

## Electronic Data Archiving and Retrieval in the Public Sector

A Frost & Sullivan White Paper Sponsored by EMC  
Analyst and Author: Jarad Carleton

*Data archiving and retrieval in city, county, and state Public Sector organizations is a task that grows in complexity with each passing year. Population growth, regulatory changes, federal funding requirements stating the need for audit trails, and private sector needs for public sector data have made the job of state and local government data archival and timely retrieval a daunting task that challenges even the most efficient public sector agencies.*

*This white paper discusses the business challenges faced by public sector organizations in a demanding era of constrained IT budgets and the expectation that data should be available on demand at Internet speed 24 hours a day. Current business practices and problems are discussed as well as one possible solution with the potential to decrease operational costs for the Public Sector and enable valuable resources to be used elsewhere.*

### About Frost & Sullivan

Based in Palo Alto, California, Frost & Sullivan is a global leader in strategic growth consulting. This white paper is part of Frost & Sullivan's ongoing strategic research into the Information Technology industries. Frost & Sullivan regularly publishes strategic analyses of the major markets for products that encompass storage, management, and security of data. Frost & Sullivan also provides custom growth consulting to a variety of national and international companies.

The information presented in this publication is based primarily on interviews and therefore is subject to fluctuation. Frost & Sullivan takes no responsibility for any incorrect information supplied to us by manufacturers or end users.

This publication may not be downloaded, displayed, printed, or reproduced other than for non-commercial individual reference or private use within your organization, and thereafter it may not be recopied, reproduced or otherwise redistributed. All copyright and other proprietary notices must be retained. No license to publish, communicate, modify, commercialize or alter this document is granted. For reproduction or use of this publication beyond this limited license, permission must be sought from the publisher.

For information regarding permission, write:

Frost & Sullivan  
2400 Geng Rd., Suite 201  
Palo Alto, CA 94303-3331, USA

## Introduction

The public service sector in the U.S. is a market that has been forced to cope with large and rapidly expanding databases of information on everything from property records, to corporate and personal taxes, highway projects, health services, welfare, crime statistics, weather, harbor depths, and countless other topics and issues common to state and local governance. In some cases the electronic data collected by public sector agencies is temporary and is not subject to any type of retention requirements, in other cases it must be retained for long-term use by state and local authorities for differing time periods based upon individual agency needs as mandated by state and federal statutes.

In addition to having to cope with a plethora of different data retention and destruction guidelines, public sector organizations have been forced to utilize storage technology that is in many cases obsolete or at a minimum, several years behind the curve when compared to private sector organizations. The result is that it is still common, for example, for county recorder offices across the country to utilize optical storage systems to record all property data in a county despite the fact that optical storage technology is outdated and obsolete. In another example, many state and county governments continue to utilize high-speed tape storage systems for long-term archiving despite the difficulty in maintaining the integrity of the data on tape over several years. Although some might consider archiving data on tape to be antiquated, some agencies continue to store original paper documents in warehouses as a last resort backup for electronic data. The hard truth is that the majority of states are restrained from modernizing IT storage infrastructure due to continuing budgets cuts that have forced state CIO's to insist upon a positive ROI within the same fiscal year of any new storage implementation.

## Storage Problems and Solutions in the Public Sector IT

### Public Sector Market Segmentation

Market segmentation in the public sector falls into 4 generalized categories that include: federal government, state government, county government, and city government. The more complex side of public sector market segmentation lies in the countless number of agencies within each one of those generalized silos, each of which have distinct storage needs that in some cases could be mutually exclusive such as the needs of a state department of forestry and those of a state gambling commission. In each instance an agency would have a business need to archive data over a period of several years, but a department of forestry isn't subject to the same retention and destruction rules as a gambling commission and may have little if any need to keep archived data online and readily accessible. Conversely, a gambling commission would have good reason to keep archived data online for law enforcement agencies, tax boards, and other agencies that monitor gambling in a state.

