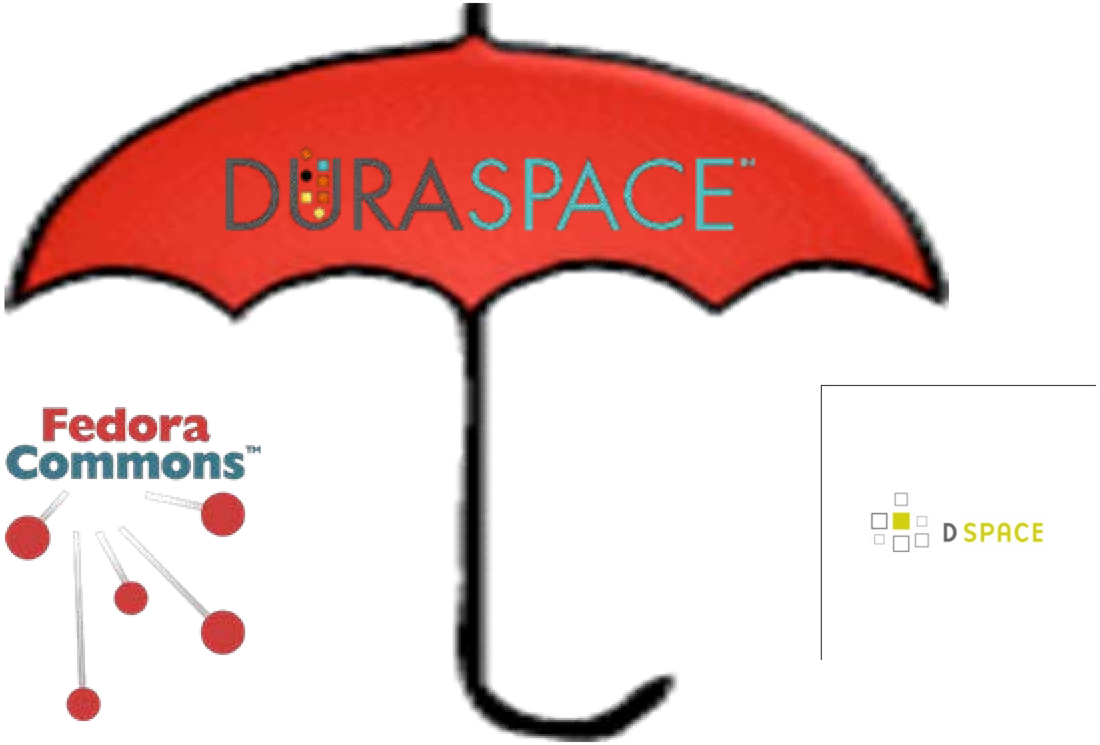
DuraCloud Pilot Program: Experiences

Bradley McLean
CTO, DuraSpace

Not for Profit Organization



DURACLOUD

DURACLOUD™

DuraCloud Platform

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

Services and Capabilities

Replication

Image Viewing

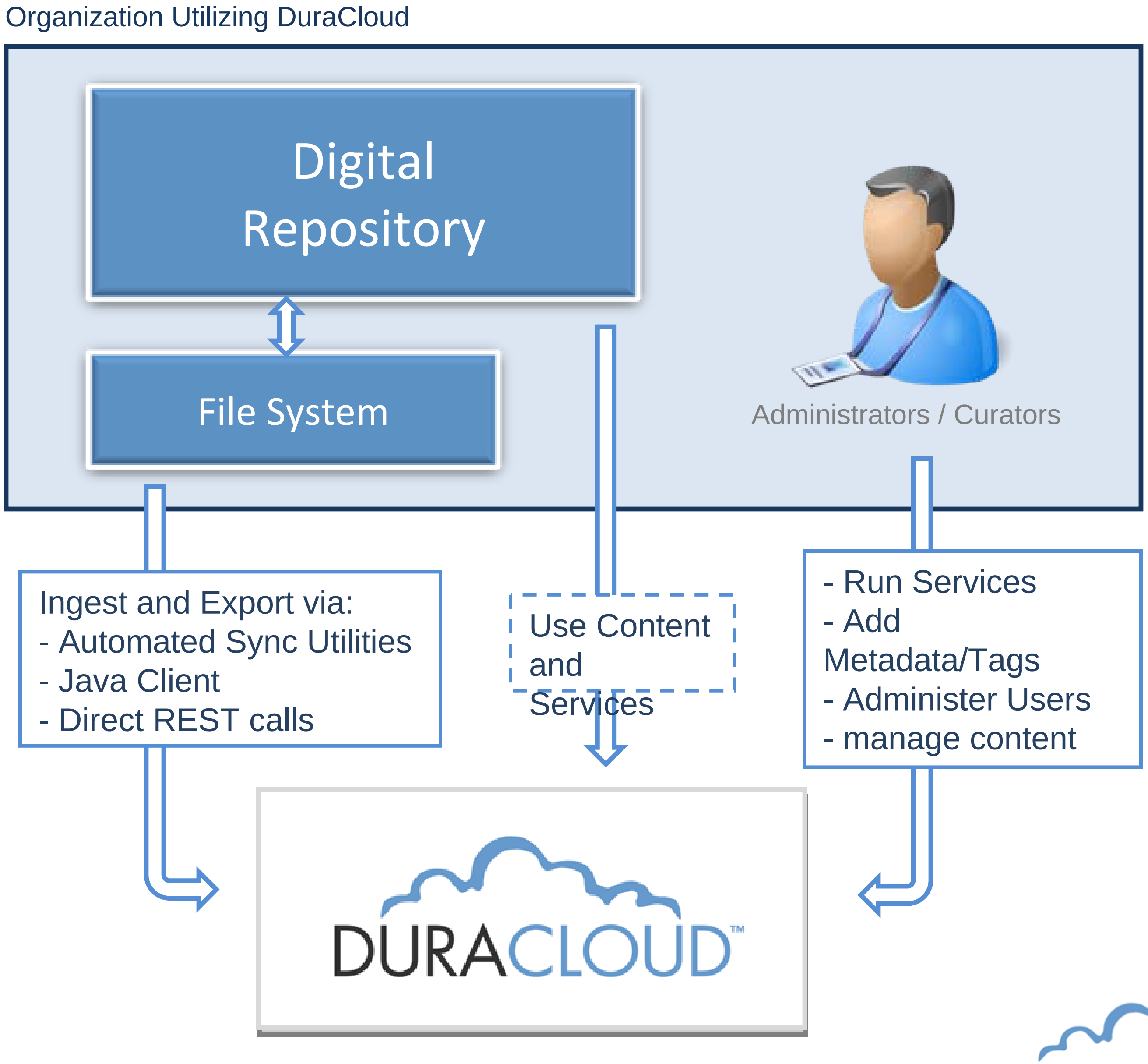
Image Transformation

Media Streaming

Bit Integrity Checking

General Compute Services

DuraCloud In Context



Organization Utilizing DuraCloud

Digital Repository

File System

Administrators / Curators

Ingest and Export via:
- Automated Sync Utilities
- Java Client
- Direct REST calls

Use Content and Services

- Run Services
- Add Metadata/Tags
- Administer Users
- manage content

Pilot Partners

University	Use Case	Repository
Rice U	Preservation	DSpace, meta archive
Hamilton College	Access/international collaboration	Fedora
Northwestern U	Preservation books, audio, image	Fedora
U of PEI	Integration	Fedora/Islandora
Cornell U	NSF data for Vet. School	Fedora
ICPSR	Access and Preservation	Fedora
SUNY Buffalo	Preservation	DSpace
IUPUI	Preservation	DSpace
Rhodes College	Image Access	DSpace
North Carolina State U	Preservation	DSpace
CARL	Preservation and Services	Fedora
Orbis Cascade Alliance	Preservation and Services	DSpace
MIT	Preservation	Dspace
NYPL	Preservation and Services	Fedora
WGBH	Access and Preservation	DAM

Timeline

- Begin pilots– September 2009
- DuraCloud Alpha Pilot release- Oct 2009
- Pilot data loading and testing – Fall 2009
