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

Michele Kimpton
Project Director, DuraCloud
NDIPP Partner meeting
July 21, 2010
DuraSpace not for profit

DSpace
Fedora
DuraCloud
Overview

- What is DuraCloud
- Results of survey
- Pilot program
- Use cases
- Future direction
Future Direction, Disruptive Technologies

Interoperable
Web enabled
Distributed
Collaborative
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
Services and Capabilities

- Replication
- Image Viewing
- Image Transformation
- Media Streaming
- Bit Integrity Checking
- ...more on roadmap
- File format validation
- Parallel processing
DuraCloud Shared Services Vision

Services Registry

Verified Services
- Replication
- Bit Integrity Checking
- Image Viewing

Public Services
- Geo Tagging
- Twitter Publisher
- Data Mining
- Drupal
- Recollections

Private Services
- MyApp 1
- MyApp 2
- MyApp 3

Hosted services

Private services

Submit
Publish
# Key Advantages Cloud

completed 1/22/2010

145 participants higher ed

<table>
<thead>
<tr>
<th>Most Impactful Advantages</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electronic Survey</td>
<td></td>
</tr>
<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 Cloud

completed 1/22/2010
145 participants higher ed

<table>
<thead>
<tr>
<th>Key Challenges Electronic Survey</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</td>
<td></td>
<td>51.3%</td>
</tr>
<tr>
<td><strong>US Institutions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institution Size</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><strong>US Institutions</strong></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><strong>Public/Private</strong></td>
<td>Public</td>
<td>46.9%</td>
</tr>
<tr>
<td></td>
<td>Private</td>
<td>59.3%</td>
</tr>
</tbody>
</table>
### Preservation support

11. Please identify the elements of preservation support that are of most value to your institution.

<table>
<thead>
<tr>
<th>Element</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replication</td>
<td>59.5%</td>
<td>91</td>
</tr>
<tr>
<td>Multiple copies managed under multiple administrations</td>
<td>41.2%</td>
<td>63</td>
</tr>
<tr>
<td>Online access</td>
<td>75.2%</td>
<td>115</td>
</tr>
<tr>
<td>File migration</td>
<td>51.0%</td>
<td>78</td>
</tr>
<tr>
<td>File validation</td>
<td>50.3%</td>
<td>77</td>
</tr>
<tr>
<td>Bit integrity checking</td>
<td>47.1%</td>
<td>72</td>
</tr>
<tr>
<td>File repair</td>
<td>33.3%</td>
<td>51</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td></td>
<td>10</td>
</tr>
</tbody>
</table>

answered question 153

skipped question 71
Partners and Pilots

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

- Selected 3 initial pilot partners
  - WGBH
  - Biodiversity Heritage Library
  - The New York Public Library
# 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>
MIT DuraCloud use case, preservation support

- Retrieval of lost files (admin error)
- Replacement of damaged files

Preservation services:
- Monitoring
- File retrieval
- Access
- Auditing
- Replication
WGBH Access Services utilizing DuraCloud

- Digital Access Management system
- Open Vault
- Fedora Repository
- Access file streaming
- WGBH DuraCloud Instance

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

Ingest

Amazon
EMC
RackSpace
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 can be time intensive
• Internet Latency is high
  • Minimize transactions across the wire
  • Data should be close to compute
• Storage more mature than compute
• Market still developing
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-style license
- Architecture to create cloud networks
  - Public clouds
  - Private clouds
  - University consortia
- Partner implementations/Integrations
Thank You

For more information:
Come to our Pilot Partner panel at 2:45 pm today

DuraSpace organization: www.duraspace.org
Wiki: wiki.duraspace.org/display/duracloud/
DuraCloud project page: duracloud.org

**DuraCloud demonstration:** [https://demo.duracloud.org/durastore/movie-archive/duracloud-movie-large-take-5.m4v](https://demo.duracloud.org/durastore/movie-archive/duracloud-movie-large-take-5.m4v)

DuraCloud open source: [wiki.duraspace.org/display/duracloud/DuraCloud](https://wiki.duraspace.org/display/duracloud/DuraCloud)