White Paper

Building the Optimal Environment for Big Data Management

Complementary Technologies to Lower TCO and Increase Business Efficiencies

Table of Contents

You're in a business that generates a staggering amount of data. Maybe your organization is mapping genome sequences, analyzing satellite images, scanning medical images, or producing block-buster feature films.

Digital content management for production and for archives is crucial. Terabytes of data is involved in delivering a finished project. The content must be captured or scanned, browsed, edited, analyzed, managed, securely stored and easily retrieved. But today, you need to know where the data is located, information about the data and its metadata and the levels of access to it at any given time. So it's an operational challenge, to say the least.

The solution is a combination of systems that make it possible for you to better manage your business and your resources. The right combination – a mix of technologies that perform complementary functions – will result in a workflow management system with archival capabilities that enables you to manage your media from ingest to archive and actively access that content at any point in the pipeline.

Three products available through SGI, the global leader in high bandwidth storage solutions, work together to deliver the capabilities you need to manage large data holdings and significant dynamic data turnover. Any one of these technologies stands alone as a boon to digital content management; together they work information lifecycle magic. They are SGI InfiniteStorage™ Data Migration Facility (DMF), DataFrameworks DFX and Hitachi Content Application Platform (HCAP). These three applications interfaced to the appropriate SGI InfiniteStorage SAN, NAS or hybrid production infrastructure provide seamless migration of content from the desktop to the archive; higher level management of the storage and filesystem; and monitoring and authentication of the archived data content.

1.0 Asset Management: Managing Storage in Production Environments

For those who create content and generate data, as well as those responsible for managing the infrastructure in which the data resides, data management can create a number of complex problems. Locating a file in a traditional enterprise storage system requires users to know where the data physically resides – on fhat file server, in what file system, in what directory and in which file. DataFrameworks is an example of asset management software that eliminates this requirement.

Designed for provisioning, monitoring and reporting on data, DataFrameworks software forms the central hub for data management for most any type of workflow. It is designed to standardize, simplify and enable seamless scaling of data and the storage infrastructure through automation. Its data-centric abstraction layer, DFX, provides a view of data that is common to all data stakeholders – including both business and IT users – so that users with no knowledge of a file's physical location can easily navigate the data system.

