS Series storage arrays can help overcome common BURA challenges and enable cost-effective, scalable, and simplified disk-based data protection.**

By Pete Caviness

**Related Categories:**

Backup, recovery, and archiving (BURA)

Data consolidation and management

Data Storage Group

Dell EqualLogic storage

Storage

Visit DELL.COM/PowerSolutions for the complete category index.

Data backup, recovery, and archiving (BURA) systems can be among the most time-consuming and costly aspects of systems management in enterprise IT environments. Protecting and retaining data is a critical function, but the tape-based processes that many organizations use today—copying and storing the same data over and over again, week after week—is difficult to scale, and can become increasingly costly and complex as organizations contend with exponential data growth.

Disk-based backup has been available for years, but many organizations have either viewed it as too costly to implement or restricted its use to a temporary step in a disk-to-disk-to-tape process. Legacy BURA systems, meanwhile, often treat disk as sequential access tape, wasting much of their capacity by storing redundant data. For a disk-based BURA system to be viable, it must address disk capacity costs, storage scalability, management overhead, and efficient off-site protection.

Data de-duplication provides a key way for organizations to overcome the problems of traditional BURA systems and create cost-effective, scalable, and simplified disk-based data protection. By combining Data Storage Group ArchiveIQ software with flexible storage platforms such as Dell EqualLogic PS Series Internet SCSI (iSCSI) storage area network (SAN) arrays, organizations can help streamline data management while controlling costs and enabling their BURA systems to scale to meet their needs.

## COMMON BURA CHALLENGES

As organizational data growth continues to accelerate, legacy processes of copying and storing the same data every week become an increasingly inefficient way to protect and archive data. Administrators must contend with a variety of challenges related to BURA as their data needs increase:

- **Growing backup windows:** As data grows, the server and network impact of backup can no longer be isolated to nights and weekends. The amount of time between backup images may also grow, increasing the chances of data loss.
- **Time-consuming recovery:** Data recovery from tape media is notoriously time-consuming and complex compared with recovery from disk-based systems. Organizations are trying to keep an increasing number of recovery points on disk and readily available to enhance service to both internal and external users.
- **Need for remote office data:** Remote office data that is located across low-bandwidth connections typically requires dedicated backup systems and administration. Creating consistent, reliable backups of remote data can be expensive and difficult—and

keeping these backups at a centralized location can be even more challenging.

- **Large, active files:** Large files that are constantly changing—such as Microsoft® Office Outlook® .pst files, databases, and virtual machine disk files—are a significant problem for many BURA systems. Protecting these files can require a significant amount of storage capacity.
- **Unreliable backup validation:** Legacy backup images are typically validated only once, when they are first created. If an image later develops problems, administrators are typically unaware of those problems until they try to recover data—by which point the original data is gone, and it is too late to correct the problems.
- **Need for regulatory compliance:** Retaining backup images for several years to meet regulatory requirements can be inefficient and expensive. Amendments to the Federal Rules of Civil Procedure now require organizations that operate within the United States to manage their electronic data so that it can be retrieved in a timely and complete manner, and significant penalties may be levied against organizations that cannot comply.
- **Compatibility with hardware upgrades:** Backup images created on tape media are useless without the corresponding tape drive to read them. Should existing tape drive technology become obsolete, administrators could be faced with huge media conversion problems.

As data growth continues to accelerate, these common problems will likely become increasingly difficult to handle using traditional BURA technology alone—requiring organizations to pursue ways of streamlining and simplifying their processes.

## BASICS OF DATA DE-DUPLICATION

Data de-duplication has become a popular approach to BURA, one that can help organizations overcome many typical BURA problems in enterprise IT environments while helping reduce the related management time and costs. It can provide immediate and substantial savings compared with legacy BURA systems. Understanding the basics of de-duplication can help organizations understand what to expect from de-duplication software and evaluate how it can help solve both common problems and problems specific to particular environments.

The concept of de-duplication is simple: identifying and removing redundant data helps dramatically reduce the size of a backup image, enabling organizations to use existing storage capacity more efficiently than they can with legacy BURA systems. For example, when a presentation file is distributed within an organization, many people may save a personal copy in their home directory. One key de-duplication technique, single-instance storage (SIS), is designed to identify, transfer, and store only a single copy of this file. From that point on, when the same file data is found, it does not need to be transferred or stored again.

