RJ10351 (A0507-014) July 12, 2005
Computer Science

IBM Research Report

Holistic Information Management Solutions

Ying Chen, Shauchi Ong
IBM Research Division
Almaden Research Center
650 Harry Road
San Jose, CA 95120-6099

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Holistic Information Management Solutions

Ying Chen, Shauchi Ong
Storage Systems Department
IBM Almaden Research Center
San Jose, CA 95120 USA
{yingchen,ong}@us.ibm.com

July 11, 2005

Executive Summary

The growing pain of managing information for business advantages has led to the emergence of the concept of Information Lifecycle Management (ILM). According to the definition provided by Storage Networking Industry Association (SNIA), ILM represents a set of processes and policies that intend to align *business value of information* to *the most cost-effective IT infrastructure* in order to reduce Total Cost of Ownership (TCO) and maximize Return On IT Investment (ROI) [11].

Over the past few years, major vendors in both storage and business applications such as database and content management systems are chiming in with their respective ILM visions that would manage information throughout their lifecycles to meet business goals while maximizing the IT resource utilization. Although grand and inspiring, the state-of-the-art technologies in storage and business applications have proved that such visions are far from realities. The key reasons for such gaps are obvious: In order to manage information for business advantage, a holistic solution is needed to link bits and pieces of information to appropriate business functions at proper time. Without a comprehensive means of linking business rules and activities with different pieces of information and taking data management actions according to the changing business situations, simply counting on technological advances in point products alone, *e.g.*, business applications and storage products, are not going to work. The widening gaps among them will only lead to skyrocketing IT cost and diminishing return on IT investment.

To address these issues, we have developed a brand new approach: *business-semantic-aware storage solution (BSA)* for holistic information management. The brain of BSA is embedded in the business-semantic-aware storage software layer residing in between the business applications and storage devices. BSA incorporates a new type of intelligence that is commonly missing in the traditional storage, yet instrumental to successful holistic information management. BSA identifies crucial business semantics hidden in business applications in form of workflow events and data classifications, links them to data management policies, and takes data management actions in accordance to business conditions for the overall business solution streamlining and global optimization. By leveraging the workflow events and data classifications, BSA is able to determine exactly what data management actions to take on what data and under what condition.

BSA plays a critical role of automating the information management activities for a wide variety of business reasons, *e.g.*, regulatory compliance and productivity enhancements. It becomes the focal point and core component in advanced storage solutions such as compliance and ILM solutions. It bridges the gaps between business semantics and storage information management policies. Moreover, it links data, the value of data, and management of data throughout information lifecycle phases, and brings the control over information management to a new level, *i.e.*, a holistic level.

1 Why Holistic Information Management?

The rapid growth of voluminous information poses new challenges to businesses in managing them to meet business goals, such as cost, performance, reliability, and availability, etc. The mounting pressure from regulatory compliance requirements, e.g., Sarbanes-Oxley, SEC, HIPAA, and DOD [2, 3, 1, 5], which mandate corporations to maintain *fixed-content reference data* safely for years, imposes additional complexity on information management. According to [14, 13, 15], corporate data volumes are estimated to be expanding at a rate exceeding 40% annually, and petabyte-sized databases are expected to be a reality in commercial environment by 2009. Such information are vital to business operations and productivity. Cost-effective information management throughout their lifecycles hence is critical.

The outcry from customers for better information management solutions has stimulated the storage industry to seek for new strategic directions. Information Lifecycle Management is the outcome of such an effort. To date, all major storage vendors have pitched in their respective ILM stories that would utilize new storage infrastructures and technologies to intelligently map diverse business information to the right storage devices at the right time. Yet in reality not many current solutions go beyond the traditional Hierarchical Storage Management (HSM), a concept developed almost a decade ago [9, 19, 12], which moves data across tiered storage, *e.g.*, disks and tapes, using some predetermined and static policies.

