Demystifying Data De-Duplication: Choosing the Best Solution

Selecting a data de-duplication solution that enables cost effective, high performance, and scalable long-term data storage

Abstract: Data de-duplication has been described as the next evolutionary step in backup technology. Many vendors lay claim to the best data de-duplication approach, leaving customers to face the difficulty of separating hype from reality. A number of key factors must be considered in order to select a data de-duplication solution that actually delivers cost-effective, high-performance, and scalable long-term data storage. This document provides the background information required to make an informed data de-duplication purchasing decision.
Introduction

In an August 2006 report from the Clipper Group entitled ‘The Evolution of Backups – Part Two – Improving Capacity’, author Dianne McAdam writes: “De-duplication is the next evolutionary step in backup technology.” Eliminating duplicate data in secondary storage archives can slash media costs, streamline management tasks, and minimize the bandwidth required to replicate data.

Despite the appeal of this data de-duplication concept, widespread adoption has been slowed by the high costs of processing power required to identify duplicate data, index unique data, and restore compacted data to its original state. The path has only recently been cleared for a proliferation of data de-duplication solutions as processing power has become more cost-effective.

Many vendors lay claim to the best data de-duplication approach, leaving customers to face the difficulty of separating hype from reality and determining which factors are really important to their business. With some vendors setting unrealistic expectations by predicting huge reductions in data volume, early adopters may find themselves ultimately disappointed with their solution.

Companies must consider a number of key factors in order to select a data de-duplication solution that actually delivers cost-effective, high-performance, and scalable long-term data storage. This document provides the background information required to make an informed data de-duplication purchasing decision.

Data De-Duplication is Becoming an Operational Requirement

Because secondary storage volumes are growing exponentially, companies need a way to dramatically reduce data volumes. Regulatory changes such as the Federal Rules of Civil Procedure magnify the challenge, forcing businesses to change the way they look at data protection. By eliminating duplicate data and ensuring that data archives are as compact as possible, companies can keep more data online longer, at significantly lower costs. As a result, data de-duplication is fast becoming a required technology for any company wanting to optimize the cost-effectiveness and performance of its data storage environment.

Although compression technology can deliver an average 2:1 data volume reduction, this is only a fraction of what is required to deal with the data deluge most companies now face. Only data de-duplication technology can meet the requirements companies have for far greater reductions in data volumes.

Data de-duplication can also minimize the bandwidth needed to transfer backup data to offsite archives. With the hazards of physically transporting tapes well-established (damage, theft, loss, etc.), electronic transfer is fast becoming the offsite storage modality of choice for companies concerned about minimizing risks and protecting essential resources.

Key Criteria for a Robust Data De-Duplication Solution

There are eight key criteria to consider when evaluating data de-duplication solutions:

1. Focus on the largest problem
2. Integration with current environment
3. Virtual tape library (VTL) capability
4. Impact of de-duplication on backup performance
5. Scalability
6. Distributed topology support
7. Real-time repository protection
8. Efficiency and effectiveness

1. Focus on the largest problem
The first consideration is whether the solution attacks the area where the largest problem exists: backup data in secondary storage. Duplication in backup data can cause its storage requirement to be many times that which would be required if the duplicate data could be eliminated.

The following graphic, courtesy of the Enterprise Strategy Group (ESG), illustrates why a new technology evolution in backup is necessary. Incremental and differential backups were introduced to decrease the amount of data required compared to a full backup, as depicted in Figure 1.

However, even within incremental backups, there is significant duplication of data when protection is based on file-level changes. When considered across multiple servers at multiple sites, the opportunity for storage reduction by implementing a data de-duplication solution becomes huge.

2. Integration with current environment
An effective data de-duplication solution should be as non-disruptive as possible. Many companies are turning to virtual tape libraries to improve the quality of their backup without disruptive changes to policies, procedures, or software. This makes VTL-based data de-duplication the least disruptive way to implement this technology. It also focuses on the largest pool of duplicated data: backups.

Solutions requiring proprietary appliances tend to be less cost-effective than those providing more openness and deployment flexibility. An ideal solution is one that is available as both software and turnkey appliances in order to provide the maximum opportunity to utilize existing resources.

3. VTL capability
If data de-duplication technology is implemented around a VTL, the capabilities of the VTL itself must be considered as part of the evaluation process. It is unlikely that the savings from data de-duplication will override the difficulties caused by using a sub-standard VTL. Consider the functionality, performance, stability, and support of the VTL as well as its de-duplication extension.
4. Impact of de-duplication on backup performance
It is important to consider where and when data de-duplication takes place in relation to the backup process. Although some solutions attempt de-duplication while data is being backed up, this inline method processes the backup stream as it comes into the de-duplication appliance, making performance dependant on the single node’s strength. Such an approach can slow down backups, jeopardize backup windows, and degrade VTL performance by as much as 60% over time.

