

NCHC-自由軟體實驗室

Crawlzilla-輕鬆打造你的專屬搜尋引擎

安裝及使用說明



【安裝前的小提醒】

- **硬體規格**建議：記憶體 1.5G 以上
- **作業系統**建議：[請參考測試成功的作業系統](#)
- **系統環境**設定：
 - **修改主機名稱** (建議不要用 localhost 作為主機名稱)
 - 若系統為 debian 或 ubuntu (底下的 {hostname} 請修改成您要的主機名稱)

```
$ su -
```

```
# cat {hostname} > /etc/hostname
```

```
# /etc/init.d/hostanme restart
```

```
// debian 則為 /etc/init.d/hostanme.sh restart
```

```
// 建議登出再登入，讓修改的主機名稱生效
```

- 若系統為 Fedora 或 CentOS (底下的 {hostname} 請修改成您要的主機名稱)

```
$ su -
```

```
# vim /etc/sysconfig/network
```

```
HOSTNAME={hostname}
```

```
# hostname {hostname}
```

```
// 建議登出再登入，讓修改的主機名稱生效
```

- **確認 /etc/hosts 的主機名稱 和 IP 位址 也是正確的對應**

```
# cat /etc/hosts
```

```
127.0.0.1 localhost
```

```
140.100.X.X {hostname}
```

【安裝時需要的套件】

- 系統會自動檢查 **openssh, openssh-server, Sun 6 Java** 套件是否有安裝，並試著自動安裝 (ubuntu, debian)，若其他系統無法自動安裝，使用者需自行手動安裝

- 由於 Sun 6 Java 是版權軟體，因此某些 linux 版本預設無法安裝

- Ubuntu 10.04 安裝可參考以下指令

```
sudo add-apt-repository "deb http://archive.canonical.com/
```

```
lucid partner"
```

```
sudo apt-get update
```

```
sudo apt-get install sun-java6-jdk sun-java6-plugin
```

```
sudo update-java-alternatives -s java-6-sun
```

- Fedora 和 CentOS 再安裝 **crawlzilla** 過程中會自動安裝

【Crawlzilla 單機安裝步驟】

- 安裝 Crawlzilla 在一台電腦上運作，功能與穩定性不會比安裝在多台電腦上少，兩者差別僅於分析大型網站的效率而已。因此，安裝 Crawlzilla 在單機上是使用者或體驗者建議的選項。安裝的過程非常簡單，只需彈指的四步間即可完成（已安裝 sun-java-6 的前提下）。
- 如果你要讓 Crawlzilla 安裝於兩台以上的叢集系統，請看 **Crawlzilla 叢集安裝步驟**

【Step 1. 取得安裝檔】

- 至 <http://sourceforge.net/p/crawlzilla/home/> 取得 crawlzilla 最新安裝檔

【Step 2. 解壓縮並執行安裝程式】

- 參考指令如下：

```
tar zxvf Crawlzilla-0.2*.tar.gz  
./Crawlzilla_Install/install
```

ps：此指令會切換成 sudoer ，因此有可能會要您的 sudoer 密碼

【Step 3. 設定密碼及確認網路資訊】

- 此步驟將會在系統中新建一組 user 帳號-crawler，系統服務及叢集間的溝通將會已此一帳號密碼作為執行身份。
- 設定密碼並確認網路狀態資訊後，等候完成安裝即可。
- 畫面如下：

```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級, 0 個新安裝, 0 個將移除且 13 個不會升級。
需要下載 0B 的歸檔檔案。解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

系統有 Sun Java 1.6 以上版本
系統已有 ssh.
系統已有 ssh Server (sshd).
系統已有 dialog.
歡迎使用Crawlzilla, 此安裝程序會為您新建一個crawler帳號並協助您設定密碼
請輸入欲設定的crawler密碼:
password:

請再輸入一次確認密碼:
password:

Master網路IP位址為: 140.110.138.186
Master的MAC為: 08:00:27:99:4d:09
請確認上述的安裝資訊: 1.正確 2.不正確
1
```

- 待出現"恭喜您完成 Crawlzilla 安裝,按 Enter 鍵離開..."即表示單機環境已安裝完成！安裝完成後開啟網頁將會顯示畫面如下：



【註解】

- 單機版安裝程序完成後，系統將會自動開啟 tomcat 服務及 hadoop 中的 namenode 及 jobtracker，若要執行網頁 crawl 功能需自行透過系統管理介面開啟 datanode 及 tasktracker。

【Crawlzilla 叢集安裝說明】

用一台普通 pc (core 2 CPU , 2G Mem) 安裝了 Crawlzilla 的主機，去爬取總資料量約 200MB 的任務，需要 6 個小時。因此，如果你的搜尋需求更大，即時性更急迫些，你可以考慮用叢集運作的方式來平行分散工作到多台電腦去運算。