The amount of data reduction that de-duplication can provide depends on the techniques used to locate redundancies. SIS is a standard method in many de-duplication applications, but it should not be the only one used.

### Additional de-duplication techniques

ArchiveIQ includes two additional techniques in its de-duplication process: *advanced data compression* and *sub-file data reduction* (see Figure 1). Advanced file compression helps further reduce capacity requirements for backup data, while sub-file data reduction helps efficiently store large, active files. For example, SIS-level data reduction typically provides no benefits for Outlook .pst files, because the very act of reading one of these files modifies its content and forces a full version of the entire file to be copied and stored. In this situation, the de-duplication software needs to look at the sub-file level to identify redundant data and store only unique or changed information.

A fourth technique, *data chunking*, breaks data into small chunks and identifies redundancy at the chunk level. This technique is designed to identify the most redundancies, but it can come at the cost

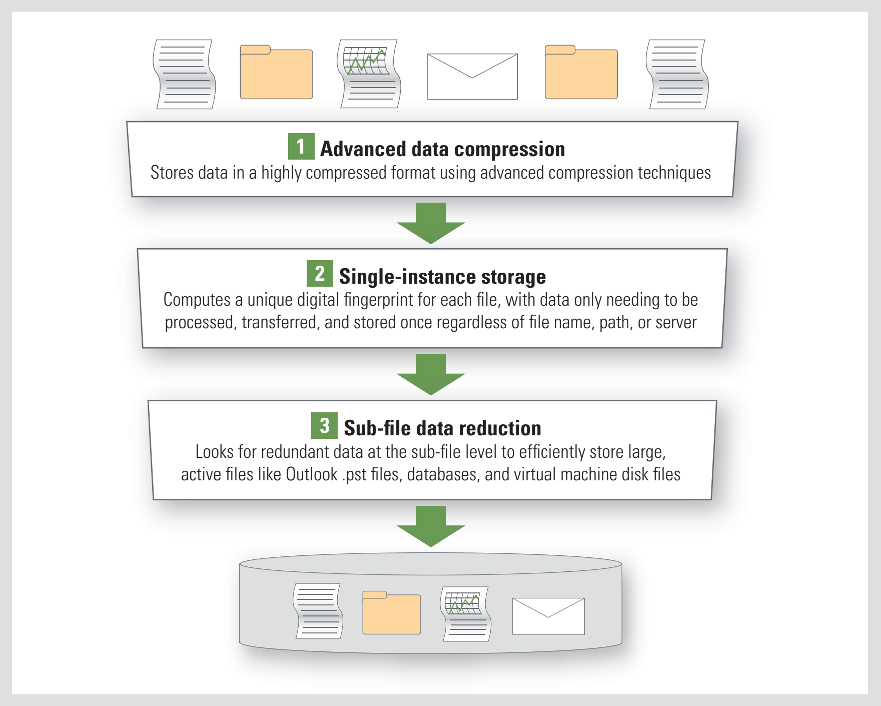
1 Advanced data compression
Stores data in a highly compressed format using advanced compression techniques
2 Single-instance storage
Computes a unique digital fingerprint for each file, with data only needing to be processed, transferred, and stored once regardless of file name, path, or server
3 Sub-file data reduction
Looks for redundant data at the sub-file level to efficiently store large, active files like Outlook .pst files, databases, and virtual machine disk files

***Figure 1.*** *Using multiple de-duplication techniques helps efficiently identify and eliminate redundant data*

of significant processing and recovery time to reassemble the de-duplicated data. In addition, the master index that maps the chunks together can become very large over time. For most organizations, the overall data reduction from this technique is not worth the system and recovery impact.

### De-duplication as part of the backup process

Another major difference between data de-duplication applications is when and where the de-duplication takes place: different applications may use *post-process*, *in-line*, or *source-based* de-duplication (see Figure 2).

Post-process de-duplication typically requires third-party backup software to perform a full backup to a disk cache and then process the data after it has been stored. This method can accelerate backups when the cache is not full, but usually requires additional disk space to cache the backup jobs.

In-line de-duplication processes full backup jobs as the data arrives at the de-duplication system. This method is generally slower than post-process de-duplication, but also typically uses less disk capacity because it does not require a cache.

