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

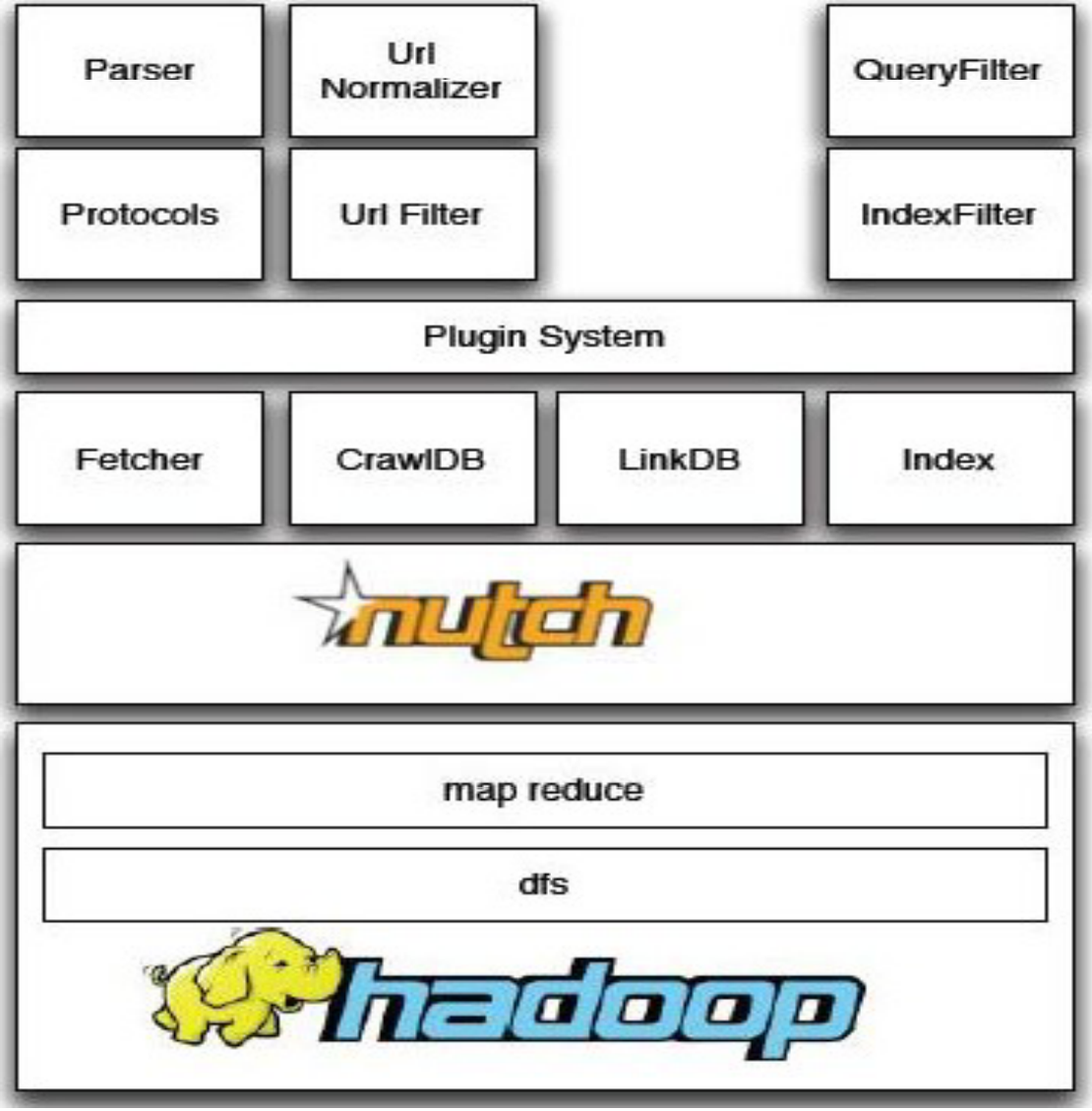
Please sort by name alphabetically

- AskAboutOil is a vertical search portal for the petroleum indust
- Baynote provides free hosted Nutch search for businesses.
- BeThere BeSquare is an Event Search Engine for the San Fra category and get details about events in 4 different views.
- Bigsearch.ca uses nutch open source software to deliver its se
- 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 langu events.
- Circuit Scout is a search engine for electrical circuits.
- Comtec Search is a search engine for UK Tour Operator Pack
- Coder-Suche.de searchs for coding stuff like apis, documenta english.
- Cornell University Library is collaborating with the research gro pages based on Nutch. The nutch-based search engine is near t
- Creative Commons is a search engine for creative commons l
- Dadi360 Usee nutch search engine for providing search of Ch
- Ecolhub Web Search an E. coli specific search engine based thereby reducing the number of spurious hits. Searches can be o 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)

