A Proof of Concept
Cloud Based Solution

Mark Evans – Tessella Inc

PASIG – Austin, TX - January 13th 2012
Agenda

• Background to Tessella and Safety Deposit Box

• Primary drivers

• Our Journey to a proof of concept

• Considerations

• Next steps
Background into Tessella

- Software services company
- Formed in 1980
- Offices throughout Europe and North America
  - Abingdon (UK)
  - Boston
  - Washington DC
- 10 years of experience in digital preservation
- Dedicated global archiving division
Experience

• Development of tools - DRIOD / PRONOM / DIOSCURI


• Active in collaborative research in EU and North America

• Award winning technology platform
What is Safety Deposit Box (SDB)

• Flexible and extensible digital preservation platform
• Built in alignment with the OAIS reference model principles
  • Functional perspective - including preservation planning
  • Information model perspective
• Fourth generation
  • Experience / knowledge gained from research efforts
• Workflow driven
  • Modular approach (set of cohesive services)
  • Provides flexibility and extensibility
• Content, metadata, policy and storage neutral
Solution challenges

Digital preservation is a universal problem

“To solve the problem you need money”

Sadly very few institutions have large pots of money

Therefore the barrier to entry has to be lowered

How do we do this?
Options to Lowering the Barrier

• Build using open source
  • Requires expertise - “People”
  • This is beyond the reach of many

• Shared Infrastructure approach
  • Still requires a capital investment
  • Still requires someone to operate and manage the solution

• Consider a service / hosted solution
  • Many might have an operational budget to do this
More and more organizations are considering the cloud.
Many are already there.
Lots of activity within US government.
Can scale up and down elastically.
Many cloud offerings:
  - Public / Private / Hybrid
  - Storage, SaaS, PaaS, IaaS etc
Many cloud providers:

Rise in Popularity of the Cloud
Three Primary Questions

Can Tessella offer Preservation as a Service?
Can this lower the barrier?
Will it be attractive and adopted?
1) Multi-tenancy approach
   - Single deployment with multiple tenants
   - Each tenant has their own workflow, storage, policy, UI branding
   - Adopted in Austria and Finland (~150 tenants)

2) Cloud storage adapter
   - Content pushed into Amazon S3
   - Application stack resides locally
   - Potential high cost associated with fixity checking

3) Tessella hosted solution
   - NDIPP multi state pilot
Goals

- Move entire SDB4 stack into the cloud
  - Selected Amazon as the cloud provider
  - Amazon Web services (EC2 / S3 / RDS)

- Provide seamless integration with SIP creation tools
- Understand / develop a cost and pricing model
- If proven – launch a service offering
After investigation of compute options, settled on an East Coast “Large” Linux instance

7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform

Associated compute cost - $0.34c an hour ~ $3000 per year
Cloud Architecture

SDB4 Amazon Web Service Deployment

- SDB User
- AWS S3 Storage
- EC2 Instance
- RDS Instance
- Reverse Proxy
- Jobqueue tomcat port 9090
- SDB tomcat port 8080
- 22 (SSH)
- 80 (HTTP)
Creating and Submitting Content

To create and submit content, you can use the SIP Creator tool. Here's a screenshot of the tool with a focus on the uploading of an S3 Submission Package:

1. **Select Files**: Navigate to your files and select them for uploading.
2. **Input Details**: Fill in the required fields such as AWS S3 Account Key, AWS S3 Secret Key, and S3 Bucket Name.
3. **Manifestation Type**: Choose the appropriate Manifestation Type for your submission.
4. **Submit**: Click the Create New SIP button to complete the submission process.

The SIP Summary section displays the total collections, deliverable units, and files included in the submission.
Ingesting Content 1

Workflow Definition: Amazon S3 Ingest Workflow (Manual Selection)

Name: Amazon Ingest (Mark E)

Description: Ingests SIPS that have been deposited to Mark Evans

Email Address: evam@tessella.com

Send email notification on error (on)

Send email notification on completion (off)

Automatically terminate on unrecoverable error (off)

**Parameters marked with a * are mandatory:**

- **S3 Access Key:** AKIAIMFAZII327EACMIA
- **S3 Secret Key:** 
- **S3 Bucket Name:** PASIGTEST

Workflow Trigger:
- Manual only (on)
- Scheduled (off)
Ingesting Content 2

SDB Digital Archive: Ingest

Select S3 SIP for Ingest

<table>
<thead>
<tr>
<th>Deliverable Unit Titles</th>
<th>Collection</th>
<th>Files</th>
<th>Size</th>
<th>Date Created</th>
<th>Created By</th>
<th>S3 Bucket</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Files from PASIG</td>
<td>PASIG - Examples</td>
<td>5</td>
<td>194 KB</td>
<td>13.01.12 05:49:20</td>
<td>evam</td>
<td>com.tessella.evam</td>
</tr>
</tbody>
</table>

