DuraCloud Pilot Program: utilizing cloud infrastructure as part of your preservation strategy

Michele Kimpton
Project Director, DuraCloud
NDIPP Partner meeting
July 21, 2010
Overview

- What is DuraCloud
- Basic architecture
- Pilot program experiences
- Future direction
DuraSpace not for profit

DSpace
Fedora
DuraCloud

DURASPACE™
Implications for our future work

- more distributed
- More collaborative
- more web-oriented
- more open
- more interoperable
Cloud Infrastructure

A style of computing where massively scalable IT-related capabilities are provided “as a service” using Internet technologies to multiple external customers. (Gartner, 6/08).
DuraCloud Platform

Open technology and hosted service for utilizing cloud infrastructure for preservation support and access services

Architectural Features:
- Interoperable across multiple cloud providers
- Web enabled
- Built on highly scalable, flexible shared infrastructure
- Open API’s for easy integration
DuraCloud High Level Interaction

UI
Administration

API
Storage Management

Service Management

Services

Amazon S3
Rackspace CloudFiles
EMC Atmos
Other Clouds
DuraCloud Storage Architecture

DuraAdmi

UI

StoreClient

DuraStore

REST API

Storage Mediation

Storage Provider Interface

Amazon Storage Adapter

Rackspace Storage Adapter

EMC Storage Adapter

Other Storage Adapter

Amazon S3

Rackspace CloudFiles

EMC Atmos

Other Clouds

Updated: July 2010 | http://www.duracloud.org
Services and Capabilities

- Replication
- Image Transformation
- Bit Integrity Checking
- Image Viewing
- Media Streaming
- ...more on roadmap
- File format validation
- Parallel processing
Purpose of Pilot Program

• Engage with users whom have concrete use cases and want to test the software
• Real data at scale
• Uncover the obstacles
• Engage community response and assessment
• Learn and assess benefits and limitations of the cloud
Partners and Pilots

- Selected initial cloud providers

- Selected 3 initial pilot partners
NYPL DuraCloud Pilot Goals

**Preservation**
- Migration of service files from an unsupported format to a supported format – Mr. Sid to JP2000
- Data integrity checking of new format
- Ingest of new data streams associated with existing objects in Fedora Repository

**Access**
- Reduce number of service files
- Reduce number of services used in delivering service files
- Provide reliable and dependable service and access to those service files
### Key Advantages Cloud provides NYPL

<table>
<thead>
<tr>
<th>Most Impactful Advantages Electronic Survey</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalability</td>
<td>79</td>
</tr>
<tr>
<td>Remote, Off Campus Storage of Digital Assets</td>
<td>64</td>
</tr>
<tr>
<td>Ease of Implementation</td>
<td>54</td>
</tr>
<tr>
<td>Flexibility</td>
<td>53</td>
</tr>
<tr>
<td>Don't Have to Staff Locally</td>
<td>39</td>
</tr>
<tr>
<td>Cost</td>
<td>33</td>
</tr>
<tr>
<td>Elasticity</td>
<td>26</td>
</tr>
<tr>
<td>Pay for Use</td>
<td>14</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
</tbody>
</table>
NYPL Preservation services utilizing DuraCloud

NYPL DuraCloud Instance

Digital Repository Storage System

Local NYPL Fedora Repository

Tiff file

JP2000 service file checksum

Amazon

EMC

RackSpace
Outcomes

• Loaded 10 TBs of TIFFS to cloud
• Were able to view, convert, and serve a subset of those files via Adore Djatoka image service
• We were able to download and verify end products of conversion process with the DuraCloud API and JHOVE
• We were able to demo the process of chunking and storing multi-gigabyte media files with the DuraCloud sync tool
• Delays in hardware did not allow for provisioning additional services from DuraCloud
WGBH Preservation using DuraCloud

**Ingest**

- **Digital access management system (DAM)**
- **Open Vault Fedora Repository**

**Preservation support**:
- File validation
- Replication management
- Administrative access
- Error checking
- File migration/transformation
- Monitoring

**Cloud Providers**:
- Amazon
- EMC
- RackSpace
**WGBH Access using DuraCloud**

**Access services:**
- Streaming
- File format transformation
- File access collaboration

**Diagram:**
- Ingest
- Digital Access Management system
- Open Vault Fedora Repository
- Access file
- Access file streaming
- WGBH DuraCloud Instance
- Amazon
- EMC
- RackSpace

