



財團法人國家實驗研究院

國家高速網路與計算中心

NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING

# Nutch 簡介

王耀聰 陳威宇

Jazz@nchc.org.tw

waue@nchc.org.tw

國家高速網路與計算中心 (NCHC)

# What's Nutch

- Nutch 是一個 open source ，以 Java 來實做的搜索引擎，它提供了架設自己的搜索引擎所需的全部工具。
- 利用 Lucene 為函式庫
- **架構於 Hadoop 之上**

# Nutch's goals

- 每個月抓取幾十億網頁
- 為這些網頁維護索引
- 對索引文件進行每秒上千次的搜索
- 提供高質量的搜索結果
- 以最小的成本運作

# Why Nutch ?

- 透明
  - Opensource ， 資訊不隱藏
- 擴充
  - 有各種函式庫應用於分析不同檔案
- 隱私
  - 可應用於搜尋專屬資料
- 客製化
  - 可以之為基礎設計自己的 data mining 工具

# Who use Nutch

## Public search engines using Nutch

Please sort by name alphabetically

- [AskAboutOil](#) is a vertical search portal for the petroleum industry.
- [Baynote](#) provides free hosted Nutch search for businesses.
- [BeThere BeSquare](#) is an Event Search Engine for the San Francisco category and get details about events in 4 different views.
- [Bigsearch.ca](#) uses nutch open source software to deliver its search results.
- [BusyTonight](#): Search for any event in the United States, by key from original source Web sites.
- [Central Budapest Search](#) is a search engine for English language events.
- [Circuit Scout](#) is a search engine for electrical circuits.
- [Comtec Search](#) is a search engine for UK Tour Operator Pack.
- [Coder-Suche.de](#) searches for coding stuff like apis, documents in english.
- [Cornell University Library](#) is collaborating with the research group pages based on Nutch. The nutch-based search engine is near top level.
- [Creative Commons](#) is a search engine for creative commons images.
- [Dadi360](#) Use nutch search engine for providing search of Chinese.
- [Ecolhub Web Search](#) an E. coli specific search engine based on Nutch thereby reducing the number of spurious hits. Searches can be on more resources getting added.
- [Epivista](#) is a search engine of epilepsy related web sites.
- [eroscanner](#) is a search engine for german adult stuff. (Watching NSFW)

.....more

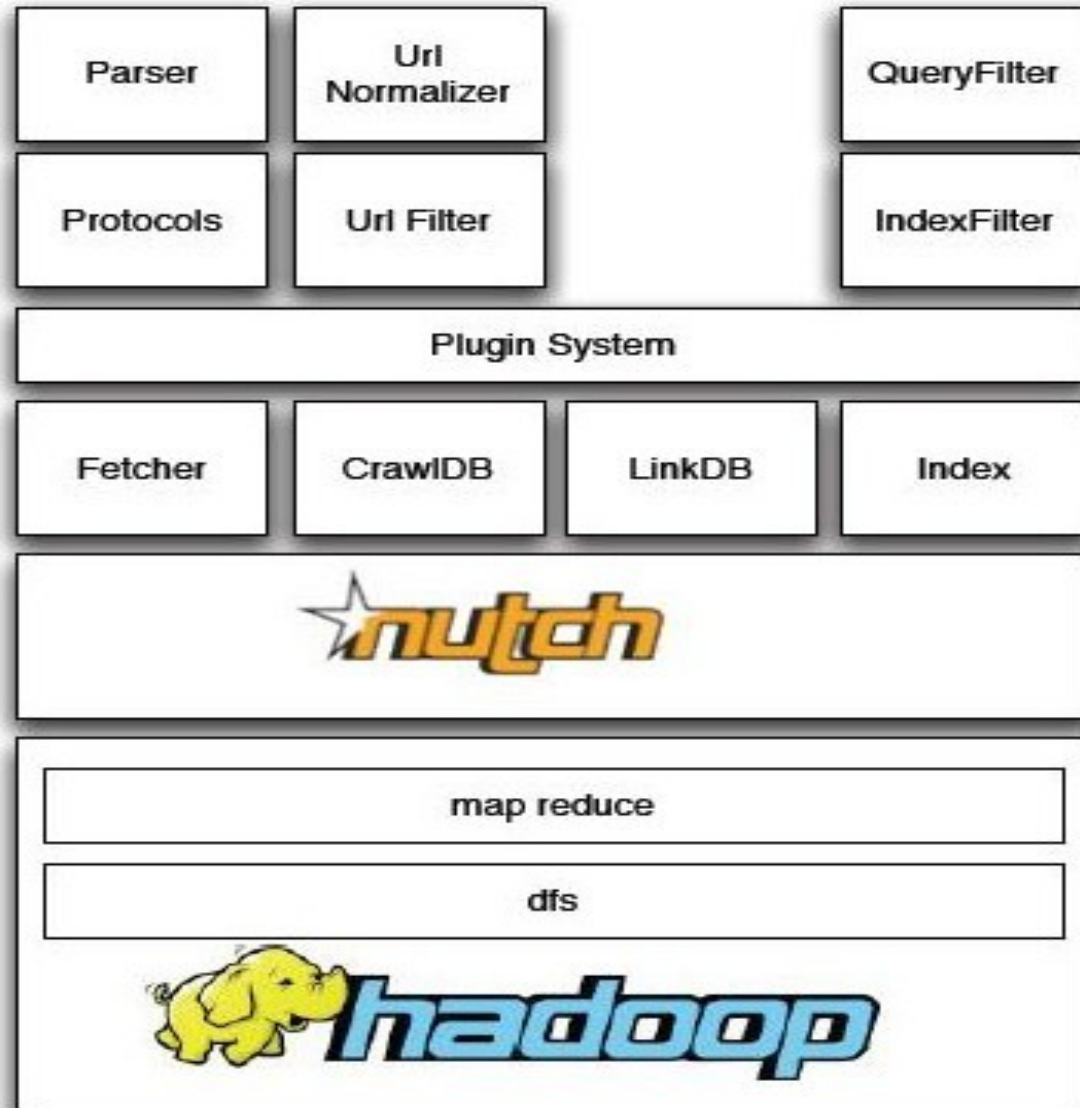
(<http://wiki.apache.org/nutch/PublicServers>)

