A Library to Manage Web Archive Files in Cloud Storage

Yinlin Chen[1,2], Zhiwu Xie[2], and Edward A. Fox[1]

[1]Department of Computer Science, [2]University Libraries

Virginia Tech, Blacksburg, VA 24061

{ylchen, zhiwuxie, fox}@vt.edu

ABSTRACT

When web archive data are not being actively used, it is usually beneficial to ingest them into a digital library for curation. However, it becomes a challenge when the volume of the data grows beyond the size of a typical repository. We propose to augment the digital library with external mass storage. More specifically, we developed a Java library to bridge the Fedora Commons repository with cloud storage services. In this demonstration and lightening talk we will demonstrate how a web archive library interacts with the cloud storage and manages remote files in the digital repository. We also will discuss scenarios suitable for using this library and what benefits it brings. This Java library (fcrepo-cloud-tool) is available as Open Source software.

1. PROJECT DESCRIPTION

The goal of this open source project is to provide an easy way to manage Fedora Commons repository [1] files with cloud storage services. It will address the common problem when a local repository needs to manage a lot of files that exceed its physical storage limit. When a local digital repository runs out of storage for new incoming archive data, they need to purchase new hardware and upgrade the system. It could take hours or days to finish a system upgrade. One solution is to put the entire system in the cloud environment, but that involves infrastructure redesign in order to fit into a particular cloud service and may not reduce cost [2]. Another approach is using the filesystem in userspace (FUSE) [3] technique to mount cloud storage as a local folder. However, this approach brings many other issues, for example, a user needs to properly configure Fedora's file block size. Further, different cloud providers have their own limitations (e.g., number of file allowed in a container). Moreover, some FUSE software keeps an in-memory cache of the directory structure which is not able to support large filesystems. Our approach is to enable a repository to work with cloud storage, and move files from local storage to cloud storage so that the repository can take advantage of the benefits from cloud providers and extend its own capability to manage many large files.

We developed this library to provide a generic way to manage files in the Fedora Commons repository. Through the APIs, a Fedora client can be implemented to move any Fedora Commons repository file to cloud storage. The library takes care of all the underlying complicated operations. These operations are: 1. upload a local file to the cloud storage; 2. create a Linked Data Platform (LDP) container with file information and a user defined field indicating the URL of that file in the cloud storage; and 3. delete a local repository file. When a Fedora client wants to download a file which is uploaded to the cloud storage, they will receive a Fedora response that contains the URL address of that file and download it directly from the cloud storage. A Fedora client can also use APIs to restore a file from the cloud storage back to the local repository. These operations are 1. download a file from the cloud storage; 2. ingest a file into the local repository and create an LDP Non-RDF source; and 3. delete or keep that file in the cloud storage.

Using this approach to manage files in a Fedora Commons repository can yield many benefits from the cloud services, making them secure, durable, low cost and highly-scalable. Depending on various file usages and scenarios, a librarian can decide whether to put frequently or infrequently accessed files into the cloud storage. For example, the infrequently accessed files can be stored in the cloud storage (Amazon S3) and further archive these files in the Amazon Glacier to reduce cost.

This library is highly customizable and currently supports Amazon S3 and will be extended to support multiple cloud environments, such as Microsoft Azure, Google Cloud Storage, and Rackspace.

JCDL '16 Newark, NJ USA

DOI: 10.1145/1235

2. ACKNOWLEDGMENTS

This work is partially supported by Amazon AWS Research Grants.

3. REFERENCES

[1] DuraSpace, "Fedora Commons Repository," *http://fedorarepository.org/*.

[2] D. S. Rosenthal and D. L. Vargas, "LOCKSS boxes in the cloud," Report to Library of Congress, September 2012, http://www.lockss.org/locksswp/wp-content/uploads/2012/09/LC-final-2012.pdf.

[3] M. Szeredi, "Fuse: Filesystem in userspace. 2005," *http://fuse.sourceforge.net*.