要完成安裝 Crawlzilla 的叢集模式一點都不複雜，只要注意安裝步驟即可。概念是，安裝的第一台為 Master，此台也可以獨立執行運作全部的服務，方法與 **Crawlzilla 單機系統安裝說明** 內容一模一樣；第二台以上的電腦，則利用 Master 所產生出來的 **Client 安裝包** 來完成安裝；之後就可以用 Crawlzilla 的管理工具輕鬆的**動態**[新增 \ 移除]運算節點囉！

你會發現 **A 安裝 Master Server** 與單機安裝的步驟方法是一模一樣的，因此如果你已經操作了 Crawlzilla 單機系統安裝說明，之後又想要再加入第二、三.... 多個節點，可以直接跳到【 Step B . 安裝 Slave 節點】分別執行即可

【安裝前的小提醒】

- **硬體規格**建議：記憶體 1.5G 以上
- **作業系統**建議：

請至 http://code.google.com/p/crawlzilla/wiki/Support_Distribution 參考測試成功的作業系統

- **系統環境**設定：
 - **修改主機名稱** (建議不要用 localhost 作為主機名稱)
 - 若系統為 debian 或 ubuntu (底下的 {hostname} 請修改成您要的主機名稱)

```
$ su -
```

```
# cat {hostname} > /etc/hostname
```

```
# /etc/init.d/hostanme restart
// debian 則為 /etc/init.d/hostanme.sh restart
// 建議登出再登入，讓修改的主機名稱生效
```

- 若系統為 Fedora 或 CentOS (底下的 {hostname} 請修改成您要的主機名稱)

```
$ su -
# vim /etc/sysconfig/network
HOSTNAME={hostname}

# hostname {hostname}
// 建議登出再登入，讓修改的主機名稱生效
```

- 確認 /etc/hosts 的主機名稱 和 IP 位址 也是正確的對應

```
# cat /etc/hosts
127.0.0.1 localhost
140.100.X.X {hostname}
```

【安裝時需要的套件】

- 系統會自動檢查 **openssh, openssh-server, Sun 6 Java** 套件是否有安裝，並試著自動安裝 (ubuntu, debian)。若其他系統無法自動安裝，使用者需自行手動安裝
 - 由於 Sun 6 Java 是版權軟體，因此某些 linux 版本預設無法安裝
 - Ubuntu 10.04 安裝可參考以下指令

```
sudo add-apt-repository "deb http://archive.canonical.com/
lucid partner"
sudo apt-get update
```



```
sudo apt-get install sun-java6-jdk sun-java6-plugin  
sudo update-java-alternatives -s java-6-sun
```

- Fedora 和 CentOS 再安裝 crawlzilla 過程中會自動安裝

【A 安裝 Master Server】

此一安裝過程將假設欲安裝 crawlzilla 單機版於 PC1 中

【Step A1. 取得安裝檔】

- 至 <http://sourceforge.net/p/crawlzilla/home/>取得 crawlzilla 最新安裝檔

【Step A2. 解壓縮並執行安裝程式】

- 參考指令如下：

```
tar zxvf Crawlzilla-0.2-100813-Shell.tar.gz  
./Crawlzilla_Install/install
```

【Step A3. 設定密碼及確認網路資訊】

- 此一步驟將會在系統中新建一組 user 帳號-crawler，系統服務及叢集間的溝通將會已此一帳號密碼作為執行身份。
- 設定密碼並確認網路狀態資訊後，等候完成安裝即可。
- 畫面如下：

```

檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級, 0 個新安裝, 0 個將移除且 13 個不會升級。
需要下載 0B 的歸檔檔案。解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

系統有 Sun Java 1.6 以上版本
系統已有 ssh.
系統已有 ssh Server (sshd).
系統已有 dialog.
歡迎使用Crawlzilla, 此安裝程序會為您新建一個crawler帳號並協助您設定密碼
請輸入欲設定的crawler密碼:
password:

請再輸入一次確認密碼:
password:

Master網路IP位址為: 140.110.138.186
Master的MAC為: 08:00:27:99:4d:09
請確認上述的安裝資訊: 1.正確 2.不正確
1

```

- 待出現"恭喜您完成 Crawlzilla 安裝,按 Enter 鍵離開..."即表示單機環境已安裝完成！安裝完成後開啟網頁將會顯示畫面如下：



【Step B 安裝 Slave 節點】

【Step B1. 透過 PC1 取得安裝提示】

- 於 client 端執行"ssh PC1"，並執行 "crawlzilla" 指令，找到"client 安裝步驟"，如下圖所示：



- 相關提示字元範例如下：

```
$ scp crawler@PC1:/home/crawler/crawlzilla/source/client_deploy.sh .
```

```
$ ./client_deploy.sh
```

