Workshop Objectives

Create a shared vision of architecture for curating collections as data at scale

- Explain the architecture
- Show an example of it in use
- Solicit Feedback
- Discuss a path to implementation
About the Presenters

Will Thomas: Digital Curation Innovation Center (DCIC), University of Maryland

Ben Galewsky: National Center for Supercomputing Applications, University of Illinois
Architecture

Cluster Orchestration

Search and Discovery

Analytics Engine

Global Data Transformation Service

Scalable Repository

Cluster Orchestration
# Solution Components

<table>
<thead>
<tr>
<th>Role</th>
<th>Grant Supported Block</th>
<th>Open Source Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search and Discovery</td>
<td>Clowder, Project Blacklight</td>
<td>Elasticsearch, Hue</td>
</tr>
<tr>
<td>Analytics Engine</td>
<td>National Data Service Workbench</td>
<td>Spark, Jupyter, RStudio, etc.</td>
</tr>
<tr>
<td>Global Data Transformation Service</td>
<td>Brown Dog</td>
<td>ImageMagick, Stanford NLP, etc.</td>
</tr>
<tr>
<td>Scalable Repository</td>
<td>DRAS-TIC</td>
<td>Cassandra</td>
</tr>
<tr>
<td>Cluster Orchestration</td>
<td></td>
<td>Kubernetes</td>
</tr>
</tbody>
</table>
Massive Horizontal Scaling of Repository Data

- DRAS-TIC stack
  - Web interface
  - Standard storage APIs
  - Highly scalable Cassandra database cluster
  - Runs on cluster of commodity servers

- DCIC Performance Tests
  - 100 Million Files
  - Almost 100 Terabytes of data
Massive Horizontal Scaling of Repository Data
Massive Horizontal Scaling of Repository Data
Data Ingestion and Computation at Scale

● Potential Use Cases
  ○ Bulk data loading
  ○ Machine Learning
  ○ Complex searches

● What does computation at scale mean to you?
  ○ What are possibilities?
  ○ What are the stumbling blocks?

● Architecture
  ○ Apache Spark running on the Cassandra Clusters
  ○ National Data Service Workbench for notebooks and in-cluster development
Data Ingestion and Computation at Scale

Kubernetes Cluster

Worker Node1
Spark Executor
Cassandra Node
Local Storage
Worker Node2
Spark Executor
Cassandra Node
Local Storage
Spark Driver
R-Studio
Jupyter

Kubernetes Cluster
Active Curation of Large Collections

Global Data Transformation Service using DIBBs Brown Dog

- Curated collection of reusable data transformation tools
  - Low-loss file format conversion
  - Metadata extraction
  - Workflow orchestration
- Tools can be deployed locally
- Complex or proprietary tools hosted as a remote service
- Questions:
  - What types of metadata extraction would be useful to you?
  - What data formats requires conversion?
Active Curation of Large Collections

DATA WORKFLOW

Local Data Transformation Service

On Premise

Common Transformation Service

NCSA

Proprietary Transformations
Scalable Search, Discovery, and Social Curation

- **Blacklight**
  - Basic discovery interface for searching. Provides search box, facet constraints, stable document urls
  - http://projectblacklight.org

- **Clowder**
  - Social Curation interface to allow peers to organize and tag datasets
  - https://clowder.ncsa.illinois.edu

- **Hue**
  - Notebook based query tool for constructing, sharing, and executing complex Spark queries
  - http://gethue.com
Crowdsourcing at Scale: The George Meany Memorial Archives

- 20,000 linear feet of records
- Largest single UMD collection
- Communities of interest:
  - Labor History
  - American History
  - African American History
  - Women’s History
  - Political History
Crowdsourcing at Scale: The George Meany Memorial Archives

- Focus on the Civil Rights Department
- Documents union involvement in
  - *Brown v Board of Ed*
  - March on Washington
  - March from Selma to Montgomery
  - Voting Rights Act of 1965
  - Affirmative Action,
  - school desegregation (“busing”)
March 29, 1945
The Rev. Dr. Martin Luther King, Jr.
President, Southern Christian Leadership Conference
224 Auburn Avenue, N. E.
Atlanta, Georgia 30303

Dear Dr. King:

I have your telegram of March 30th concerning our participation in a boycott of materials going in and out of the state of Alabama.

Although my stand on the subject of civil rights as well known to you and the people of our beloved country, however, the legal involvements arising from our contractual relationships with the hundreds of carriers going in and out of Alabama, containing no-strike clauses, as well as the prohibitions contained in the fair-sharing and nondisclosure acts, which were directed mainly against us on secondary boycotts and picketing, and which have already resulted in law suits in excess of $1.00 million dollars having been filed against us, makes it absolutely necessary that I be guided by the opinion of our legal staff, which is at this very moment examining this proposition.

