Realtime publishers

The Essentials Series: The Evolving Landscape of Enterprise Data Protection

Lowering Costs of Data Protection through Deduplication and Data Reduction

sponsored by

syncsort

by Dan Sullivan

Copyright Statement

Lowering Costs of Data Protection through Deduplication and Data Reduction

Information technology (IT) professionals are living with something of a paradox—their colleagues see ads for 1TB drives selling for less than $150 and cannot understand why storage, backup, and disaster recovery services cost so much. The simple answer is that generating and storing data is easy and sometimes inexpensive. Protecting it over the long term is not.

The *Essential Series: The Evolving Landscape of Enterprise Data Protection* will examine several of the key issues in backup and disaster recovery while illustrating how to optimize services with virtual environments and existing infrastructure. This first article addresses how to lower the costs of backup and disaster recovery by taking advantage of deduplication and data reduction technologies. Not surprisingly, there are different ways of accomplishing this goal and each has their advantages and disadvantages. We will consider how these ultimately impact the return on investment (ROI), which will help lead to selecting the proper solution for a particular set of requirements.

Common Data Protection Requirements

Overall IT requirements will vary widely between businesses, but data protection requirements fall into a few common categories. Onsite backups are required for rapid restoration of data. Production application systems are of primary interest but shared network drives for client devices and development systems also require reliable backup services. Offsite backups are essential for full protection in the case of disaster. From isolated incidents, such as a building fire, to widespread natural disaster, such as Hurricane Katrina, unexpected, large-scale damage can leave production systems and backups decimated. Businesses must have a method for creating reliable and protected copies of critical data. Furthermore, these methods should be optimized to

- Minimize storage required for backups
- Minimize recovery time
- Minimize CPU, disk, and network overhead

As noted earlier, the cost of raw storage may seem inexpensive, but the growing volumes of data in many businesses quickly and significantly increase those costs. Thus, onsite and offsite backups should require as little storage as necessary. The adage "time is money" comes to mind when systems are down. A good backup solution that optimizes for storage space but entails a prolonged recovery process is not such a bargain. Finally, a backup solution should not overtax CPU, disk, and network resources. In the past, a conventional full backup was done at night when system and network loads were low; this is no longer the case. Today, some incremental backups run frequently during the business day in order to minimize the amount of data lost, and these operations should not adversely impact the performance of other systems. Fortunately, all these key objectives are met by reducing the size of backup data.

Methods to Reduce Data Protection Costs

There are three primary methods for reducing the size of backup sets relative to the size of source data:

- Reducing redundant copies of data
- Compressing data
- Deduplicating data and copying only changed data

Each of these methods provides a different combination of benefits.

Reducing Redundant Copies of Data

The ease of copying data has lead to a proliferation of duplicate data. Consider how quickly the average user can create multiple copies of data by saving multiple versions of documents as personally managed backups or by emailing attachments to multiple recipients. From the perspective of backup requirements, systems administrators do not need to save all copies of a document but they do need to be able to restore all those copies.