A blossom of startups that claim to deliver ILM solutions ranging from Email archiving (*e.g.*, [20]), database archiving (*e.g.*, [16]), to compliance solutions (*e.g.*, [4]) serve as another indicator on the demand and desperation from the user community for information management remedies and the uprising importance of ILM. However such solutions often fall short on various aspects, *e.g.*, compliance, cost, and performance when we consider the totality of a business solution. It is extremely common to see email or database archiving solutions use single set of storage policies for all data, regardless what the data is and what lifecycle stage the data is in.

So why is such exciting ILM vision far from reality? A closer look at the existing ILM technologies and the trend of technology advances reveals the key reasons: Today's business operations no longer depend on a single piece of software or hardware for their effectiveness. Most business functions are delivered through working with a wide range of software packages and hardware devices as a whole. For instance, delivering email services requires at least email client and server software, email archiving and records management software, in addition to email servers, storage software and hardware. Merely adding email servers, upgrading software, or increasing storage capacity has proved to be inadequate to meet corporate email service objectives. However, no new holistic technologies exist to ensure the stack of software and hardware components can work coherently and synergistically for the benefits of the overall solution.

Such issues also motivated SNIA to define a reference ILM solution architecture that includes a layered stack of software and hardware devices as shown in Figure 1. Such a stack typically includes business applications (*e.g.*, SAP), information management systems (*e.g.*, Content Management Systems), data and file management systems (*e.g.*, databases and file systems), and storage devices (*e.g.*, highend Fibre Channel storage, midrange Serial ATA (SATA) storage, and lowend tape storage). Yet to date no promising technologies have been proposed or seen in the marketplace to deliver holistic information management solutions. This is simply because end users' business goals are typically far-fetched by storage. While business applications also pay little attention on storage's role in the overall business solution.

For example, over the past few years, we have watched new storage compliance technologies being rolled out, mostly emphasizing on long term data archiving, but with little attention on how higher layer compliance applications such as Records Management Systems can be seamlessly integrated to satisfy the end user compliance requirements. Similarly, business applications continued to advance with features such as Business Process Management systems (BPM) and workflow supports such as those in FileNet [8] and EMC Documentum [6], to automate tedious business and improve business productivity and efficiency. Yet they pay virtually no attention on how such automated business processes would drive the information man-

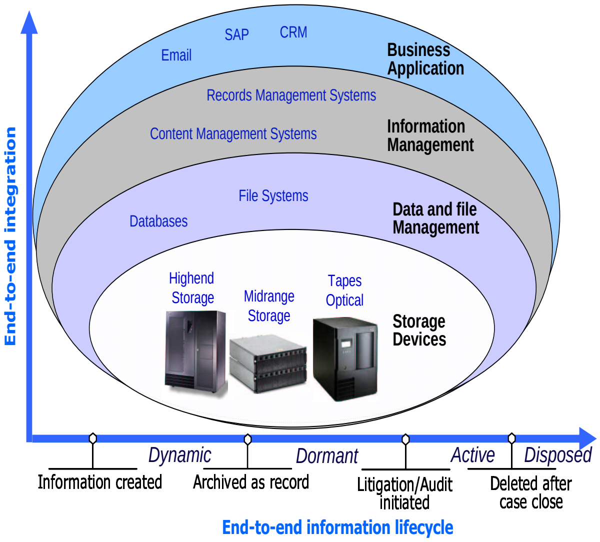

Figure 1: The key system layers in an ILM solution stack.

agement throughout their lifecycles. As a result, today, storage systems typically manage data according to *storage system conditions*, *e.g.*, capacity and bandwidth, rather than *business activities* or *business value of information*, such as those embedded in business workflows and business applications. The enhancements in storage functionality, capacity, performance, reliability and availability may not realize their full potential. Worse yet, sometimes they may even be misused, leading to non-compliance, high cost, and low performance for the overall business solution.