Further examination of segmentation in the public sector revealed a wide variety of end users of stored data ranging from government workers to the private sector. Within the private sector, citizens rely on the integrity of the data stored by public sector agencies to ensure state and local services continue without interruption for programs such as, vocational training for the unemployed, worker compensation, welfare, child protective services and more. Additional uses of government data in the private sector include citizens looking up road conditions, weather, lottery results, voting information and more on state and local web sites such as:

- State of California – <http://www.ca.gov>
- State of Florida - <http://www.myflorida.com>
- State of Minnesota – <http://www.state.mn.us>
- County of Dallas – <http://www.dallascounty.org/>
- Miami-Dade County – <http://miamidade.gov/wps/portal>
- City of Minneapolis – <http://www.ci.minneapolis.mn.us>
- City of San Francisco – <http://www.ci.sf.ca.us>

Although there are many users of stored data created in the public sector, each type of data is put through a risk assessment that enables government officials to classify information by level of importance, which has a direct correlation to the determining if procedures to ensure data integrity and secure audit trails are necessary or superfluous.

### Common Business Practices in Public Sector

Interviews with state CIO's, state archivists, and state storage infrastructure specialists in several states revealed a disturbing lack of clarity within the public sector about data retention statutes at the state and local level. In fact, the only clarity regarding any statute whatsoever was in regard to federal HIPAA (Health Insurance Portability & Accountability Act) guidelines and the need of welfare and health departments to comply with the act.

Although all public sector officials are aware that there must be state and local statutes regarding long-term data retention, accessibility, audit trails, nonrepudiation, and destruction, none were able to refer to specific laws currently in effect. Rather, there is a nationwide pattern of public sector agencies pointing fingers at other agencies and refusing to take responsibility for ensuring archived data has audit trails, is properly retained for mandated periods, and is destroyed according to standardized record destruction policies.

The resulting chaos has forced IT departments in states such as Florida, Texas, and Michigan to create guidelines for state and local agencies that puts the onus of responsibility back onto each organization that stores data in a shared state or local data center. This in turn has created a situation in which data center operators, CIO's, archivists, and storage specialists rely

upon each agency to manage its own data retention and audit trail requirements by providing detailed individual business needs to the appropriate public sector IT official in technical rather than legal terms.

Technologies commonly in use within the public sector to meet the needs of agencies requiring long-term data retention and archiving are optical storage systems, high-speed tape storage, and SAN's, although most archived data continues to be stored on tape and vaulted away from data centers. Long-term data integrity issues aside, offsite vaulting of data on tape has contributed to slow response times by government agencies when archived data is needed for audit purposes, civil litigation, criminal cases and more. In addition, not a single state mandate could be found regarding data accessibility timeframes that other than those stating that data should be available within "a reasonable" period of time, thereby introducing significant room for interpretation at each state and local agency subject to data retrieval requests.

### New and Potential Future Record Retention Trends

One of the biggest drivers towards state and local modernization of long-term data retention and retrieval has been the federal government's e-government initiative (<http://www.whitehouse.gov/omb/egov/>). Through the e-government initiative, the federal government encourages states to find ways to standardize the creation, archiving, and sharing of public sector data between state and federal entities, providing an indicator that further IT infrastructure and application centralization initiatives at state and local levels can be expected over the next decade. This is important due to the fact that state and local level agencies that receive federal funding are subject to federal data retention, accessibility, destruction and nonrepudiation statutes passed into law by Congress. In addition, all state and local agencies receiving federal funding are subject to surprise inspections by federal auditors that have the ability to fine and/or withhold federal funds until data storage and security measures comply with federal law.

Despite the fact that several states understand the need for centralization and pooling of storage and other IT resources on a state and local level, very little has been accomplished in the regions examined which include the Western, Southwestern, Southeast and Midwest states. Examination of the business benefits of pooling IT storage infrastructure as well as other IT resources is currently underway in Texas with the ACE initiative (<http://www.dir.state.tx.us/ace/index.htm>) that is focused on finding ways to enable the government of Texas to use IT more effectively and efficiently.

