

# 輕鬆入手的叢集式搜尋引擎 -



**CRAWLZILLA**

Crawlzilla Develop Team  
Free Software Lab @ NCHC



TAIWAN

[www.nchc.org.tw](http://www.nchc.org.tw)

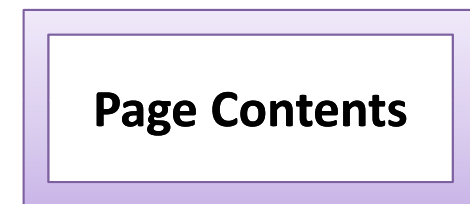
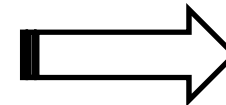
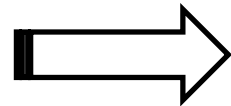


National Applied  
Research Laboratories



# 搜尋引擎運作原理 – Phase1

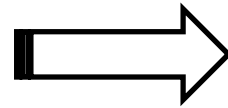
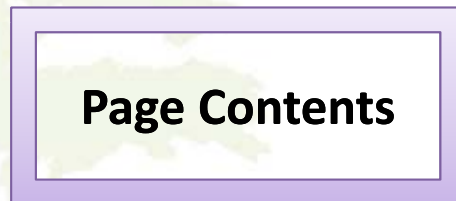
- Crawling the Web



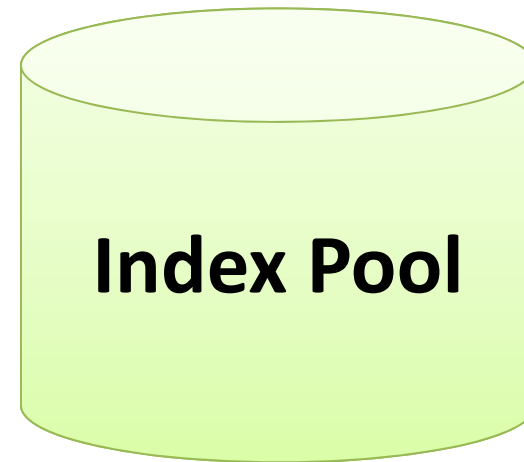
Crawler visits the web pages of the links

# 搜尋引擎運作原理 – Phase2

- Building the Index Pool

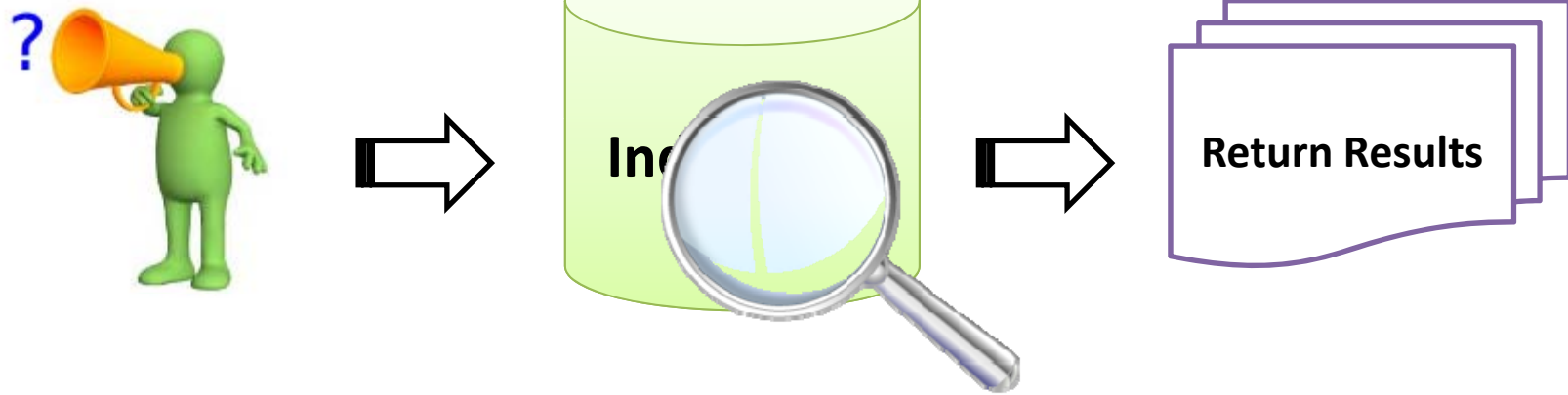


Parse Contents



# 搜尋引擎運作原理 – Phase3

- Serving Queries



User Sent a Query

Search from Index Pool

Return Results

# What is Crawlzilla?

- Crawlzilla 簡介

- 於2009推出實驗版
- Crawlzilla 於2010更名並延續實驗版開發更多新功能
- 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

