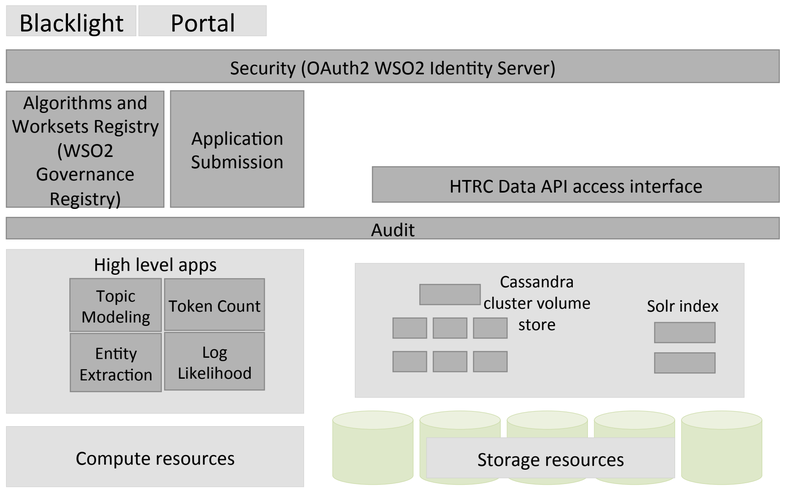
Architecture to Enable Large-scale Computational Analysis of Millions of Volumes

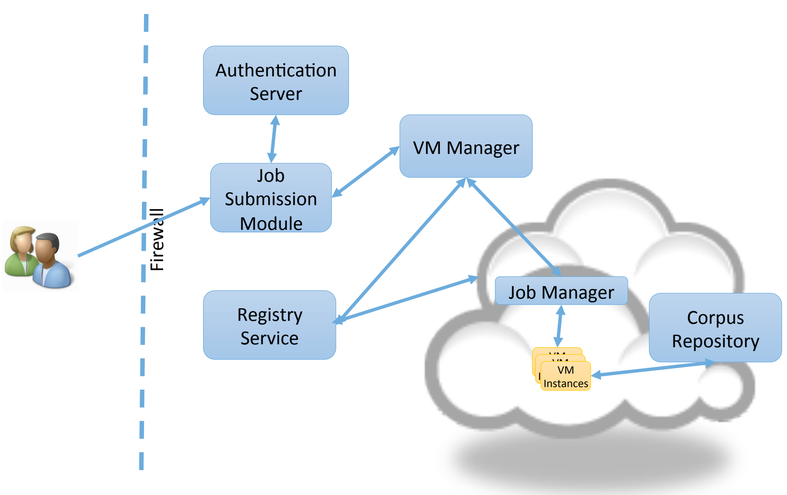
Yiming Sun, Stacy Kowalczyk, Beth Plale, J. Stephen Downie, Loretta Auvil, Boris Capitanu, Kirk Hess, Zong Peng, Guangchen Ruan, Aaron Todd, Jiaan Zeng

HTRC System Architecture



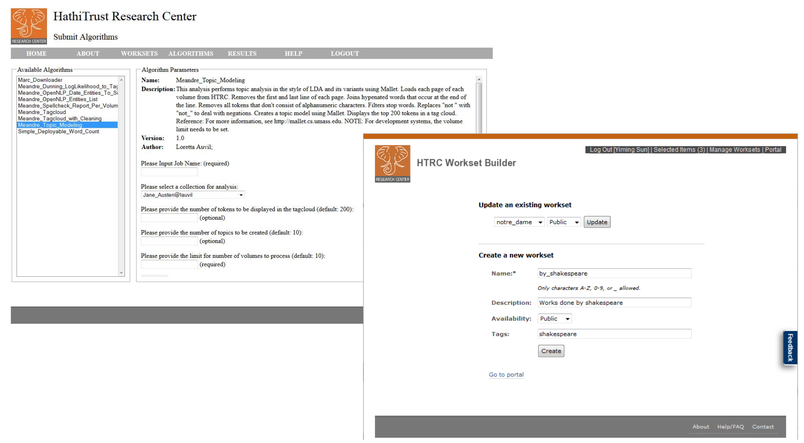
- 2.7 million volumes in corpus
- 3 stacks: production, development, sandbox
- HTRC-provided non-consumptive algorithms

Prototype Sloan Cloud



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

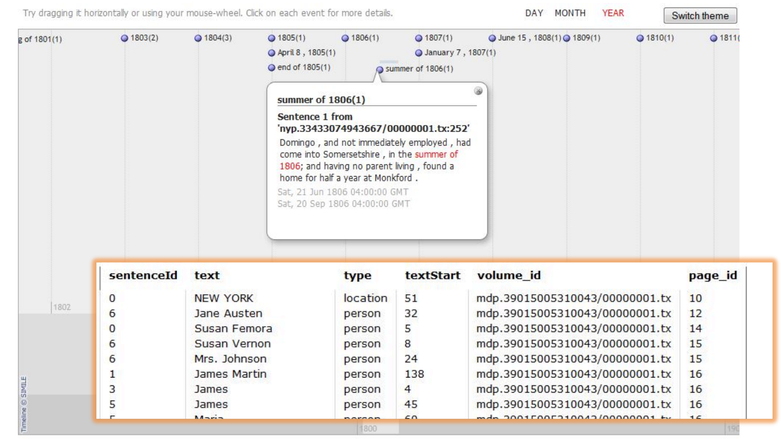
HTRC Portal & Blacklight



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



sentenceId	text	type	textStart	volume_id	page_id
0	NEW YORK	location	51	mdp.39015005310043/00000001.tx	10
6	Jane Austen	person	32	mdp.39015005310043/00000001.tx	12
0	Susan Femora	person	5	mdp.39015005310043/00000001.tx	14
6	Susan Vernon	person	8	mdp.39015005310043/00000001.tx	15
6	Mrs. Johnson	person	24	mdp.39015005310043/00000001.tx	15
1	James Martin	person	138	mdp.39015005310043/00000001.tx	16
3	James	person	4	mdp.39015005310043/00000001.tx	16
5	James	person	45	mdp.39015005310043/00000001.tx	16

- 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)