Clearly, solid holistic information management technologies connecting business logic to cost-effective storage infrastructure are critical. Such technologies promise to bridge the gaps between business applications and storage and encourage global optimization and compliance for the benefits of overall business solution. Specifically, we define holistic information management as *a set of technologies that identify critical business semantics that are essential to drive cost-effective data management, associate them with data management policies, and act on data according to business changes*. This is precisely what a *business-semantic-aware storage solution* (BSA) is about. A BSA solution works across the application and storage domains. It leverages useful business semantics, *e.g.*, workflow events and the business value of information, delivered by business applications to devise new data management optimizations and policies. It further drives runtime data management activities in accordance to business events for improved performance, reliability, and availability with reduced cost.

Without such technologies, the ILM solutions will remain to address isolated issues *in* an overall business solution but not necessarily *for* the overall solution. This is true despite the speedy technological advances in the areas of storage hardware and software, and business applications such as databases and content management middleware, and other end user applications.

2 Information Management Challenges

The need for holistic information management infrastructure across organizations of all sizes has surfaced up due to several daunting challenges in the enterprises:

Challenge One: High level semantic gaps The high growth rates in the worlds of business applications, databases, content management middleware systems, and storage have been impressive over the past few

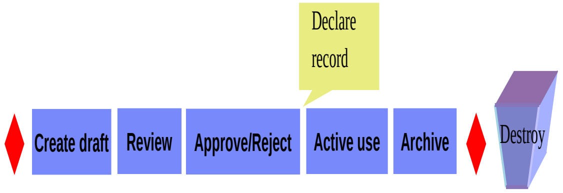

Figure 2: An analyst research report processing workflow process.

decades. Yet most of the growths occur within the specific software and/or hardware sectors themselves, *e.g.*, databases, content management, and storage. Little is done across different sectors to bring about seamlessly integrated solutions. Even professional system integrators, VARs, or ISVs typically do no more than simple assembling and packaging due to limitations in different system layers as well as the lack of innovative integration technologies. The divergence of storage and business applications appears to continue. This leads to high level semantic gaps across system layers and poor knowledge of information residing in various storage devices. Often the high level knowledge of business information in form of emails, contracts, or purchase orders available in business applications lose their meanings as they traverse from one system layer to another. At the storage layer, the notions of emails and contracts are essentially nonexistent. This often leaves storage with no other option but to use single-storage-policy-for-all scheme, resulting in inefficient resource usage or high cost for the overall solution.

Such high level semantic gaps are even more pervasive with the proliferation of workflow and BPM supports in business applications. Workflows often generate and/or process large volume of data, and provide strong implications on how information must be managed throughout their lifecycles to meet business goals and regulatory compliance requirements. Unfortunately, such dynamic workflow events are never utilized by today's storage. Data management activities done at the storage layer remain to rely on the storage observed usage patterns or the age of the data, ignoring the effects of business events on information altogether. No wonder why storage servers are often blamed for their unresponsiveness and poor performance.

Challenge Two: High risk and low trust The high level semantic gaps and poor integration across system layers increase the risk of security breaches, malicious tampering, and illegal disclosure of confidential information. Meanwhile they reduce the ability for the organizations to respond to regulatory compliance requests, such as record preservation and litigation discovery, in a timely fashion. Many system layers can be attacked at different points of time without leaving a trace. Even if traces are available, tools that work across system layers to provide meaningful insights to end users have not been available. In most customer environments, managing risks and improving overall business solution trustworthiness remain to be a wishful thinking rather than a reality.

Figure 2 illustrates such issues using an analyst research report processing workflow as an example. According to business policies, when a report is approved or rejected, it must also be declared as a *corporate record*. Under Security Exchange Commission (SEC) regulation [3], corporate electronic records are required to be stored on Write-Once-Read-Many (WORM) storage for compliance guarantees. In essence, the workflow event, *declare record*, dictates exactly when the document must be migrated to the WORM storage. Yet no known storage solutions are able to act according to such business events for data migration and retention. Data is often solely migrated by age or last access time. Consequently, the corporate records may not be stored on the WORM storage until days after the record declaration event, leaving significant security holes in the system.