- 由於此步驟需以 crawler 的身份 ssh 至 PC1，因此過程中約需輸入 1~2 次 crawler 密碼

【Step B2. 於 PC2 執行上述之提示字元】

- 取得提示執行後輸入主機之 crawler 密碼兩次並確認網路資訊即可自動完成安裝

執行畫面如下：

- 確認安裝資訊

```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
shunfa@ubuntu-187:~$ ls
client_deploy.sh  hosts          下載  圖片  文件  模板
examples.desktop ubuntu10.04-sun-java.sh  公共  影片  桌面  音樂
shunfa@ubuntu-187:~$ ./client_deploy.sh
checking ssh ... found
checking sshd ... found
crawler@140.110.138.186's password:
client_deploy.sh          100% 516      0.5KB/s   00:00
client_install            100% 2882     2.8KB/s   00:00
client_install_func.sh   100% 10KB     10.5KB/s  00:00
client_remove             100% 4833     4.7KB/s   00:00
CrawlzillaForClientOf_140.110.138.186.tar.gz 100% 54MB    26.8MB/s  00:02
lang_en_US                100% 16KB     15.5KB/s  00:00
lang_zh_TW                100% 15KB     15.4KB/s  00:00
log.sh                    100% 996      1.0KB/s   00:00
README.txt                100% 192      0.2KB/s   00:00
/home/shunfa/crawlzilla_client_install/client_install
[sudo] password for shunfa:
Sorry, try again.
[sudo] password for shunfa:
/home/shunfa/crawlzilla_client_install/client_install
身份是 root
Master 的 IP位址: 140.110.138.186
資料是否正確 (yes/no): yes
```

- 輸入密碼

```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成
正在編輯延伸狀態訊息... 完成
沒有套件將會被安裝、升級或移除。
0 個套件升級, 0 個新安裝, 0 個將移除且 30 個不會升級。
需要下載 0B 的歸檔檔案。解裝後將用去 0B。
正在編輯延伸狀態訊息... 完成
正在讀取套件清單... 完成
正在重建相依關係
正在讀取狀態資料... 完成
正在讀取延伸狀態檔案
初始化套件狀態... 完成

check_sunJava
Crawlzilla 需要 Sun Java JDK 1.6 以上的版本
系統有 Sun Java 1.6 以上版本
系統已有 ssh.
系統已有 ssh Server (sshd).
系統已有 dialog.

請輸入 Master 上 crawler 使用者的密碼:
請再輸入一次: 
```

- 安裝完成

```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
nutch/conf/tika-mimetypes.xml
nutch/conf/smb.properties
nutch/conf/hadoop-default.xml
nutch/conf/nutch-default.xml
nutch/conf/hadoop-default.xml.bek
nutch/conf/configuration.xsl
nutch/conf/subcollections.xml
nutch/conf/context.xsl
nutch/conf/hadoop-env.sh
nutch/conf/commons-logging.properties
nutch/conf/common-terms.utf8
nutch/conf/httpclient-auth.xml
nutch/conf/custom-fields.xml
nutch/conf/regex-urlfilter.txt
nutch/conf/prefix-urlfilter.txt
nutch/conf/hadoop-site.xml
nutch/logs
nutch/nutch-1.0.jar
nutch/default.properties
nutch/build.xml
網路IP位址是 140.110.138.187
網卡MAC位址是 08:00:27:9a:7b:83
Crawlzilla 已完成安裝此一Client端
press any key to continue...
```

【Step B3. 驗證是否安裝成功】

- 於 PC1 執行指令-"crawlzilla"，出現以下畫面後選擇 "檢查 Cluster 狀態"，畫面如下：

```
檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
[管理功能選項]
請選擇：
cluster_status  檢查 Cluster 狀態
cluster_setup   設定 datanode & tasktracker
server_setup    設定 namenode & jobtracker
tomcat_switch   啟動/停止/重新啟動 Tomcat
tomcat_port     更改 Tomcat port
lang_switch     更換語言
client_install  Client 安裝步驟
exit           結束
< 確定 >      < 取消 >
```

- 若出現 2 個運算節點表示安裝成功！

```

檔案(F) 編輯(E) 檢視(V) 終端機(T) 求助(H)
= [Crawlzilla 管理介面] ~by NCHC =
                                [Cluster 狀態]
[IP]           [Hostname]       [Network]       [Dtatnode & Tasktracker]
-----
140.110.138.186  ubuntu-186         online          running
140.110.138.187  ubuntu-187         online          shutdown
                                < 離開 >
                                100%

