Lab White Paper

The Cost of Managing Unstructured Data

By Kerry Dolan, Lab Analyst and Brian Garrett, Vice President, ESG Lab

May 2014

This ESG White Paper was commissioned by Hewlett Packard and is distributed under license from ESG.

Contents

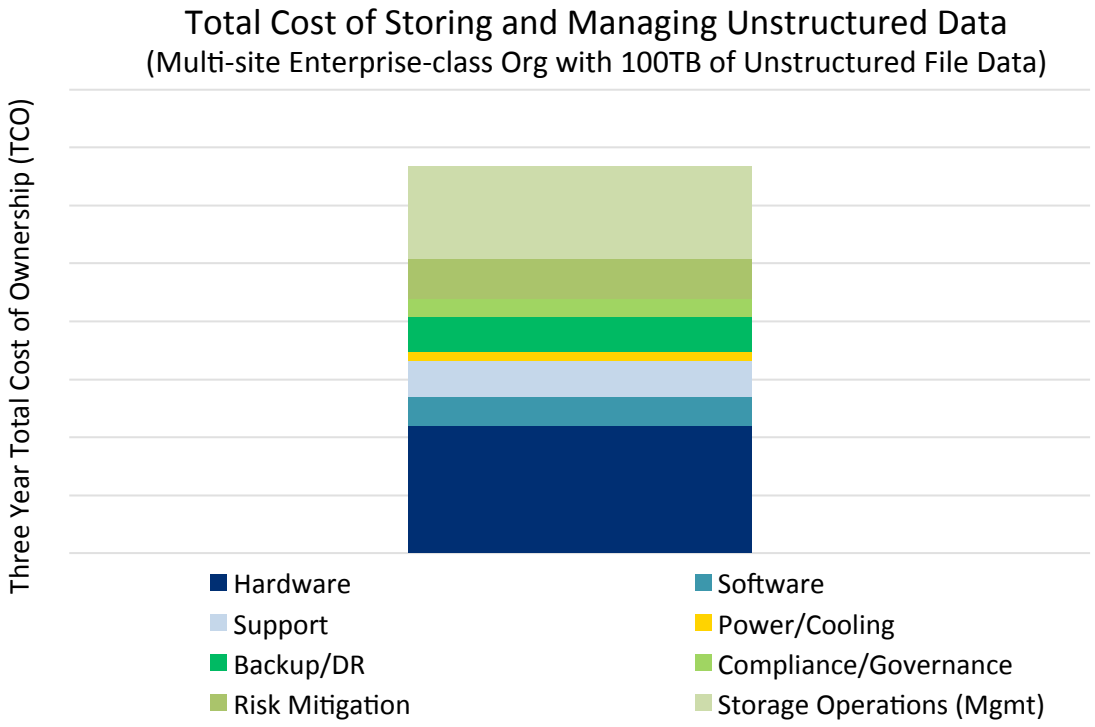
Executive Summary

Many in the storage industry talk about the cost of managing unstructured file data, but there are few measurements of it. Based on decades of research and experience with storage total cost of ownership (TCO) models, ESG has developed a model for estimating the cost of managing unstructured data to give customers a more concrete idea, rather than a vague concept. The cost of managing 1GB of unstructured data will vary quite a bit by industry and by company, and is affected by many factors: configurations, business policies, compliance and governance regulations, etc.

> *ESG found that the costs vary anywhere from $4/GB to $100/GB, but $25/GB is a good rule of thumb for a typical multisite, enterprise-class organization.*

Where do these costs come from? The chart in Figure 1 shows the cost model for storing and managing unstructured data in a traditional way. We'll be taking a closer look at the assumptions and costs with this model later in this paper, but let's focus first on the blue bars toward the bottom for acquisition costs: hardware, software, and support. Many organizations compare only these costs when selecting solutions, ignoring what tend to be the heavier costs, shown in green bars: backup/DR, compliance and governance, risk mitigation, and storage management. As you can see, the costs associated with *managing* unstructured data can be 2X or more than the cost of storing that data. Organizations that neglect to consider these costs are likely to make decisions based on insufficient information, often leading to higher costs to meet the true needs of the business.

