Archival Data-at-Rest

September 21, 2012

**Concept**

NARA’s vision of “archival data-at-rest” is simply ensuring that Federal electronic records that are created and stored in the Cloud can also be archived and made publicly available in the Cloud. As Federal agencies transition administrative systems and data storage to Federally-approved Cloud environments through FedRAMP, increasing volumes of Federal records will be generated in these hosted environments. As those records age, they will transition from “active” records to “inactive” records (infrequently referenced or changed) and will eventually be either archived or disposed of. Inactive and archived electronic Federal records will become “data-at-rest”, which must be retained, managed, and preserved in an efficient and cost-effective manner.

**Technical Approach**

Over time, Federal agencies will establish a number of Federal Cloud environments established which will appear as isolated cloud instances, much like today’s Federal data centers are isolated.

These Clouds need to connect and be interoperable. Fortunately, early work on the Federal Enterprise Architecture (FEA) helped connect Federal agencies with the expectation of agencies being able to easily share data and services. Much of this early work is evidenced in the growth of shared services.

This work continues today under the guidance of the Office of Management and Budget. Activities now need to take place to ensure our multiple Federal Clouds securely connect and are interoperable. Doing so will allow data in Federal Clouds to be treated and managed as if they were in a common location, which will be critical for continued expansion of shared services.

Shared services are now being offered by the Departments of Agriculture, Defense, Interior, Health and Human Services, Transportation, Treasury, and the General Services Administration. These shared services are primarily in the categories of human resources and financial management.

The Federal CIO Council is in the process of establishing Lines of Business, and one being considered is a Records Management line of business, which could result in a Records Management shared service – RMaaS – sponsored and managed by NARA.

The archival data-at-rest model adapts Federal recordkeeping requirements and re-conceptualizes the traditional records lifecycle to function in cloud architectures:

* *Active Records.—*Federal agencies are already creating new Federal records in a cloud environment and are moving born-electronic and digitized records into cloud storage. NARA envisions an environment where these records are appraised and scheduled in the Cloud. In this model, commodity IT shared service providers retain and manage Federal records as functionality provided within business and administrative applications. A well-architected Federal Cloud environment is critical to supporting this effort.
* *Inactive Records.—*Federal agencies require secure and cost-effective solutions for storing electronic records that still require some access, but on an infrequent basis. Today, agencies frequently store their inactive analog records in fee-for-service NARA Federal Records Centers, but there is currently no electronic equivalent. NARA must develop robust alternatives for agency storage of inactive records that do not require the physical transfer of these files.
* *Archival Storage.—*Once cloud-based electronic records have reached disposition age, NARA must provide a mechanism to transfer those records to the legal custody of the National Archives without physically moving the data. NARA envisions having procedures, tools, and a processing environment that would allow for electronic archival records to be transferred to NARA’s legal custody, processed (including adding metadata and redacting personally-identifiable information), preserved, and made publicly available in machine-readable formats without being physically transferred or leaving a hosted environment.

**Economic Benefits**

The archival “data-at-rest” model replaces current practices, which largely pass electronic records on portable storage devices through the same physical transfer process used for analog records in boxes. Electronic records are physically moved as they transition to inactive records and then permanent, archival records: At each step, physical transfer devices are acquired, physical devices are transported, and a new storage environment is acquired, configured, maintained, and refreshed at the host location.

Developing an archival data-at-rest model offers the potential for significant savings to the Government. In one recent example, the Census Bureau transferred the 2010 Census to the physical custody of NARA. About 350 terabytes of data was transferred to NARA from Census. In the world today, this is not a large amount of data, but it does pose a difficult challenge if you need to *move* this data. The Census Bureau collected and stored the 2010 census data in their data center. Once all the data was captured and secure on their infrastructure, the physical hardware was moved to NARA on 17 racks of IT equipment that required two large trucks to transport.

Once received by NARA, the equipment needed to be installed and data processing had to take place to ensure that all the data arrived intact. This process took NARA 10 months to complete, utilizing two technical employees approximately 40% of their time. Additionally, the hardware and software on the transport equipment now needs to be maintained and updated, which costs approximately $1 million per year. This estimate does not include the costs of refreshing storage hardware as it ends its useful life or the costs of updating records formats so the Census data remains readable after current formats are updated or become obsolete.

Utilizing a data-at-rest approach would eliminate the need to procure redundant hardware to physically move electronic data. It would also eliminate the need to prepare data for movement (Census took 8 months to complete the preparation of the data to move), and eliminate the need to check the data once it is received (NARA took 10 months to do this).

Reassigning access rights to data throughout the lifecycle of records will streamline the process and eliminate costs. Additionally, from a practical perspective, as data sets continue to grow in size, it will become impossible (in actuality, it already has) to physically move data and prepare it for archiving.

A common working definition of Big Data is as follows: Datasets whose **size** is beyond the ability of typical software tools to **capture**, **store**, **manage**, and **analyze** within a tolerable elapsed time. We are dealing with Big Data today, and need to develop a Federal Cloud architecture to work effectively with this data, keeping data-at-rest and processing this data in place. This will allow the Federal government with NARA’s support and leadership to establish RMaaS.