Social News Search

SFBay Lucene/SOLR September Meetup

Sammy Yu
Digg, Inc
September 3, 2009

Agenda

- Overview
- Architecture
  - Performance
  - Relevance
- Demo
- Future
- Conclusion

digg

Overview

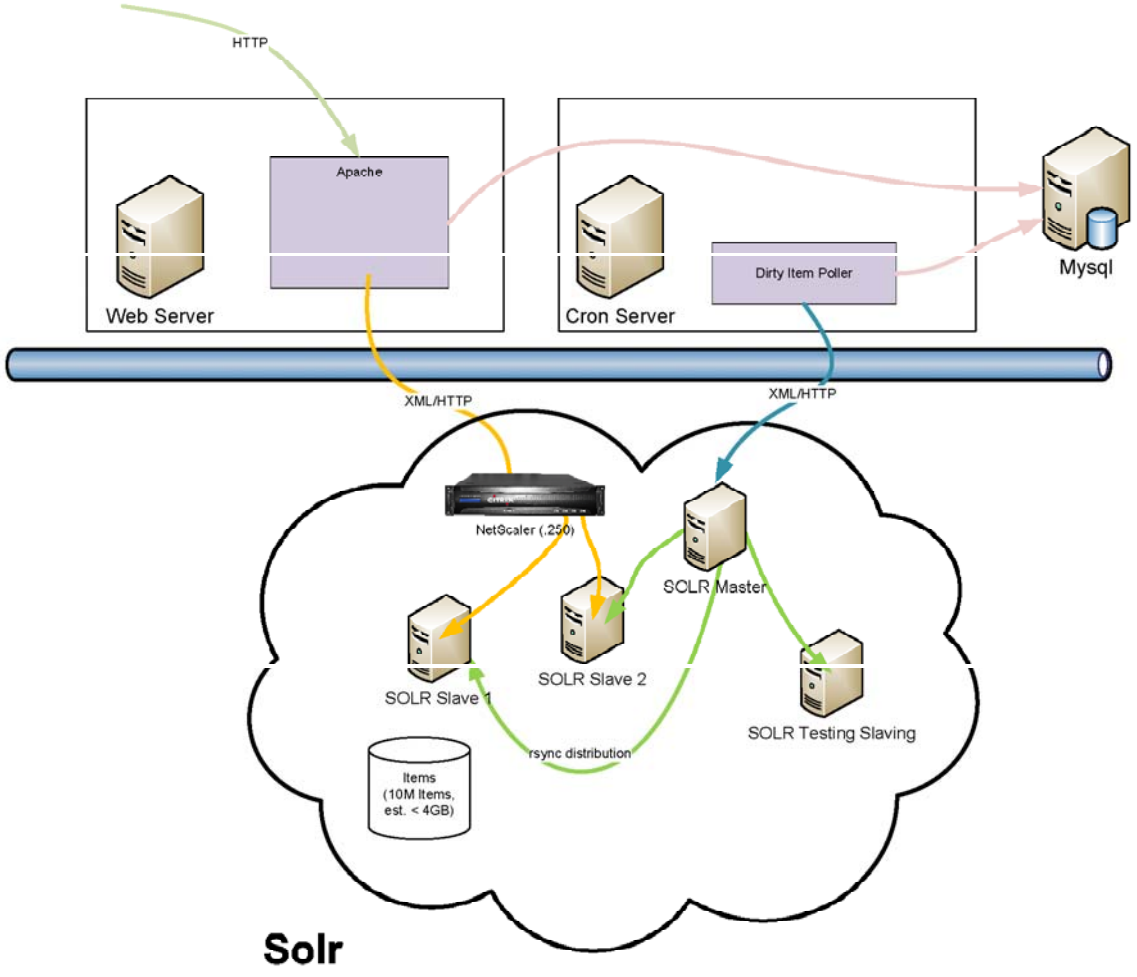
- Digg.com is a social news website
  - Community comes to discover and share news.
  - Stories can come from traditional news site such as nytimes.com as well as individual blog sites
  - Quality of stories are judged by diggs and buries (and others)

Architecture

- Lucene/SOLR
  - Highly-scalable, performant, customizable IR
- Search as a platform
  - Related By Keyword
  - Related By Source
  - Related By Search
  - Dupe Check
  - API

digg

Architecture



Architecture

- SOLR 1.3/Jetty 6.1
- 1 Master/10 Slaves
- 4.8 million queries/day
- 13 million documents/8 GB on disk
- 26 fields/documents

digg

Performance

- Reduce Index Sizes
  - Go through every field and determine if we really need to tokenize, index, and store it.
  - Heterogeneous Documents
- Use SOLRJ for indexing
- Date Fields (1.3 unique cardinality issue)
  - Coarser Granularity (specialized Year, Quarter, Month, Day fields)

digg

Relevance

- Understanding the domain
- Social news
  - "Freshness" is a critical factor
  - Heavily weigh stories that have showed up on the first page
  - Functional query against collaborative filter attributes such as digg counts

digg

Relevance

- Recognize different search types:
  - specific story
  - category
  - source
- Facets
  - Allows user to refine their search
  - Enable discover of new content

digg

Demo

- http://digg.com/search
- http://digg.com/search?s=iphone
- http://digg.com/search?s=nytimes.com
- http://digg.com/apple/Mac_OS_X_10_6_Snow_Leopard_the_Ars_Technica_review

digg

Future

- Dynamic Facet via Carrot2
- Performance Improvements
  - Facet improvements
  - Trie-field Support
- Document Duplication Detection
- Distributed Search
  - Partition By time
  - Partition By field values

digg

Conclusion

- Thank You
- Comments, Questions?

digg