DuraCloud

Open technologies and services for managing durable data in the cloud

Michele Kimpton, CBO DuraSpace
“Repositories in the Cloud” Seminar, Feb 2010
mkimpton@duraspace.org
Open Source Portfolio

Fedora Commons™

mulgara™

DURA CLOUD
Implications for our future work:

- More distributed
- More collaborative
- More web-oriented
- More open
- More interoperable
Challenges (from survey 1/22/2010)

Preservation support is hard to implement consistently

“Our preservation support is collection based where we have had grants or specific initiatives. There is no system effort.”

“Where it is prioritized as mission critical, it is being done well. It is not being done well where it is not mission critical.”

“We have not invested enough to make it a service of which we are proud…”

“Collection development and storage are more important than computing. “
### Key Advantages
completed 1/22/2010
145 participants higher ed

<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>
## Key Challenges

completed 1/22/2010
145 participants higher ed

<table>
<thead>
<tr>
<th>Key Challenges</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trusting Third Party to Manage Critical Assets</td>
<td>64</td>
</tr>
<tr>
<td>Long-term Reliability of Solution</td>
<td>52</td>
</tr>
<tr>
<td>Data Security</td>
<td>51</td>
</tr>
<tr>
<td>Performance and Bandwidth Concerns</td>
<td>37</td>
</tr>
<tr>
<td>Loss of Control</td>
<td>34</td>
</tr>
<tr>
<td>Administrative Burden of SLAs</td>
<td>17</td>
</tr>
<tr>
<td>Transparency of Solution</td>
<td>16</td>
</tr>
<tr>
<td>Concerns about Data Lock-in</td>
<td>16</td>
</tr>
<tr>
<td>Less Customizable</td>
<td>10</td>
</tr>
<tr>
<td>Other</td>
<td>12</td>
</tr>
</tbody>
</table>
Likely to use cloud services in next 12 months

Percentage of electronic survey respondents noting it is “very likely” or “likely” they will use cloud compute or cloud storage services to manage, store or provide access to digital collections in the next twelve months.

<table>
<thead>
<tr>
<th>Category</th>
<th>Subcategory</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-US</td>
<td></td>
<td>47.7%</td>
</tr>
<tr>
<td>US Institutions</td>
<td>Institution Size</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Large, very large</td>
<td>47.2%</td>
</tr>
<tr>
<td></td>
<td>Medium</td>
<td>68.8%</td>
</tr>
<tr>
<td></td>
<td>Small, very small</td>
<td>42.9%</td>
</tr>
<tr>
<td>US Institutions</td>
<td>Enrollment Profile</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RU/VH</td>
<td>52.1%</td>
</tr>
<tr>
<td></td>
<td>RU/H, DRU</td>
<td>50.0%</td>
</tr>
<tr>
<td></td>
<td>Master's S, M and L</td>
<td>46.2%</td>
</tr>
<tr>
<td></td>
<td>Bac and Assoc</td>
<td>57.1%</td>
</tr>
<tr>
<td>Public/Private</td>
<td>Public</td>
<td>46.9%</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>59.3%</td>
</tr>
</tbody>
</table>
## Institutional needs: managing digital collections

<table>
<thead>
<tr>
<th>Service Area</th>
<th>Importance</th>
<th>Extent Need is Met</th>
<th>Difference</th>
<th>Likelihood to Use Cloud Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remote secondary storage of digital collections</td>
<td>3.54</td>
<td>2.60</td>
<td>0.94</td>
<td>3.09</td>
</tr>
<tr>
<td>Preservation support</td>
<td>3.35</td>
<td>2.17</td>
<td>1.18</td>
<td>2.88</td>
</tr>
<tr>
<td>Intra-institution shared collections</td>
<td>3.11</td>
<td>2.47</td>
<td>0.64</td>
<td>2.69</td>
</tr>
<tr>
<td>Inter-institution shared collections</td>
<td>2.72</td>
<td>2.07</td>
<td>0.65</td>
<td>2.67</td>
</tr>
<tr>
<td>Compute services</td>
<td>2.80</td>
<td>2.25</td>
<td>0.55</td>
<td>2.54</td>
</tr>
<tr>
<td>Online primary storage</td>
<td>3.51</td>
<td>2.97</td>
<td>0.53</td>
<td>2.29</td>
</tr>
</tbody>
</table>
Services in the cloud for durable digital content

DuraCloud Platform: Allow organizations to utilize cloud infrastructure easily offering data storage, data replication, preservation support and access services
Preservation Services

- ability to replicate content to multiple providers and locations
- ability to synchronize backup with primary store or repository system
- access to content through web based interface
- ability to do bit integrity checking
- ability to do file format transformations
Partners and Pilots

- Selected initial cloud providers
  - Amazon Web Services
  - Rackspace Cloud
  - EMC

- Selected 3 initial pilot partners
  - WGBH
  - Biodiversity Heritage Library
  - New York Public Library
NYPL pilot

Digital Gallery Collection

Use case: back up online preservation copy to Fedora, file format transformation

- back up copy all TIFF images (10 TB data)
- transformation from Tiff to JPEG 2000 using Imagemagick
- run J2k image server in cloud
- Push JPEG 2000 back into Fedora Repository
BHL pilot

BioDiversity Heritage Library

Use case: Find the best cost competitive solution for keeping multiple copies in multiple geographies, easily accessible.

- back up copy entire corpus (40 TB data- JPEG, Tiff)
- have multiple copies including Europe
- Run J2K image server in cloud
WGBH Media Library and Archives

Use case: Provide backup preservation for video files from repository and other sources, and create derivative files for access and streaming.

• Archive large video files
• Provide public access to streaming versions
• Transcode files in cloud
• Edit files where appropriate to sell clips
• Give third party access to cloud store for processing and access
Challenges

- Provisioning bandwidth at local institution to transfer data
- Transferring large files over the wire (over 5 GB is rejected, found issues in transfer over 1 GB)
- Consistency of operation of 2\textsuperscript{nd} tier providers (EMC, RackSpace)
- Enabling others to easily build on platform
- Best process for integration of 3\textsuperscript{rd} party applications into hosting service
- Cost effective bit integrity checking
- Balancing ease of use and more sophisticated functionality
Advantages of hosted platform

- Strategic partnerships with cloud providers
- Better pricing
- Transparency
- Early notification

- Ease of implementation for end user
- Multiple copies in multiple geographies/administrations through one interface
- Access to broad number of services relevant to the repository community
Timeline

- Begin pilots – September 2009
- DuraCloud Alpha Pilot release - Oct 2009
- Pilot data loading and testing – Fall 2009
- Beta for repository community – Q2 2010
- Pilot testing with software services Q2 2010
- Cloud partner evaluations complete - Q3 2010
- Hosting service pricing and SLA’s complete - Q3 2010
- Report pilot results – Q3 2010
- Code available open source - Q3 2010
- Launch production service Q4 2010
Next Steps (Feb-April)

- V.2 release complete
  - Replication, web access and viewing, file format conversion, J2K image server, bit integrity checking
  - Launch Fedora and DSpace plug ins
  - V.3 release primary features
    - Synchronization with local repository (Fedora and DSpace)
  - Expand pilot in April to include 15 new users, to connect with current repositories
  - Continue to test robustness and performance of commercial cloud partners
Thank You

For more information:

DuraSpace Organization: http://duraspace.org
Wiki:
http://www.fedora-commons.org/confluence/display/duracloudpilot/
DuraCloud project page: http://duracloud.org
Mkimpton@duraspace.org