Digital Preservation Solutions in Google Cloud

Designing Storage Architectures
Library of Congress
Law Library Archive Migration

- Migration completed between 2019 - 2020
- Data center decommissioning project
- Maintain NDSA level2 compliance
- File provenance records & fixity checks
Google Cloud Storage

**Usability:** A consistent API, latency, and speed across storage classes simplifies development and experience.

**Reliability:** Storage in geo-redundant locations offers 99.99% availability with transparent failover.

**Costs:** Cold storage as low as $0.0012 per GB/month. Still accessible in ms and with geo-redundant options.

**Security:** Data is encrypted at rest and in transit through various mechanisms; governed using Identity & Access Management.
Cloud Vision API

● Detect image characteristics, including dominant colors and crop hints
● Detect labels, landmarks and logos within images
● Read, extract, and translate text in documents, including images and handwritten text
● Perform content analysis in videos and extract metadata
$ md5sum archive/data/bp1892-0001.pdf
fcea3e2968c9a192e4d228668cc32256

$ bagit archive/
archive
  data
    bp1892-0001.pdf
    bp1892-0002.pdf
    bp1892-0003.pdf
    manifest-md5.txt
  bagit.txt

Traditional File System
<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Type</th>
<th>Created</th>
<th>Storage class</th>
<th>Last modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>bp1892-0001.pdf</td>
<td>0 B</td>
<td>application/pdf</td>
<td>Mar 10, 20...</td>
<td>Archive</td>
<td>Mar 10, 20...</td>
</tr>
<tr>
<td>bp1892-0002.pdf</td>
<td>0 B</td>
<td>application/pdf</td>
<td>Mar 10, 20...</td>
<td>Archive</td>
<td>Mar 10, 20...</td>
</tr>
<tr>
<td>bp1892-0003.pdf</td>
<td>8 B</td>
<td>application/pdf</td>
<td>Mar 10, 20...</td>
<td>Archive</td>
<td>Mar 10, 20...</td>
</tr>
</tbody>
</table>

gs://19th-cent-pamphlets/1892/data/bp1892-0003.pdf:
  Creation time:          Thu, 10 Mar 2022 21:26:07 GMT
  Update time:            Thu, 10 Mar 2022 21:26:07 GMT
  Storage class:          ARCHIVE

...  
Hash (crc32c):          okryTg== ✓
Hash (md5):              hRM8VqzPTuDfc1ljwL757A== ✓
Automated Approach

“Fixity, in the preservation sense, means the assurance that a digital file has remained unchanged, i.e. fixed.”
Best Practices for NDSA Level 2 Compliance

- **Geographic Redundancy**: Store files in multiple geographic locations across the United States.

- **Audit & Access Logs**: Record each user action, including object upload and retrieval with Access Logs.

- **Checksums**: Checksums are generated for each file and can be retrieved via API.

- **Object Versioning**: Supports retrieval of objects that have been deleted or overwritten.