Figure 1. The Total Cost of Storing and Managing Unstructured File Data



The bottom line is that enterprise organizations need a way to measure not just the acquisition costs of storing unstructured data, but the costs of managing it as well. There are costs associated with redundant/obsolete/trivial data (ROT), dark data that could be valuable or detrimental, and hidden data that could impact risk and compliance objectives. HP ControlPoint can help identify this data so organizations can either use it for the business or delete it to reduce costs. HP can help customers identify the TCO of their specific storage environments to jump-start the process of taking back control of data.

Measuring *All* Costs Is Essential to Saving Money

Digital infrastructure is a critical component of every organization today, defining an organization's ability to communicate, analyze trends, build new products, and enter new markets. Data storage is a significant piece of the puzzle that provides the essential foundation for every application, document, database, analysis, and process in today's business environment.

The storage infrastructure required for both structured data (such as databases) and unstructured data (such as e-mail, business documents, images, video, etc.) can be expensive. But where are the true costs, and how can you bring them down? The challenge is that much of the cost is not easy to ascertain, and you cannot improve what you haven't measured. Because hardware costs are easy to find (and tend to be high), many organizations make decisions based only on this information. However, as the price of disk has dropped, CapEx takes up less than half of the costs—OpEx consumes the rest with factors like storage management and administration. But other less visible costs can have significant impact, such as the processes required to minimize risk and maintain compliance, and the lack of visibility into data. These costs recur with every business process unless addressed.

What customers need is a way to measure both hard costs (CapEx and OpEx) and soft costs (such as failure to proactively manage data) for their own organizations. If organizations can discover exactly what is costing them the most, mapped to their specific infrastructures and business processes, they can reduce the costs that are most burdensome. It's easy to get caught up in the current hype, make decisions based on it, and assume all is well. A good example is the recent proliferation of solutions focused on reducing power consumption—such as solid-state disk and disk drives that automatically spin down when not in use. These are cool ideas, but if power is not one of the highest costs in your particular situation, why spend money on it? Wouldn't your cash be better spent on solutions that rein in the specific costs that are slamming your budget?

> *Ultimately, what matters is not how much your storage costs, but gaining the maximum value from your data with the lowest TCO.*

How much does it cost to collect, retain, protect, and use your information? You can easily calculate the cost per 1GB for acquiring hardware. If you can also figure out how much it costs to *manage* 1GB of your unstructured data, you can plug that into your TCO calculations and make better decisions for your business.