Other states that have conducted studies to increase IT efficiencies are Washington, California, and Florida. Washington enlisted the help of PriceWaterhouseCoopers to conduct an audit with the intent of finding additional potential areas of IT improvement for the state that could result in cost savings. California approached the issue after a new governor came to office with an interest in running the state more like a business. The result was the creation of a 21 member CPR (California Performance Review – <http://cpr.ca.gov/>) panel tasked with examining all

areas of California state government including its IT operations and suggesting changes to positively enhance productivity and reduce operational costs, which continues to be an ongoing process. As opposed to the other states previously mentioned, Florida has already passed through a review process called the Enterprise Information Layer study that was tasked with finding ways to enhance government responsiveness for the Department of Children and Families through IT consolidation and pooling of resources. The goal was to reduce costs over the long-term and ensure that the state didn't fail in its mission to help children in need. Unfortunately the plan was rejected by the state legislature as too costly despite ROI projections of millions of dollars in savings in the first 3-5 years.

Attempts to pool data storage and other IT resources are underway however, in both Minnesota and Michigan. Minnesota's storage resources pooling initiative has been focused exclusively on law enforcement in the state and has led to a firestorm of controversy over the lack of secure audit trails for data stored by the CriMNet initiative. Subsequently, the Minnesota legislature has been working to develop legislation to address the lack of focus statewide on maintaining data security and integrity through various revisions of House Bill HF2800 (<http://www.leg.state.mn.us/leg/legis.asp>).

Conversely, Michigan's drive to centralize IT infrastructure has received support from the highest levels of government and although not complete, has been considered a success in terms of operational cost savings already achieved by the centralization push. The initiative to pool state IT resources that relies upon state and local agency cooperation has been so successful that it was recognized as the winner of the 2004 Digital States Award by the National Governor's Association in July (<http://www.centerdigitalgov.com/center/04sustained.php>).

### **Frost & Sullivan Opinion on the Centralization of Storage in the Public Sector**

This white paper has established the nature of the business challenges as it pertains to long-term public sector data storage and retrieval. Despite the fact that the majority of state and local governments have structures and legal separations that make it extremely difficult to work together to achieve cost savings for IT infrastructure, the fact remains that single best way to achieve business-like efficiencies in government is to standardize the technologies used and to pool IT resources.

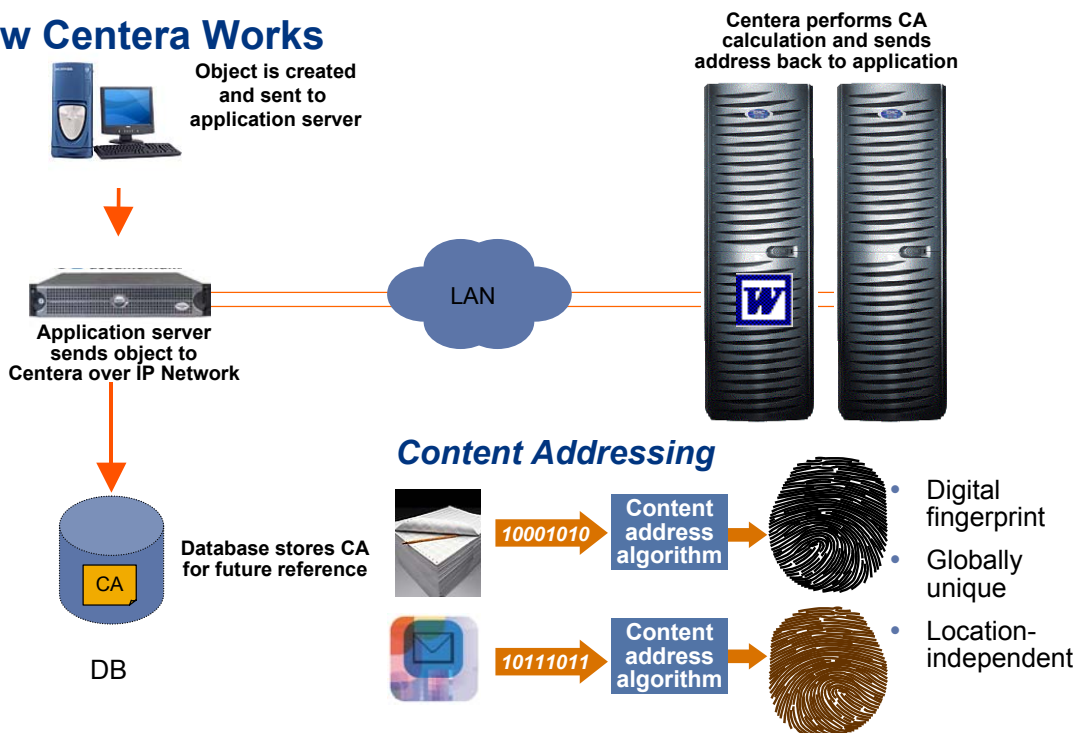
Technologies available today provide the ability to maintain a structural and legal separation between state and local governments while also allowing those governments to reap the financial benefits of pooling storage infrastructure as well as other technologies necessary to running government operations. In light of the dire need for cost savings in the public sector, the remainder of this paper is devoted to presenting a reliable solution with a proven ability to meet regulatory scrutiny for data storage today and in the future.

