

# Profiling Data —

A Method for Understanding and Justifying Data Storage Expenditures

An Executive Overview



**SUN**™ Copyright 2003 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, CA U.S.A. All rights reserved.

This product or document is protected by copyright and distributed under licenses restricting its use, copying, distribution, and decompilation. No part of this product or document may be reproduced in any form by any means without prior written authorization of Sun and its licensors, if any. Third-party software, including font technology, is copyrighted and licensed from Sun suppliers.

Parts of the product may be derived from Berkeley BSD systems, licensed from the University of California. UNIX is a registered trademark in the U.S. and other countries, exclusively licensed through X/Open Company, Ltd.

THE NETWORK IS THE COMPUTER, Sun, Sun Microsystems, the Sun logo, and Sun StorEdge are trademarks, registered trademarks, or service marks of Sun Microsystems, Inc. in the U.S. and other countries. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. in the U.S. and other countries. Products bearing SPARC trademarks are based upon an architecture developed by Sun Microsystems, Inc.

The OPEN LOOK and Sun™ Graphical User Interface was developed by Sun Microsystems, Inc. for its users and licensees. Sun acknowledges the pioneering efforts of Xerox in researching and developing the concept of visual or graphical user interfaces for the computer industry. Sun holds a non-exclusive license from Xerox to the Xerox Graphical User Interface, which license also covers Sun's licensees who implement OPEN LOOK GUIs and otherwise comply with Sun's written license agreements.

RESTRICTED RIGHTS: Use, duplication, or disclosure by the U.S. Government is subject to restrictions of FAR 52.227-14(g)(2)(6/87) and FAR 52.227-19(6/87), or DFAR 252.227-7015(b)(6/95) and DFAR 227.7202-3(a).

DOCUMENTATION IS PROVIDED "AS IS" AND ALL EXPRESS OR IMPLIED CONDITIONS, REPRESENTATIONS AND WARRANTIES, INCLUDING ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NON-INFRINGEMENT, ARE DISCLAIMED, EXCEPT TO THE EXTENT THAT SUCH DISCLAIMERS ARE HELD TO BE LEGALLY INVALID.

**SUN**™ Copyright 2003 Sun Microsystems, Inc., 4150 Network Circle, Santa Clara, Californie 94303 Etats-Unis. Tous droits réservés.

Ce produit ou document est protégé par un copyright et distribué avec des licences qui en restreignent l'utilisation, la copie, la distribution, et la décompilation. Aucune partie de ce produit ou document ne peut être reproduite sous aucune forme, par quelque moyen que ce soit, sans l'autorisation préalable et écrite de Sun et de ses bailleurs de licence, s'il y en a. Le logiciel détenu par des tiers, et qui comprend la technologie relative aux polices de caractères, est protégé par un copyright et licencié par des fournisseurs de Sun.

Des parties de ce produit pourront être dérivées des systèmes Berkeley BSD licenciés par l'Université de Californie. UNIX est une marque déposée aux Etats-Unis et dans d'autres pays et licenciée exclusivement par X/Open Company, Ltd.

THE NETWORK IS THE COMPUTER, Sun, Sun Microsystems, le logo Sun, et Sun StorEdge sont des marques de fabrique ou des marques déposées, ou marques de service, de Sun Microsystems, Inc. aux Etats-Unis et dans d'autres pays. Toutes les marques SPARC sont utilisées sous licence et sont des marques de fabrique ou des marques déposées de SPARC International, Inc. aux Etats-Unis et dans d'autres pays. Les produits portant les marques SPARC sont basés sur une architecture développée par Sun Microsystems, Inc.

L'interface d'utilisation graphique OPEN LOOK et Sun™ a été développée par Sun Microsystems, Inc. pour ses utilisateurs et licenciés. Sun reconnaît les efforts de pionniers de Xerox pour la recherche et le développement du concept des interfaces d'utilisation visuelle ou graphique pour l'industrie de l'informatique. Sun détient une licence non exclusive de Xerox sur l'interface d'utilisation graphique Xerox, cette licence couvrant également les licenciés de Sun qui mettent en place l'interface d'utilisation graphique OPEN LOOK et qui en outre se conforment aux licences écrites de Sun.

