



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

# Hadoop應用

## - Nutch 簡介

王耀聰 陳威宇 楊順發

jazz@nchc.org.tw

waue@nchc.org.tw

shunfa@nchc.org.tw

國家高速網路與計算中心(NCHC)



自由軟體實驗室

# Outline

- What is Nutch
- Why Nutch
- Nutch's Details
- Let's go

# 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

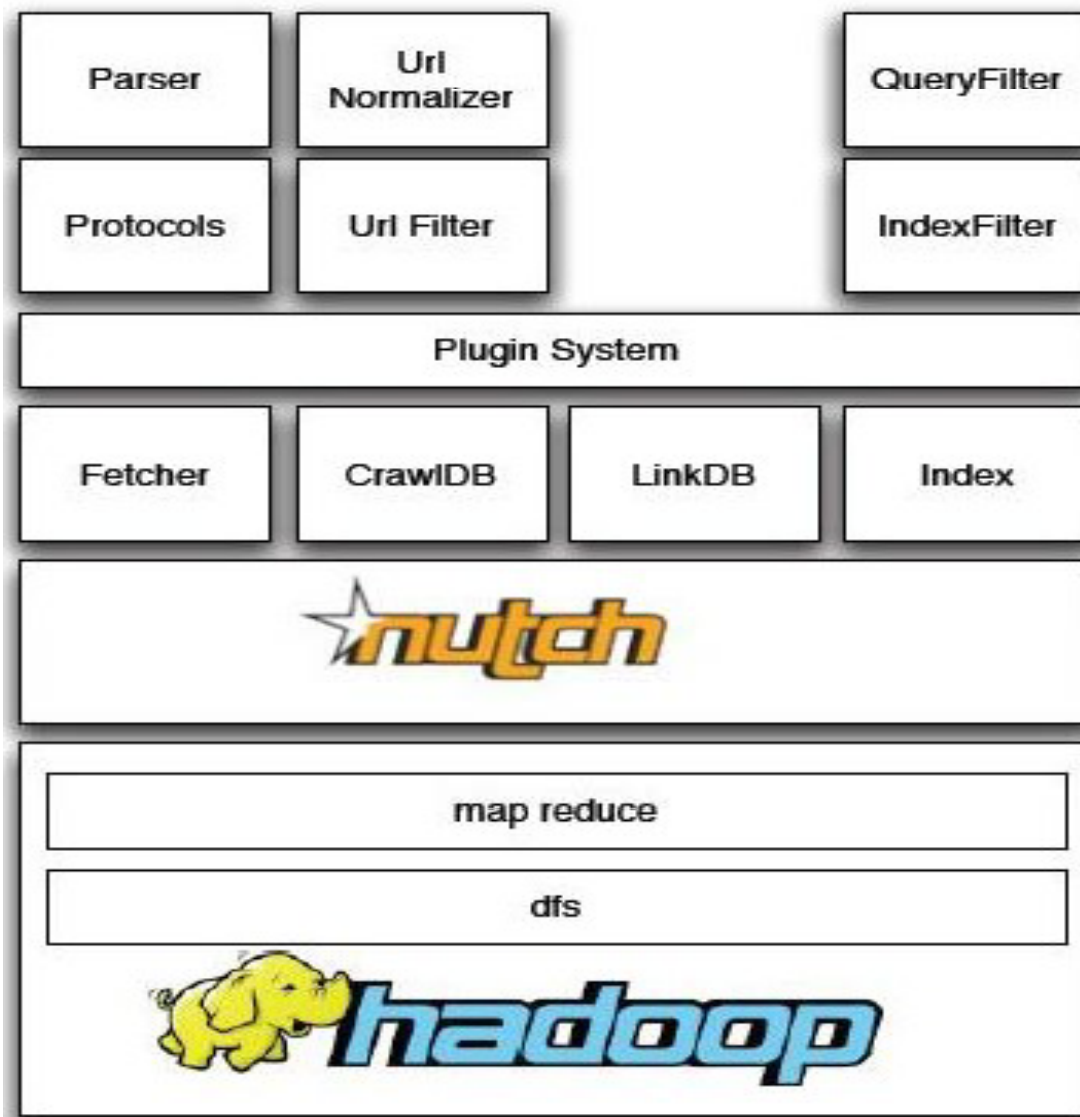
- [AskAboutOil](#) is a vertical search portal for the petroleum industry.
- [Baynote](#) provides free hosted Nutch search for businesses.
- [BeThere BeSquare](#) is an Event Search Engine for the San Francisco category and get details about events in 4 different views.
- [Bigsearch.ca](#) uses nutch 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, documentation in english.
- [Cornell University Library](#) is collaborating with the research group pages based on Nutch. The nutch-based search engine is near final.
- [Creative Commons](#) is a search engine for creative commons license.
- [Dadi360](#) Use nutch search engine for providing search of China.
- [Ecolhub Web Search](#) an E. coli specific search engine based on Nutch thereby reducing the number of spurious hits. Searches can be on E. coli. 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>)