However, I would very much like to discuss with you your know how deeply I feel, not only about the death of Mrs. Viola Liuzzo, but also all those who have died before her in the cause of freedom and the right to organize for all citizens of our nation.

With all best wishes, I remain
Sincerely yours,

James R. Hoffa
General President

- Subset already digitized
- 100s of GBs of TIFF and JPEG
- Multiple types of documents:
  - Letters
  - Telegrams
  - Slips
  - Reports
  - Newsletters
  - Clippings
  - Pamphlets
  - Magazines
  - Booklets
  - Email (not bad for the 1970s)
Crowdsourcing at Scale: The George Meany Memorial Archives

- Crowdsourced digitization:
  - Can be captured ad hoc
  - May be at lower resolution
  - Likely will be skewed
  - May have blurring
  - May not be optimally lit
  - Won’t respect des fonds
Mr. William E. Pullard
Director
Department of Civil Rights
American Federation of Labor and
the Congress of Industrial
Organization
405 Eleventh Street, N.W.
Washington, D.C. 20006

Dear Bill:

I am herewith receipt of your letter of February 10, 1935 and the
attached copy of a letter from George B. Gable, North Georgia
Building and Construction Trades Council.

Our compliance review of construction contractors and sub-con-
nactors on the World War II rail building program, and the
examination of the construction labor unions, particularly the
CIO of construction workers, has led us to believe that additional
training programs are necessary for successful implementation of
the World War II rail building program. Accordingly, we have
authorized our staff to correspond with the appropriate unions
and organizations to develop plans for additional training
programs.

We are aware of the difficulties faced by the construction
industry in providing adequate training for its workforce and
recognize the importance of such training in ensuring the
success of the rail construction program. Therefore, we are
willing to provide support in any way possible to help ensure
the successful completion of the program.

Sincerely,

[Signature]

CC: North Georgia Building and Construction Trades Council

[Handwritten note:]

Mr. John Smith, President
North Georgia Building and Construction Trades Council

110 Main Street
Atlanta, Georgia 33303

Dear Mr. Smith:

We appreciate your letter of February 10, 1935 and the
attached copy of a letter from George B. Gable, North Georgia
Building and Construction Trades Council.

Our compliance review of construction contractors and sub-con-
nactors on the World War II rail building program, and the
examination of the construction labor unions, particularly the
CIO of construction workers, has led us to believe that additional
training programs are necessary for successful implementation of
the World War II rail building program. Accordingly, we have
authorized our staff to correspond with the appropriate unions
and organizations to develop plans for additional training
programs.

We are aware of the difficulties faced by the construction
industry in providing adequate training for its workforce and
recognize the importance of such training in ensuring the
success of the rail construction program. Therefore, we are
willing to provide support in any way possible to help ensure
the successful completion of the program.

Sincerely,

[Signature]

CC: North Georgia Building and Construction Trades Council

[Handwritten note:]
Crowdsourcing at Scale: The George Meany Memorial Archives

Dear Bill:


Dear Chairman, Wright:

For more than two years the North Georgia Building Trades Council has been ready to implement a minority recruitment and training effort designed to make certain that MARTA's construction would fully reflect the racial composition of metro Atlanta. This plan, with a governing board composed of contractors, unions, the affected political jurisdictions, the Urban League, and members of the minority community, has been endorsed by the Mayor of Atlanta and City Council, and Congressman Andrew Young.

- More than describing individual records
- Reasoning about connections between them
- Highlighting archival silences
Dear Bill:


Dear Chairman, Wright:

For more than two years the North Georgia Building Trades Council has been ready to implement a minority recruitment and training effort designed to make certain that MARTA's construction would fully reflect the racial composition of metro Atlanta. This plan, with a governing board composed of contractors, unions, the affected political jurisdictions, the Urban League, and members of the minority community, has been endorsed by the Mayor of Atlanta and City Council, and Congressman Andrew Young.

- More than describing individual records
- Reasoning about connections between them
- Highlighting archival silences
Dear Bill:


Dear Chairman, Wright:

For more than two years the North Georgia Building Trades Council has been ready to implement a minority recruitment and training effort designed to make certain that MARTA's construction would fully reflect the racial composition of metro Atlanta. This plan, with a governing board composed of contractors, unions, the affected political jurisdictions, the Urban League, and members of the minority community, has been endorsed by the Mayor of Atlanta and City Council, and Congressman Andrew Young.

- More than describing individual records
- Reasoning about connections between them
- Highlighting archival silences
Crowdsourcing at Scale: The George Meany Memorial Archives

- Ingesting images via RESTful interface
- Handle multiple simultaneous ingests
- Update researchers on progress
- Horizontally scalable image repository
- Deep learning at scale to:
  - Recognize text
  - Extract linked data
- Database to:
  - Ingest JSON
  - Infer relationships
  - Georeference/Match to authority records