



# REPORT

## **Why CIOs Should Look To Data Deduplication**

**By Lauren Whitehouse & Brian Babineau**

**May, 2009**

# Table of Contents

<b>Table of Contents</b> .....	<b>i</b>
<b>Introduction</b> .....	<b>1</b>
<b>Data Growth Conundrum</b> .....	<b>1</b>
Relentless Information Growth.....	1
Information Costs .....	1
<b>Data Protection’s Complicity</b> .....	<b>2</b>
The Multiplier Effect .....	2
Compounding the Problem .....	3
A Difficult Balancing Act.....	3
<b>Controlling Secondary Storage Costs</b> .....	<b>3</b>
Do Not Store the Same Data Twice.....	3
Data Reduction Addresses Top IT Priorities.....	5
<b>Summary</b> .....	<b>7</b>

All trademark names are property of their respective companies. Information contained in this publication has been obtained by sources The Enterprise Strategy Group (ESG) considers to be reliable but is not warranted by ESG. This publication may contain opinions of ESG, which are subject to change from time to time. This publication is copyrighted by The Enterprise Strategy Group, Inc. Any reproduction or redistribution of this publication, in whole or in part, whether in hard-copy format, electronically, or otherwise to persons not authorized to receive it, without the express consent of the Enterprise Strategy Group, Inc., is in violation of U.S. copyright law and will be subject to an action for civil damages and, if applicable, criminal prosecution. Should you have any questions, please contact ESG Client Relations at (508)482-0188.

# Introduction

IT executives barely have time to keep track of the technology in their own environments; finding the hours in the day to stay abreast of all the new technology solutions in the marketplace is next to impossible. At some point, however, those in the data center corner office are bound to hear about solutions that can deliver near immediate payback and benefit other IT initiatives.

CIO Insight, a leading IT executive publication and accompanying Web site, recently published its *Top CIO Priorities for 2009* research, which found that 38% of its 200+ respondents have “cutting costs” as the top business priority for 2009, up from 29% in 2008. Thirty-seven percent of respondents stated that “reducing ROI costs” is the top management priority in the upcoming year while 34% said “improving the ROI in IT investments” is their biggest management initiative.<sup>1</sup> It is pretty clear that IT will spend very wisely in 2009—with investment paybacks being measured in months rather than years.

An investment in data deduplication products is easy to rationalize. These products offer benefits to more than just the storage group because they cut data protection capital and operating expenses, facilitate consolidation of distributed backup operations, and slash server virtualization-related storage costs. Data deduplication offerings should, therefore, be on every CIO’s project short list in 2009.

## Data Growth Conundrum

### Relentless Information Growth

ESG estimates that database data is growing at 25% per annum, with unstructured data increasing at two to three times that rate.<sup>2</sup> This growth is fueled by a dependence on digital assets to conduct business and the need to support an increasingly mobile workforce. Collaboration, Web 2.0 applications, and use of messaging systems also contribute to information growth.

Retention policies dictated by corporate and regulatory mandates exacerbate primary information growth as organizations are required to save data for longer periods of time (see Figure 1). Additionally, companies are saving historical information in an effort to improve business intelligence processes as decision support functions benefit from having access to more data.