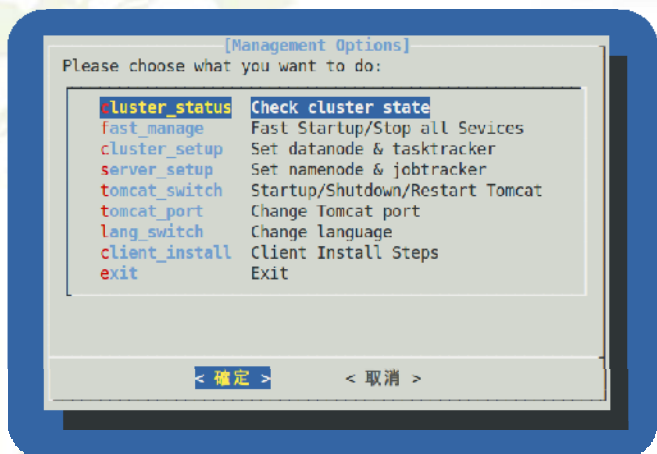
# Why Crawlzilla?

- 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

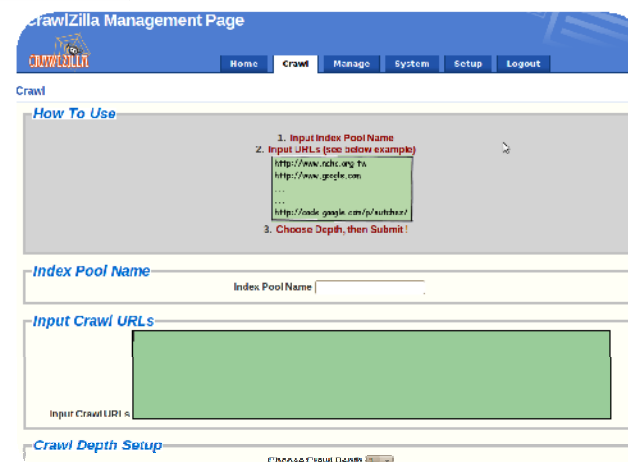
# Crawlzilla 操作介面特色

```
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 account 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 information of above : 1.Yes 2.No
```

