National Applied Research Laboratories

National Center for High-Performance Computing

快速佈署叢集式搜尋引擎

陳威宇 郭文傑 楊順發

{ waue, rock , shunfa }@nchc.org.tw

國家高速網路與計算中心

自由軟體實驗室

August 14~15, 2010

COSCUP 2010

建立專屬的搜尋引擎?

- 我的需求是...?
  - 公司內部資訊網站為避免機密資料外流，需要建立內部搜尋引擎
  - 個人文件可以透過自己的搜尋引擎建立索引
  - 厭倦了商業搜尋引擎的搜尋結果(ex：廣告資訊)
  - 某些情況下需花費許多時間過濾無意義的資訊
  - 自己建一個搜尋引擎好像蠻酷的

建立專屬的搜尋引擎?

- 工具選擇
  - 商業軟體 V.S. 自由軟體

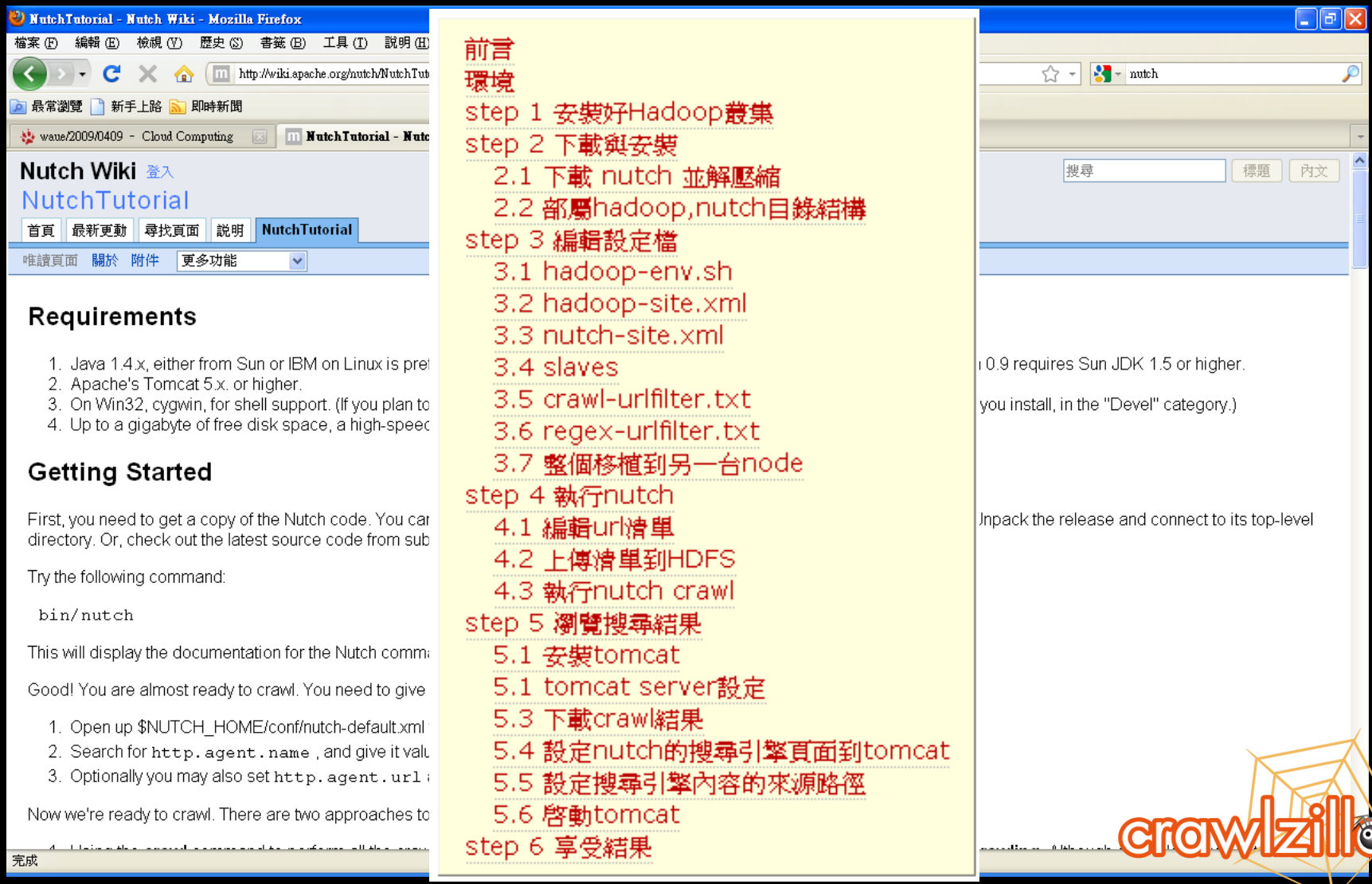
	Goole企業版(客戶)	自由軟體
收費標準	18,000美金	免費!!!
程式碼	不提供	免費公開
系統操作技術門檻	較低	稍高
客製化	需洽相關技術人員	需自己修改程式碼
系統維護成本	年費另計	成本較低

使用Open Source 建搜尋引擎

- 優點
  - 透明
  - 擴充
  - 隱私
  - 客製化

免費!!!