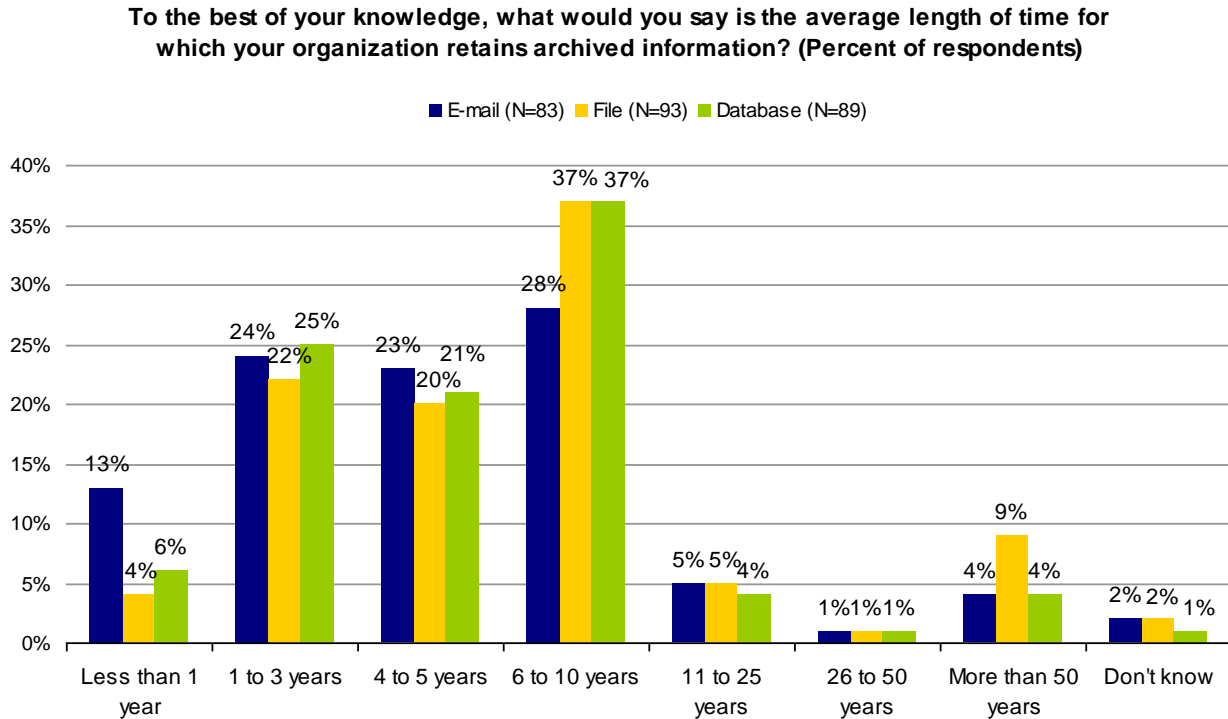
### Information Costs

IT executives must factor annual data growth rates into all areas of the budget. With more information being created, more primary storage capacity is required. The increase in capacity requirements may affect the primary storage systems’ footprint in the data center and potentially require the company to secure or rent additional floor space. Further, storage operating costs such as the associated power and cooling requirements, additional networking infrastructure, redundancy components, and resource management software licensing will also grow when more storage is bought. An increase in primary storage capacity, in turn, triggers an increase in secondary storage capacity (disk and/or tape), media management servers, backup software licensing, backup reporting software licensing, and offsite media expenses. IT must staff appropriately to manage the entire infrastructure, which represents another expense that sometimes gets overlooked in the cost of information growth.

<sup>1</sup> <http://www.cioinsight.com/c/a/IT-Management/Top-CIO-Priorities-for-2009/>.

<sup>2</sup> ESG Research Report, *Database Archiving Survey*, December 2007.

**FIGURE 1. AVERAGE RETENTION PERIOD FOR ARCHIVED FILE, E-EMAIL, AND DATABASE DATA**



Source: Enterprise Strategy Group, Digital Archiving Survey, December 2007

Because primary information growth also occurs outside of the data center, most notably at remote and branch offices (ROBOs), organizations must also budget for infrastructure and staff to handle distributed data. Sometimes, these expenses fall under the CIO’s purview; in other instances, the remote office must fund IT expenses out of its operating budget. With a global economy, employees who can work almost anywhere, and outsourcing partnerships, ROBO IT costs are quickly impacting the corporate bottom line.

## Data Protection’s Complicity

### The Multiplier Effect

Primary data growth is expensive, but the biggest contributors to the “cost of information” are all the copies made for data protection purposes. When ESG asked nearly 400 IT decision makers what their greatest data protection challenge was, the top response was “keeping pace with the capacity of data to protect.”<sup>3</sup> IT organizations have standard practices in place to protect all digital data records within the organization. Typically, that means IT makes a copy of a volume, LUN, or file(s) at one or more points in time during the day and saves the copy—locally for operational recovery and at an offsite location for disaster recovery (DR).

The problem is that data protection operations can be inefficient—backup applications make many backup copies of the same (or slightly modified) file when only a small amount of the data within the file has actually changed. Dozens of copies of the same data may be made and stored for lengthy periods of time—even when the file is not changing or has lost its usefulness to the organization. Consider the following example:

<sup>3</sup> Source: ESG Research Report, *Data Protection Market Trends*, January, 2008.