## Technology Capabilities

### EMC Centera™ Content Addressed Storage (CAS) System

Content Addressed Storage places stored data within a storage array based upon a Content Address (CA) rather than the information's physical or logical placement in the array. EMC Centera™ is the industry's first implementation of Content Addressed Storage (CAS). Centera stores information objects based on a 128-bit globally unique address that is derived from the object's binary representation. With a Content Address derived from the content itself, Centera eliminates the storage of multiple copies of identical information, regardless of how many requests to store a piece of content are made. For business continuance purposes, Centera stores the content and protects it using content mirroring or content parity protection within the same Centera array. The addressing and encryption functions are similar to a public key infrastructure (PKI) or digital fingerprints ensuring, security, authenticity and nonrepudiation.

### How Centera Works



When content is stored in Centera, the unique content address for the stored object and the metadata describing the object are inserted into an XML file. The XML file is referred to as a C-Clip™ Descriptor File (CDF), which also has a unique content address calculated for it. The CDF is then protected (in the same way as the object itself) and is the mechanism used by an application to retrieve an information object. Retrieval of content within Centera is based entirely upon the content addresses rather than through the use of a centralized directory, pathnames, or URLs. Using a content address to access fixed content makes the management of physical and logical location of the information unnecessary, which results in a dramatic

reduction in system/storage management and related costs. In the event that fixed content is altered and stored again, Centera computes a different content address for the altered content and stores it in the array. Original fixed content is not overwritten, ensuring an intact audit trail and assurance that fixed content remains in its original state.

As an integral part of maintaining data integrity and audit trail in the event of a hardware failure in one part of the array, Centera will self-heal by detecting the fault and generating a new copy of the content objects. As this process takes place, the affected disk drive or storage node is isolated from the rest of the system until it can be replaced. Lastly, due to the fact that applications don't have knowledge of the physical placement of fixed content within Centera, components can be replaced and Centera software upgraded without disruption, demonstrating a solution architected to easily scale up to one petabyte and beyond.

Centera can and has been configured to help meet the most stringent requirements of regulated environments. Specifically, Centera enforces application-based retention periods within its microcode. End users have the ability to lengthen the retention periods, but cannot shorten them.

Just as important as data retention in a regulatory environment however, is the importance of automatically deleting unwanted information using U.S. Department of Defense data destruction standards, overwriting information multiple times with random characters, complimentary values, ones, and zeros, when data reaches its expiration date, which has proven to be of importance in both private and public sector organizations due to best practices guidelines as well as federal regulations. This not only frees IT personnel from doing a low level maintenance task, but also creates a standard data destruction policy within an organization eliminating potential legal liabilities, and negates any ability to recapture deleted information using disk management tools. Also essential to any data management policy, is the ability to immediately suspend data destruction in event of litigation. Centera's central policy management allows immediate extension of data retention periods as needed for any eventuality.