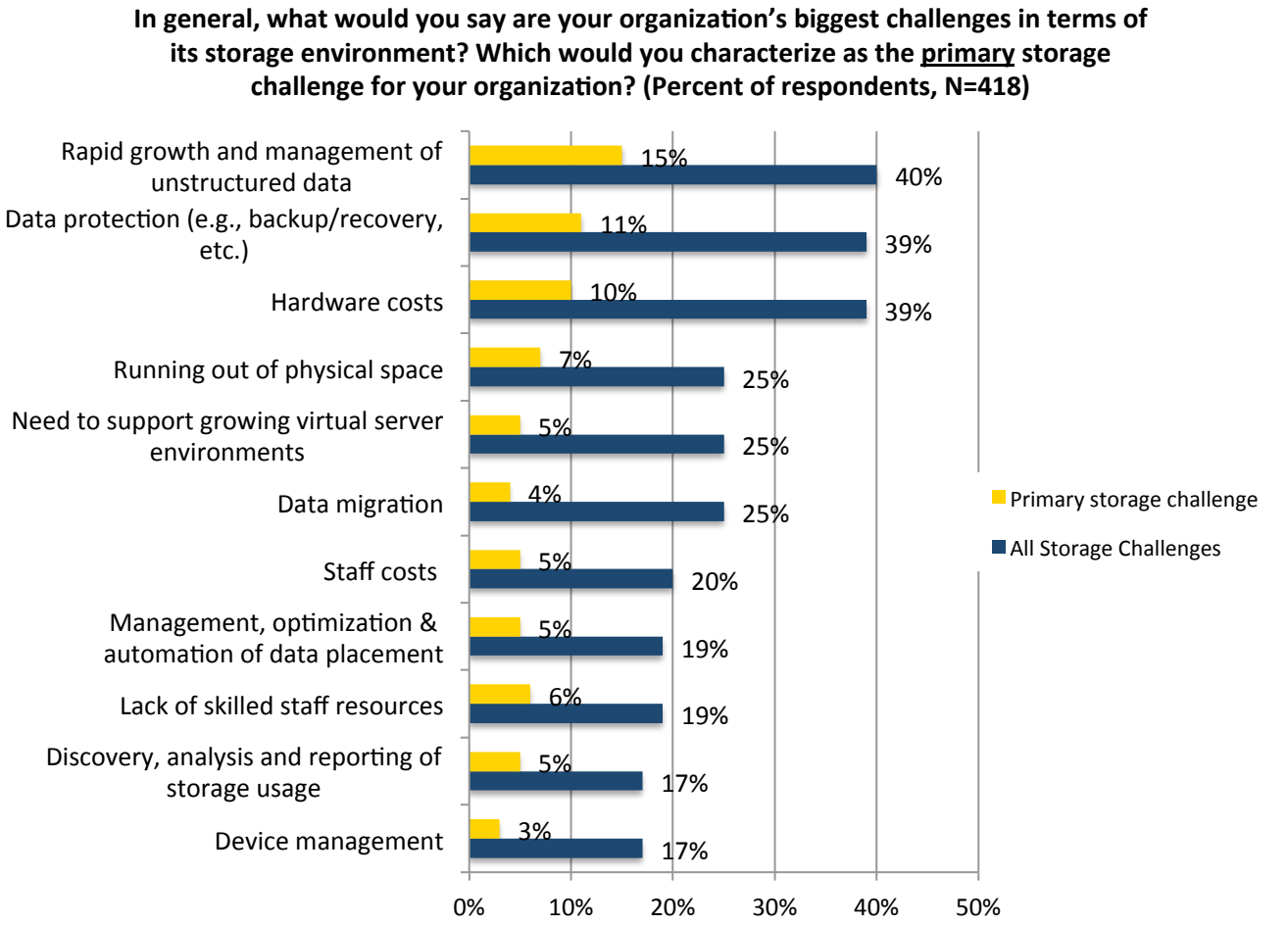
Data Growth Remains a Constant Challenge

An easy place to start looking at cost is in the continual onslaught of growing data volumes, which add to both CapEx and OpEx. According to ESG's annual IT Spending Intentions Surveys, managing data growth has been among the top four most important IT priorities reported by respondents for the past four years, so it's clearly a concern.[1] How important is it? You can tell by the company it keeps: Its place in the top echelon is shared with such critical priorities as server virtualization and information security initiatives.

In particular, the growth and management of unstructured data is wreaking havoc on IT operations and budgets. Key locations of unstructured data typically include file shares, plus Microsoft Exchange and SharePoint repositories. According to ESG research, when respondents were asked to name their organizations' biggest challenges pertaining to the storage environment, the most cited reason (provided by 40% of those surveyed) was the rapid growth and management of unstructured data (see Figure 2).[2]

[1] Source: ESG Research Report, *2014 IT Spending Intentions Survey*, February 2014.
[2] Source: ESG Research Report, *2012 Storage Market Survey*, November 2012.

Figure 2. Storage Challenges



Source: Enterprise Strategy Group, 2014.

How Is Your Data Costing You Money?

The overall cost picture starts with the basics: primary systems used for production data, and secondary systems for backups and long-term archiving. Accompanying these systems are costs for administration, data center floor space, power for operating and cooling systems, and, for some, infrastructure hosting services.