Source-based de-duplication equally distributes the de-duplication process among the production servers being protected. Compared with the post-process and in-line methods, source-based de-duplication can help solve more of the problems described in the "Common BURA challenges" section in this article. Once the unique data has been processed, validated, transferred, and stored, it should never require processing again—with the result that large, active files can be efficiently processed at the source, where only changed parts of the files are identified and stored. This method can help dramatically reduce the backup window, increase backup reliability, and enable remote office data protection without requiring a dedicated backup system or an administrator at that location.

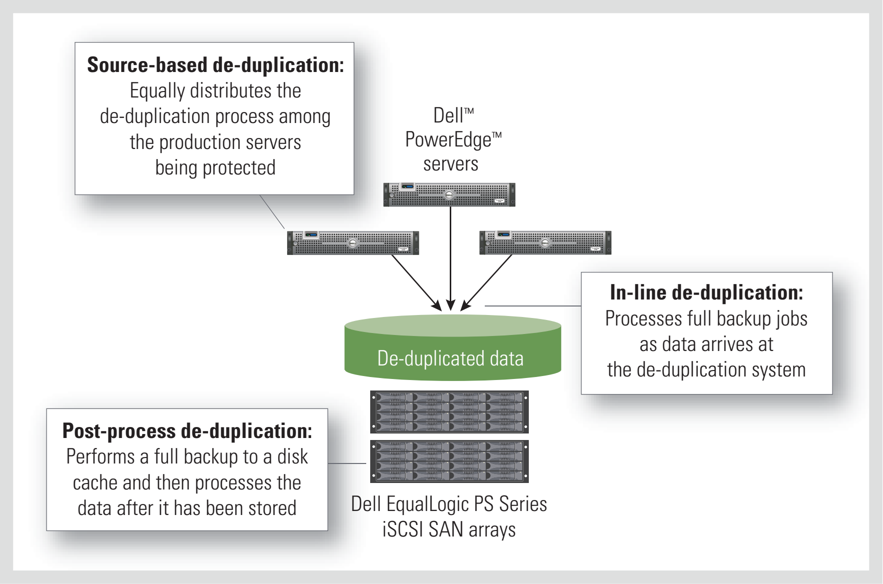
Source-based de-duplication:
Equally distributes the de-duplication process among the production servers being protected
Dell™ PowerEdge™ servers
In-line de-duplication:
Processes full backup jobs as data arrives at the de-duplication system
De-duplicated data
Post-process de-duplication:
Performs a full backup to a disk cache and then processes the data after it has been stored
Dell EqualLogic PS Series iSCSI SAN arrays

***Figure 2.*** *De-duplication may occur at different stages of the backup process*

### Other features

Several other features are also integral to a comprehensive de-duplication system:

- **Reliable data validation:** Compared with legacy BURA approaches, de-duplication has a major advantage in terms of data validation. De-duplication creates and validates a digital "fingerprint" of backup data when the backup is first stored—meaning that a backup image can be validated without accessing the production data. The most reliable data validation process performs a virtual recovery of the content to help ensure the data can be reconstructed exactly to its original state.
- **Flexible data retention policies:** Data retention periods are not the same for all organizations, but one common concern is that this time frame will unexpectedly increase. A comprehensive de-duplication system should support retention policies that allow administrators to manage specific parts of the de-duplicated data independently to service potential litigation requests.
- **Simplified data recovery:** Efficiently storing data is only part of the problem; comprehensive de-duplication should also offer quick and innovative ways to recover data. De-duplicated data is typically on disk and readily available, helping simplify recovery of individual files and folders.
- **Hardware independence:** De-duplicated data depends on the techniques used by specific software, but should not be dependent on specific hardware.

## ARCHIVEIQ AND DELL EQUALLOGIC PS SERIES SAN ARRAYS

The advanced de-duplication technology in ArchiveIQ and the flexibility and features available in Dell EqualLogic PS Series iSCSI SAN arrays serve complementary purposes in enterprise IT environments. Deploying these products together can help organizations create a cost-effective, scalable, simplified BURA system to help efficiently protect data.