CETTE PUBLICATION EST FOURNIE "EN L'ETAT" ET AUCUNE GARANTIE, EXPRESSE OU IMPLICITE, N'EST ACCORDEE, Y COMPRIS DES GARANTIES CONCERNANT LA VALEUR MARCHANDE, L'APTITUDE DE LA PUBLICATION A REpondre A UNE UTILISATION PARTICULIERE, OU LE FAIT QU'ELLE NE SOIT PAS CONTREFAISANTE DE PRODUIT DE TIERS. CE DENI DE GARANTIE NE S'APPLIQUERAIT PAS, DANS LA MESURE OU IL SERAIT TENU JURIDIQUEMENT NUL ET NON AVENU.



Please  
Recycle



Adobe PostScript

## Introduction

In today's technology-driven economy, it is a common understanding that data is a company's most valuable asset, making it imperative that IT managers protect that data. However, because there are so many different vendor approaches, the task of finding solutions to data storage management problems can be difficult and perplexing. IT managers usually seek these solutions to address increasing needs for primary data storage for users, file server storage, or backup. Some of the alternatives include: more hard disk space, RAID systems, manual or automatic backup, and archiving solutions. The quest for any solution should really begin long before these specific solutions are even considered — it should start with an analysis of the data in use by the enterprise and the nature of that data so that the correct and most cost-effective solution can be justified, purchased, and implemented.

Managing constantly ballooning data stores is becoming more expensive — in actual dollar costs for equipment and media, and in the time spent by administrative personnel performing day-to-day backup and archival tasks. When a site uses only conventional manual backup, expenses can escalate exponentially as the amount of data increases. Automating archiving functions with solutions such as Sun StorEdge™ Utilization Suite (SAM-FS) software can help reduce the costs of this activity.

The first step toward efficiently automating an archiving facility is understanding the difference between dynamic and static data. Only by understanding the data differences can a site administrator provide the most cost-effective and reliable service.

## Understanding Data Differences

There are three different types of data that need to be identified: abandoned static, important static, and dynamic data. Classifying data into one of these three categories is often the most important step in identifying areas for improving efficiency.

### Static Data

Static data is data that does not change or changes infrequently. Static data can be divided into two basic categories: *important* data and *abandoned* data. Abandoned data is defined as data that is the product of an intermediate computing step or output, which after initial use is no longer needed. In either case, it can be discarded.

Separating important static data from abandoned data defines which data should be placed on secure media. By only keeping multiple copies of the important data, and not the abandoned data, companies can save disk space and expenditures. This also minimizes the bulk of the copies stored off-site as part of a disaster control plan.

A file that has not been modified can be considered a static file. If a file has not been accessed it might also be considered an abandoned file. Examples of static data that are not modified and are not abandoned are data such as released engineering drawings, fingerprints, and medical images. Similarly, important static files that are not accessed might be those that are kept for regulatory reasons, i.e., financial records or aircraft maintenance records. The criteria for defining static, abandoned, and dynamic data vary by site.

### Static Application Data

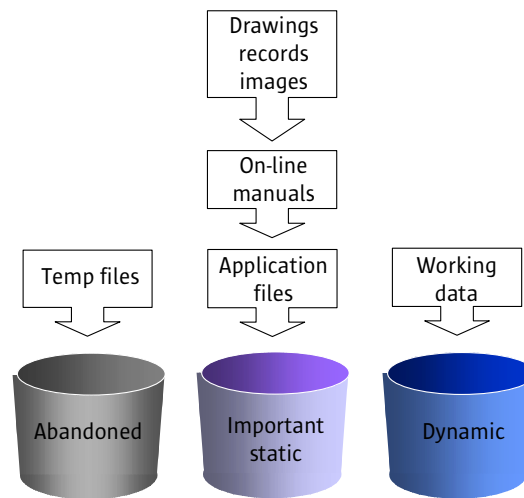
Application files that are also frequently accessed are more readily available when the data resides on magnetic disk. However, with applications distributed throughout the system, the same files often reside on multiple workstations on the network. For example, on-line manuals are a classic example of static data that may not necessarily be accessed often. When the manuals are needed the user requires fast access that is best accomplished when the data is on-line. An analysis of how many copies of on-line manuals are repeated on each workstation (250 MB repeated on only 10 workstations equals 2.5 GB) can be enlightening. A great deal of on-line storage can be made available for other uses by eliminating duplicate sets of data. Perhaps equally important, this is a good example of static data that *should not be backed up* on each workstation week after week.