- A file is created and backed up the same day.
- The file is continually updated and backed up with incremental backup strategies over the course of a week.
- The file is then e-mailed to a group of people and is backed up anew as part of the e-mail application backup.
- One or more of the recipients modifies the file slightly (changes the date on the cover page of a presentation, for example), which is backed up again in the next incremental backup.
- A copy of the file is made under a new name and is selected for backup again.
- In the meantime, every on-premises copy of a backup is replicated offsite, doubling the copy instances.

In this scenario, it is easy to see the level of inefficiency in the backup process. Highly redundant backup files clog LANs, WANs, and SANs and consume on- and off-premises storage capacity. Therefore, the data protection process and secondary storage systems contribute significantly to the capacity glut problem and present the most glaring opportunity for optimization.

### Compounding the Problem

In some instances, organizations are adding to the data protection capacity problem by implementing new technologies to solve other IT-related problems. For example, many CIOs are driving data center consolidation and 'green' projects by deploying server virtualization solutions. These solutions allow customers to run multiple servers on a single piece of hardware, which drives up utilization. However, ESG research found that more than one-third of organizations that have implemented server virtualization technology have seen an increase in the total amount of data to back up.<sup>4</sup> Since virtual machine disk images contain the operating system, applications and data, there is a high degree of redundant information across virtual machines on a single physical server. The .vmdk files for ten virtual machines running Windows will contain ten very similar binaries, patches, and auxiliary applications.

### A Difficult Balancing Act

As the capacity of data grows and regulatory mandates dictate longer retention periods, the amount of data under management may exceed the time allocated for backup. In an effort to reduce backup times, IT organizations are deploying disk in their backup processes at an increased rate. However, ESG found that the cost of storage systems is another top concern, creating a conundrum for IT organizations. How can IT provide adequate service level agreements for data protection, keep pace with data growth, and keep spending in check?

## Controlling Secondary Storage Costs

Data capacity growth is not going to abate. Data protection processes, such as backup and replication, magnify capacity growth. Therefore, it makes sense for organizations to employ tactics and technology to optimize this environment first—without sacrificing performance or introducing risk with recovery practices. Data deduplication has emerged as a compelling technology to control storage capacity and costs.

### Do Not Store the Same Data Twice

Data deduplication identifies and eliminates redundant data. It can be performed at the file, block, or byte level. The opportunity to find and eliminate redundancy becomes greater with more granular examination. In secondary storage processes, such as backup, data is initially seeded on the secondary storage device and all subsequently written data is examined for redundancy. Replicate data is not stored twice; instead, a pointer to the stored duplicate data is written (which takes up significantly less space).

