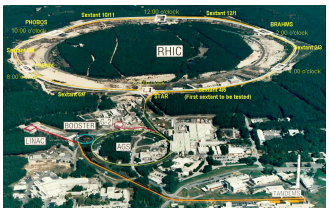
Improving PHENIX search with Apache Solr, nutch and Drupal

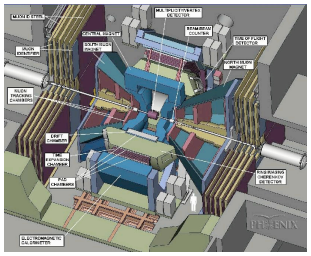
I.V. Sourikova, D.P. Morrison for the PHENIX collaboration

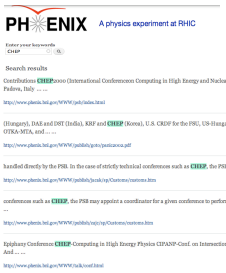
BROOKHAVEN NATIONAL LABORATORY

The Phenix experiment at RHIC



- 13 countries, 70 institutions
- 20 years of R&D, construction and operation
- Terabyte of proprietary documents hosted on many servers around the world
- Many web areas closed for commercial search engines like Google
- Legacy search infrastructure did not scale well with the fast growing document base, produced results inadequate in both precision and recall
- Good search critical for fighting institutional memory loss

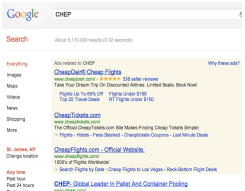




PHENIX results for 'CHEP'

Why custom search? Why not Google? Why not Google Search Appliance?

- We know what we are searching for and Google does not.
- PHENIX web is a collection of intranets with many areas closed for commercial crawlers
- Google Search Appliance (GSA): costly to support several web servers, no built-in support for using internal knowledge about data
- Apache Nutch and Solr provide a good mix of functionality and flexibility, support sub domain searching



Google results for 'CHEP'

Crawling with nutch

- Multi-protocol, multi-threaded, distributed crawler
- Full-text indexer
- Collections typically 1 – 200 million documents
- Crawling from seed urls is a perfect fit for crawling a few web servers
- Highly flexible, easily extensible
- Robust crawling frontier controls
- Uses Tika as a parser
- Uses Solr as a search front-end
- Large user base, good support

Searching with Apache Solr

- Standalone full-text search server
- Input via XML, JSON or binary over HTTP
- Query via HTTP GET and receive XML, JSON, or binary results
- Rich document handling
- Faceted search out of the box
- Hit highlighting, query spelling suggestions

Web Interface with Drupal

- Open source CMS
- Functionality via thousands of add-on modules
- Very popular, used by such sites as
  - The White House, The Economist, Examiner.com
- Provides Apache Solr Search Integration and Nutch modules for integration with Solr and Nutch

nutch Apache Solr Drupal putting it all together

- Nutch 1.4.dev Solr 3.4.0 Drupal 7.12
- All on a single server
- Solr, Nutch and Drupal share an expandable data schema defined in schema.xml and present in all three config areas
- Solr and Nutch resolve schema name differences via solrindex-mapping.xml in the Nutch config area
- Out of the box configuration did not work – Nutch 1.4 schema was lacking Drupal version 7 modifications
- Custom Drupal module and help from developers (http://drupal.org/node/708886#comment-5083502) solved schema incompatibility problem

Making logical collections with Nutch

Subcollections plug-in:
- limit search to tagged collections of urls
- activated in nutch-site.xml, defined in subcollections.xml:

```
<subcollection>
  <name>Logbook</name>
  <id>Logbook</id>
  <whitelist>
    logbook.phenix.bnl.gov:7815/
  </whitelist>
</subcollection>
```

- improves search precision

Automatic discovery of static pages

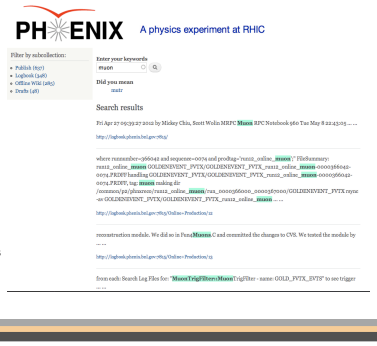
AdaptiveFetchSchedule crawling:
- recrawl schedule depends on the frequency of page update

Challenges of crawling legacy data

- Thousands of old private directories with broken links and file permission problems are the source of http errors
- Nutch effectively stops on urls with certain types of errors by going to an infinite loop of retries (Jira issue NUTCH-1245, a critical bug)
- A few problem urls can be filtered out by regex-urlfilter but filtering too many of them slows down the crawler significantly.

Nutch Issues: 30 Day Summary. 35 created and 22 resolved

Waiting for the bug fix to complete the crawl

Summary

- Nutch, Solr and Drupal provide a complete set of tools to build a federated search for heterogeneous collection of 10 million documents
- Nutch crawls multiple PHENIX domains in parallel
- Internal data knowledge used for high-precision search of logical document collections
- Solr aggregation of multiple index sources supports easy incorporation of custom parsers
- Adaptive Fetch Schedule helps to avoid re-crawling of static pages
- Configuration is time consuming but provides high levels of customization in return
- Maintenance remains an issue since all three components are in active development and there are many security updates that cannot be ignored