Nutch search engine overview

- Some of the open source search engines are:
  - Nutch
  - Sphinx
  - Lucene
  - Namazu
  - Wikia
- In this presentation, we will discuss details about Nutch and its architecture.

- Nutch is an open source search engine implemented in Java
- Nutch implements “Map Reduce” distributed processing model.
- Nutch installations typically operate at one of three scales: *local filesystem*, *intranet*, or *whole web*
- Nutch is built on top of Lucene, which is an API for text indexing and searching.

Difference between Nutch and Lucene

- Use Lucene if a web crawler is not needed.
- A common scenario is that you have a web front end to a database that you want to make searchable. The best way to do this is to index the data directly from the database using the Lucene API, and then write code to do searches against the index, again using Lucene.
- Nutch is a better fit for sites where you don't have direct access to the underlying data, or it comes from disparate sources.

Nutch architecture

- Nutch is divided into two pieces: the crawler and the searcher.
- The crawler fetches pages and turns them into an inverted index.
- This inverted index is used by the searcher to resolve user's queries.
- Searcher and crawler components can be scaled independently of each other.

- The crawler system is driven by:
  - Nutch crawl tool.
  - Several types of *data structures*, including the *web database*, a set of *segments*, and the *index*.
- The *web database*, or *WebDB* stores two types of entities: *pages* and *links*.
- A page represents a page on the Web, and is indexed by its URL and the MD5 hash of its contents.
- A *link* represents a link from one web page (the source) to another (the target).

- A segment is a collection of pages fetched and indexed by the crawler in a single run.
- The fetchlist for a segment is a list of URLs for the crawler to fetch, and is generated from the WebDB.
- The fetcher output is the data retrieved from the pages in the fetchlist. The fetcher output for the segment is indexed and the index is stored in the segment.
- The index is the inverted index of all of the pages the system has retrieved, and is created by merging all of the individual segment indexes. Nutch uses Lucene for its indexing

Steps performed by the crawler…

1. Create a new WebDB.
2. Inject root URLs into the WebDB.
3. Generate a fetchlist from the WebDB in a new segment.
4. Fetch content from URLs in the fetchlist.
5. Update the WebDB with links from fetched pages.
6. Repeat steps 3-5 until the required depth is reached.
7. Update segments with scores and links from the WebDB.
8. Index the fetched pages.
9. Eliminate duplicate content (and duplicate URLs) from the indexes.
10. Merge the indexes into a single index for searching.