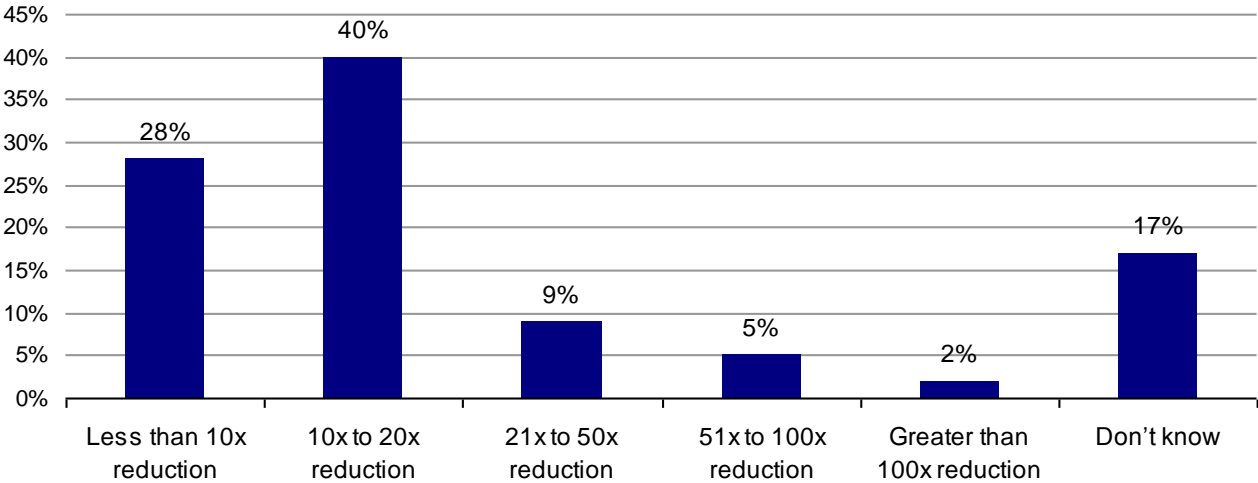
---

<sup>4</sup> Source: ESG Research Report, *The Impact of Server Virtualization on Storage*, December 2007.

Regardless of the implementation method, deduplication delivers measurable results. One of the key measures is the degree of capacity reduction, or “reduction ratio.” A “10x,” “10:1,” or 10 times reduction indicates that an organization was able to reduce the size of a backup, 500 GB for example, to just 50 GB. As shown in Figure 2, among data protection survey respondents, 48% of those using deduplication reported a 10-20x reduction and 18% reported reductions ranging from 21x to more than 100x. While data deduplication ratios will vary based on the type of data, frequency of full backups, retention, inter-file and inter-application redundancy, local or global deduplication, deduplication algorithm and more, ESG Lab’s testing has found a 20:1 reduction ratio to be broadly achievable.

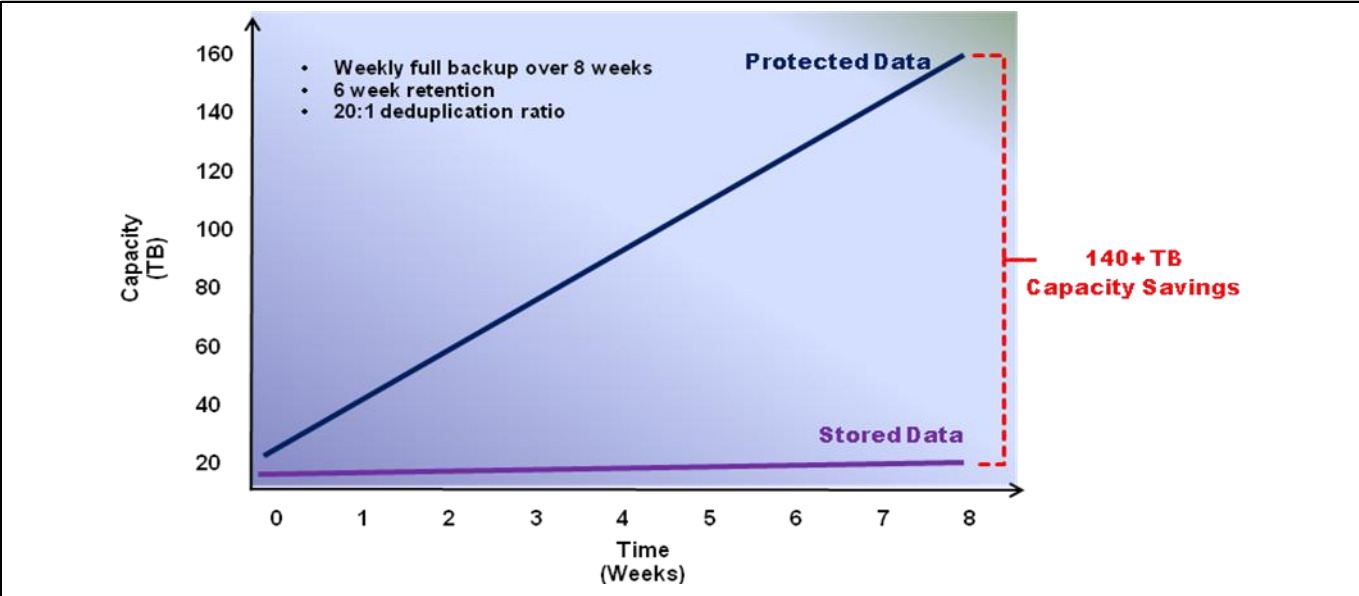
**FIGURE 2. REDUCTION RATIOS**

**What degree of capacity reduction has your organization experienced by using data de-duplication technology? (Percent of respondents, N = 58)**



Source: Source: ESG Research Report, Data Protection Market Trends, January 2008

**FIGURE 3. DEDUPLICATION IMPACT**



Deduplication ratios of 20:1 can produce significant capacity savings. In the example in Figure 3, weekly backup of 20 TB of data would normally balloon to 160 TB of backup capacity over an 8-week period (with a 6-week retention setting). Applying deduplication with a 20:1 deduplication ratio would deliver a savings of over 140 TB of capacity, requiring less than 18 TB of deduplicated storage capacity.

