University of California
CDL
California Digital Library

UNIVERSITY
OF
CALIFORNIA

Implementing an Ingest Workspace

Using AWS S3, FSx for Lustre and Lambda to Stage Content at Scale

DPC Workflow Webinar Series – 2025-02-05

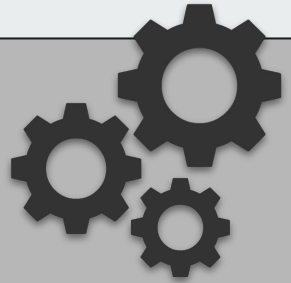
Eric Lopatin, Digital Preservation Service Manager
University of California, California Digital Library

Problem definition

- Across UC campuses, many libraries and departments interested in digital preservation have limited resources to apply toward archiving their content.
- Though they have control over catalog records and associated metadata, many tend to work with vendors or other library units to acquire digitization services.
- Resulting digitized content may be relayed to CDL on HDDs or is copied to an S3 bucket or another local staging volume.

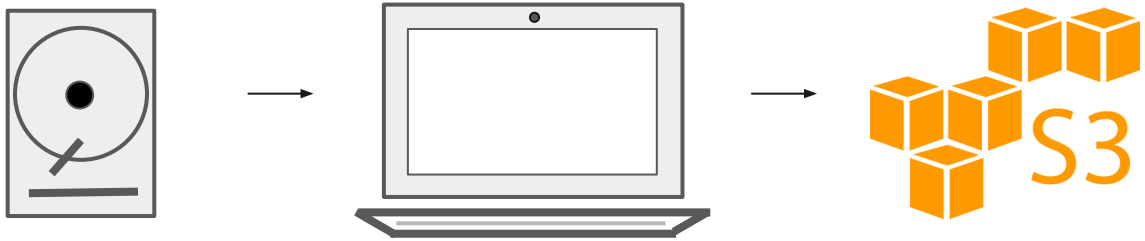
Context and requirements

- To enable acquisition from devices and have a space to work on preparing content for ingest, we discussed a dedicated workstation, use of individuals' laptops, etc. Understandably, internal IT pushed back on these approaches due to cost and maintenance.
- Connectivity: The staging volume and/or bucket must be accessible to the repository's Ingest service.
- Lifecycle: We needed an environment that could be created when a project is active, and then destroyed post-project, to make it economically feasible.

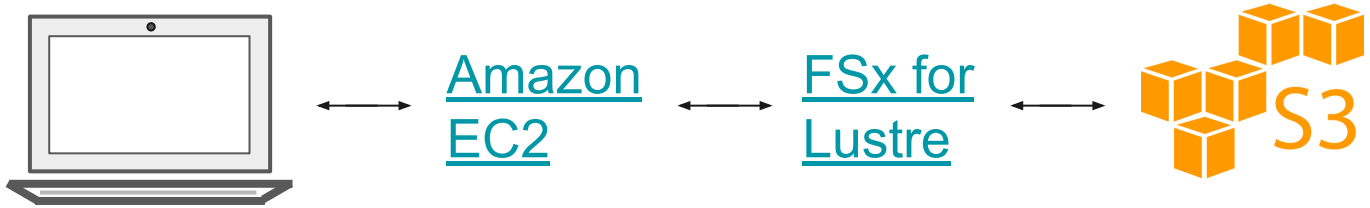


Solution – S3-based Workspace

- We created an “ingest workspace” which serves as a bridge between a linux shell (running on an EC2 instance) and an S3 bucket.
- Users are able to push content to S3, interact with files via the shell and incorporate metadata into batch submissions.
- On submission, content is pulled from S3, combined with metadata via manifest files and ingested into the repository.



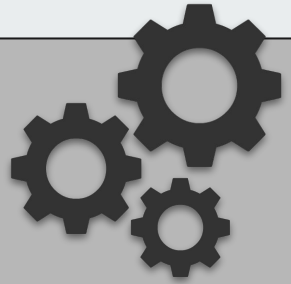
A: Authenticate via AWS Systems Manager in shell, then push content from HDD to S3 – or use (e.g.) VS Code to drag and drop.



B: Interact with S3 through connection provided by EC2 and Amazon FSx for Lustre File representation using POSIX semantics. S3 object store becomes transparent.

Underlying resources

- Configuration management with Sceptre
  - Drives CloudFormation tasks
  - Quickly instantiate a workspace instance
- AWS Lambda
  - Lists content of S3 Bucket
  - Generates pre-signed URL's to download S3 content
  - ALB provides web access to Lambda
  - Enables merging metadata with object manifests



Demo!

Using zsh, Excel & Visual Studio Code

Questions/comments? Have you built a similar solution?

Eric Lopatin
eric dot lopatin at ucop dot edu
Digital Preservation Services Manager

University of California Curation Center (UC3)
California Digital Library

UNIVERSITY OF CALIFORNIA

University of California CDL California Digital Library