(1) Easy to Deploy Crawling Cluster Environment



(2) Easy to Manage



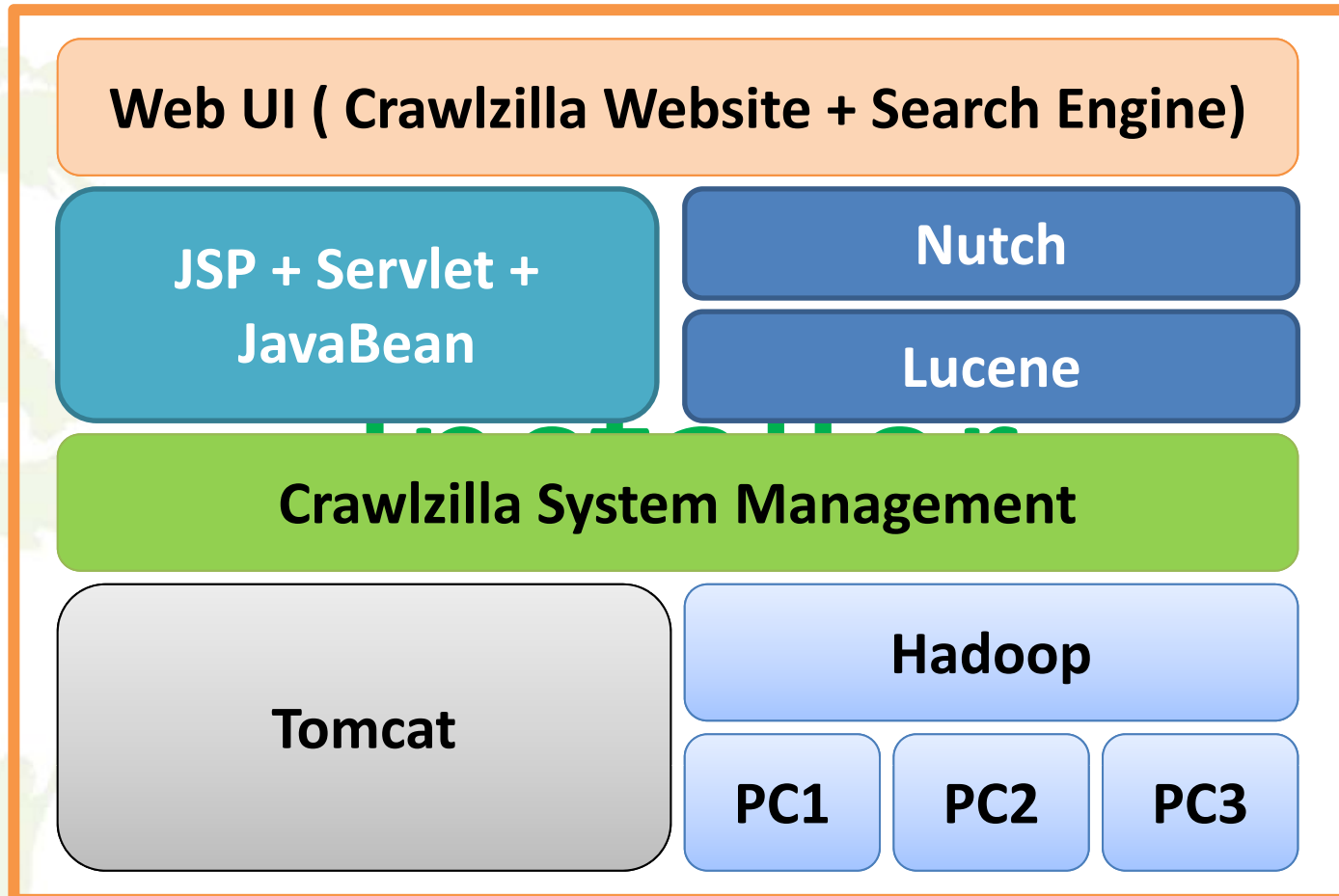
(3) Easy to Use

# Crawlzilla 系統功能

- 支援叢集運算及顧全安全性
- 支援中文分詞功能
- 支援多工網頁爬取
- 支援多重搜尋引擎
- 即時瀏覽資料庫資訊
- 解決中文亂碼及中文支援
- 支援多國語言
- 網頁管理



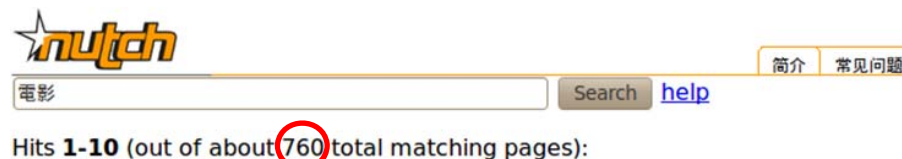
# 系統架構



# 搜尋引擎加入中文分詞功能

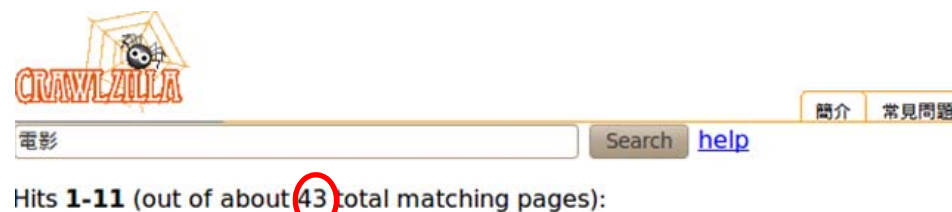
- 索引資料庫會以中文字詞為基本單位建立索引
- 加入中文分詞針對同一網站爬取進行搜尋
  - 搜尋引擎**無**中文分詞功能時，搜尋關鍵字 - 電影

- **760** 筆搜尋結果



- 搜尋引擎**加入**中文分詞功能時，搜尋關鍵字 - 電影

- **43** 筆搜尋結果



- 可提高搜尋的精準度

# Crawlzilla - 叢集環境需求

- 如果你覺得...
  - 一台電腦無法滿足你的運算需求
  - 閒置電腦太多
  - 解：讓多台電腦分工運算
- 但是...
  - 架設叢集環境很麻煩!?
  - 解：Crawlzilla 提供叢集安裝模式，只要三分鐘即可建立叢集式搜尋引擎!!!

# Resources

- **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>