The amount of data stored—either due to a greater frequency of full backups or longer retention times—tends to increase data deduplication ratios. This provides more incentive for organizations to leverage deduplication solutions wherever possible because the capacity and associated budget savings are likely to improve, while also improving the likelihood that data can be recovered from disk.

### **Data Reduction Addresses Top IT Priorities**

The capacity savings seen in the example are just the tip of the iceberg. Improvements may be realized in many places in the secondary storage environment, addressing several of IT's top priorities or initiatives, including:

- Optimizing secondary storage environment and processes
- Supporting Green IT initiatives
- Cutting costs

### ***Better, Lower-Cost Data Protection***

Deduplication changes the economics of disk-based data protection. First, it makes the transition from tape- to disk-based protection more palatable as it drives the total cost of ownership for disk-based backup closer to that of a tape-based strategy. Capital cost savings associated with replacing a tape-based approach may encompass: tape infrastructure (hardware and software licenses), tape media acquisition, and disaster recovery costs. Second, deduplication optimizes disk-based backup environments as companies can replicate more data for disaster recovery more efficiently. With duplicate data being removed, companies do not have to buy as much disk capacity at the remote site and the replication process does not require such significant network bandwidth infrastructure.

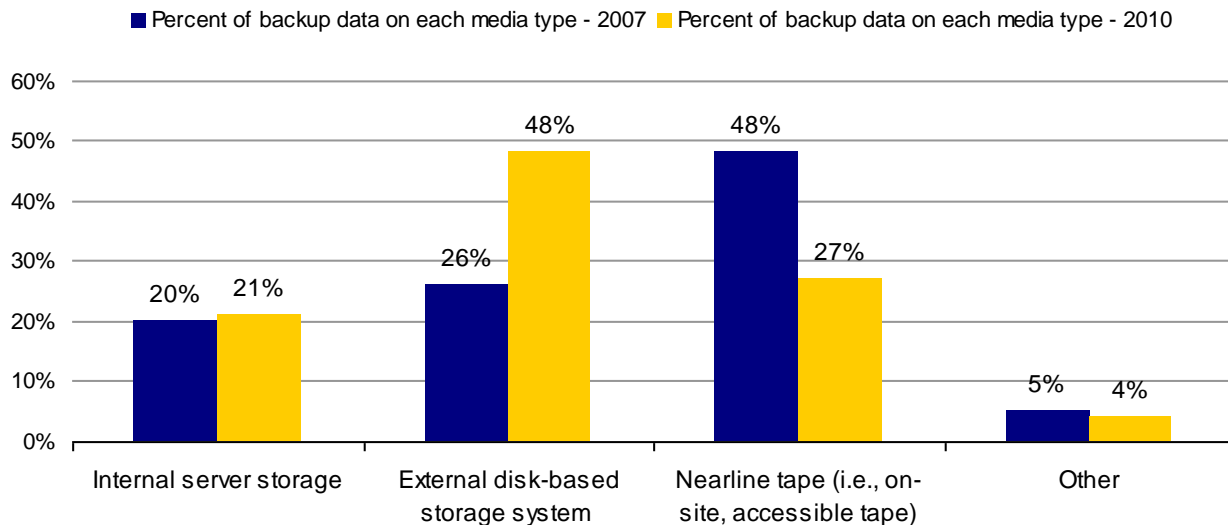
The reduction in backup data as a result of deduplication allows companies to consolidate more backups on fewer devices. Organizations can also choose to increase retention policies for data, which helps make compliance and electronic discovery recoveries go smoother since the information is more accessible (when compared to the data being saved on tape). Most importantly, organizations can reduce backup windows and improve restoration times by using disk versus tape. Recovery time objectives (RTOs) will improve because data can be recovered from disk. With more capacity available, IT may choose to increase the frequency of backups conducted during the day, improving recovery point objectives (RPOs).