我要選擇開源軟體建搜尋引擎 !!!



開始安裝

- 費了一番功夫後…
  - 技術文件有看沒有懂
  - 照著安裝步驟後安裝完了
  - 接下來呢…?
  - 如何驗證是否安裝成功?
  - 如何使用?
  - 怎麼解決錯誤訊息?
  - ……
  - ……
  - ……

糗大了 因為...

- 我想自己建搜尋引擎 又不想花錢 可是…
  - 技術文件看不懂
  - 不會寫程式
  - 系統維護是個問題

三個願望 一次滿足

Crawlzilla !?

- **Crawlzilla 簡介**
  - 於2009推出實驗版，僅提供單機版本及基本功能
  - 感謝約2,800次下載量的支持並提供寶貴意見
  - 於2010加入叢集版本並命名為 Crawlzilla
  - 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
  - 提供索引庫瀏覽功能，即時了解爬取結果

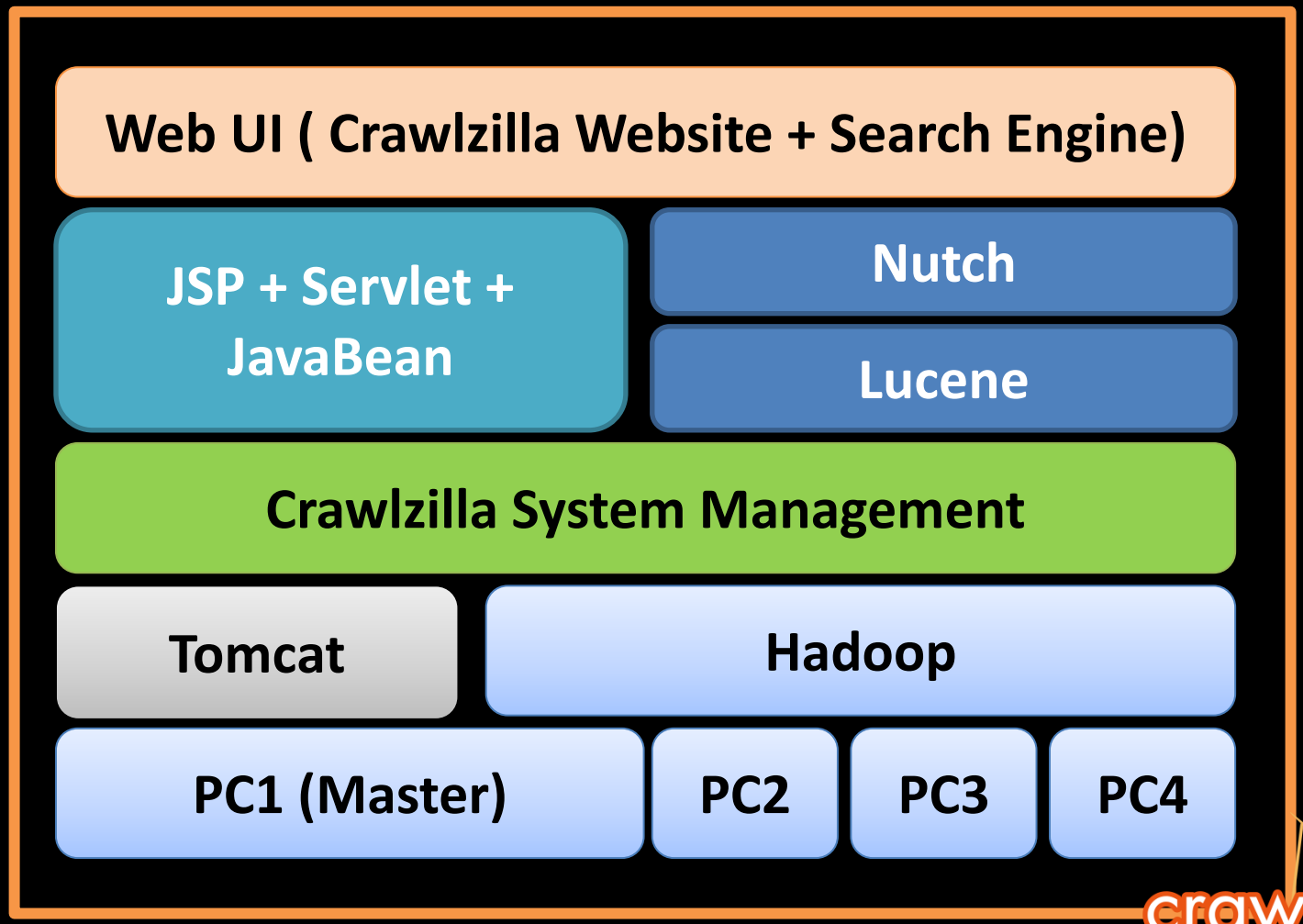
Crawlzilla

- **Crawlzilla 提供以下特色功能：**
  - 簡易安裝
  - 支援叢集運算及顧全安全性
  - 支援多工網頁爬取
  - 網頁管理
  - 支援同時存在多個搜尋引擎
  - 即時瀏覽資料庫資訊
  - 支援中文分詞功能
  - 解決中文亂碼及中文支援
  - 多種語言
  - 適用於泛Linux平台