By comparison, data de-duplication solutions that run after backup jobs complete or concurrently with backup avoid this problem and have no adverse impact on backup performance. This post-process method processes the backup data by reading it from the VTL storage after backups have been cached to disk, which ensures that backups are not throttled by VTL or storage limitations. An enterprise-class solution that offers this level of flexibility is ideal for organizations looking for a choice of de-duplication methods.

For maximum manageability, the solution should allow for granular (tape- or group-level) policy-based de-duplication based on a variety of factors: Resource utilization, production schedules, time since creation, and so on. In this way, storage economies can be achieved while optimizing the use of system resources.

5. Scalability
Because the solution is being chosen for longer-term data storage, scalability, in terms of both capacity and performance, is an important consideration. Consider growth expectations over five or more. How much data will you want to keep on disk for fast access? How will the data index system scale to your requirements?

A de-duplication solution should provide an architecture that allows economic “right-sizing” for both the initial implementation and the long-term growth of the system. For example, a clustering approach allows organizations to scale to meet growing capacity requirements – even for environments with many petabytes of data – without compromising de-duplication efficiency or system performance. Clustering enables VTL to be managed and used logically as a single data repository, supporting even the largest of tape libraries. Clustering also inherently provides a high-availability environment, protecting the VTL and de-duplication nodes by offering failover support (Figure 2).

6. Distributed topology support
Data de-duplication is a technology that can deliver benefits throughout a distributed enterprise, not just in a single data center. A solution that includes replication and multiple levels of de-duplication can achieve maximum benefits from the technology.

For example, a company with a corporate headquarters, three regional offices, and a secure disaster recovery (DR) facility should be able to implement de-duplication in the regional offices to facilitate efficient local storage and replication to the central site. The solution should only require minimal bandwidth for the central site to determine whether the remote data is contained in the central repository. Only unique data across all sites should be replicated to the central site and subsequently to the DR site, to avoid excessive bandwidth requirements.

7. Real-time repository protection
Access to the de-duplicated data repository is critical and should not be vulnerable to a single point of failure. A robust data de-duplication solution will include mirroring to protect against local storage failure as well as replication to protect against disaster. The solution should have failover capability in the event of a node failure. Even if multiple nodes in a cluster fail, the company must be able to continue to recover its data and respond to the business.
8. Efficiency and effectiveness

File-based de-duplication approaches do not reduce storage capacity requirements as much as those that analyze data at a sub-file or block level. Consider, for example, changing a single line in a 4MB presentation. In a file-based solution, the entire 4MB file must be stored, doubling the storage required. If the presentation is sent to multiple people, as presentations often are, the negative effects multiply.

Most sub-file de-duplication processes use some sort of “chunking” method to break up a large amount of data, such as a virtual tape cartridge, into smaller-sized pieces to search for duplicate data. Larger chunks of data can be processed at a faster rate, but less duplication is detected. It is easier to detect more duplication in smaller chunks, but the overhead to scan the data is much higher.

If the “chunking” begins at the beginning of a tape (or data stream in other implementations), the de-duplication process can be fooled by the metadata created by the backup software, even if the file is unchanged. However, if the solution can segregate the metadata and look for duplication in chunks within actual data files, the duplication detection will be much higher. Some solutions even adjust chunk size based on information gleaned from the data formats. The combination of these techniques can lead to a 30-40% increase in the amount of duplicate data detected. This can have a major impact on the cost-effectiveness of the solution.

Summary: Focus on the Total Solution

As stored data volumes continually increase due to the demands of business applications and regulatory requirements, data de-duplication is fast becoming a vital technology. Data de-duplication is the only way to dramatically reduce data volumes, slash storage requirements, and minimize data protection costs and risks.

Although the benefits of data de-duplication are dramatic, organizations should not be seduced by the hype sometimes attributed to the technology. No matter the approach, the amount of data de-duplication that can occur is driven by the nature of the data and the policies used to protect it.

In order to achieve the maximum benefit of de-duplication, organizations should choose data de-duplication solutions based on a comprehensive set of quantitative and qualitative factors rather than relying solely on statistics such as theoretical data reduction ratios.

For more information, visit www.falconstor.com or contact your local FalconStor representative.