### Dynamic Data

Dynamic data is working data. The media selected for it should be readable and writable since it will be reused. In general, this media will be different from the media noted above, since the data is accessed and updated frequently (usually it is on a fast hard disk). As this data ages, it may change into static data and be separated into important or abandoned categories, as shown in Figure 1, and then may be “migrated” to another media. This is the principle behind hierarchical storage management (HSM) — the different data and media form a “hierarchy” according to cost, security, speed, or other characteristics.

It is possible to achieve considerable savings on hard disk space, as well as increased data integrity if data storage is managed according to its importance to the enterprise.

**Figure 1:** Separating data into categories is an important step toward an efficient data management strategy.



## Managing the Data

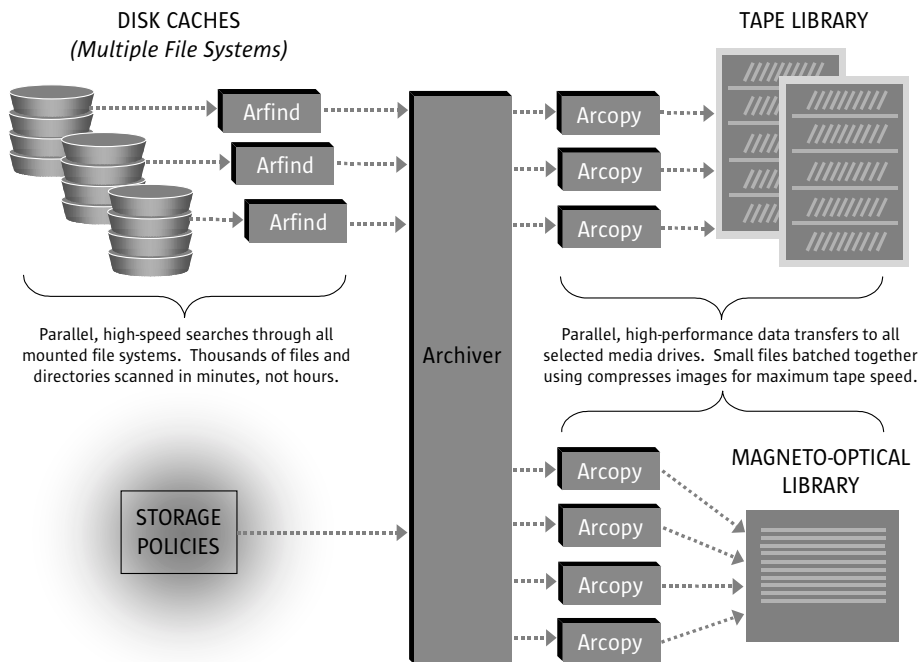
As mentioned above, the data at computational facilities can be classified into two major categories — static and dynamic — although many facilities do not separate these types of data. A majority are not aware of which data in the organization falls into each category. In most cases all of the data, whether static or dynamic, is backed up, even though typically only 15 to 25 percent of the overall data is dynamic. This means that up to 85 percent of the backups performed today are of the same data, week after week. Likewise, backups that are recorded to media in a robotic unit can create many copies of the same static data. This waste of resources highlights the need for the site administrator to profile the data.

Static data is an obvious candidate for an automated archiving facility because it can be removed from the daily backup process. Dynamic data remains a candidate for regular backups. Managed correctly, dynamic data can also be selectively removed from the backup process over time, according to usage patterns or life span requirements. This management process can greatly reduce the load on backup systems and forms the foundation of a disaster control plan.