```

【註解】

- 叢集版安裝完成後，需回 PC1 執行系統管理介面開啟運算服務後才可加入 crawl 運算分派資源中。
- 第三個運算節點以上的安裝方式，則是重複步驟 B 即可

【Crawlzilla 網頁執行介面】

管理介面預設網址為：<http://localhost:8080> 或 <http://ServerIP:8080>

- 登入後首頁如下：



【設定網頁管理者密碼】

- 首次進入網頁介面時，必須先重設管理者密碼（預設密碼為：crawler），設定密碼點選送出並重新登入後就可執行系統。



- 當然之後若要在更改密碼，也可以透過右方的連結 "更改網頁密碼" 以進行更改

【1. 建立第一個搜尋引擎】

【Step 1.1.開啟所有運算服務】

由於執行 Crawl 必須透過 Hadoop 運算，因此執行 Crawl 前請先依序確認以下服務是否已開啟，若為關閉狀態，請依序開啟這些服務。

- Namenode and Jobtracker
- Datanode and Tasktracker(需開啟全部的運算節點)

若不熟悉開啟步驟，請參考系統管理介面操作說明?

【Step 1.2.至 Crawl 網頁中設定爬取項目】

- 依序填入：索引庫名稱，欲抓取的網址 (可多行，如圖所示) 及設定爬取深度即可送出



- 送出後如圖所示，等候時間需視視每台主機的運算速度而定。



【Step 1.3.瀏覽網頁爬取進度】

- 透過系統狀態頁面，可即時了解網頁爬取進度



The screenshot shows the CrawlZilla web management interface. The main navigation bar includes '首頁', '爬網設定', '索引庫管理', '系統狀態', '使用者設定', and '登出系統'. The '系統狀態' (System Status) section is active, displaying the '索引庫狀態' (Index Library Status) for 'tracCloud_and_nhcTW_3' with a 'crawling' status and a 'Delete' button. Below this, the 'Jobtracker 工作排程器狀態 (New Window)' section shows 'Running Jobs' as 'none' and 'Completed Jobs' as a table with one entry:

Jobid	Priority	User	Name	Max Con
job_201008181050_0001	NORMAL	crawler	inject tracCloud_and_nhcTW_3/urls	100

- 待出現"Finish"表示索引庫已建立，並可將此一訊息刪除



The screenshot shows the CrawlZilla web management interface. The '索引庫狀態' (Index Library Status) for 'tracCloud_and_nhcTW_3' now shows a 'finish' status and a 'Delete' button. The 'Jobtracker 工作排程器狀態 (New Window)' section shows 'Running Jobs' as 'none' and 'Completed Jobs' as empty. A filter box is visible above the 'Running Jobs' section with the text: 'Filter (Jobid, Priority, User, Name) Quick Links' and 'Example: 'user:smith 3200' will filter by 'smith' only in the user field and '3200' in all fields'.

- 完成此一步驟，第一個搜尋引擎已建置，右側快速連結中的 "tracCloud_and_nhcTW_3" 即為此次所建立的搜尋引擎。

【Step 1.4.測試搜尋引擎功能】

- 點選右側快速連結中的 "tracCloud_and_nhcTW_3" 進入搜尋引擎後，輸入一組關鍵字測試搜尋結果，下圖為輸入 "nchc" 為例：



- 搜尋結果：



【2.索引庫管理】

- 索引庫管理頁面中將會顯示目前已建立的所有索引庫，管理者可於此頁面進行瀏覽，刪除及提供網頁嵌入語法，如下圖所示：



【2.1 索引庫瀏覽】

進入索引庫管理頁面後，在欲瀏覽的索引庫欄位點選"preview"即可瀏覽此一索引庫的資訊，目前提供瀏覽的資訊包括：

- 爬取網址
- 爬取文字數
- 爬取文件數
- 相關索引排名

如下圖所示：

檔案 (E) 編輯 (E) 檢視 (V) 歷史 (S) 書籤 (B) 工具 (T) 說明 (H)

http://140.110.138.186:8080/crawlzilla/Statistics.do?fileName=n

Google

CrawlZilla

索引庫管理

索引庫名稱	建立時間	刪除索引庫	預覽統計資料	嵌入搜尋引擎到網頁的語法
nchc-en_3	2010-08-24 16:16:14	Delete	Preview	embed code
nchc-tw_3	2010-08-24 15:22:48	Delete	Preview	embed code