These costs generally apply across companies and industries, but other costs vary. A good example is the cost of configuring and managing your environment to comply with industry regulations and corporate governance. In government organizations as well as in the healthcare, legal, and financial industries, there are strict requirements for minimizing the risk of data loss, keeping data highly available for discovery, retaining data over the long term, ensuring privacy, etc. These requirements add to both the hard and soft costs.

Costs Depend on Workloads

The storage options for unstructured data include direct-attached storage (DAS) and networked storage (SAN or NAS), and this is an illuminating place to begin to understand TCO. In terms of acquisition, DAS is usually less expensive than networked storage, but that is only the high-level view. Let's say this storage is for your Exchange deployment that handles e-mail for the entire organization, so the environment must perform fast and remain available at all times. While DAS hardware costs less, the only way to deliver on SLAs for high performance and availability is to over-configure hardware, make additional data copies, and move data around to ensure that it's always available. The extra capacity and management required to deliver on the business requirements drives the cost of DAS much higher, making networked storage a better option. The problem is that many organizations don't

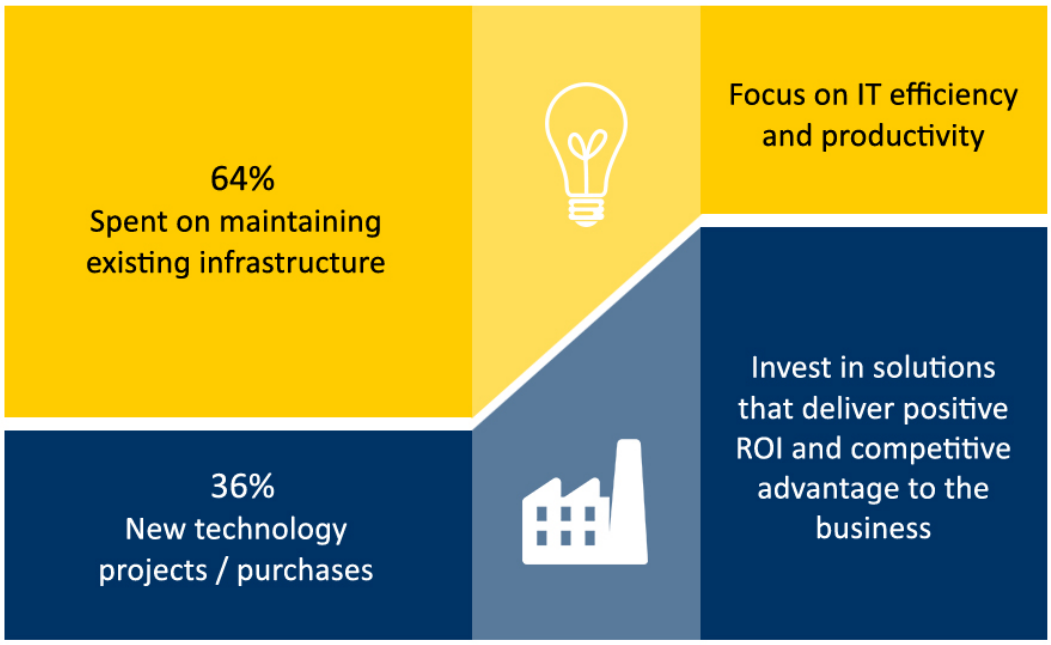
make decisions with these scenarios in mind; they look only at vendor pricing, and miss the critical details that could save them money.

Lack of Understanding about Your Data Is Expensive

Unfortunately, business decisions are made every day with insufficient understanding around data and its costs. Many organizations simply don't know what unstructured data they have, how many copies there are, what they need to keep, or what they need to remove. Employees may be creating and copying documents, downloading video and pictures, and generally growing data volumes of which IT is unaware. As a result, a lot of data may be clogging your network pipeline, taking up storage space, and consuming backup and archiving cycles unnecessarily. This can cause interruptions to production operations, such as when backups grow beyond the time allotted to execute them.

