Title: Cloudstor: GLAMming up the ecosystem

Affiliation: AARNet, Australia's Academic and Research Network

Authors: Guido Aben guido.aben@aarnet.edu.au ,
Hilary Goodson hilary.goodson@aarnet.edu.au

Introduction

CloudStor, AARNet's ownCloud-based synch&share platform[1] was initially conceived as a data movement and synchronisation platform to operate at AARNet line speed (~10Gbps) and synchronise across the Australian continent; it was initially pitched at individual users. Indeed, its interface and client platform compatibility were designed to resemble other household names in commodity cloud storage solutions ; except in the case of CloudStor, the backend was meant to deal with science-size data volumes (Gigabytes to Terabytes, and directories with multiple thousands of files).

It appears this product met a demand, and usage has grown to 43.000 accounts and a PB of active data under management. Over time, we began to realise that these volumes meant we had well and truly entered the realm of Metcalfe's Law – we noticed Cloudstor was beginning to be used as a research hub, not merely a synch app.

By corollary, we had an opportunity to cater for project- and collections level functionality; and cater to research managers and collections managers.

Fortunately, the system we built is fitted out with APIs and connectors, which we could use to extend the system and make it more powerful. So, we embarked on a strategic mission to make CloudStor "collections level relevant".

A Research informed Roadmap

Given that the strong point of CloudStor is the digital enablement of mid- tail users and their research groups (as opposed to early adopters; i.e., typically the hard sciences), we focused our engagement on the Galleries, Libraries, Archives and Museums sector ("GLAM"). This is a sector with vast stores of digitized, scientifically and scholarly relevant content, who more often than not are struggling to understand how to tackle accessibility[2] and findability (can be as simple as an indexed search engine…) within and between their collections; often, the collections aren't even "on the Internet" in the first place. Beyond digitization, generally funded on a cap-ex model, there was a gap in addressing post-digitization, including ingest of rapidly growing borne-digital objects. Cases exist where such collection owners threw up their hands and "just copied the entire collection over to YouTube" or Google Gallery; a fine fix tactically perhaps, but with significant strategic ramifications. There was a need for scalability and interoperability to enable long-term research capabilities; the known but not executable (collaboration across collections, use of datasets able to be found and drawn from across multiple institutions) and the unknown, or 'translational' capabilities, when able to draw together components of metadata to compute facilities.

In an effort to take our "simple, democratised" tools for individuals to the collections level, we've begun trialling, with selected partners, the following add-on technologies on top of CloudStor, which this proposed talk will cover.

---

[1] cf.TNC15 http://slideplayer.com/slide/10380308/],
[2] https://sites.google.com/site/glaminnovationstudy/

Direct transfer of oversize datasets

AARNet partnered with the Australian Institute of Aboriginal and Torres Strait Islander Studies to migrate an on-premise storage of 195,000 digitized artefacts. The existing Synch and Share paradigm enabled by Cloudstor worked well, but the initial migration (the aim of which was to shut off any on premise-storage) there was a backlog of data that simply needed to be uploaded, never downloaded back to the instrument). Deployment of an S3[3] gateway (Minio) directly into CloudStor backend storage, plugging into the institution's current CMS (Alfresco)_ has been successful. We are currently exploring deploying a CERN-based open source model (Invenio) to replace the current CMS

Archiving and Repository services

We have also developed a full tape backup service straight from collections held inside CloudStor; so while the user side interface retains the simplicity of a synch&share service, the collections still benefit from enterprise grade long-term digital archiving and retrieval. Given this capability to cater for long term archival, we are now investigating options around repository services; at the time of writing, a proof-of-concept is being formulated with the Tasmanian Museum and Art Gallery, as lead agent with the Tasmanian Archives and the Tasmanian State Library, to develop a network-hosted CMS and repository functionality which is meant to deliver a whole of-state Tasmanian "Digital Access to Collections" service. This centralised service will equip the Tasmanian agencies with capabilities to preserve and retain access to digitised and born-digital content in as FAIR a manner as possible.

Multimedia streaming/viewing via plugin (through Kaltura):

Large multimedia holdings are being stored in CloudStor, notably the collections of the Australian Centre for the Moving Image. We have deployed Kaltura[4] as the mechanism to extend media management (including streaming functionality) to our existing storage architecture; the resultant environment provides large scale data movement, collections management, transcoding, streaming and streaming/rights control (e.g., geofencing. The collection can now be viewed both inside the physical gallery, but it can also be made available to "online visitors"; cloudstor nodes are distributed across Australia and as a result they form a simple CDN. By virtue of the collection now living "inside the research cloud", it has become easy to give access to parts of the collection to 3rd party machine learning and analysis services (e.g., Google Cloud Video Intelligence); this opens up new avenues for collections research.

Emulation as a Service

A funding bid is under review as of the time of submission of this abstract, which puts us in proposed partnership with two Australian and one German university (Universität Freiburg) to trial Emulation as a Service[5], with the intent of providing such a capability as a cloud-based service. This research acts as a path-finder to test the feasibility and usefulness of web-presenting executables (running software; whether current or historically archived) within Cloudstor. We are building capacity for cultural and academic sector in the area of born digital heritage, to access compute environments on demand, for components of datasets.

[3] https://en.wikipedia.org/wiki/Amazon_S3#S3_API_and_competing_services
[4] https://corp.kaltura.com/deployment-options/kaltura-community-edition
[5] http://bw-fla.uni-freiburg.de/

Biographies

Guido Aben is AARNet's director of eResearch. He holds an MSc in physics from Utrecht University. In his current role at AARNet, Guido is responsible for building services to researchers' requirements, and generating demand for said services, with CloudStor perhaps the most widely known of those.

Hilary Goodson is AARNet's Manager of Strategic Engagement. She holds an MA in Philosophy from the University of Melbourne. In her current role at AARNet, Hilary is responsible for engaging with institutions, particularly the GLAM sector, to enable the deployment of services and growth of infrastructure capabilities to meet their needs.