資料總覽

起始URL	http://www.nchc.org.tw/tw/		
本機索引路徑	/home/crawler/crawlzilla/archieve/nchc-tw_3/index		
總共文字數	37095	文件檔數量	1036
索引庫更新日期	Tue Aug 24 15:22:46 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:www.nchc.org.tw	336	1	site:pcluster.nchc.org.tw	87
2	site:bioinfo.nchc.org.tw	66	3	site:www.narl.org.tw	57
4	site:edu.nchc.org.tw	53	5	site:service.nchc.org.tw	35
6	site:accta.nchc.org.tw	28	7	site:colife.nchc.org.tw	14
8	site:wlanrc.nchc.org.tw	13	9	site:elib.nchc.org.tw	13
10	site:www.medicalgrid.org	13	11	site:volunteer.nchc.org.tw	9
12	site:www.stpi.org.tw	7	13	site:noc.twaren.net	7
14	site:ecogrid.nchc.org.tw	6	15	site:www.sipa.gov.tw	3
16	site:asp.104ehr.com.tw	3	17	site:viml.nchc.org.tw	3
18	site:www.ym.edu.tw	2	19	site:www.tnu.edu.tw	2
20	site:www.usc.edu.tw	2	21	site:www.svs.tp.edu.tw	2
22	site:www.smelearning.org.tw	2	23	site:ecocam.nchc.org.tw	2

完成

由於有加入中文分詞功能，因此可以明顯看出索引庫的建立是以"中文字詞"作為基本單位

檔案 (E) 編輯 (E) 檢視 (V) 歷史 (S) 書籤 (B) 工具 (T) 說明 (H)

http://140.110.138.186:8080/crawlzilla/Statistics.do?fileName=n

Google

CrawlZilla

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:text/html	989	1	type:html	989
2	type:text	989	3	type:application	47
4	type:application/pdf	34	5	type:pdf	34
6	type:xml	10	7	type:application/xml	10
8	type:msword	3	9	type:application/msword	3

出現次數前五十名的字彙:

排序	內容	引用次數	排序	內容	引用次數
0	content:網	805	1	content:路	777
2	content:國	758	3	content:中心	750
4	content:計	744	5	content:資	742
6	content:與	740	7	content:訊	734
8	content:頁	712	9	content:電	705
10	content:學	698	11	content:算	696
12	content:家	692	13	content:的	684
14	content:關	676	15	content:議	674
16	content:統	666	17	content:1024	665
18	content:768	664	19	content:系	662
20	content:高速	648	21	content:一	643
22	content:號	636	23	content:區	635
24	content:站	633	25	content:導	632
26	content:解析	628	27	content:建	627
28	content:會	627	29	content:解析度	624
30	content:務	622	31	content:覽	618
32	content:講	614	33	content:發	614
34	content:體	608	35	content:上	607

完成

【2.2 索引庫刪除】

- 在欲刪除的索引庫中點選刪除，確認後即完成刪除索引庫




【3.在網頁中嵌入搜尋引擎】

若企業內部有需要將 Search Bar 整合於企業首頁中，則可使用此一功能，方式如下：



【Step 3.1.開啟索引庫管理頁面】



【Step 3.2.點選 embed code】



```
<form name="search"
action="http://140.110.138.186:8080/nchc-en_3/search.jsp" method="get"><input
name="query" size=15></form>
```

【Step 3.3.複製後貼在欲整合 Search Bar 的頁面】

- embed code 範例：

```

<form name="search" action="http://140.110.138.186:8080/nchc-en_3/search.jsp"
method="get">
<input name="query" size=15>
</form>
```

【4.系統操作】

將現在的工作、叢集運算與儲存系統整合於一頁

CrawlZilla 網頁管理介面

[首頁](#) [爬網設定](#) [索引庫管理](#) [系統狀態](#) [使用者設定](#) [登出系統](#)

系統狀態

索引庫狀態
索引庫名稱 抓取狀態 爬取時間 刪除狀態

Jobtracker 工作排程器狀態 ([New Window](#))

[Quick Links](#)

State: RUNNING
Started: Thu Oct 14 19:11:24 CST 2010
Version: 0.19.1, r745977
Compiled: Fri Feb 20 00:16:34 UTC 2009 by ndaley
Identifier: 201010141911