Organizations are quickly realizing the benefits of using disk-based backups for onsite data protection and are commencing tape replacement projects. ESG research indicates that nearly 50% of onsite backup data will be stored on disk in 2010, up from 26% in 2007 (see Figure 4).

The impact on operational budgets may be seen in a number of ways. For tape replacement scenarios, operational overhead for tape handling, troubleshooting, and manual intervention in the backup process can be eliminated, as can power charges, tape hardware and software maintenance fees, and media storage costs. For companies using disk in their backup schemas already, deduplication can drive consolidation, which should reduce power costs as well as minimize data center floor space consumption.

**FIGURE 4. EXPECTED INCREASE IN ON-SITE EXTERNAL DISK SECONDARY CAPACITY BY 2010**

Approximately what percentage of your organization's total on-site backup data is currently stored on each of the following storage media types? Please also indicate what you expect these percentages to be in 2010? (N = 364)



Source: Source: ESG Research Report, Data Protection Market Trends, January 2008.

### Backup Consolidation

Another benefit of capacity optimization will be realized with network bandwidth. Less data means less network traffic. This benefit enables consolidation of backup data from distributed sites, such as ROBOs, to a central site. Tape-based backup infrastructure, processes, and tape handling overhead can be eliminated at distributed sites when deduplication is added to the local backup process. Remote site disk-based backup performed locally can provide operational recovery while replicating the deduplicated backup store to a central data center can provide disaster recovery.

### Support for Multiple IT Initiatives

IT staffs continuously look for ways to improve resource utilization, drive efficiencies, and generate better service levels. Many of the measurable benefits of deduplication assist with these objectives, with much of the positive impact taking place in the storage environment. For example, lowering capacity requirements can impact sustainability efforts. As previously discussed, capacity optimization can postpone additional capacity purchases as well as reduce power consumption and required data center floor space. In the case of tape elimination, the associated facility and environmental costs of the tape infrastructure may create negligible power and cooling savings versus a disk-based backup system with deduplication. Seventy percent of business executives measure the success of corporate green initiatives by tracking reductions in energy costs.<sup>5</sup> If IT executives want to align with business priorities, cutting power consumption via deduplication is a great start.

Another area where data deduplication supports IT initiatives is in data center consolidation. These solutions reduce the number of storage systems needed to support backup and disaster recovery and help mitigate the need for IT operations at distributed locations. Data deduplication also facilitates server virtualization deployments as it eliminates much of the downside of server virtualization projects—virtual machine disk images contain highly redundant data and increase storage capacity requirements. Through server virtualization, customers can reduce the amount of servers in their environments and through deduplication, they can reduce the amount of storage capacity.

<sup>5</sup> Source: ESG Research Report, *Global Green Business and IT Initiatives*, March 2008.

## Summary

The more information there is, the more it costs to maintain. Although always an important consideration, current financial uncertainty has elevated cost control to the top of IT management's priority list, creating an imperative for IT organizations to optimize environments and processes and wring out cost savings where possible. Importantly, cost reduction and efficiency cannot be gained at the expense of providing high levels of service—especially when it comes to protecting the information that drives the business.

Deduplication is one of the few IT solutions that cuts costs quickly and improves service levels. With it, organizations can reduce storage expenses without compromising data protection. It can also help align IT with business priorities—which 61% of IT executives said was their top management priority in 2009<sup>6</sup>—by driving green and data center consolidation initiatives. Now that deduplication has proven itself out over the last few years, enterprise-class organizations are more widely adopting it. For these reasons, ESG believes CIOs and IT executives should look to deduplication for cost reduction projects in their environments.



20 Asylum Street  
Milford, MA 01757  
Tel: 508-482-0188  
Fax: 508-482-0218

[www.enterprisestrategygroup.com](http://www.enterprisestrategygroup.com)

---

<sup>6</sup> <http://www.ciainsight.com/c/a/IT-Management/Top-CIO-Priorities-for-2009/>.