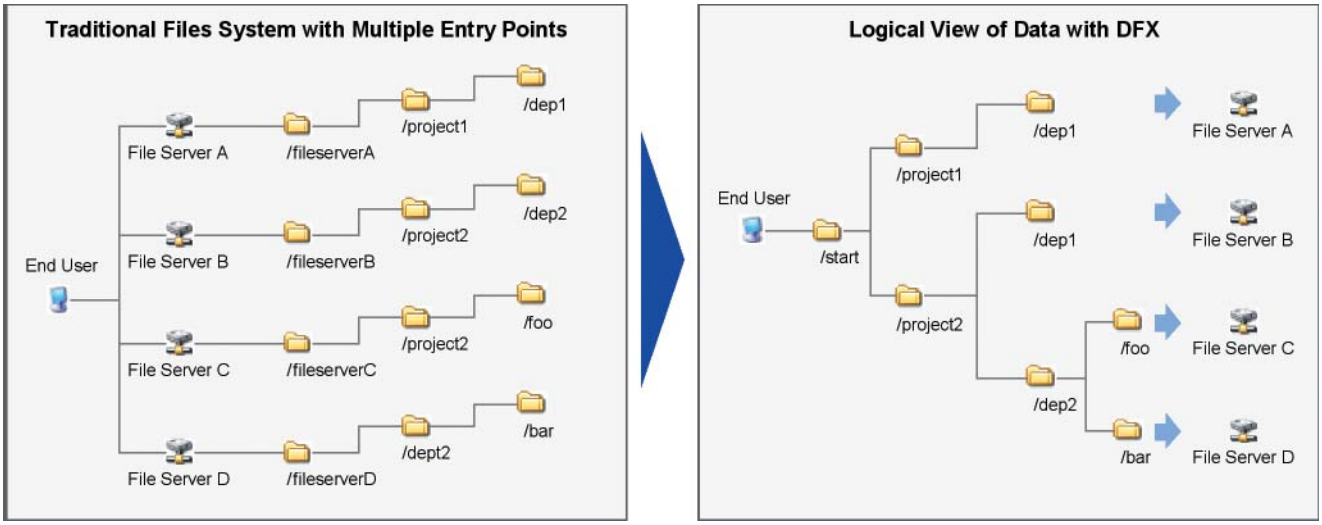

Figure I: File systems : Traditional and with DataFrameworks DFX systematic data organization schema

The comprehensive view of corporate data that DataFrameworks provides, improves the user's ability to find data and maintain a disciplined, standardized file structure. It offers systematic data organization, consistency and discipline in the production workflow. DataFrameworks also provides the tool set to assist the data management process of managing location, movement and tracking of data to ensure the right files are in the right place at the right time.

DataFrameworks works with SGI NAS Servers and SGI SAN solutions to provide high speed shared access to files, eliminating bottlenecks that hamper data intensive digital operations. The solution provides the ability to efficiently manage limited production infrastructure resources for multiple projects where everyone shares the same resources. DataFrameworks has demonstrated increases in file server utilization of up to 25% by controlling and monitoring disk space consumption, and then setting up and enforcing directory structures based upon actual utilization instead of assumed requirements.

Additionally the DataFrameworks reporting feature makes it possible for users to view current disk space consumption and historical disk space consumption trends, enabling them to track and monitor space consumption from a business perspective, not just a storage device perspective. Security, another key element in managing digital content, is also addressed by DataFrameworks reporting tools, which can be configured separately from the authorization and access levels. This allows users to see the status of the file system data without having access to the files within the production systems themselves.

DataFrameworks also provides the foundation to introduce automation that improves efficiency and minimizes human error. Data organization schemes are consistent, predictable and repeatable.

EFILM, a leading Hollywood digital intermediate facility, has integrated DataFrameworks file system management tool into its production pipeline to manage hundreds of terabytes of data across all of their SGI® InfiniteStorage CXFS™ SAN clusters.

One unique feature that DataFrameworks provides is targeted views of the project status, geared towards different department's requirements. The sales department needs a high-level view of the project status, to communicate effectively with a client. This view is quite different from that of the project's producer, the colorist or the IT manager. DataFrameworks allows for customized views, hierarchical secured access to content and other data management tools which enable everyone in the facility to be more pro-active, productive and efficient.

"DataFrameworks allows our SANs to be managed on a per project basis, providing a repeatable data hierarchy for each project and resource allocation and utilization access to all users. DataFrameworks is one important tool for users to access their managed data on the SAN." - John Bennett, Technology Manager, EFILM.

With the goal of speeding production, more efficiently managing data and ultimately limiting the number of people required to manage that data, EFILM envisions DataFrameworks as a tool that will reduce storage requirements and ultimately reduce costs.

Bob Eicholz Vice President, Corporate Development for EFILM stated, "Prior to DataFrameworks, EFILM managed its hundreds of terabytes of on-line storage manually. As volumes of data grew, EFILM needed a robust data management solution. We found that solution in DataFrameworks."

2.0 File Virtualization: Maximizing Capacity and Minimizing Cost

In the early stages of a project, new content is often accessed intensively. Later, it may be accessed less frequently and may even remain untouched for many months or years, with only occasional brief periods of renewed intense access. To ensure that data is on hand when you need it, but not taking up valuable primary storage space when you don't, storage management environments need to flexibly and economically adapt to changing data access patterns without forcing end-users to understand how the underlying storage environment has been optimized. For that, they need file visualization capability like that procided by SGI InfiniteStorage Data Migration Facility (DMF).