Cluster Summary

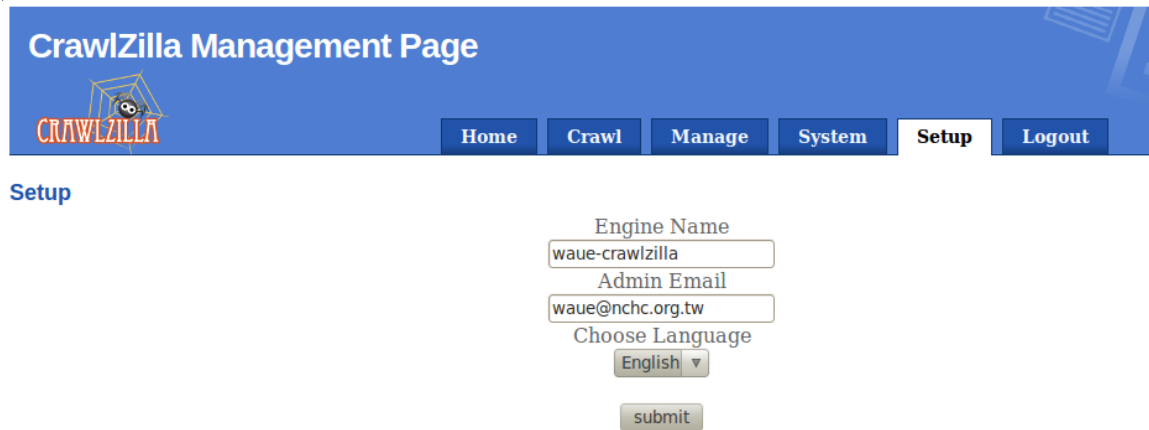
Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node
0	0	0	1	2	2	4.00

Scheduling Information

Namenode 空間管理員狀態 ([New Window](#))

【5.管理者設定】

可以設定 搜尋引擎的名稱，管理者 Email，以及欲使用的語言（中/英）



The screenshot shows the 'CrawlZilla Management Page' with a navigation menu containing 'Home', 'Crawl', 'Manage', 'System', 'Setup', and 'Logout'. The 'Setup' section is active and contains the following form fields:

- Engine Name:
- Admin Email:
- Choose Language: (dropdown menu)
-

【系統管理介面操作說明】

- 此一說明頁面操作環境為兩台已安裝好之叢集環境
- 單機版操作方式皆同。

【系統運算架構】

- 已熟悉 Hadoop 架構之使用者可略過此一段落
- 由於底層的運算是交由 Hadoop 作運算，相關的運算原理可參考 **Hadoop 官方網頁說明**
- 若無法了解上述運算架構及原理，僅需記住以下啟動順序即可。

Step1. 啟動 Namnode & Jobtracker

Step2. 啟動 Datanode & Tasktracker

- 系統安裝完成時，運算節點預設為關閉，需透過系統管理介面開啟服務，執行網頁爬取前，請先確認 Hadoop 相關運算服務已開啟，否則將無法順利執行。

【系統管理介面功能】

在 PC1 終端機中輸入指令-"crawlzilla"即可進入系統管理介面如下圖：



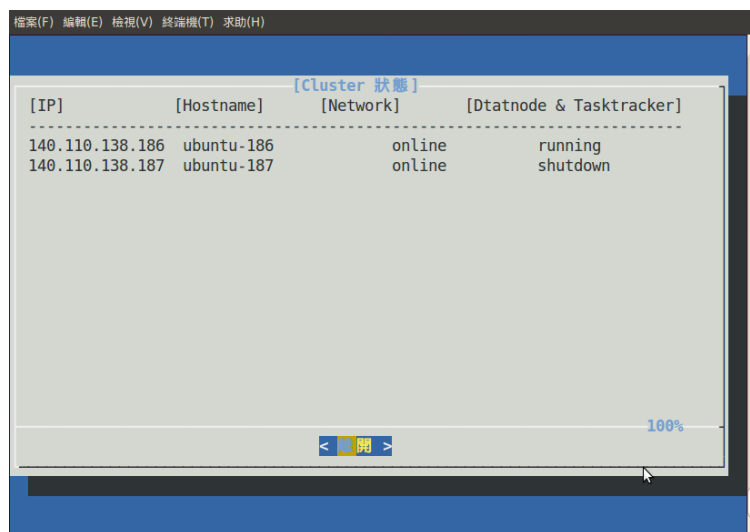
此一管理介面提供數個系統管理功能，於下列子段落一一說明。

【1. 檢查 Cluster 狀態】

此一功能主要顯示叢集中的電腦狀態，主要顯示欄位分別為：

【 IP 】	【 Hostname 】	【 Network 連線狀態 】	【 Datanode & Tasktracker 狀態 】
--------	--------------	------------------	-------------------------------