Challenge Three: Ineffective information management policies A lack of knowledge about the business semantics of data, coupled with the inability to take data management actions according to the business events renders the ineffectiveness of information management policies in traditional storage solutions. Currently, the most immediate and critical data management policies concern data retention, HSM in tiered storage, replication, backup, and restoration. Without proper integration with business applications such as Records Management Systems, storage may still use a single retention rule for all data, even if records are categorized according to business rules defined in the Records Management Systems.

As for HSM, traditional solutions manage data movements in tiered storage using a fixed migration path and according to usage patterns and/or the age of data. For instance, a typical HSM policy may move a piece of data from the highend storage (*e.g.*, EMC Symmetrix [18]) to the midrange Serial ATA-based storage (*e.g.*, Network Appliance NearStore [10]) after the data is 30-day old, and then to the lowend tape storage (*e.g.*, StorageTek tape systems [17]) after another 60 days. Although such policies may work for certain situations, in practice, the rigidity of those policies are becoming increasingly insufficient, mainly due to their unresponsiveness to business changes.

In the analyst research report processing example as shown previously, a *rejection* event on an analyst report may inherently indicate that the report will most likely expect little subsequent use. Therefore, it can be archived to the lowend tape storage right away without waiting for the age threshold to occur or going through the fixed migration path. The *published* reports, on the other hand, should be cached in the highend storage as much as possible since they are the source for business revenues. It is important to ensure fast accesses to such reports. Clearly, without the awareness of such business activities, existing HSM policies may mistakenly migrate published reports to slow storage while unnecessarily keeping rejected reports on the highend storage, leading to inefficient resource utilization and degraded performance.

With the increased stringency in regulatory compliance requirements, today HSM solutions not only have to manage the tiered read-writable storage, but also tiered WORM storage, *e.g.*, disk-based WORM storage and WORM tapes and/or WORM optical storage. This further complicates the HSM policies: data may have to be moved from read-writable storage devices to WORM storage when they are declared as *corporate records*. They may be moved to read-writable storage or deleted when the retention periods expire. Yet today's HSM solutions not only fail to make appropriate or optimal decision on *where* to place data when it involves multiple storage types, but also fail to decide exactly *when* to move data to the right place.

Different data protection schemes, *e.g.*, replication and backup, incur different cost and provide different levels of data loss and recovery time guarantees. In practice, different classes of data may require different protection schemes. Such data classification may be done by business applications. For instance, an email archiving software can classify email into *confidential, compliant, irrelevant to compliance* categories. Such email categorization implies how emails should be retained and protected. Unfortunately, such data classification semantics available at the business application layer is mostly lost as data reaches storage. Hence, the use of single-data-protection-policy-for-all scheme is prominent, but it is far from satisfactory.

Challenge Four: Inefficient resource utilization Poor storage resource utilization is a common theme in many enterprises and a universal refrain for IT spending. Enhancing storage utilization has becoming a hot button in almost all IT organizations. Albeit the promises from storage vendors in delivering ILM solutions that can place right data in the right devices at the right time, the realization of those goals has been challenging. This is simply because storage resource utilization enhancements typically do not come from storage layer alone. Instead, they depend on the business knowledge above the storage and the derived insights on how to precisely map different pieces of business information to their respective storage resources under proper business situations. For instance, the data classification information available in business applications could be well utilized to drive storage resource assignment.

3 Business-Semantic-Aware Solution for Holistic Information Management

To address such monumental issues in the current IT organizations, we introduce a new holistic information management solution, called *business-semantic-aware storage solution*, which promises to close the gaps between business applications and storage. It further incorporates new levels of intelligence to remedy many of the shortcomings of the existing information management solutions, including risk exposure, poor resource utilization, and high cost.