### Key features of ArchiveIQ

ArchiveIQ is an innovative software application that runs on the Microsoft Windows Server® OS and helps eliminate the high costs associated with using disk media for BURA. By removing redundant data and only storing truly unique data over time, it enables organizations to cost-effectively protect and archive many years' worth of data using only a fraction of the storage capacity typically required by traditional BURA approaches. For example, although disk capacity may cost 10 times as much per gigabyte as tape capacity, ArchiveIQ is designed to keep significantly more backup data on disk than legacy BURA tape-based products—effectively enabling a lower cost per gigabyte for disk compared to tape.

ArchiveIQ uses multiple de-duplication techniques to achieve high levels of data reduction, including SIS, advanced data compression, and sub-file data reduction. Its source-based de-duplication approach enables it to identify and remove redundant data at the source, before it is transferred across the network—helping protect remote office data and dramatically reduce backup windows and recovery point objectives. Administrators can configure ArchiveIQ to use post-process de-duplication if the source server is not running a Microsoft Windows® OS.

ArchiveIQ is designed to simplify and accelerate file recovery. Its file name index can support wild card searches across several months of de-duplicated data, and administrators can quickly explore each recovery point like a normal file share. Full folder recovery is designed to be a simple drag-and-drop process. Finally, because ArchiveIQ does not chunk data into small pieces, recovery jobs can be carried out at full disk speeds.

Several other key features also help organizations implement powerful, flexible BURA systems:

- **Automated data validation:** Recovery points are continuously validated based on administrative policies—helping the system identify unexpected problems with the storage media early and automatically repair itself from the source data.
- **Automated data retention:** Administrators can specify how long recovery points should be retained, and the system can automatically identify and remove de-duplicated data that does not meet the defined retention policy. This process helps make efficient use of available storage capacity and helps limit litigation and compliance liability.
- **Flexible capacity management:** Administrators can optionally increase available storage capacity on Windows Server–based file servers that are running out of space. ArchiveIQ transparently "stubs" inactive file data and frees the associated storage capacity for new files and active data. If a user or application needs to access the stubbed data, this data can be transparently cached back from the ArchiveIQ server.
- **Hardware independence:** Administrators can use existing server and storage capacity or purchase new capacity based on considerations such as replication, expansion, migration, and price. As long as the storage platform supports NT File System (NTFS) volumes, ArchiveIQ can use it to store de-duplicated data.

On installation, ArchiveIQ allows a fully functional 30-day evaluation period. Using the software in their specific environment is the best way for organizations to understand data de-duplication and measure its benefits.

### Key features of Dell EqualLogic PS Series SAN arrays

Dell EqualLogic PS Series iSCSI SAN arrays provide a scalable storage platform for ArchiveIQ de-duplication. The virtualized iSCSI storage helps simplify scaling and upgrading storage capacity over time by enabling the transparent distribution of data across one or more arrays. When administrators add a PS Series array to the storage pool, data can be automatically redistributed to take advantage of the additional storage and processing capacity. The same automatic process can take place when administrators need to repurpose or retire a legacy PS Series array. This simplicity and flexibility helps avoid the need for time-consuming management of data expansion or migration.

PS Series arrays are also designed for reliability and avoiding single points of failure. Critical hardware components have a redundant partner to help eliminate unexpected downtime caused by a hardware failure—one of the most common causes of downtime and data loss. In addition, the advanced data replication features included with PS Series arrays at no additional cost can efficiently replicate data off-site, enhancing disaster recovery systems and helping protect data against local events such as fire or flooding.

The thin provisioning feature of PS Series arrays can help increase storage utilization and reduce administrative overhead across an organization—two key ways that shared storage can help reduce costs. This innovative feature is designed to make storage capacity available, but not truly allocate that capacity until it is used. By allocating capacity on demand, administrators can help avoid over-allocating storage and reduce the time required to recapture capacity from production servers.

## SCALABLE, COST-EFFECTIVE, SIMPLIFIED DATA PROTECTION

Accelerating data growth exacerbates common problems with traditional tape-based BURA approaches. By taking advantage of the complementary features of ArchiveIQ data de-duplication software and Dell EqualLogic PS Series arrays, organizations can consolidate resources and implement cost-effective, scalable, simplified data protection.

**Pete Caviness** is director of marketing at Data Storage Group, Inc.

MORE ONLINE
DELL.COM/PowerSolutions

QUICK LINKS

**Data Storage Group ArchiveIQ:**
www.archiveiq.com

**Dell EqualLogic PS Series:**
DELL.COM/EqualLogic
DELL.COM/PSSeries