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 - Crawlzilla 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 Crawlzilla
- Why Crawlzilla
- Crawlzilla's Details
- Let's go

What's Crawlzilla

- 為一Opensource專案為使用者建立客製化搜尋引擎軟體
- 提供簡單安裝及操作管理介面，輕鬆建立搜尋引擎的套件工具
- 提供索引資料庫瀏覽功能，搜尋引擎資料庫資訊一目了然
- 利用Lucene為函式庫
- **架構於Hadoop之上**

版本演進

版次/項目	NutchEZ	Crawlzilla v0.0.3	Crawlzilla v1.x.x
特色	• 為crawlzilla前身 • 需自行設定執行環境 • 無提供叢集安裝介面 • 終端機操作	• 安裝簡單 • 快速佈署hadoop環境 • 叢集安裝容易 • 友善的操作環境 • 詳細的搜尋引擎資訊 • 單一建立使用者	• 延續v0.0.3特色 • 開放多人使用，不需架設多台相同環境 • 系統排程功能 • 索引庫自動更新 • 軟體自動更新 • 即時體驗：http://demo.crawlzilla.info/
發行時間	2009.09	2010.08	2011.07

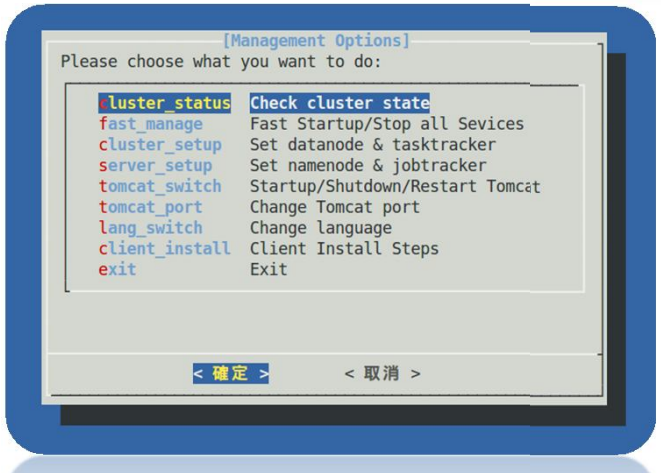
Why Crawlzilla ?

- 開放式搜尋引擎不適用於企業內部網站
- 使用Opensource建立搜尋引擎的技術門檻太高
- 叢集環境架設不易
- 使用Crawlzilla優點
  - Opensource專案，使用者可依自己的需求修改源始碼
  - 使用簡單，可輕鬆建立叢集環境
  - 友善的操作環境，節省適應系統時間
  - 支援中文分詞，提高搜尋精準度

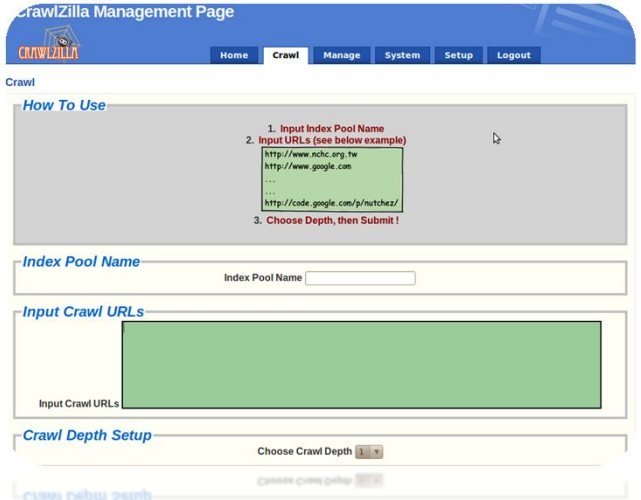
Crawlzilla 操作介面特色

```
check_sunJava
Crawlzilla need Sun Java JDK 1.6.x or above version
System has Sun Java 1.6 above version.
System has ssh.
System has ssh Server (sshd).
System has dialog.
Welcome to use Crawlzilla, this install program will create a new accunt and to
assist you to setup the password of crawler.
Set password for crawler:
password:

keyin the password again:
password:

Master IP address is:  140.110.138.186
Master MAC address is:   08:00:27:99:4d:09
Please confirm the install infomation of above : 1.Yes 2.No
```

