Defining Big Data

Volume
Variety
Velocity

(Douglas, 2001)
DL Curated Collections

Authors
Genre
Publications
Material Type
Original Material Organization

Funding
Technology
Unique requirements
Standards

Original Material Organization
Digital Libraries as Big Data

• The “massive digital library” (Crane, 2006)
  – Scale
  – Heterogeneity of content
  – Granularity of content
  – Noise
  – Audience
  – Collections and distributors
Massive Digital Libraries

Google books

HATHI TRUST
digital library
Moving to Big Data Analytics

- A collaborative effort of Indiana University and the University of Illinois Urbana Champaign, along with the HathiTrust
- Meets the technical challenges of dealing with massive amounts of digital text
- Creating rational research access to the digitized books of libraries.
Challenges in Big Data for the Humanities in the HTRC

• Data Storage and access
  – Text, bib data, external data
  – Granularity of data
  – Wide variety of sizes of volumes
    • 20 Vol set (10,058 pages / 3,516,895 words)
    • 50 Vol Set (26,398 pages / 24,020,622 words)
    • 100 vol set (34,279 pages / 12,570,832 words)

• Allocating computational resources
• Mining and DH methodology algorithm management
• Persistent results management
• Metadata inconsistencies
Direct programmatic access (by programs running on HTRC machines)

Discovery And Access

Portal Access
- Discovery

Services Management

Agent
- Interface API
- Registry API
- Job Submission

Registry
- Applications
- Collections
- Result Sets

Compute resources

Data Management

Data API access interface
- Audit
  - Cassandra cluster
  - volume store

Solr Proxy
- Solr index

Storage resources

Applications

Collections

Result Sets

Security (OAuth2)
NoSQL Methodology

• Allows us to manage flexible schemas
• Key-value based column store
• Offers good scalability, redundancy, and performance
• Use of Cassandra enabled HTRC to share content over a commodity based Cassandra cluster of virtual machines
Cassandra Schema

- Each row represents a volume
  - Row key is the volume ID
  - Each row contains many columns
  - First column contains metadata attributes about the volume
  - Each subsequent column family is a page, key is page ID
  - Page-specific columns contain page contents and metadata about the page
Data API

• An abstraction between the analysis application and the data
• Allows secured data access to trusted 3\textsuperscript{rd} parties
• Simplifies programming
• Access data by
  – Volumes
  – Pages
  – Metadata (METS and bib data)
  – SOLR indexes
Registry

• Organizes and manages components
  – Collections
  – Algorithms, analyses, and processes
  – Input parameters
  – Results

• Uses the WS02 governance registry open source software
Agent

• Controls session and access permissions
• Submits and manages jobs
  – Determines computational resources
  – Monitors job status
  – Communicates to portal
  – Updates registry
More on the HTRC

• Wednesday
  – HathiTrust Research Center: Computational Access for Digital Humanities and Beyond Poster

• Thursday
  – Introduction to the HathiTrust Research Center: A Demo and Hands-On Session
  – White River Ballroom C

• HTRC UnCamp
  – September 8-9, 2013
  – http://www.hathitrust.org/htrc_uncamp2013