Architecture to Enable Large-scale Computational Analysis of Millions of Volumes

Yiming Sun, Stacy Kowalczyk, Beth Plale, J. Stephen Downie, Loretta Auvil, Boris Capitanu, Kirk Hess, Zong Peng, Guangchen Ruan, Aaron Todd, Jiaan Zeng

**HTRC System Architecture**

- **Blacklight**
- **Portal**

<table>
<thead>
<tr>
<th>Algorithms and Worksets Registry (WSO2 Governance Registry)</th>
<th>HTRC Data API access interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security (OAuth2 WSO2 Identity Server)</td>
<td></td>
</tr>
<tr>
<td>Application Submission</td>
<td></td>
</tr>
</tbody>
</table>

**Audit**

- High level apps
  - Topic Modeling
  - Token Count
  - Log Likelihood
- Cassandra cluster volume store
- Solr index

**Compute resources**
- VM Manager
- Registry Service
- Job Manager
- Corpus Repository

**Storage resources**
- HTRC Job Submission Module
- HTRC Job Manager

**Prototype Sloan Cloud**

- Authentication Server
- VM Manager
- Job Submission Module
- Registry Service

**HTRC Portal & Blacklight**

**Algorithm: Topic Modeling**

- LDA-style topic analysis using Mallet
- First and last lines of each page removed
- End of line hyphenated words are joined
- Tokens containing non-alphanumeric characters removed
- Stop words filtered, with “not” replaced by “not_”
- 12 topics, 100 tokens per topic over Jane Austen collection

**Algorithm: Entity Extraction**

- Date entities extracted and displayed using SIMILE
- Location, person, time, date, money, organization, percentage can be extracted and displayed in tabular form
- Entities extracted from Jane Austen collection

**Algorithm: Dunning Log Likelihood**

- Words more frequent in Charles Dickens collection (analysis) than in Jane Austen collection (reference)

- 2.7 million volumes in corpus
- 3 stacks: production, development, sandbox
- HTRC-provided non-consumptive algorithms

- Support for large-scale non-consumptive algorithms
- Allow researchers to submit algorithms and worksets to run as Map/Reduce jobs