A BSA solution leverages innovations in the areas of business semantic preservation, discovery and delivery in order to create key enablements for building seamlessly integrated business solutions. Meanwhile it incorporates a new generation of storage policy management and optimizations that are fully aware of the business semantics, as opposed to business-semantic-ignorant. When deployed in front of storage hardware devices, a BSA solution links business applications to storage coherently to achieve the overall goal of minimizing TCO and maximizing ROI. BSA ensures useful business semantics to be *preserved and passed* across system stack and well *utilized* for cost-effective storage optimizations and policy management for the overall solutions. This is not only a significant leapfrog in comparison to traditional storage solutions, but also a critical step toward realizing the ILM goal – aligning business value of information to cost-effective IT infrastructures.

The overall BSA solution takes a three-step approach to achieve the holistic information management goals: First, BSA leverages the business application knowledge to identify critical business semantic information useful for holistic information management. A business application often contains a wide variety of business semantics. However, not all semantics are relevant or useful for information management. The ability to link right high level semantics to data management policies is essential to the success of holistic information management. Once the desired semantics are identified, such information needs to be preserved and delivered across the system layers to allow BSA to take advantage of them for data management activities close to storage. This requires cooperation among business applications and storage in an integrated ILM solution. Finally, a new generation of business-semantic-aware data management optimizations and policies must be devised to fully realize the potentials of such high level semantic information in managing information throughout their lifecycles.

3.1 Crucial business semantics identification

In storage, most of the policies and optimizations are centered around making decisions about *when* to take *what actions* on *what data*. The most effective policies and optimizations are those that can take *right actions* at the *right time* on the *right data*. The storage actions may differ from one category of storage policies to another. For example, a set of possible migration actions can include moving data in a fixed migration path, *i.e.*, from the highend storage to midrange and then to the lowend storage, or taking a shortcut, *i.e.*, from the highend storage directly to the lowend. A set of caching actions, on the other hand, may including caching based on access frequency or recency. For each category of storage policy, the right actions can be easily formulated once storage knows *when* to act and *what data* to act upon.

As we discussed earlier, we have observed that today business processes and activities as defined by workflows often drive the information lifecycle changes. Such changes in turn imply how information must be managed accordingly. Workflow processes are often defined by business users. They describe how the business activities must be carried out in order to accomplish a given business task. Hence they often have an *activity-centric* view as shown in Figure 3. In this figure, a sample workflow for analyst research report processing with six activities is presented.

Not all events in a workflow trigger information lifecycle phase changes. Some, however, may change information from one state to another and require data management policy enforcement. For instance, a workflow event that informs some account representative to send acknowledgment letters to customers

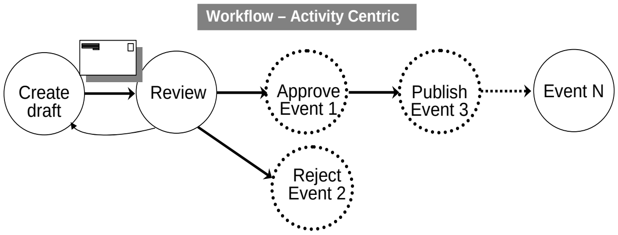

Figure 3: An analysis research report processing workflow with six activity nodes.

when their credit card applications are received do not necessarily impose any storage activities. However, the rejection of a credit card application may mean that the electronic version of the credit card application must be retained on WORM storage according to the company's retention rules. In Figure 3, we show that activity nodes *approve, reject* and *publish* trigger information lifecycle phase changes. Others do not.

In an integrated business solution that contains BSA, as workflow processes are being defined by business users, some events in that workflow will be identified as *data-related* events. Such events refer to those that trigger runtime information lifecycle phase changes. Once identified, users can also define what data management policies or optimizations they would like to trigger when those events occur. For instance, as shown in Figure 4, a migration and a retention policy may be triggered when a report is rejected. Similarly, when a report is approved, a retention policy is triggered to ensure that the report is retained for appropriate amount of time. In this manner, the data management policies are linked to workflow events.