(1) Easy to Deploy Crawling Cluster Environment

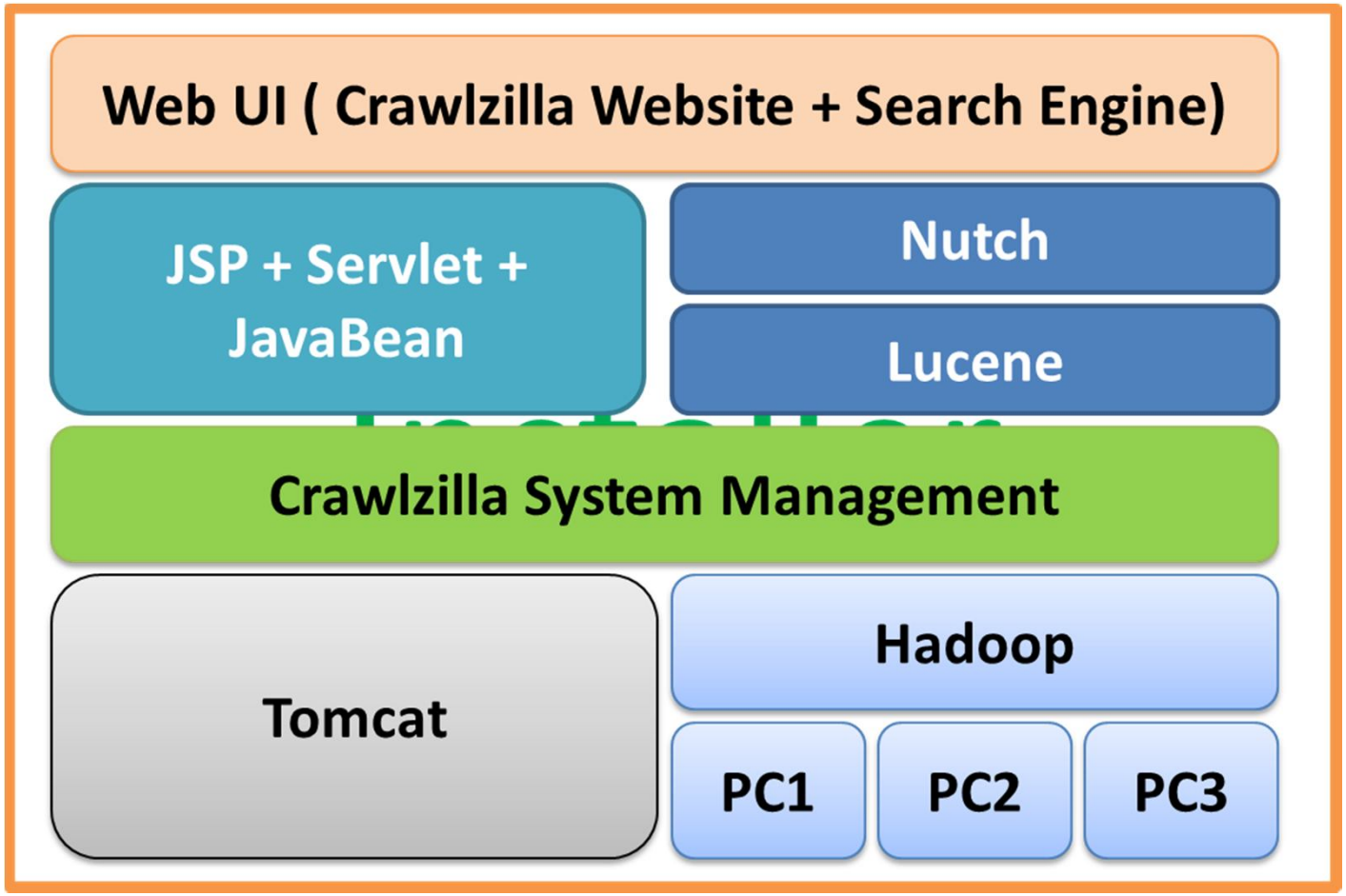


(2) Easy to Manage



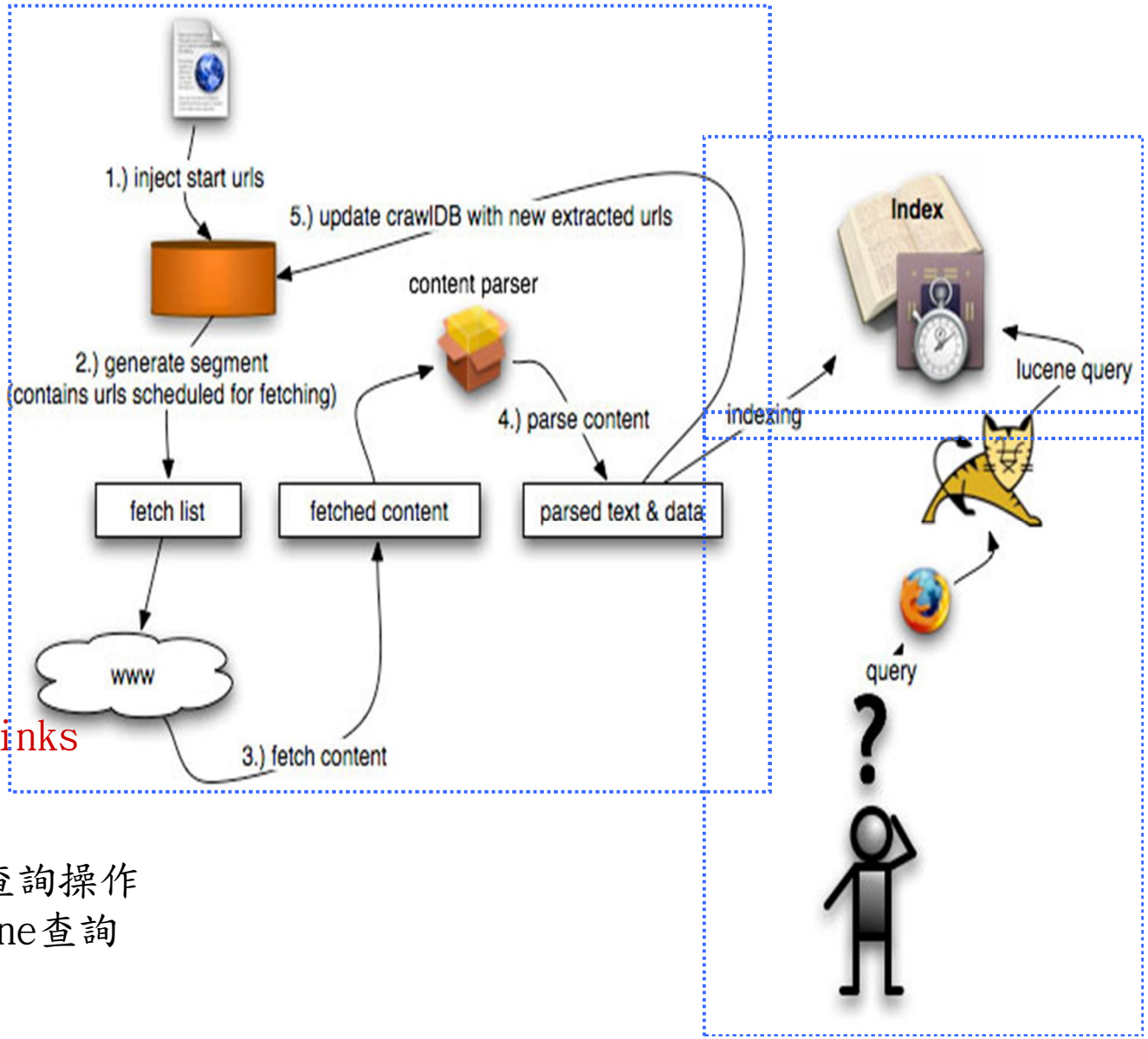
(3) Easy to Use

架構


Web UI (Crawlzilla Website + Search Engine)
JSP + Servlet + JavaBean
Nutch
Lucene
Crawlzilla System Management
Tomcat
Hadoop
PC1
PC2
PC3

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

References..

- **Crawlzilla @ Google Code Project Hosting (中文說明頁)**
  - **http://code.google.com/p/crawlzilla/**
- **Crawlzilla @ SourceForge(英文說明頁)**
  - **http://sourceforge.net/p/crawlzilla/home/**
- Crawlzilla User Group @ Google
  - **http://groups.google.com/group/crawlzilla-user**
- **NCHC Cloud Computing Research Group**
  - **http://trac.nchc.org.tw/cloud**