Another result is that organizations end up spending the largest portion of their technology budgets on "keeping the lights on" rather than on using technology to innovate and move the organization forward. According to ESG research, respondent organizations expected to spend nearly two-thirds (64%) of their 2013 IT budgets on maintaining existing infrastructure, and only 36% on new technology projects (see Figure 3).[3]

Figure 3. Budget Allocation



Source: Enterprise Strategy Group, 2014.

Forward-thinking organizations search for ways to spend more of their budgets on innovative solutions that provide strategic value and free up IT resources to focus on driving business value rather than maintaining infrastructure. It's hard to do that when you can't identify what data you have and what it's costing you.

Defining ROI for Unstructured Data: Hard costs

Organizations must figure out what their hard and soft costs are, with goals of 1) measuring them, and 2) reducing them. The following outline of key components can help you identify costs in your organization. We'll start with the "hard costs" that are fairly visible, if not always easy to measure.

[3] Source: ESG Research Report, *2013 IT Spending Intentions Survey*, January 2013.

Hardware

The cost of acquiring servers, networking gear, storage arrays, and disk capacity, including upgrades, maintenance, and depreciation are relatively easy to find as they are priced by vendors. These include:

- Primary: Production systems. Servers, filers/arrays, gateways, disk, software, networking, and security. Key contributing factors are data size, ongoing growth, and data copies.
- Secondary: Backup/archive systems. Backup servers, local and remote filers/arrays, disk/tape media, software, networking, and security. Size and growth are key here, plus the frequency of copying and the retention time. The more copies that are created and the longer data is retained, the higher the cost.

Data Center Costs

Data center floor space, energy for power and cooling, and hosting fees are all important factors. In addition, it's worth looking at utilization rates for production and non-production systems, particularly storage. High utilization indicates that the purchase is being used fully and not wasted, while low utilization indicates that an organization is incurring a cost without getting back value. Consolidated storage can greatly improve utilization rates and reduce management complexity.

Labor/Administrative Costs

Outside of salaries, these costs are much harder to define. They include operating and managing the infrastructure for all activities, across all applications, and across primary and secondary systems. These include:

- Implementation/deployment costs
- Ongoing provisioning and management
- Data migrations
- Upgrades

The less you know about your data, the higher all of these costs will be. Streamlined processes executed on the right data at the right time will result in lower costs; conversely, organizations that are provisioning server and storage resources and meeting SLAs for large volumes of data that don't need to be retained are absolutely losing money. The essential realization is this: IT productivity will increase as you gain visibility into data categories and can match services to requirements in the most efficient and granular way.

Data Copies Slam Storage and Management

The number of data copies an organization creates and retains has an impact not just on the cost of storage capacity, but the cost of management as well. Typically, structured data includes the onsite production copy, a local copy for backup, a remote copy for disaster recovery, plus copies for test/dev (often numerous copies for multiple developers) and to load data warehouses to perform analytics. An organization may have eight copies of a data set, all of them getting backed up over and over again. This is where data deduplication features can make a big difference.

Not all organizations need bulletproof DR capability for unstructured data, although some do. So file shares might have a production copy plus a backup copy, while Exchange and SharePoint repositories may have more copies. *But the bigger problem with unstructured data is that IT often has no idea how many copies there are*. An employee has a video on her home directory, she sends it to another employee, and now they both have it on home directories plus there's a copy on the corporate share. The same video file is getting backed up on different shares, but if IT is only viewing at the volume level, they don't recognize the inefficiency. If the HR department sends out a document to all 5,000 employees in a company, 5,000 are copies out there in different file shares. It brings up the question: What else doesn't IT know about? Is there data out there that should be protected but isn't? Are there copies that should be deduplicated? Is there data that shouldn't be stored in any repository at all, but instead should be removed?

Defining ROI for Unstructured Data: Soft Costs

The soft costs are a different problem altogether, of which many organizations are simply unaware. Three common soft costs are redundant/obsolete/trivial data (ROT), dark data, and risk/compliance.