### Sun StorEdge Utilization Suite Software

Sun StorEdge Utilization Suite software enables administrators to manage data according to its business value. The software automatically backs up work in progress and creates file copies from the source that can be migrated to any storage media. It enables administrators to set automatic archiving policies that determine when, where, and how data is stored. As organization needs evolve, the management policies can be changed without adversely affecting previously archived files. In the event of a disruption, the business can return to full productivity quickly, because users have access to files as soon as metadata is restored, rather than waiting for all data to be restored before coming back online. Figure 2 illustrates how Sun StorEdge Utilization Suite software archiver technology to reduce the time required to complete backups.

**Figure 2:** Sun StorEdge Utilization Suite's archiver functionality enables administrators to set automatic archiving policies that determine when, where, and how data is stored.



## Recording Format

For a complete analysis, data should be profiled over a period of several months. The administrator should become familiar with the data life cycle management of their data to determine the best archiving method or strategy. Critical data — preserved longer than two or three years — should not be in a notation linked to a particular CPU architecture or vendor because it will most likely outlive the hardware technology. Users and those who purchase backup technology should be made aware of this problem because migrating the data to new platforms can be quite expensive.

Proprietary storage formats are an even bigger problem for the site. Many IT departments have purchased “open” architectures, and are confident in the ability to move forward in an independent manner. However, many have not asked the important questions:

- How is the data stored on the media? Is it open?
- Are ANSI standard labels recorded?
- How are the files recorded?
- Who can read this data?

If only the recording vendor's product can read the data, the site is totally dependent on this vendor. This is not an open solution because data conversions may need to be performed for changes in vendor or operating platform. This can cause an integrity problem if the vendor is no longer supporting a particular product or platform.

### Regulatory Compliance

Another factor that can affect the manner in which companies manage data is the need to comply with increasing regulatory scrutiny for “fixed content” — unchanging digital assets that are retained for active reference and long-term value. Examples of fixed content include: X-rays, MRIs, electronic business documents, e-mail archives, check images, electronic statements, and completed CAD/CAM designs.

Sun StorEdge Utilization Suite software writes to a variety of near-line media based on the time value of the data — in an open non-proprietary format that does not require special Application Programming Interfaces (APIs). Writing files in an open format keeps the company in control of its data. This is especially important with fixed content that may need to be restored in the event of a legal dispute or audit in order to comply with regulations. Managing fixed content with flexible policy-based software that automates storage administration founded on data life cycle management can offer more options for protection and recovery of data to adhere to records retention regulations. Table 1 lists some of the regulatory issues, descriptions, and how Sun StorEdge Utilization Suite software can address the issues.

**Table 1:** Regulatory issues and how Sun StorEdge Utilization Suite software addresses them

Regulatory Issue	Description	Solution
Retention enforcement	Helps enable compliance officers to set hardened retention periods on electronic records, to satisfy regulations such as SEC Rule 17a-4	Sun StorEdge Utilization Suite software policy-based administration automates the length of time a file is stored, media targets (disk, tape, or optical), user access, and more...
Enhanced disposition, or “shredding”	Helps ensure that deleted data cannot be recovered using disk scanning tools, and complies with regulations such as Department of Defense 5015.2	The optimal strategy for compliance is a combination of best practices that include a data life cycle management and flexible policy-based storage administration to help ensure that files that are removed from the system cannot be recovered.
Application access security	Permits system operators to establish access security and authorized activities at the application or server level, and helps ensure the privacy of sensitive records for regulations such as the Health Insurance Portability and Accountability Act (HIPAA)	Solaris provides a Role-Based Access Control (RBAC) function that can be further enhanced by the Sun StorEdge Utilization Suite software file system Access Control List (ACLs) function. Again, the combination of best practices, data life cycle management, and policy-based storage management can assist with compliance.