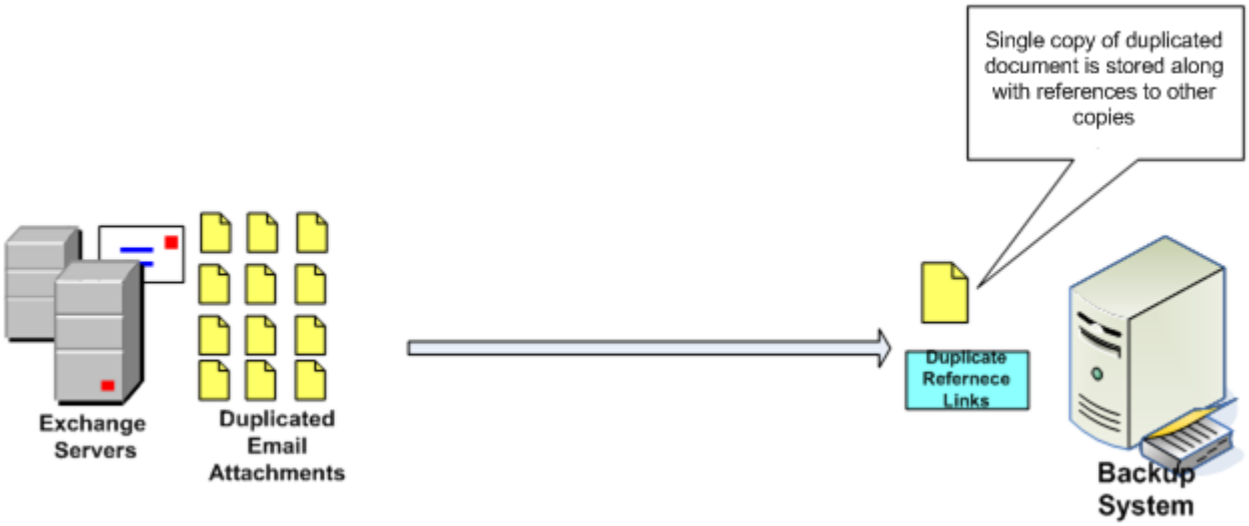

Figure 1: Multiple copies of the same data need not be copied to the backup device; a single copy plus reference data to recreate other copies is sufficient.

The advantage of this approach is that it reduces the amount of storage required on the backup devices. However, duplicate data detection is performed on the backup system, so the duplicate data is still copied over the network. The result is that there is no reduction in the overhead on the source system when backups are made; in addition, there is no reduction on network resources, and the time required to complete backup operations is not reduced.

Compressing Data

Data compression techniques can significantly reduce file sizes. Lossless compression algorithms are designed to represent the original data in more compact representation and to reproduce it accurately. Different lossless compression algorithms exist, and you can select an algorithm according to how it optimizes for three attributes: compression speed, compression ratio, and decompression speed. The algorithm in the popular PKZIP and gzip programs, for example, are optimized for ratio and decompression speed at the cost of slower compression. Lossless compression is required when data must be reproduced exactly as the original; for example, application data and word processing documents. In the case of media data, such as video and audio files, some loss of fidelity would be imperceptible to humans, so other algorithms, known as *lossy compression techniques*, may be used.

The advantages of compression are that size reduction can be significant for files with highly redundant data, which can be the case for some types of office documents and denormalized databases. Another advantage of this approach is that compression can be performed on either the source or target device.