### Meeting Public Sector Record Retention Needs with EMC Centera™

Long-term data reliability and security are of the utmost concern to Health Departments, Law Enforcement, Welfare agencies, Departments of Highways and Transportation, Tax Boards, Environmental agencies, and any organization that receives state or federal funding. Data reliability, within Centera is assured as a result of its RAIN architecture that eliminates all single points of failure within the platform and enables non-disruptive servicing of the system. Centera itself is composed of independent nodes with one terabyte of raw storage capacity, and is interconnected to all other nodes in the cluster via CentraStar™ software (Centera's operating environment) and a private LAN.

Additional Centera features of critical importance to the public sector are:

- 128 bit globally unique CAS
- Secure audit trails that reveal identities of persons that alter data
- Content is stored and mirrored once despite being used for numerous email attachments
- Data is mirrored in the same array and can be replicated to another array when needed
- Permits management of retention periods from one GUI for multiple DB applications

Centera easily handles access to fixed content on demand via LAN connectivity to application servers. Because each node within the cluster operates as a storage node or an access node, performance can be scaled to meet demand by non-disruptively adding additional front-end access nodes to augment bandwidth to application servers. This enables Centera to work within a growing database-driven environment that is common in the public sector. Centera makes this possible by allowing database fields to interact with Centera's API to a content address as a pointer to specific objects in the cluster. Thus, when a database request is placed from any of a number of public sector agencies into the same array, the application will use the Centera content address to retrieve the correct information quickly and efficiently.

Centera's contribution to modern cost-conscious public sector environments is exemplified in two strikingly obvious ways. The first is the manner by which Centera ensures one copy and one replica of fixed content is stored on the system regardless of the number of times it is used. Thus, when Centera is used to store e-mail for example, it only keeps the original and protects that information via content mirroring or content parity protection, even when that document is attached to an email. Furthermore, because Centera is self managing, operational costs decline when compared with optical and tape storage systems used by the public sector in 2004.

In the private sector as well as the public sector, the efficient storing, protecting, and replication of information within Centera has been shown to substantially lower the total cost of ownership (TCO) for long-term data retention.

Centralized control of one to many Centera clusters also plays an integral role in lowering the TCO of long-term online storage. Key to lowering TCO for long-term storage of fixed content is Centera's ability to provide a single repository for multiple public sector agency applications. This translates into measurable savings in cost-per-megabyte of storage for public sector IT budgets as maintenance and training costs significantly decrease. Elimination of multiple repositories also dramatically decreases potential mistakes that lead to criminal liability or suspension of federal financial resources as a result of state and federal audits. In these situations, when an IT team has to manage several repositories, data could be unintentionally destroyed before a technician is able to modify the retention parameters on one or more

repositories. Public sector IT departments utilizing Centera are able to extend retention periods for specific information objects or categories of objects as controlled by the users application, effectively minimizing potential liability issues.

## **Conclusion**

Content Addressed Storage, as implemented in EMC Centera, offers the public sector an efficient and cost effective managed storage solution for fixed content originating from and accessed by multiple enterprise class applications. Centera's ability to satisfy the stringent regulatory requirements some state and local public sector agencies must comply with helps organizations avoid potential fines or the withholding of federal funds that can result through the use of less secure data repositories with questionable audit trails which is frequently the case with archived data stored on tape. In addition, it provides a highly cost effective networked storage solution that addresses the need to provide immediate access to information 24 hours a day which can be critical for law enforcement, health services, and numerous other agencies that regularly interface with their federal counterparts. Finally, Centera's ability to eliminate multiple repositories reduces IT management and maintenance costs and accelerates public sector return on investment when used to replace both optical and tape storage systems commonly used in the public sector in 2004. In light of continuing budgetary difficulties in the public sector, rapid and positive ROI's on new IT infrastructure implementations has become a critical measure for budget conscious counties and states across the United States and will remain a permanent fixture for all future IT purchasing assessments.