© 2010 WGBH
Sending disks to the cloud

**WGBH Delivery to DuraCloud via Hard Drive (.7 TB)**

(.7 TB transferred, 1TB drive, ext3 format)

- Collect Assets
- Create Amazon Manifest
- Chunk Assets
- Transfer to Shipping Drive
- Calculate Checksums
- Pack and Ship Drive
- Files available at DuraCloud

2 days | 1.75 days | up to 7 days
Using tubes and wires

WGBH Delivery to DuraCloud via Internet
(.7 TB transferred, 300Mbps pipe)

Collect Assets

Run DuraCloud Sync Tool

Chunk Assets

Calculate Checksums

Files available at DuraCloud

2 days

.5 days
Outcomes

- 5.5 TB of audio & video uploaded
  - Preservation and access files
- Still working with sync tool
- Streaming service works
  - But still need to integrate with Open Vault web site (for access)
Achievements during Initial Pilot

- DuraCloud integrated with 3 cloud storage providers
- Pilot partners loaded 30 TB into Duracloud
- Integrated and deployed multiple independent services
- Developed tools to overcome limitation of 5 GB file size and ease data loading
Lessons Learned

- Content transfer requires time and effort
- Internet Latency can be high
  - Minimize transactions across the wire
  - Data should be close to compute
  - Minimize data transfer with local host
- Storage more mature than compute
- Must allow for “eventual consistency”
- Market still developing
## Extended Pilot Partners

<table>
<thead>
<tr>
<th>University</th>
<th>Use Case</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rice U</td>
<td>Preservation</td>
<td>DSpace, meta archive</td>
</tr>
<tr>
<td>Hamilton College</td>
<td>Access/international collaboration</td>
<td>Fedora</td>
</tr>
<tr>
<td>Northwestern U</td>
<td>Preservation books, audio, image</td>
<td>Fedora</td>
</tr>
<tr>
<td>U of PEI</td>
<td>Image viewing/hosting</td>
<td>Fedora/Islandora</td>
</tr>
<tr>
<td>Cornell U</td>
<td>Data stream access and preservation</td>
<td>Fedora</td>
</tr>
<tr>
<td>ICPSR</td>
<td>Access and Preservation</td>
<td>Fedora</td>
</tr>
<tr>
<td>SUNY Buffalo</td>
<td>Preservation</td>
<td>DSpace</td>
</tr>
<tr>
<td>IUPUI</td>
<td>Preservation</td>
<td>DSpace</td>
</tr>
<tr>
<td>Rhodes College</td>
<td>Image Access</td>
<td>DSpace</td>
</tr>
<tr>
<td>North Carolina State U</td>
<td>Preservation</td>
<td>DSpace</td>
</tr>
<tr>
<td>CARL</td>
<td>Preservation and Services</td>
<td>Fedora</td>
</tr>
<tr>
<td>Orbis Cascade Alliance</td>
<td>Preservation and Services</td>
<td>DSpace</td>
</tr>
<tr>
<td>MIT</td>
<td>Preservation, OAIS compliance</td>
<td>Dspace</td>
</tr>
</tbody>
</table>
Key milestones

- **Release 0.1**: Cloud Storage Mediation 11/2009
- **Release 0.3**: Security & File Synchronization 05/2010
- **Release 0.5**: Open Source Release 07/2010
- **Public Beta Release**: 02/2011

- **10/2009**: Initial Pilot Program Begins
- **02/2010**: Release 0.2 Cloud-based Service Infrastructure
- **06/2010**: Release 0.4 Media Streaming
- **09/2010**: Expanded Pilot Program Begins
DuraCloud now available open source

- Open core
  - Open API
  - Open Source
  - Apache license
- Architecture to create cloud networks
  - Public clouds
  - Private clouds
  - University consortia
- Partner implementations/Integrations
Thank You

DuraSpace organization: www.duraspace.org
Wiki: wiki.duraspace.org/display/duracloud/
DuraCloud project page: duracloud.org
**DuraCloud demonstration:** [link available at duracloud site]
DuraCloud open source: wiki.duraspace.org/display/duracloud/DuraCloud