Workflow events represent one kind of business semantic information that can help BSA to determine the *right time* to apply data management policies, *i.e.*, when events occur. To take the *right actions*, BSA must also identify the *right data* to operate on. In some cases, workflow events can provide hints on *what data* the policies should act on. For instance, the information traversing in a workflow may be the data that should be operated on. However, in general, the *right data* is often determined by the *business value of information* which are commonly hidden in business applications in form of data classifications. For instance, email archiving software sometimes classifies emails into different categories. Such classification implies the business value of different emails. In BSA, the high level information categorization represents another form of business semantics that can help determine the *right data* to operate on. The combination of *workflow events* and *business value of information* from business applications can be leveraged by BSA to determine the *right actions* at the *right time* and for the *right data*.

3.2 Business semantic preservation and runtime delivery

Once the workflow events and data classes are identified, such information must be preserved and delivered across system layers to allow effective data management close to the storage devices. At runtime, the most straight-forward method for delivering such semantics is to let applications monitor and deliver the workflow events to BSA as appropriate. Existing applications that support workflow already have the infrastructure for monitoring workflow events. For BSA solutions to work, applications must not only monitor but also deliver the *data-related* events as they occur. This capability can be easily supported by the applications themselves, or by customizing the workflow definition without changing application internals. Passing data classification information can be done by letting application tag data with special attributes, and allowing BSA to automatically extract and interpret the tagged information.

Delivery of semantic information inevitably requires extension of the existing storage API or creation of new APIs. One way to allow such semantic pass-through is to leverage the standard extended attribute API support, *e.g.*, setxattr() in NFSV4. With such a standard API, new attributes can be defined for such events and data classes. BSA understands the meanings of such attributes and can act accordingly. The popularity

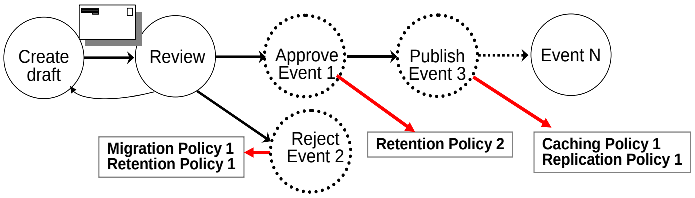

Figure 4: Sample data management policies that must be triggered as different workflow events are triggered.

Events	Data Management Policies
Approve	Retention policy: retain for 5 years Migration policy: migrate to WORM disk storage
Reject	Retention policy: retain for 3 years Migration policy: migrate to WORM tape
Publish	Caching policy: keep a cached copy on highend storage whenever possible

Table 1: Sample business-semantic-aware data management policies.

of NFS API also makes such an API more attractive than otherwise. At runtime, whenever applications or users generate a workflow events or change the data classes, an setxattr() call is made to BSA to set an extended attribute that indicates a given event or a data class change has occurred. BSA in turn determines and applies appropriate data management policies. Using existing and standard API make it easy for BSA to support a wide range of applications and simplifies the integration of applications and storage.

3.3 Business-semantic-aware policy creation and management

Once the business semantic information arrives at BSA, they are available for data management actions based on an entirely new generation of policies and optimizations. Such policies are devised and tuned based on business requirements, rather than storage capacity. Using the workflow shown in Figure 3 as an example, we defined some sample policies and listed them in Table 1. When a piece of information is *approved*, a *retention policy* can be precisely defined to ensure that the information can no longer be changed or deleted in a specific retention period. Furthermore, a *migration policy* may also be defined to store a copy of information to the WORM disk storage, such as EMC Centera or Network Appliance's Snaplock, to ensure non-erasability and non-rewritability, while leaving a read-writable copy on highend storage for fast accesses.

When a document is *rejected*, it implies that the document is probably of little value to the business. For archival and compliance reasons, corporations may be mandated to retain the rejected documents for a specific time frame. However, there is no reason to keep such documents on highend storage. They can be moved directly to the WORM tape storage. A migration policy can be defined for such reports and events. Finally, when an analyst research report is *published*, it may be desirable to cache such published reports on highend storage whenever possible. A caching policy can be defined accordingly. Clearly, BSA provides a brand new sets of opportunities for data management optimizations. This is also one of the most compelling aspects of BSA. The ability to act according to business logic changes dynamically have significant implications – seamlessly integrated business solutions can now be built and data management is no longer isolated from business processing.