DMF transparently moves files between different tiers of storage - from online storage to near-line storage based on user-defined criteria, such as time of last access. At the same time, DMF keeps directory data online, making the combination of online and near-line storage appear as one very large virtual storage device. This eliminates the need to either maintain a mountain of inactive data on expensive disk storage or undertake the resource-intensive activity of archiving that data to tape, where it may be difficult or even impossible to access.

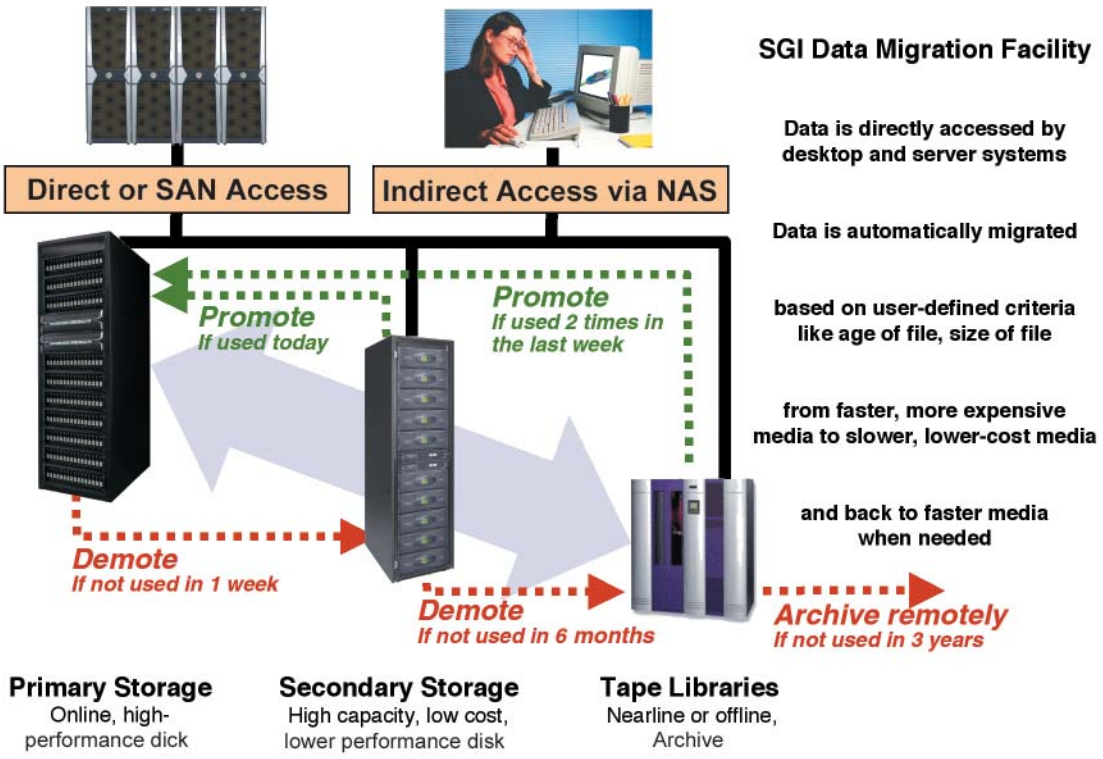

Figure 2: SGI Data Migration Facility – transparently migrating data to optimize performance and minimize cost

The result is that a nearly infinite amount of data be cost-effectively maintained and managed without sacrificing accessibility. Users never need to know where the data is stored; it is transparently recalled to primary storage when accessed, appearing just as it would if it had been residing in online storage. The impact is that customers need to purchase less online storage and can maximize the utilization of the storage that they do purchase. For customers with 200TB or more of total storage requirements, DMF can reduce both storage related capital and operational expenses by 50%.

DMF can be used in conjunction with a variety of storage solutions to provide file virtualization capabilities that allow users to transparently access data on multiple tiers of storage. It performs a key function in working with DataFrameworks, helping to manage assets and the workflow process by recognizing when to demote data to a lower performance, more cost effective tier so that space can be allocated to another project.