- Expanded pilot for community – Q2 2010
- Pilot testing with software services Q2 2010
- Code available open source-Q3 2010

- Cloud partner evaluations complete-Q4 2010
- Report pilot results – Q4 2010
- Launch hosted service Q1 2011

Pilot Datasets

- ~ 10 TB each from 3 partners
  - NYPL: Tiff images to convert to JPG2K
    - Direct from NYPL to cloud.
  - BHL: 10-13 TB of varied types & sizes
    - Harvested from Internet Archive
  - WGBH: Video for distribution
    - Both via disk and via network

Lessons #1

- File Size Limits (E.G. 5GB), requires:
  - Chunking & Stitching
  - Compute support
- Per server bandwidth limits
  - 42 processes across 6 servers to complete in a few days.
  - 10x bandwidth allocation difference between small and medium servers

Lessons #2

- Large Files Challenging
  - Checksum error rates of several percent eventually reduced to 0.2%
  - Difficult to resolve with simple streaming APIs.
- Widely variable performance
  - “Brownouts” during transfers
- Naming matters with many files
  - Key distribution affects performance

Data Under Management

- ~ 30 TB during pilots
- ~ 20 TB today
- ~ 50 TB 1Q11 to ?? over 2011

Storage API Requirements

- Today, we work via basic webservices storag APIs
- We'd love to have:
  - Efficient periodic physical checksums
  - In place updates
  - Bucket to Bucket transfers.

DURA SPACE

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
BMclean@duraspace.org

DURACLOUD™