Nutch Issues: 30 Day Summary.

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

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 setup and support.
- Google Search Appliance does not support sub domain searching.
- Nutch, Solr and Drupal provide a good mix of functionality and flexibility, support sub domain searching.

Apache Nutch and Solr:

- Good search critical for fighting institutional memory loss.
- Legacy search infrastructure did not scale well with the fast growing Terabyte of proprietary documents hosted on many servers around the world.
- Many web areas closed for commercial search engines like Google.
- We know what we are searching for and Google does not.
- Multi-protocol, multi-threaded, distributed crawler.
- Full-text indexer.
- Collections typically 1 - 200 million documents.
- Crawling from seed urls in 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.

PHENIX results for 'CHEP' and 'I.V. Sourikova, D.P. Morrison for the PHENIX collaboration.'