Having the flexibility to store data in the appropriate level of the hierarchy yet access it on demand results in a number of benefits, including reduced cost of capacity and administration, improved productivity, optimized data management and – since there are fewer opportunities for human error in this automated system – a reduced risk to data retention.

A number of weather forecasting centers around the world, including KNMI and GFDL use DMF for file virtualization to simplify operations and contain costs. Weather forecasting centers regularly tune their models by looking for statistical bias over many different forecasts, and by examining large forecast anomalies. Both types of analysis require input, forecast and observed data to be stored for up to 10 years, but with only a small fraction being accessed at any point in time. Keeping this data on-line storage would exceed cost and energy budgets and add complexity to the on-line environment. DMF's file virtualization capability allows these centers to do reanalysis as if the data was on-line, but migrate the majority of it to low-performance disks and then off to near-line tape archives.

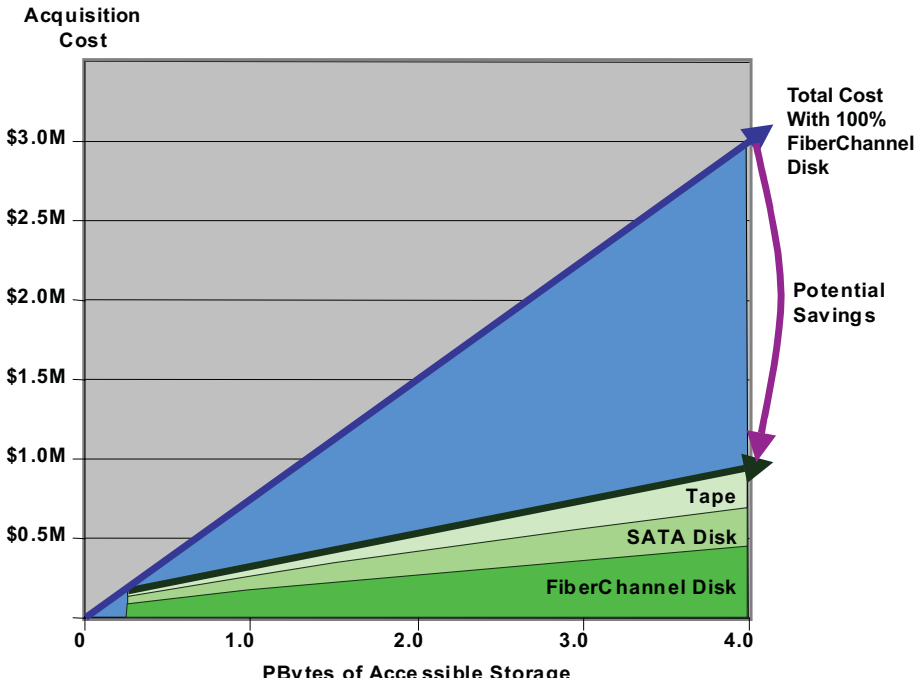

Figure 3: Potential Reductions in Acquisition Costs with File Virtualization

3.0 Digital Archiving: Ensuring Data Availability and Searchability

In many environments, critical data must immediately be archived in a secure repository where it can be rapidly searched and accessed but not modified. Many of these have specific government regulations requiring the preservation of financial, design and product safety data for a fixed amount of time. This is where Hitachi Content Application Platform (HCAP) comes in.

Until recently, there were no true online digital archiving solutions. Archiving with magnetic tape systems lacks the ability for rapid search and recovery of archived assets. An alternative, archiving with disks lacks Write Once Read Many (WORM) protection, authentication guarantees, and access control.

With the introduction of HCAP, many of these issues can be mitigated. Recognizing that the simplest and most cost-effective means of providing continuous longer-term access to digital records is to preserve them digitally, HCAP software that allows customers to store, protect and manage fixed-content data, such as documents, emails, database, images, and audio files, in an open and operationally efficient online environment. This enables immediate access to archive data, allowing users to search for files by name, file attributes, metadata and even content.