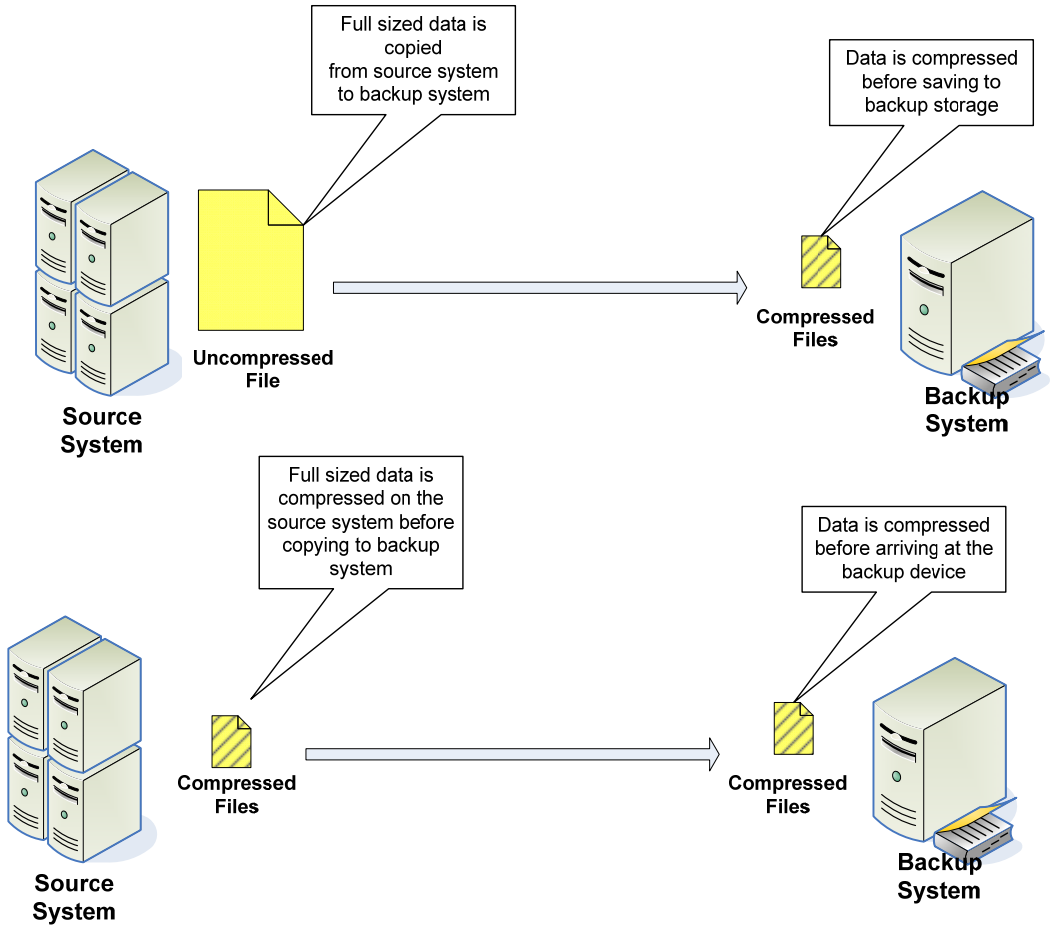

Figure 2: Data can be compressed either on the source or the backup system. Data compressed on the backup system is transmitted in uncompressed form from the source system, saving CPU cycles on the source system but requiring greater network bandwidth.

There are, however, several disadvantages to compression:

- Time required to compress and decompress files
- Compression is a CPU-intensive operation
- Compressing on the source system can reduce throughput of other applications
- Compressing on target system requires copying all data from the source system

These disadvantages stem, in part, from the fact that the operation deals with a large amount of source data. We are, in effect, trading the reduced load on of network bandwidth to transmit compressed files and reduced storage on the backup device for the cost of compute cycles to perform the compression.

Deduplicating Data and Copying Only Changed Data

One way to significantly reduce the amount of compute, network, and storage resources required for backups is to reduce the amount of data that is backed up. You can do so by focusing on only data that has changed since the most recent previous backup. For many business applications, data grows incrementally over time. We edit documents, send emails, and add records to databases. Even the most prolific email users create a small fraction of new messages in one day compared with what is stored in their long-term folders. Similarly, on any given day, most of the data in a database existed on the prior day. If there is a complete backup of yesterday's data, all that is required today is a backup of the changed data.

To be most effective, this type of incremental backup should occur at the block level. A large database file should not be completely copied when only a small fraction of the total blocks used have actually changed. In addition, unallocated blocks on a disk should not be copied either.

This approach offers several important advantages:

- Reducing the amount of data backed up, typically to levels less than achievable by deduplication or compression alone
- Backup times are reduced because less data is subject to copying and compression
- CPU requirements are reduced because smaller volumes of data are compressed
- More frequent backups are possible, providing more frequent recovery points
- Small backups increase backup success rates
- When block-level backups are performed, the process bypasses file system overhead and limitations, such as problems backing up open files
- With minimized overhead, backups are more cost effective for remote offices and offsite disaster recovery sites

The technical advantages of deduplicating data translate into a single, well-understood business advantage: maximized ROI. This result is due to reduced storage costs; you no longer need to maintain multiple full backups or file-based incremental backups. Also, frequent backups increase the number and recency of recovery points, thus leading to less data loss.

Summary

Backups are like insurance—you never want to be in a position to need it and you do not want to pay a lot for it, but when you do need it, you are glad it is there. Business data needs to be protected. We have backup and disaster recovery strategies in place to ensure business operations can continue in the case of adverse events, but these IT tasks are subject to the same business constraints as other business operations—they need to be cost justified and effective. The keys to successful data protection programs are focusing on meeting business objectives while lowering the costs of meeting those objectives. Fortunately, by adopting technologies and processes that deduplicate data in data protection operations you can improve ROI on those initiatives.