ROT

One of the most insidious challenges is ROT: Data that is redundant, obsolete, or trivial that takes up storage capacity, clogs networks, lengthens the backup window, and consumes management cycles. Organizations are spending money to store, manage, and copy it for data protection over and over, often without even knowing it's there. Examples include:

- Redundant: duplicate copies of documents and files, e-mails records, or database information in file shares, on SharePoint sites, or in mail systems
- Obsolete: data that is out of date or no longer being used
- Trivial: file types that offer no content value, such as executable files, system files, and thumbnails

If you can find ROT data, then you can choose what to do with it—in most cases, the best strategy is to delete it and stop spending hard-earned cash on storing, managing, and protecting it.

Dark Data

Similarly, there is what's known as "dark data." Dark data is unstructured data that is not indexed or categorized, often stemming from employees downloading audio and video, participating in social media, etc. Dark data is collected but not used, taking up hardware capacity and management cycles without contributing to the business at all. Some of this data may have real business value, but because it's hidden, it is not being leveraged for business benefit. Other dark data could be detrimental, but because no one knows about it, it stays in your system like a ticking time bomb instead of being removed. More than a few organizations have found themselves in hot water because of data they didn't know about that turned up the legal heat when they were under subpoena.

Risk and Compliance

Most enterprise-class organizations must comply with stringent regulations regarding data availability, protection, retention, and privacy. These include HIPAA in the healthcare sector, FISMA for government dealings, Sarbanes-Oxley for publicly traded organizations, PCI DSS for handling credit cards, and GLBA for financial industry dealings. Not every organization must meet these regulations, but those that do are taking tremendous risks if they don't have a handle on all their data. They face potentially enormous regulatory fines, legal exposure, and lengthy periods of low productivity while they try to recover needed files. Numerous organizations have been bankrupted by improper compliance and risk procedures. Some organizations go to the other extreme, providing all their data with fully protected, highly available environments so they are covered for compliance, but this is a huge waste of money for data that doesn't need to be governed this way.

Failure to proactively manage these information categories can be very costly. However, managing them manually can be cumbersome, time consuming, and equally expensive. HP ControlPoint can automate management to streamline processes and reduce costs.

HP ControlPoint: Policy-based Information Management

HP ControlPoint offers a path through the maze of siloed repositories of unstructured data that IT struggles to identify and properly manage, enabling organizations to optimize information for business benefit and reduce TCO.

ControlPoint helps customers to identify, analyze, and control content across multiple repositories, and apply policies to data to ensure that it is managed optimally—for both functionality and cost—now and in the future. Using HP's Intelligent Data Operating Layer (IDOL) and connectors (such as for Microsoft Exchange, SharePoint, and file shares), ControlPoint helps organizations to classify and categorize data and apply policies to the content. Organizations gain a centralized console for governing all connected data sources, and can define and apply policies regardless of the data format or location. In addition, it brings risk, compliance, legal, and other managers together to enforce policies in a single system. HP ControlPoint includes graphical visualization of information clusters for easy identification of trends, themes, and concepts.

The key to ControlPoint's value is that it automates categorization and policy application so they remain consistent across file formats. Training the content categories using IDOL relieves organizations from manually creating or mapping categories. Storage management, deletion prevention, and content disposition policies are automatically enforced, with deduplication across repositories to minimize storage costs and reduce discovery times.

Not Just Dedupe, but Policy-based Governance

Storage deduplication is effective for reducing storage volumes and management for redundant data copies. But it's not always that simple to identify, and that's where policy-based governance with HP ControlPoint can help. While the concept of ROT is clear, actually *defining* what is ROT in any organization is not. Is the data redundant if both HR and Finance need their own copies of your personnel file? When does data become obsolete: based on the date it was created or the date it was last modified? And what is trivial to one department might be critical to another. The business processes and methodologies around identifying what is ROT and what to do with it can consume a huge amount of management time, often by high-level personnel. Clerical employees are usually not in a position to determine what is ROT; it may take lawyers focused on discovery or senior managers focused on financial data. It's a granular process that takes a lot of people and a lot of time.