The screenshot shows the Krugle search engine interface. At the top, there are navigation tabs for "Open Source Code", "Open Source Projects", and "SCM Comments". Below these is a search bar with the text "nutch" entered. To the right of the search bar are options for "Language" (set to "All") and "Found in" (set to "Any area"). There is also a "Project" dropdown menu with the placeholder text "Enter project name".

The search results section is titled "Code Search for nutch" and indicates "Code Files 1-10 (out of about 1849 matching files)". The results are listed as follows:

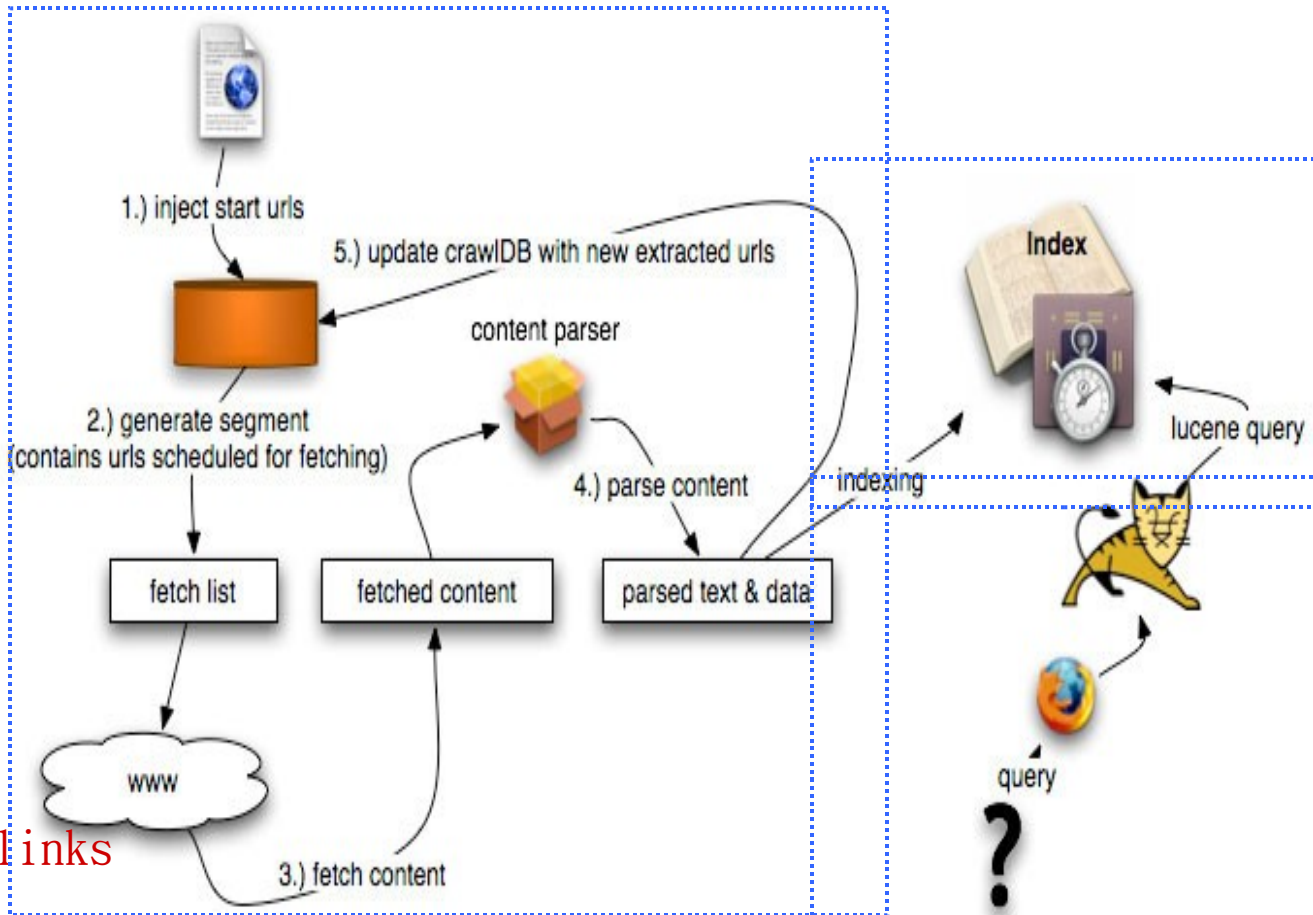
- Nutch.java** | Creative Commons Tools | Apache-2.0  
26 \* @author J&eacute;acutecirc;me Charron  
27 \*/  
28 public interface **Nutch** {  
30 public static final String ORIGINAL\_CHAR\_ENCODING =  
31 "OriginalCharEncoding";
- Nutch.java** | Nutch | Apache-2.0  
26 \* @author J&eacute;acutecirc;me Charron  
27 \*/  
28 public interface **Nutch** {  
30 public static final String ORIGINAL\_CHAR\_ENCODING =  
31 "OriginalCharEncoding";
- NutchConfiguration.java** | Nutch | Apache-2.0  
31 /\*\* Utility to create Hadoop (@link Configuration)s that include **Nutch**-specific  
32 \* resources. \*/  
33 public class **NutchConfiguration** {  
35 private final static String KEY = **NutchConfiguration**.class.getName();  
37 private **NutchConfiguration**() {} // singleton
- NutchJob.java** | Nutch | Apache-2.0 | Show Clones  
21 import org.apache.hadoop.mapred.JobConf;  
23 /\*\* A (@link JobConf) for **Nutch** jobs. \*/  
24 public class **NutchJob** extends JobConf {  
26 public **NutchJob**(Configuration conf) {  
27 super(conf, **NutchJob**.class);