Especially beneficial in meeting requirements for corporate governance, legal and federal compliance mandates, HCAP offers not only the ability to enforce WORM requirements, but delivers a definable retention period for WORM files and a means to audit and prove that none of the data has been corrupted, changed or deleted WORM files. The system stores content in an immutable format with the ability to set file level retention. This prevents file deletion before the retention period expires. Additionally, users can set a Retention Hold on any file. HCAP offers the highest level of data protection, ensuring a specified number of replica copies are maintained to tolerate simultaneous points of failure. The system can be set to maintain one to four internal copies, depending on the value of the data. Additionally, the system offers unprecedented authentication, guarding against corruption or tampering by periodically computing the digital signature and comparing it to the value stored when the file was archived. On the opposite end of the spectrum, for files with sensitive data which must be destroyed, HCAP offers shredding capabilities to ensure that no trace of a file which must be permanently and irrevocably deleted is recoverable.

The ability to scale to accommodate the exponentially growing amount of data in the business environment is key. The HCAP architecture is optimized to handle petabytes of content, with all objects stored in a single, archive-wide global name space, and SGI InfiniteStorage NAS and SAN solutions are uniquely designed to manage these vast amounts of data. Users can add capacity to their SGI storage infrastructure and then integrate it into the HCAP domain quickly and without reconfiguring the archive. In addition, HCAP's ability to rapidly search metadata in new ways, a key attribute of archiving, enables organizations to obtain maximum value from their digital content over the maximum amount of time.