架構

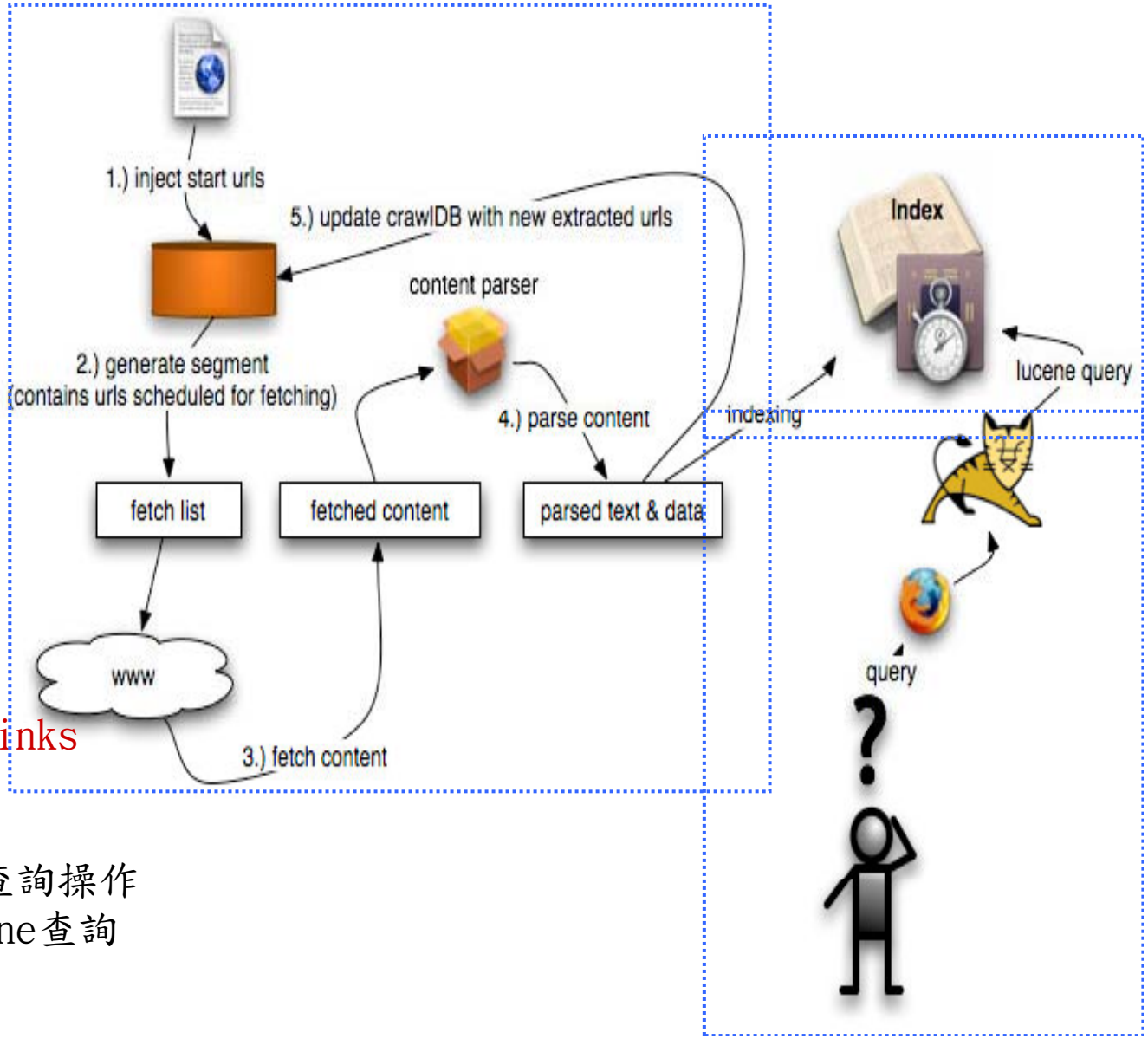

Parser
Url Normalizer
QueryFilter
Protocols
Url Filter
IndexFilter
Plugin System
Fetcher
CrawlDB
LinkDB
Index
nutch
map reduce
dfs
hadoop