Exit  |  | Terminate |  |  |  | Continue |

Copyright © 2011 Tessella
### Ingesting Content 3

#### Workflow Details

<table>
<thead>
<tr>
<th>Workflow Context</th>
<th>Amazon Ingest (Mark Evans)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Workflow Definition</td>
<td>Amazon S3 Ingest Workflow (Manual Selection)</td>
</tr>
<tr>
<td>Workflow ID</td>
<td>147</td>
</tr>
<tr>
<td>Workflow State</td>
<td>Completed</td>
</tr>
<tr>
<td>Date Started</td>
<td>13.01.12 05:50:33</td>
</tr>
<tr>
<td>Date Finished</td>
<td>13.01.12 06:19:14</td>
</tr>
<tr>
<td>Number of Files</td>
<td>5</td>
</tr>
<tr>
<td>Total Size</td>
<td>202 KB</td>
</tr>
<tr>
<td>Collection Code</td>
<td>PASIG</td>
</tr>
<tr>
<td>Submission name</td>
<td>PASIG - Examples</td>
</tr>
<tr>
<td>Top Level Record</td>
<td>Test Files from PASIG</td>
</tr>
</tbody>
</table>

#### Step Progress

<table>
<thead>
<tr>
<th>State</th>
<th>Name</th>
<th>Progress</th>
<th>Started</th>
<th>Finished</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select</td>
<td>Select</td>
<td></td>
<td>13.01.12 05:50:33</td>
<td>13.01.12 06:18:33</td>
<td></td>
</tr>
<tr>
<td>Import from S3</td>
<td>Import from S3</td>
<td></td>
<td>13.01.12 06:18:33</td>
<td>13.01.12 06:18:35</td>
<td></td>
</tr>
<tr>
<td>Virus Check</td>
<td>Virus Check</td>
<td></td>
<td>13.01.12 06:18:35</td>
<td>13.01.12 06:18:38</td>
<td>View</td>
</tr>
<tr>
<td>Fixity Check</td>
<td>Fixity Check</td>
<td></td>
<td>13.01.12 06:18:38</td>
<td>13.01.12 06:18:41</td>
<td></td>
</tr>
<tr>
<td>Metadata Integrity</td>
<td>Metadata Integrity</td>
<td></td>
<td>13.01.12 06:18:41</td>
<td>13.01.12 06:18:44</td>
<td></td>
</tr>
<tr>
<td>Content Integrity</td>
<td>Content Integrity</td>
<td></td>
<td>13.01.12 06:18:44</td>
<td>13.01.12 06:18:47</td>
<td></td>
</tr>
<tr>
<td>Characterise</td>
<td>Characterise</td>
<td></td>
<td>13.01.12 06:18:47</td>
<td>13.01.12 06:18:56</td>
<td>View</td>
</tr>
<tr>
<td>Store Files</td>
<td>Store Files</td>
<td></td>
<td>13.01.12 06:18:56</td>
<td>13.01.12 06:18:59</td>
<td></td>
</tr>
<tr>
<td>Store Metadata</td>
<td>Store Metadata</td>
<td></td>
<td>13.01.12 06:18:59</td>
<td>13.01.12 06:19:02</td>
<td></td>
</tr>
<tr>
<td>Delete from S3</td>
<td>Delete from S3</td>
<td></td>
<td>13.01.12 06:19:02</td>
<td>13.01.12 06:19:05</td>
<td></td>
</tr>
<tr>
<td>Store Metadata File</td>
<td>Store Metadata File</td>
<td></td>
<td>13.01.12 06:19:05</td>
<td>13.01.12 06:19:08</td>
<td></td>
</tr>
<tr>
<td>Update Search Index</td>
<td>Update Search Index</td>
<td></td>
<td>13.01.12 06:19:08</td>
<td>13.01.12 06:19:11</td>
<td></td>
</tr>
<tr>
<td>Thumbnail Creation</td>
<td>Thumbnail Creation</td>
<td></td>
<td>13.01.12 06:19:11</td>
<td>13.01.12 06:19:14</td>
<td>View</td>
</tr>
</tbody>
</table>
Complete Web Services API can be exposed from AWS
Possible to create mobile device integration

Basic Android app
Browse SDB holdings
View Metadata
Download and view content
We have proved we can do it technically.....But

We still have many things to consider including:

- Choice of cloud provider(s)
- Cost and Pricing models
- Service Levels
- Security
- Privacy
- Locality
- Ingest bandwidth
- Additional functionality
Next Steps

- Investigate and understand the considerations
- Potential pilot with early adopter
- Aim for launch later in 2012
Thank you for your attention

Contact:
mark.evans@tessella.com
http://www.digital-preservation.com