4 Where Does BSA Work the Best?

BSA is a software layer interfacing between storage devices and business applications. It can present a storage API, such as CIFS/NFS or POSIX file system API, to communicate to applications. Hence to business applications, it looks like a normal storage backend, *e.g.*, a file system, with a new level of intelligence. It can communicate with a wide range of backend storage devices using any kinds of storage API that it supports, ranging from CIFS/NFS to special archiving solution API, *e.g.*, CAS supported by EMC Centera. This allows for large potential for deep integration between storage and applications. To storage, it is an extended storage front-end with advanced data management capabilities that traditional storage devices do not have.

BSA monitors data activities for data movement, retention, and replication activities. The monitored information are much smaller then actual data. Hence it has little overhead on the underlying storage. The new business-semantic-aware data management policies are designed with special care to avoid performance intrusion to normal data path. For instance, it is designed to avoiding scanning the entire file set for migration purposes while still ensuring quick identification of migration candidates. This allows BSA to scale up to large configurations easily.

BSA is designed to be the vehicle for constructing seamlessly integrated intelligent information management solutions. In particular, the boiling demand for compliance and ILM solutions, the immaturity of the existing solutions, and the unique characteristics and requirements of such solutions make BSA a perfect candidate for making significant contributions in those areas. Note that both kinds of solutions depend on end-to-end knowledge and coherent integration and management. This is exactly what BSA is designed for. We briefly describe how BSA can fit in to help building successful compliance and ILM solutions below.

Compliance solutions Compliance is prominent yet sophisticated in enterprises. In today's customer environments, it is not surprising to see various compliance related products scattering across the organizations, *e.g.*, email archiving software, Records Management Systems, Content Management Systems, WORM storage, *etc.*. Without careful integration among such components, various pieces of information can easily be leaked out unwantedly.

By leveraging BSA, records management retention rules defined in Records Management Systems can be mapped appropriately to WORM storage retention rules, to guarantee policy enforcement. The data classes can be linked to different storage device types of partitions, to improve storage resource utilization and reduce the overall cost. Data management actions, such as moving data from read-writable storage to WORM storage, can be triggered according to business needs and the effects are immediate and precise. This ensures information transparency and aligns data management activities much closer to enterprise goals that are established at the business-level than any existing methods.

ILM solutions ILM promises to reduce TCO and improve ROI by aligning business goals with cost-effective IT infrastructure through appropriate storage policy management and optimizations. Yet the very missing component in most of the ILM attempts is the ability to align business semantics with storage infrastructure. BSA is designed to create such linkage between business applications and storage. Moreover, BSA goes beyond simply linking – it further incorporates intelligence to utilize the business semantics in applications to create a new generation of data management policies that traditional ILM solutions were not able to. Such policies and optimizations respond promptly to business events and recognize the value of information according to the classification done by business applications, as opposed to the conventional approaches that ignore the business semantics completely. This allows for fine-grained data management and precise control, which are key to successful resource utilization optimizations. Without BSA, most of the ILM solutions will continue to manage data on the granularity of blocks and device partitions, leaving little room for optimizations.

BSA is poised to bridge the gaps in the existing ILM solutions and establish an end-to-end intelligent data management infrastructure that can take right data management actions on the right data at the right time. We believe that BSA can become a decision point for customers to consider whether or not to deploy an ILM solution. BSA can also be leveraged as the core technology in building ready-to-go fully integrated ILM solutions. In both cases, the difference that BSA makes to an enterprise in comparison to other existing technologies is going to be clear: The data management is much better aligned with business goals. The overall solution is much easier to manage and use. Information is protected and managed in an end-to-end fashion. The storage optimizations and policies are carried out for the benefit of the total solution. Without BSA, ILM is much more difficult to realize.