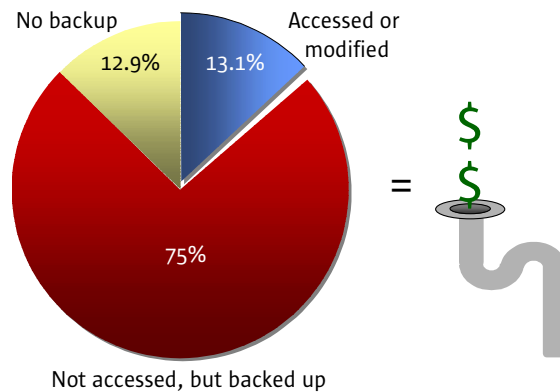
### Customer Examples

To illustrate the points in this document, the following is data from an actual site. A large computation development facility has 1.8 terabytes (millions of megabytes) of data consisting of 3 million files, of which only 7.7 percent were *modified* within 30 days. This amounts to only 9.5 percent of the data on the system. Only 14.1 percent of the 3 million files were *accessed* within 30 days, which equals 13.1 percent of the data.

This particular site used conventional backup methods (i.e., incremental backups every day with full backups every weekend). Over a span of a month (including 4 weekends), 1.6 terabytes (.905 x 1.8 terabytes = 1.6 terabytes) of the same data is repeated 4 times (full backups) in its backups. This is equal to four thousand 3490 tapes (400 megabytes per tape) x 4,

or 16,000 tapes holding the same data. This is a quite unnecessary and inefficient considering that only 7.7 percent of the files were modified, as illustrated in Figure 3. Moving a majority of this data to an archival facility can greatly reduce the time and resources required to manage backups.

**Figure 2:** Percentage of accessed or modified data to data that is backed up unnecessarily



### Earth Satellite Corporation

Earth Satellite Corporation (EarthSat), Rockville, MD, specializes in the application and development of remote sensing and geographical information systems (GIS) for the exploration, development, monitoring, and management of the earth's resources. Historically, EarthSat allocated entire disks as needed for specific projects. Users were required to manage the space and archive their data. "User A was not archiving data; User B did not have enough space; User C had data spread over several disk volumes. The big advantage with Sun StorEdge Utilization Suite software is that our 60-plus application specialists no longer worry about storage allocation and archiving, leaving them to concentrate on their specialty," says Christopher J. Peterson, Principal Scientist.

EarthSat deploys Sun StorEdge Utilization Suite software as a virtual infinite disk solution. Given the volumes of data that the company is processing, throwing more and more disk at the problem would be unmanageable and cost-prohibitive. Sun StorEdge Utilization Suite software lets EarthSat store data on lower-cost media while the files remain available on-demand to end users. Often, EarthSat's team of talented engineers are working on multiple projects at the same time. Sun StorEdge Utilization Suite software allows the engineers to multi task by staging files from tape while processing other images. The software helps EarthSat create a single high-speed logical volume that is accessible to virtually all users.

### Perlegen Sciences, Inc.

Perlegen Sciences, Inc. is a private company that uses high-density, whole-water micro array technology in combination with new approaches to scan entire human genomes. Perlegen expects that 120 terabytes of fresh data will be produced in 18 months — amounting to more than 2 terabytes per week. Greg Bandeau, Perlegen Sciences CIO, explains, "The data only need be online disk for several weeks, then it is stale — but, not worthless." Although the genome data does not need to be stored on high performance disk after the initial weeks of study, all 120 terabytes of data must be preserved and easily accessible from lower-cost media.

Sun StorEdge Utilization Suite software allows Perlegen's engineers to access data from and write data to tape without administrative intervention — keeping costs down by utilizing lower-cost media and minimizing administrative man-hours. It also enables Perlegen to set policies to automate the system to make three backup file copies across different media and geographic locations for added protection, and supports Perlegen's existing investments in tape storage technology.

### **An Ivy League University**

The Functional Magnetic Resonance Imaging (fMRI) Data Center at an Ivy league university is making critical breakthroughs in the storage and analysis of functional neuro-imaging data sets with the goal of furthering understanding of basic cognitive processes. Brain scans are an important tool for medical science, basic research, and education, but this expensive technology is often out of reach for many institutions. Novel methods are required to warehouse and organize large-scale data sets.