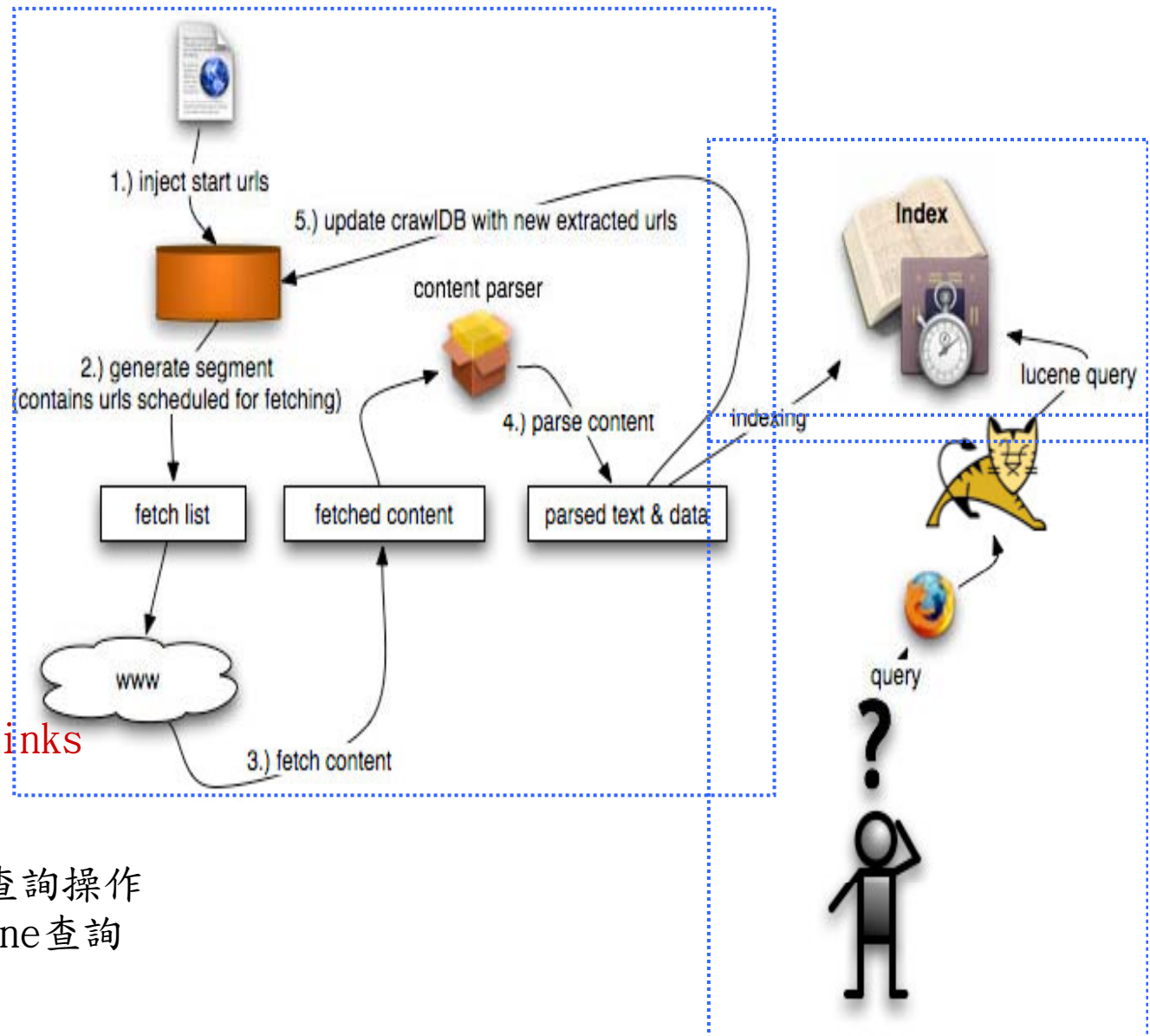
The screenshot shows the Krugle search engine interface. At the top, there are tabs for 'Open Source Code', 'Open Source Projects', and 'SCM Comments'. Below these are search filters: '[Clear Filters]', '[Advanced Search]', 'Language: All', 'Found in: Any area', and 'Project: Enter project name'. The search term 'nutch' is entered in the search box. The results section is titled 'Results' and 'Code Search for nutch'. It shows 'Code Files 1-10 (out of about 1849 matching files)'. The first result is 'Nutch.java' from 'Creative Commons Tools' and 'Apache-2.0'. The code snippet shows a public interface Nutch with a static final String ORIGINAL\_CHAR\_ENCODING. The second result is 'Nutch.java' from 'Nutch' and 'Apache-2.0', showing a similar code snippet. The third result is 'NutchConfiguration.java' from 'Nutch' and 'Apache-2.0', showing a class NutchConfiguration. The fourth result is 'NutchJob.java' from 'Nutch' and 'Apache-2.0', showing a class NutchJob that extends JobConf.

# 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](#)

# 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.html>

# Start

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

# Let's Go

# Stepssssssssssssssssssss!

前言

環境

step 1 安裝好Hadoop叢集

step 2 下載與安裝

2.1 下載 nutch 並解壓縮

2.2 部署hadoop,nutch目錄結構

step 3 編輯設定檔

3.1 hadoop-env.sh

3.2 hadoop-site.xml

3.3 nutch-site.xml

3.4 slaves

3.5 crawl-urlfilter.txt

3.6 regex-urlfilter.txt

3.7 整個移植到另一台node

step 4 執行nutch

4.1 編輯url清單

4.2 上傳清單到HDFS

4.3 執行nutch crawl

step 5 瀏覽搜尋結果

5.1 安裝tomcat

5.1 tomcat server設定

5.3 下載crawl結果

5.4 設定nutch的搜尋引擎頁面到tomcat

5.5 設定搜尋引擎內容的來源路徑

5.6 啟動tomcat

step 6 享受結果

**The Other Choose is...**

**Crawlzilla!!!**