> *Typical example: If it takes 2,000 hours to determine data policies using attorneys, compliance officers, and director-level staff, and their average cost per hour is $250, that's $500,000 just to figure out what to delete and what to put to better use.*

HP ControlPoint can help identify ROT so organizations can take action on it; even more valuable is its ability to illuminate the business context of unstructured content, dramatically reducing the time required for policy definition. And what can you do with the valuable data you find? HP ControlPoint supports a number of options. IT can manage it in place or move it to a repository such as HP Records Manager for safekeeping. Data that has business value can be used to gain business insight and inform decisions.

HP ControlPoint Benefits

HP ControlPoint provides visibility into your data, bringing hidden data into the light so that it can be used or deleted. This enables organizations to gain more value from data they've collected, and eliminate risks they didn't know they had. In addition, it enables organizations to reclaim storage capacity—production, backup, and archiving—as well as reducing licensing costs, data center floor space, and energy. Equally important, by reducing data volumes, you reduce management costs.

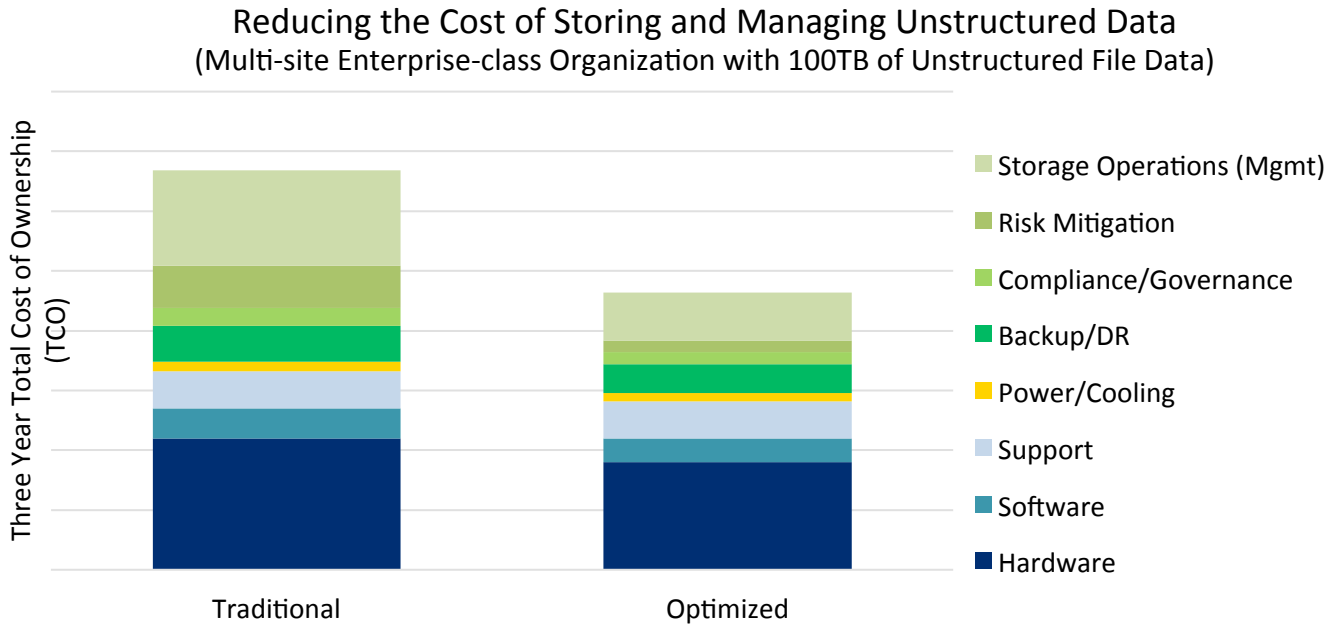
With ControlPoint, organizations gain both visibility and an understanding of their data's business value or risk, automated management, and automated policy enforcement. Organizations gain a much better handle on ROT and how to address it, and can much more easily manage legal hold, compliance, etc. By managing data better up front, ControlPoint helps to reduce the downstream burden to secondary systems (such as backup and archiving) and their management.