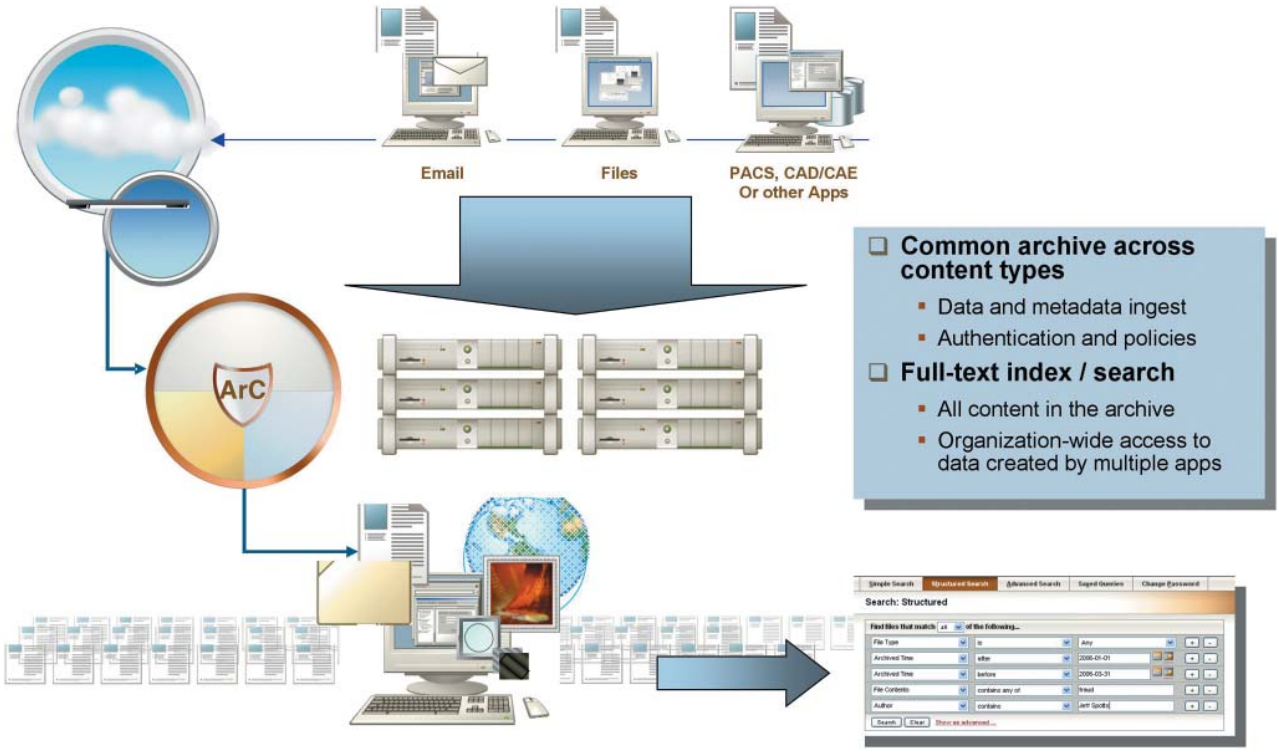

Figure 4: Hitachi Content Application Platform Architecture

As an example, the aerospace industry needs to maintain separate databases for each aircraft they manufacture, not just for each airplane design. These databases maintain key information for each individual plane "as manufactured" plus all of the associated use, inspection and maintenance information over the 40 year life of the aircraft. The immutable nature of the data stored within HCAP supports this capability so if 20 years from now an aircraft needs to be repaired at a remote site or if becomes important to determine which individual aircraft have specific parts installed, the searchable archive is accessed to determine which specific records need to be analyzed.

4.0 The Whole Greater than the Sum of Its Parts

Each of these technologies plays a vital role in managing the massive amount of information that needs to be recorded, cataloged, accessed and preserved. Using 2 or 3 of these together taps into the strengths of each and ultimately reduces production and storage costs by realizing time and cost efficiencies.

DataFrameworks allows users to provision, monitor file systems and report on data; DMF delivers file virtualization with on-demand access; and HCAP provides secure online storage and searchable access for fixed content data. Together they work to offer full coverage of the three stages of data management: initial organization of data, juggling storage to meet immediate and short term needs and secure long term storage. Imagine the combined power of all three to offer end to end management of your workflow.

Each piece of the solution contributes to a comprehensive work-flow that can be leveraged to perform at a level that stand-alone pieces cannot hope to match. The first puzzle piece organizes data in a structure that optimizes management capabilities. It's a key component of a total solution that can be made even stronger by adding a second piece – one that moves data dynamically according to priority. Adding the final piece – archiving capabilities that ensure security and compliance with federal mandates – creates a full spectrum system for accommodating your data with the fewest number of storage resources.

There are many cases where asset management, file virtualization and archiving can work together to dramatically improve workflows and equipment utilization. One example is in genome centers, where new generation of DNA sequencers can create up to 1TByte of information per day per instrument.

Using asset management technology, this information can be automatically classified into project names, asset types that describe the nature of the data, asset attributes that might identify which chromosome it comes from, and specific attributes of the individual it came from (e.g. they have a specific disease that is being analyzed). This a-priori classification allows raw image data to be found and reprocessed if better algorithms are developed, allows processed data to be reassembled or resequenced given recent discoveries, and allows statistical analysis to be run against only relevant data.

At the same time, file virtualization allows organizations to keep PBytes of information available by migrating less critical raw image data to migrate rapidly to near-line tape and to keep often accessed statistical summary data to migrate to the fastest disks available. This not only saves money because less disks are used, but also because expensive chemical reagents and laboratory personnel need not be used to recreate data that was lost or accidentally erased.

Finally, archiving ensures that records required by the FDA and other licensing agencies are preserved, and that ad-hoc data analysis can occur whenever required. This greatly simplifies compliance and accelerates the analysis of “what-if ideas” that scientists often generate.

When taken together, asset management, file virtualization and archiving can increase scientific productivity at genomics research sites, reduce operational costs, and help meet regulatory requirements – giving a site that combines these technology a competitive head start.

As your company tackles the data tidal wave with fewer resources, increased productivity and lower costs, These three heterogeneous system components improve data quality and increase process efficiency by introducing consistency to the data management process. That’s good news for the bottom line, and excellent news for creative and technical data professionals, IT managers and the monetization and preservation of content archives.

Corporate Office
1140 E. Arques Avenue
Sunnyvale, CA 94085
(650) 960-1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

4023 [08.23.2007] **J15304**