# 架構



# 運作流程

- 1) 建立初始 URL 集
- 2) 將 URL 集注入  
crawldb---inject
- 3) 根據 crawldb 建立抓取清單 ---generate
- 4) 執行抓取，獲取網頁內容 ---fetch
- 5) 用獲取到的頁面資訊更新 crawldb---updatedb
- 6) 重複進行 3 ~ 5 的步驟，直到預先設定的抓取深度
- 7) 更新 linkdb ---invertlinks
- 8) 建立索引 ---index
- 9) 用戶通過用戶接口進行查詢操作
- 10) 將用戶查詢轉化為 lucene 查詢
- 11) 返回結果



# Plugin

- 修改 `conf/nutch-site.xml` 的 `plugin.includes` 屬性
- 在 `nutch` 基本功能之上擴充其功能
  - “`parse-xx`”：加入解析 `xx` 檔案類型的能力
  - “`protocol -xx`”：加入在此協定內的檔案也處理

**parse-text**

**parse-ext**

**parse-html**

**parse-js**

**parse-mp3**

**parse-zip**

**parse-rtf**

**parse-msword**

**parse-msexcel**

**parse-pdf**

**parse-rss**

**parse-oo**

**parse-swf**

**parse-mspowerpoint**

**protocol-file**

**protocol-ftp**

**protocol-http**

**protocol-httpclient**



# International

- 已有多國語言版可選，但若還要客製化…
- the page header
  - `src/web/include/language/header.xml`
- the "about" page
  - `src/web/pages/lang/about.xml`
- the "search" page
  - `src/web/pages/lang/search.xml`
- the "help" page
  - `src/web/pages/lang/help.xml`
- text for search results
  - `src/web/locale/org/nutch/jsp/search_lang.properties`

# No ! Nutch

- 告訴網頁機器人是否允許進入爬網
- 將 robots.txt 放在 web 上
- robots.txt

```
User-agent: Nutch  
Disallow: /
```

# Home Page



[About](#)

[FAQ](#)



[help](#)



[ca](#) | [de](#) | [en](#) | [es](#) | [fi](#) | [fr](#) | [hu](#) | [it](#) | [jp](#) | [ms](#) | [nl](#) | [pl](#) | [pt](#) | [sh](#) | [sr](#) | [sv](#) | [th](#) | [zh](#)

# Start

- **23 March 2009 - Apache Nutch 1.0 Released**

# Let's Go

# References..

- Nutch Website
  - <http://lucene.apache.org/nutch/>
- Nutch wiki
  - <http://wiki.apache.org/nutch/>
- Nutch API
  - <http://lucene.apache.org/nutch/apidocs-1.0/index>