5 Summary and conclusions

Over the last few years, the customer demands and market dynamics have transformed the traditional information management practices that only deal with bits and bytes of data into the one that must understand some aspects of the business semantics of the bites and bytes of data and then act accordingly. Such a transformation requires a data management layer that is able to understand business semantics of information that is often embedded in applications as well as the storage device capabilities. BSA is positioned to span such application and storage domains and provide the core functions of such a data management layer.

BSA is a generic infrastructure for linking critical business semantics in form of workflow events and data classifications to data management policies and optimizations for storage. The workflow events and data classifications provide strong implications on when storage must act and what data to act upon. The key components of the framework consists of semantic preservation and delivery mechanisms in applications and a set of data management policies that make use of such semantic information for more advance information lifecycle management, such as migration, caching, and retention. Such an infrastructure bridges the gaps between high level business semantics and data management. It links data, the value of data, and management of data throughout information lifecycle phases. The resulting solution improves the overall system trustworthiness, allows for global system resource optimization, and ensures end-to-end seamless integration.

References

[1] Public Law 104-191, Health insurance portability and accountability act. http://aspe.hhs.gov/adnsimp/pl104191.htm, August 1996.

[2] Public Law 107-204, Sarbanes-Oxley, An Act. http://corporate.findlaw.com/industry/corporate/docs/-pub107.204.pdf, July 2002.

[3] Securities Exchange Act of 1934. http://www.sec.gov/about/laws/sea34.pdf, September 2004.

[4] EMC Corporation. Centera Universal Access Data Protection and Disaster Recovery Best Practices, May 2005.

[5] DOD 5012.2-STD. Design criteria standard for electronic records management software applications. Assistant Secretary of Defense for command, control, communications, and intelligence, June 2002.

[6] EMC. Business Process Manager. http://www.documentum.com/products/glossary/bus-_process_manager.htm.

[7] EMC. Compliant E-mail Archiving Solutions. http://www.emc.com/vertical/pdfs/financial/sec_compliant.pdf, December 2003.

[8] FileNet. FileNet Business Process Manager. http://www.filenet.com/English/Products/Business_Process_Manager.

[9] J. Menon and K. Treiber. Daisy: A virtual-disk Hierarchical Storage Manager. *SIGMETRICS Performance Evaluation Review*, 25(3):37–44, December 1997.

[10] Network Appliance. Network Appliance NearStore. http://www.netapp.com/products/nearstore/.

[11] M. Peterson. Information Lifecycle Management: A vision for the future. http://www.snia.org/tech_activities/dmf/SRC-Profile_ILM_Vision_3-29-04.pdf, March 2004.

[12] S. Powers, M. Barbeau, F. Kalmbach, and R. Vannucci. Complementing AS/400 Storage Management Using Hierarchical Storage Management APIs, 1999.

[13] Forrester Research. High-end arrays: The difference is in the details, April 2005.

[14] Forrester Research. Storage virtualization comes in many flavors, April 2005.

[15] Gartner Research. Forecast: Storage management software, worldwide, 2002-2009, executive summary, May 2005.

[16] Princeton Softek. Archiving Complex Relational Databases. http://www.princetonsoftech.com/library/rt/Archive-RelationalData-WP-us.pdf, March 2005.

[17] StorageTek. StorageTek Tape Storage. http://www.storagetek.com/products/tape_storage.html.

[18] Symmetrix System. EMC Corporation. http://www.emc.com/products/product_pdfs/ds/symm5000_I741-2.pdf.

[19] J. Wilkes, Richard Golding, Carl Staelin, and Tim Sullivan. The HP AutoRAID hierarchical storage system. *ACM Transactions on Computer Systems*, 14(1):108–136, February 1996.

[20] Zantaz. Zantaz Email and File Archiving Solutions. http://www.zantaz.com/solutions/email_file_archiving/index.php.