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

http://wiki.apache.org/nutch/PluginCentral

International

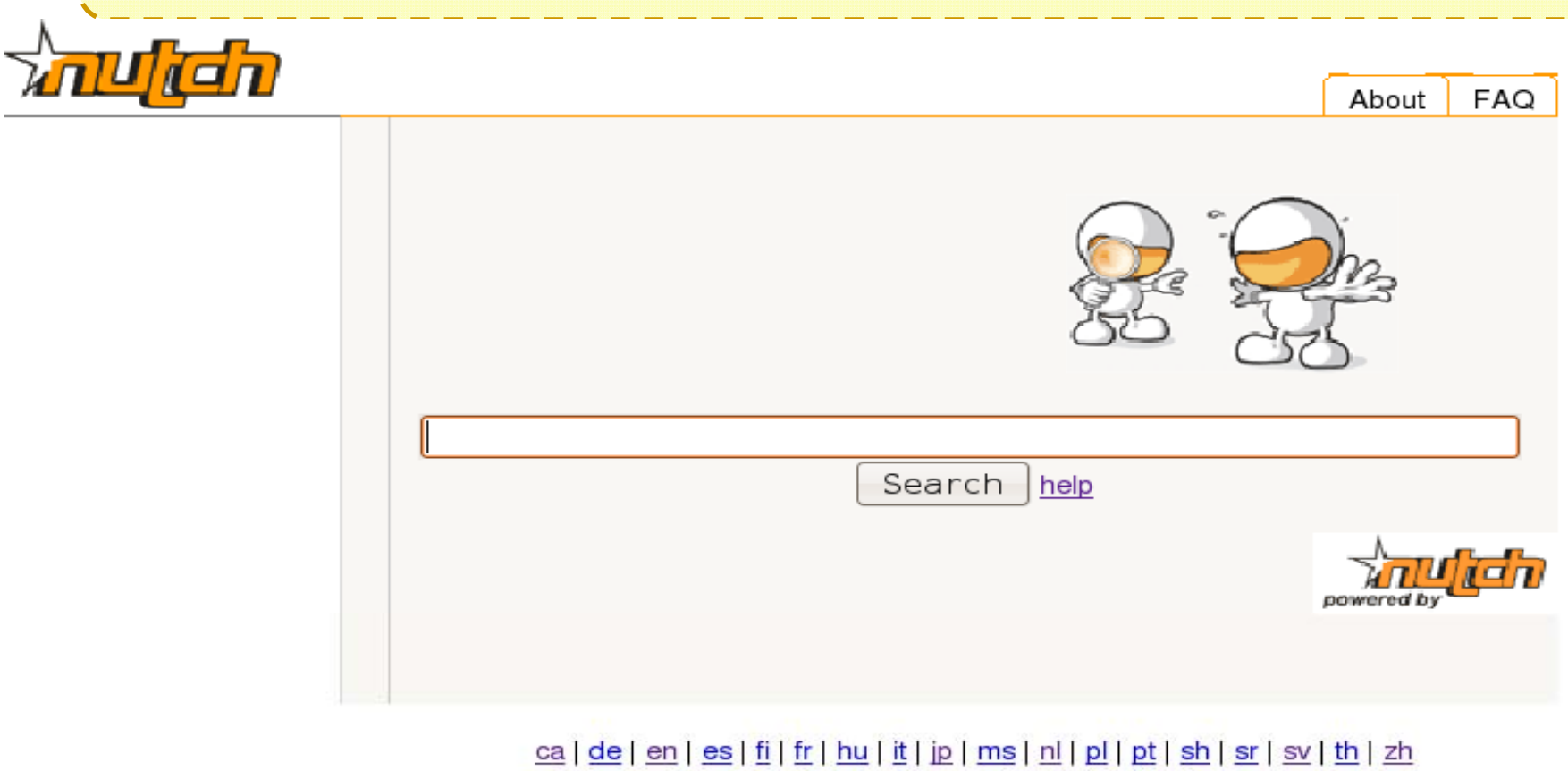
- 已有多國語言版可選，但若還要客製化…
- the page header
  - src/web/include/*language*/header.xml
- the "about" page
  - src/web/pages/*lang*/about.xml
- the "search" page
  - src/web/pages/*lang*/search.xml
- the "help" page
  - src/web/pages/*lang*/help.xml
- text for search results
  - src/web/locale/org/nutch/jsp/search_*lang*.properties

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



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

Stepssssssssssssss!

The Other Choose is…

Crawlzilla!!!