如下圖：



- 註：Namenode & Jobtracker 狀態需進入其管理頁面查詢之

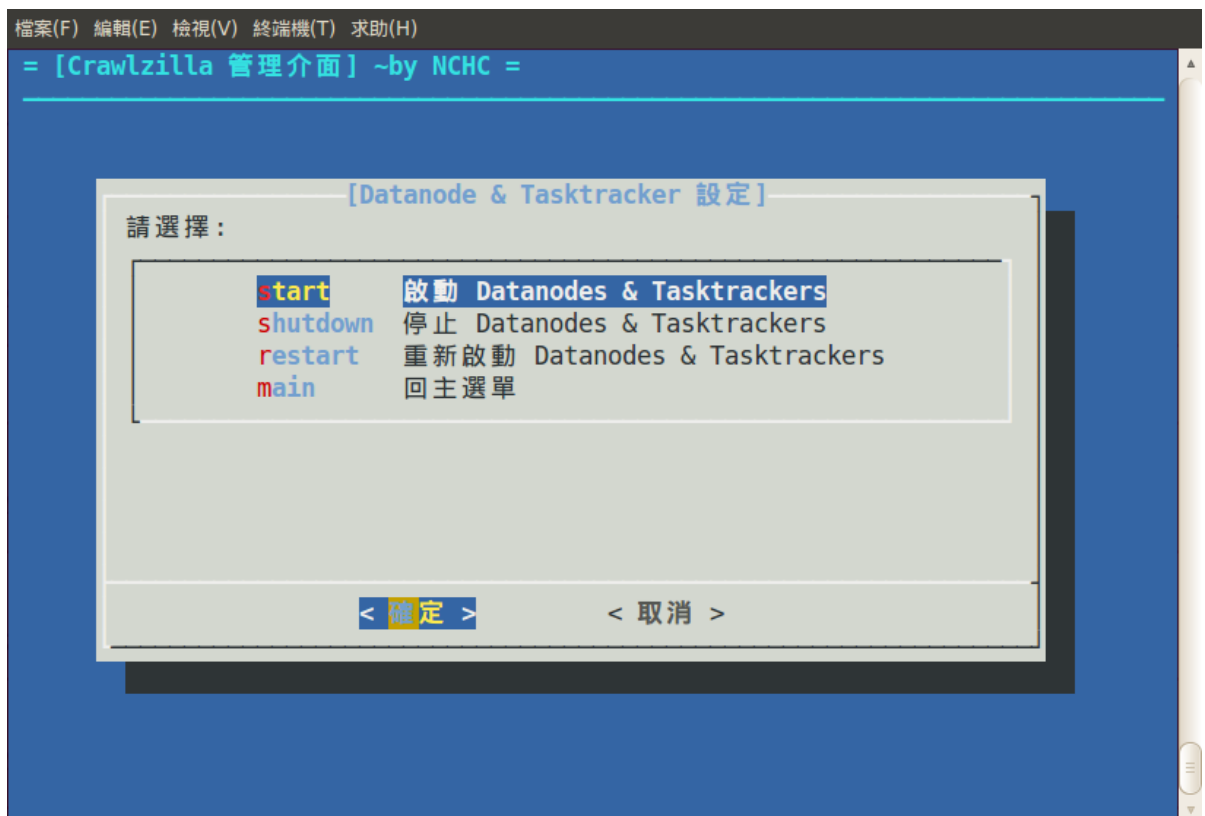
【2. 啟動 Datanode & Tasktracker】

此一功能主要管理叢集中的電腦狀態，並可分為管理全部節點或部份節點

- 主要的功能為【啟動】·【停止】及【重新啟動】：
- 選擇管理全部或部份節點：



- 進入後，選擇欲操作的功能：

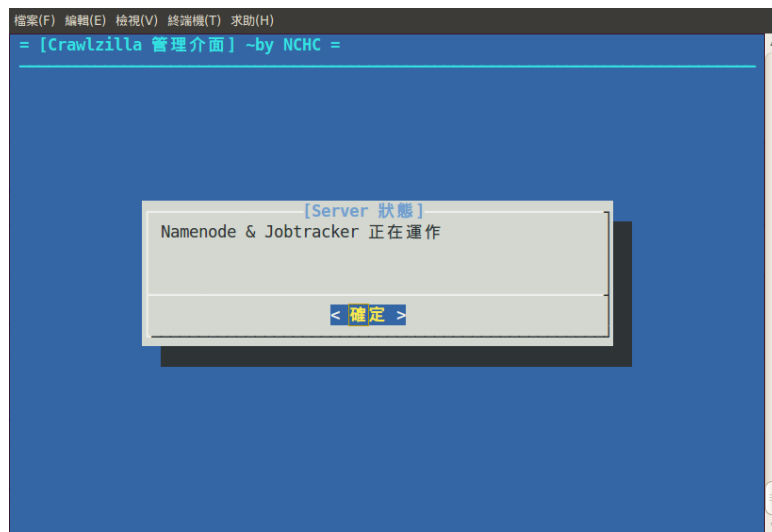


【3. 啟動 Namenode 及 Jobtracker】

此一功能主要管理叢集中的主要電腦狀態（此為 PC1），主要的功能為啟動，停止及重新啟動，操作方式如下：

【Step 3.1】

- 進入時，程式將先檢查目前的服務狀態：



【Step 3.1】

- 進入後，選擇欲操作的功能：

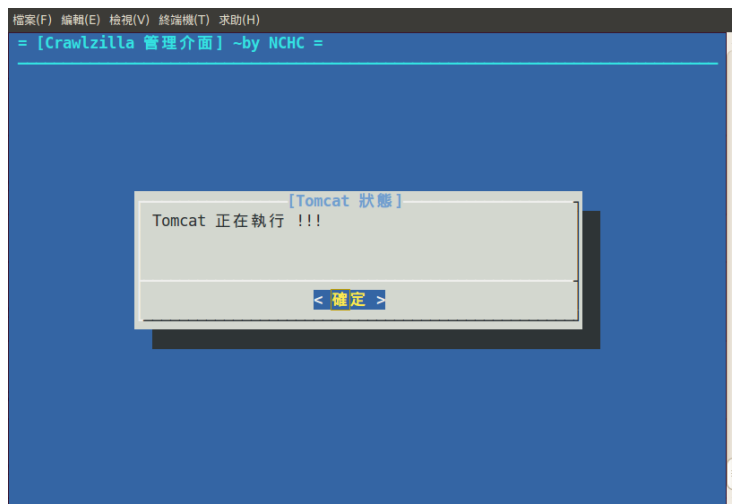


【4. 啟動/停止/重新啟動 Tomcat 服務】

此一功能主要管理叢集中的主要電腦狀態（此為 PC1），主要的功能為啟動、停止及重新啟動網頁伺服器，操作方式如下：

【Step 4.1】

- 進入時，程式將先檢查目前的服務狀態：



【Step 4.2】