The Bottom Line: Reducing Costs with HP ControlPoint

A review of five storage TCO models that had been developed by ESG and an audit of an HP ControlPoint TCO/ROI model were used to analyze the three-year total cost of unstructured data in a multisite, enterprise-class organization with 100TB of network-attached, shared file data. While the absolute numbers will vary for each customer, ESG is confident that if you take a close look at the relative costs associated with storing and managing your unstructured data, the results will look similar to those shown in Figure 4:

- Storage CapEx is less than half of the total cost.
- OpEx is dominated by storage management, risk mitigation, and compliance/governance.
- An optimized solution that minimizes ROT and simplifies management reduces *all* of the costs, both CapEx and OpEx, with the biggest bang for the buck coming from simplified management and risk mitigation.

Figure 4. Reducing the Costs of Storing and Managing Unstructured Data



The assumptions that ESG used when building the TCO model shown in Figure 4 include:

- **Storage operations (management) costs of $25/GB.** This value varies between $4/GB and $100/GB depending on the type of storage (tier-1/tier-2), the number and salaries of administrators, and the criticality of the data including whether it's subject to compliance mandates or being used for e-discovery.
- **3.5 copies, on average, of file data spread over all of the shared drives in the organization.** This estimate was based on ESG Lab testing of backup solutions with single instance storage (SIS) and deduplication with real-world file data, an analysis of a corporate file share used to create this report, and the fact that most enterprises maintain at least three copies of data for backup, disaster recovery, and test/dev.
- **Risk and compliance costs equal to the cost of storage acquisition.** While these costs vary dramatically between organizations and industries, and one big fine can blow the budget, ESG believes that this conservative assumption is a good rule of thumb for most enterprise-class organizations.[4]

[4] Examples that illustrate the budget-busting potential of a legal discovery request or a compliance violation: eDiscovery cost estimate of $940/GB to collect, $2,931/GB to process, $13,636/GB to review, $13.3M for a PCI violation in 2010, $7.5M for a HIPPA violation in 2011

The Bigger Truth

Despite macroeconomic improvements over the past few years, cost reduction remains a focus for most organizations, which ESG research confirms. When asked what business initiatives they believed would drive the most technology spending over the next 12 months, the largest number of ESG survey respondents cited cost reduction initiatives, placing it at the top of the list.[5] Unfortunately, when it comes to data and the infrastructure to support it, organizations are missing out on significant opportunities to reduce costs by focusing only on obvious acquisition costs. To get the most out of your data at the lowest TCO, organizations should gain visibility into hidden costs as well.

Using numerous TCO models, ESG was able to show the typical costs for storing and managing 100TB of network-attached, shared file data for a typical multisite enterprise. A comparison between a traditional environment and one optimized with HP ControlPoint clearly demonstrates the potential hard and soft cost savings available.

HP ControlPoint manages information instead of storage. Decisions are made from the top, based on the business perspective of data in a particular business context, instead of from the bottom based on storage configurations. And it's not focused only on hardware costs, but includes ongoing management across the data lifecycle. It includes deduplication and aging, but also helps to identify what data is mission-critical to better manage in terms of risk, compliance, and optimization. Especially valuable is ControlPoint's ability to illuminate dark data that could be leveraged, identify ROT that can be deleted, and enforce policies when data is created in order to improve downstream processes.

Customers gain better protection with less risk, lower storage and management costs, and visibility/insight to make better use of their data. Of course the impact will vary depending on each individual situation, but the potential savings are huge. Data is the lifeblood of an organization, but if it's not properly managed it can be a tremendous drain on systems and budgets. If you want to reduce the costs of managing unstructured data and make your data work better for your organization, ESG recommends that you work with HP on a TCO/ROI analysis that's customized for your business.

[5] Source: ESG Research Report, *2014 IT Spending Intentions Survey*, February 2014.

Enterprise Strategy Group | **Getting to the bigger truth.**

20 Asylum Street | Milford, MA 01757 | Tel: 508.482.0188 Fax: 508.482.0218 | www.esg-global.com