To solve the problem, the university needed to construct a framework that allows scientists easy access to raw data from published, peer-reviewed studies. The data center needed a method to handle metadata about the studies — such as how old the subject was, what stimuli or tasks were given during the study, and how often images were collected.

The resulting architecture is designed to take the *data life cycle* into account. When a study is first submitted it generates a lot of requests. As a newer study comes in, that study may not be requested as often and is moved to secondary storage.

The biggest innovation in the system is the HSM provided by Sun StorEdge Utilization Suite software that automatically handles cycling data from disk to tape and back again. This lets the university use expensive, faster disk systems for studies that are requested often and lets them automatically determine what is most important. From a total cost of ownership perspective, it lets them spend money on the right technology, giving them a unique solution to solve their problems.

## **Summary**

As an exercise, internal analysis of static and dynamic data can identify areas of waste and can suggest whether more disk, more backup, and/or an HSM strategy is the right solution. In addition, the exercise can yield some further benefits, such as forcing site administrators to answer questions that can increase productivity and reduce costs, including:

- Is the data safeguarded adequately?
- Is the site making the most efficient and cost-effective use of media?
- Does the site have the data administration process under control?
- Does the enterprise know where its most valuable data resides?
- Can the enterprise comply with records retention regulations?

Finding the answers to these questions and implementing solutions such as Sun StorEdge Utilization Suite software to address them can help companies justify new expenditures in storage technology as well as help reduce current expenses by optimizing existing resources.

For more information on Sun StorEdge Utilization Suite software and other Sun storage solutions see <http://www.sun.com/storage/>.

Sun Microsystems, Inc. 4150 Network Circle, Santa Clara, CA 95054 USA Phone 1-800-555-9SUN or 1-650-960-1300 Web [sun.com](http://sun.com)



**Sun Worldwide Sales Offices:** Africa (North, West and Central) +33-13-067-4680, Argentina +5411-4317-5600, Australia +61-2-9844-5000, Austria +43-1-60563-0, Belgium +32-2-704-8000, Brazil +55-11-5187-2100, Canada +905-477-6745, Chile +56-2-3724500, Colombia +571-629-2323, Commonwealth of Independent States +7-502-935-8411, Czech Republic +420-2-3300-9311, Denmark +45 4556 5000, Egypt +202-570-9442, Estonia +372-6-308-900, Finland +358-9-525-561, France +33-134-03-00-00, Germany +49-89-46008-0, Greece +30-1-618-8111, Hungary +36-1-489-8900, Iceland +354-563-3010, India-Bangalore +91-80-2298989/2295454; New Delhi +91-11-6106000; Mumbai +91-22-697-8111, Ireland +353-1-8055-666, Israel +972-9-9710500, Italy +39-02-641511, Japan +81-3-5717-5000, Kazakhstan +7-3272-466774, Korea +822-2193-5114, Latvia +371-750-3700, Lithuania +370-729-8468, Luxembourg +352-49 11 33 1, Malaysia +603-21161888, Mexico +52-5-258-6100, The Netherlands +00-31-33-45-15-000, New Zealand-Auckland +64-9-976-6800; Wellington +64-4-462-0780, Norway +47 23 36 96 00, People's Republic of China-Beijing +86-10-6803-5588; Chengdu +86-28-619-9333; Guangzhou +86-20-8755-5900; Shanghai +86-21-6466-1228; Hong Kong +852-2202-6688, Poland +48-22-8747800, Portugal +351-21-4134000, Russia +7-502-935-8411, Singapore +65-6438-1888, Slovak Republic +421-2-4342-94-85, South Africa +27 11 256-6300, Spain +34-91-596-9900, Sweden +46-8-631-10-00, Switzerland-German 41-1-908-90-00; French 41-22-999-0444, Taiwan +886-2-8732-9933, Thailand +662-344-6888, Turkey +90-212-335-22-00, United Arab Emirates +9714-3366333, United Kingdom +44 0 1252 420000, United States +1-800-555-9SUN or +1-650-960-1300, Venezuela +58-2-905-3800