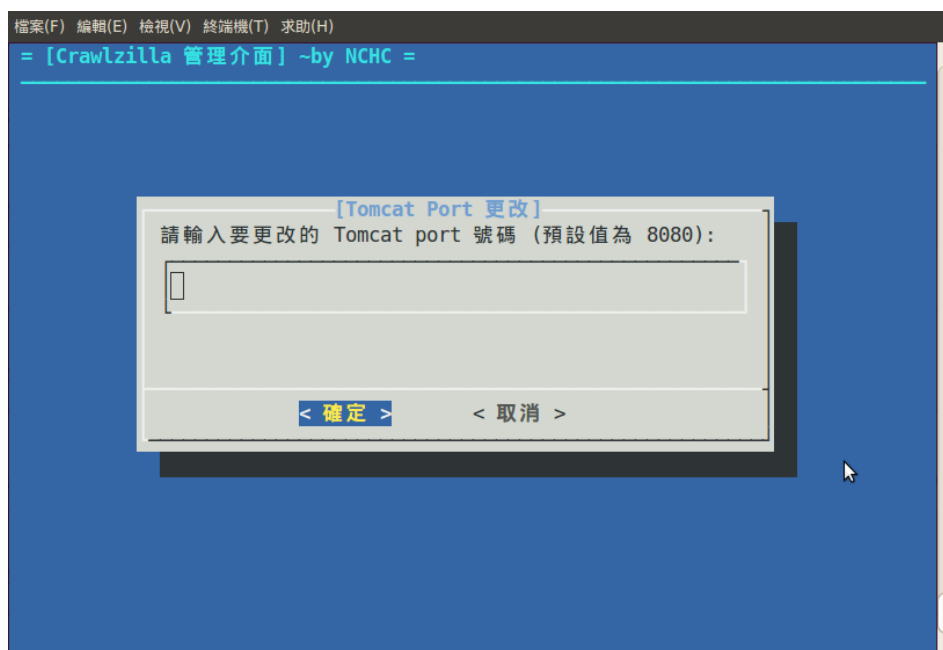
- 進入後，選擇欲操作的功能：



【5. 更改 Tomcat Port】

此一功能主要為更改網頁伺服器的 Port 號，若預設的 Port 號已被其他程序佔用，即可透過此功能進行更改，操作方式如下：

- 進入後，直接輸入欲更改的 Port 號即可。



【6. 更換語言】

提供中英文操作語言更換，進入後直接選擇操作語言即可。



【移除 Crawlzilla】

單機移除

打開終端機，用 **root** 帳號或有 **sudoer***權限的帳號輸入

```
crawlzilla_remove
```

sudoer 帳號需要輸入該帳號的密碼，等程式跑完就移除完囉！

叢集移除

導覽

- Crawlzilla 的叢集是主 (master) -從 (slave) 架構的服務模式，因此移除的時候，先將所有 slave 移除之後，最後再移除 master。
- master 與 slave 的移除都是輸入指令

```
crawlzilla_remove
```

Slave 移除方法

移除方法同 單機移除

Master 移除方法

請先確認所有的 Slave 都已經移除 crawlzilla 完畢後，再執行移除程式

```
crawlzilla_remove
```

移除方法同 單機移除