FEDTECH TECHNOLOGY INSIGHTS FOR LEADERS IN FEDERAL GOVERNMENT
BROUGHT TO YOU BY: CDWG
Published on *FedTech Magazine* (http://www.fedtechmagazine.com)

How Agencies Are Scaling Mountains of Data

Data growth is going nowhere but up, and agencies must find a way to keep up.

As steward of the nation's documents, images, audio, video and other important records, the National Archives and Records Administration is responsible for storing massive amounts of data. Much of that information is still in legacy form — paper, tapes, photos and more — and stored in archival stacks of controlled material storage in NARA's 44 facilities across the country.

NARA has several challenges when it comes to storage. First, it is working hard to digitize its existing physical inventory. At the same time, it must find the best way to store that newly digitized data, along with the born-digital data it already manages.

As if that weren't challenging enough, the amount of data NARA must store is growing exponentially. One of the largest data sets it has received electronically is the 2010 Census, more than 300 terabytes of data that includes metadata and images of completed census forms.

"The only way we could actually receive that amount of data was as 17 racks of storage delivered to us on a truck," says NARA CIO Michael Wash.

And that's only one data set. Today, NARA stores large data sets in various forms, such as geospatial data from NASA. NARA is starting to accept these large data sets for its fledgling Electronic Records Archive (ERA) program.

The agency is responsible not only for storing data, but also for ensuring that it is easy to locate and retrieve — meaning that, along with the data itself, the agency must store technical, descriptive and preservation metadata. Today, electronic records are stored on spinning disks in many cases, but Wash says that will change out of necessity as volumes grow.

Data growth is going nowhere but up, says Greg Schulz, senior adviser for the Server and Storage IO Group. That's partly because of the growth of unstructured data, such as geospatial, video and audio. But it's also because collecting information that may not have been necessary a decade ago, such as cell phone records, is so easy today. Agencies such as NARA, the U.S. Postal Service and Lawrence Livermore National Laboratory are looking at a variety of technology solutions to handle their skyrocketing storage needs.

"More data is being captured and collected, and more of it needs to be analyzed," Schulz says. "At the same time, people have begun to realize the need for unlocking the potential value of all of this information by doing things like looking at trends, doing more modeling, and producing graphs and simulations."

You've Got Mail — Lots of It

The U.S. Postal Service manages a massive quantity of data: 22 petabytes of enterprise storage for roughly 600 applications, an enterprise data warehouse for corporate reporting, and about 40TB in Microsoft Exchange.

The challenge, says John Edgar, USPS' vice president of IT, is that this data is growing as much as 30 percent per year, yet because of budget and other issues, the agency must make do with the two data centers it now operates.

"We aren't in a position to be building new data centers, so for the past several years, we have been trying to evolve our storage technologies to meet capacity and growth requirements in a way that drives down the physical footprint of the storage technology itself and reduces cooling and power costs," he says.

The first step was moving from the agency's traditional islands of information via direct-attached storage to centralized data centers. The next step was improving the storage infrastructure by tiering data and using thin provisioning when possible. Around 2009, the agency began virtualizing its network-attached storage and is currently working on virtualizing its storage area network.

330TB

The amount of data generated by the 2010 Census, which must be stored by the National Archives and Records Administration

SOURCE: NARA

"We are working toward elasticity and the ability to be able to deliver to our users quickly," says Dan

Houston, USPS' database software services manager. "Through our virtualized environment, we can very quickly provision storage either to new environments needed for particular programs or to address growth within a project."

Storage virtualization will hold the agency in good stead as it moves toward implementing a new, intelligent mail barcode system for all postal mail. As the system is applied to the billions of letters USPS processes, each one becomes a scan event, and that data must be stored and then provisioned, both internally and to external mailers.

"We're continuing to make improvements to our storage all the time," Edgar says. "In the next few years, we hope to make progress on dealing with unstructured data and be in the position to provide continuous data protection."

A Different Type of Storage Problem

While the Lawrence Livermore National Laboratory is well known as a science and technology development arm of the Energy Department, its National Ignition Facility is far less familiar. NIF specializes in conducting fusion experiments using high-energy lasers — experiments that can easily create 5 gigabytes of data in a few billionths of a second.

The high-resolution still images and video created by NIF experiments then must be stored in a way that fully preserves the data while allowing access by scientists who need to study it. Scientists tend to access the data repeatedly within short periods after an experiment to compare results, identify trends or mine the data, explains Timothy Frazier, a senior architect at NIF responsible for its IT infrastructure, including storage.

To keep data available and in formats useful to its scientific community, NIF stores much of its data in databases. The data takes the form of TIFF files (for images) and Hierarchical Data Format (for plasma data). Data files tend to be large, from 20 to 150 megabytes per file. Storage takes place on spinning disks inside NIF's data center, located on Livermore's campus.

Because the data center has a finite capacity, it was important to think through how to store data over time, given space constraints. The solution was to make an agreement with the high-performance computing division at Livermore, which now stores NIF's older data in its tape archives.

"It has minimized our need for spinning disk while allowing us to preserve our data and still access it if necessary," Frazier says.

Cloud as a Part of the Storage Equation

With continued data growth and complexity, many agencies have a plan for moving at least some of their storage to a cloud environment. By doing so, they gain agility and flexibility, along with the ability to share information.

NARA's Wash is a major proponent of the cloud for just those reasons. He says the agency is already in the process of determining how best to use the cloud for storage.

"Moving to a cloud-based storage environment will enable better access," Wash explains. "Once we get the access piece moved to the cloud, we will work on the archival store, which will employ some hierarchical storage management and possibly be a combination of onsite and cloud storage."

The cloud also is likely to become an important tool for Livermore's NIF, which is in the process of expanding its data and research opportunities outside of the Energy Department. Eventually, NIF will begin making its research available to scientists worldwide. When that happens, Frazier says the cloud storage infrastructure will become a valuable tool.

Schulz agrees that the cloud can be the way to go for storage, but warns agencies to carefully evaluate which type of cloud environment is best suited to various types of data storage.

"A hybrid cloud approach often works well for cloud storage," he says. "That way, you can put archived

inactive data into an inexpensive cloud environment with a slower response time but use a cloud service with solid-state storage attached to it for active data that you need access to more quickly."

The Future of Federal Storage?

As data grows exponentially throughout government, it is becoming clear that traditional storage methods aren't adequate. They don't make sense, they can't keep up, and they are prohibitively expensive.

A November 2011 memo from the White House on managing government records called for the federal government to better coordinate records management activities. In August 2012, a second memo, jointly authored by Jeffrey Zients, acting director of the Office of Management and Budget, and David Ferriero, the National Archives and Records Administration's chief archivist, spelled out a potential solution. The memo explains that NARA is looking into establishing a secure, cloud-based service that will store and manage unclassified electronic records on behalf of all agencies.

Today, when agencies transfer records to NARA, the data must physically reside at a NARA facility. In a cloud environment, employing a data-at-rest model, "it would just be a matter of identifying where it is and pointing NARA to the data," Wash says. "At that point, NARA would begin taking responsibility for that section of data and manage it from that point forward."

Author: Karen D. Schwartz [1]

Source URL: http://www.fedtechmagazine.com/article/2012/10/how-agencies-are-scaling-mountains-data

Links:
[1] http://www.fedtechmagazine.com/author/karen-d-schwartz