Dynamic Solutions
INTERNATIONAL

Demystifying Deduplication

Introduction

Data redundancy was once an acceptable operational part of the backup process, but in recent years the rapid growth of digital content in the data center has pushed organizations to rethink how they approach this issue.

Data deduplication technologies were introduced to assist organizations with optimizing storage capacity. There are many providers of data deduplication solutions today, and each vendor lays claim to offering the best approach, placing unrealistic expectations by predicting huge reductions in data volume—and ultimately disappointing customers. Companies must consider a number of factors in order to select a data deduplication solution that suits their needs and provides significant value with minimal disruption to infrastructures and processes.

Audience

The intended audience of this whitepaper is DSI customers, prospects, partners and others who want to learn more about aspects of deduplication technology and how to determine the best solution.

Data deduplication is an operational requirement

The proliferation of duplicated data is a result of industry-standard backup practice. In the interest of data protection, the traditional backup paradigm copies data to a safe secondary storage repository over and over again, creating a monstrous overload of backed-up information. Under this scenario, every backup exacerbates the problem.

Secondary storage volumes are growing exponentially; companies need a way to dramatically reduce these data volumes. Regulatory requirements 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 deduplication is now a required technology for any company wanting to optimize the performance, efficiency, and cost-effectiveness of its data storage environment.

Data deduplication can minimize the bandwidth needed to transfer backup data to offsite archives. With the hazards of physically transporting tapes being 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.

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. Data deduplication technology provides reductions in data volumes that customers require.

Demystifying Deduplication

Key criteria for effective deduplication

There are ten important criteria to consider when evaluating deduplication solutions:

1. **Focus on the largest problem area**
   The first consideration is whether the deduplication software can attack the area where the largest problem exists: backup data in secondary storage. Duplication in backup data can cause storage requirements to be many times what would be required if duplicate data could be eliminated. Make sure the deduplication solution is able to reach across multiple servers and multiple sites to reach maximum benefits.

2. **Easy integration with current environment**
   An effective deduplication solution should be as non-disruptive as possible. Many companies turn to virtual tape library (VTL) technology as a method of implementing deduplication and improving the quality of their backups without significant changes to policies, procedures, or software.

   Other companies leverage a disk-to-disk (D2D) backup paradigm, which requires that deduplication presents a network interface to the backup application. This process simplifies and enhances D2D backups, performing deduplication without disruption to ongoing operations.

   Solutions requiring proprietary appliances tend to be far less cost-effective than those providing more openness and deployment flexibility. An ideal solution is one that integrates with an organization's existing backup environment and is available in flexible deployment options to provide global coverage across the data center as well as branch and remote offices.

3. **VTL capabilities**
   VTL-based data deduplication is one of the least disruptive ways to implement a tape-centric environment. Capabilities of the VTL must be considered as part of the evaluation process; it is unlikely that the savings from data deduplication 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 deduplication extension. Also consider how well the VTL can emulate your existing tape environment (e.g. same libraries, same tape formats) and communicate with your physical tape infrastructure if required.

4. **Impact of deduplication on backup and restore performance**
   It is important to consider where and when data deduplication takes place in relation to the backup process. Although some organizations attempt deduplication while data is being backed up, this inline method processes the backup stream as it comes into the deduplication appliance, making performance dependent on single node strength. This approach can slow down backups, jeopardize backup windows, and degrade the deduplication performance over time.

   By comparison, deduplication that runs after backup jobs are complete—or concurrently with backup processes—avoid this problem and have no adverse impact on backup performance. The backup data is read from the backup repository after backups have been cached to disk. This method ensures that backups are not throttled by any storage limitations. An enterprise-class solution that offers this level of flexibility is ideal for organizations looking for a choice of deduplication methods.

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

Restore performance is also crucial. Some technologies are good at deduplicating data but perform much slower when it comes to rebuilding data (often referred to as “re-inflating” data). If you are testing systems, you need to know how long it will take to restore a large database or full system. Ask solution providers to explain how they can ensure reasonable restore speeds.

5. **Scalability**

Because deduplication solutions are often chosen for longer-term data storage, scalability is an important consideration, particularly in terms of capacity and performance. Consider growth expectations over five years 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?

Deduplication solutions 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 deduplication efficiency or system performance. Clustering enables deduplication technology 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 backup repository interface (VTL or file interface) and deduplication nodes by offering failover support.

6. **Beyond tape backup**

Although backup operations have primarily depended on a tape backup paradigm, D2D backup operations can be more suitable for organizations that don’t have long-term data retention requirements and are willing to eliminate or lessen the use of tape. Data deduplication can reduce storage consumption in other data management processes beyond backup, including data archiving or database dumps. An ideal deduplication solution should be able to effectively support these processes.

7. **Distributed topology support**

Data deduplication should occur throughout a distributed enterprise, not just in the data center. A solution that includes replication and multiple levels of deduplication provides maximum benefits to customers. For example, a company with a corporate headquarters, three regional offices, and a secure disaster recovery (DR) facility should be able to implement deduplication in the regional offices to facilitate efficient local storage and replication to the central site. The deduplication 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 usage.

8. **Highly available deduplication repository**

It is extremely important to create a highly available deduplication repository. Since a very large amount of data has been consolidated in one location, risk tolerance for data loss is very low. Access to the deduplicated data repository is critical and should not be vulnerable to a single point of failure. A robust deduplication solution will include mirroring to protect against local storage failure as well as replication to protect against disaster. The solution should have failover capabilities 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 maintain ongoing business operations.

9. **Efficiency and effectiveness**
   File-based deduplication 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 4-megabyte presentation. In a file-based solution, the entire 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 deduplication processes use a "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 "chunking" begins at the beginning of a tape (or data stream in other implementations), the deduplication 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 thirty-forty percent increase in the amount of duplicate data detected. This can have a major impact on the cost-effectiveness of deduplication.

10. **End-to-end backup recovery process**
    It is important to keep in mind that deduplication is only one part of a larger data protection and recovery process that likely includes some or all of the following: backup within a specified window, copying data to tape, deduplication data, replicating data, restoring data and management. Placing focus on deduplication only may provide companies with a solution that breaks down somewhere in the larger process. Before investing in deduplication, understand the entire process from backup to restore, and know how to manage it.

Conclusion

As stored data volumes continually increase, while IT spending decreases, data deduplication is fast becoming a vital technology. Data deduplication is the best way to dramatically reduce data volumes, slash storage requirements, and minimize data protection costs and risks. Although the benefits of data dedpulication are dramatic, organizations should invest in a solution based on a comprehensive set of quantitative and qualitative factors rather than relying solely on statistics and hype.