Crawlzilla

- **Crawlzilla 組成**
  - 系統
  - 管理
  - 執行
- **開發者註解**
  - 目前Crawlzilla仍處開發階段，若系統出現不正常反應屬自然現象，不需要太恐慌並將問題回報給開發者改善後即可正常使用。

Crawlzilla 系統架構



Crawlzilla – 叢集系統需求!?

- 嫌一台電腦跑不夠快!?
- 閒置電腦太多!?
  - 讓多台電腦分工合作運算
- 架設叢集環境很麻煩?
  - 啟動Crawlzilla懶人叢集模式
    - 輕鬆安裝及移除
    - 可動態增加及刪除運算主機
    - 即時檢視叢集運算主機狀態

Crawlzilla中文分詞

- 中文分詞是...?
  - 以中英文字詞結構解釋：
    - 英文以”詞”為文字的基本單位
    - 中文以”字”為文字的基本單位
  - 例：
    - 英文 → I am a good student.
    - 中文(無分詞時) → 我 | 是 | 一 | 位 | 好 | 學 | 生
    - 中文(加入分詞) → 我 | 是 | 一位 | 好學 | 學生

Crawlzilla 加入中文分詞效果

Why ik-analyzer?

	Paoding	imdict	mmseg4j	ik-analyzer
使用者自定詞庫	支援	不支援	支援	支援
每秒處理文字 (基於官方介紹)	100萬字	約26萬字	15萬字~24萬字	50萬字
相關支援文件	很少,幾乎沒有	很少, 幾乎沒有	MMSeg演算法 說明文件	提供一個詳細的使用範例及設定說明文件
其他	提供詞庫自動更新檢測	基於ictclas模組開發，分詞效率佳	實作最多分詞，但技術仍不成熟，待改善	針對Lucene全文檢索提供最佳化的查詢分析器-IKQueryParser，能提高Lucene檢索的命中率

Crawlzilla網頁管理介面

Crawlzilla網頁管理介面 – 資料庫管理

資料總覽

起始URL	http://iso.nchc.org.tw/document/		
本機索引路徑	/home/crawler/crawlzilla/archieve/4@nchc-document/index		
總共文字數	13593	文件檔數量	303
資料庫更新日期	Wed Aug 11 23:38:41 CST 2010	使用者名稱	crawler

被搜尋分析到的網址:

排序	內容	引用次數	排序	內容	引用次數
0	site:www2.nchc.org.tw	130	1	site:iso.nchc.org.tw	96
2	site:www.nchc.org.tw	39	3	site:service.nchc.org.tw	20
4	site:itf.nchc.org.tw	4	5	site:edu.nchc.org.tw	2
6	site:intra.nchc.org.tw	2	7	site:www.medicalgrid.org	1
8	site:wlanrc.nchc.org.tw	1	9	site:ecogrid.nchc.org.tw	1
10	site:volunteer.nchc.org.tw	1	11	site:www.floodgrid.nchc.org.tw	1
12	site:elib.nchc.org.tw	1	13	site:www.narl.org.tw	1
14	site:pccluster.nchc.org.tw	1	15	site:colife.nchc.org.tw	1
16	site:bioinfo.nchc.org.tw	1			

分析的文件型態:

排序	內容	引用次數	排序	內容	引用次數
0	type:application	202	1	type:application/pdf	200
2	type:pdf	200	3	type:html	101
4	type:text	101	5	type:text/html	101
6	type:msword	1	7	type:vnd.ms-powerpoint	1
8	type:application/vnd.ms-powerpoint	1	9	type:application/msword	1

Crawlzilla叢集與伺服器管理介面

Crawlzilla系統Demo

- 系統安裝 (Demo Video also on Youtobe)

Crawlzilla系統Demo

- 網頁操作
  - 執行網頁爬取任務
  - 資料庫管理
  - 搜索引擎展示
- 系統管理

Q & A

mail list

- 陳威宇 waue@nchc.org.tw
- 郭文傑 rock@nchc.org.tw
- 楊順發 shunfa@nchc.org.tw

相關資源

- **Crawlzilla @ Google Code Project Hosting**
  - **http://code.google.com/p/crawlzilla/**
- **NCHC Cloud Computing Research Group**
  - **http://trac.nchc.org.tw/cloud**