組別：201402

專題名稱：雲端分散式爬蟲之索引-以專利局公告資料為例

一、**指導老師：**江茂綸老師

二、**組　　員：**林文耀（10030078）、林穎廷（10030082）、吳彥青（10030088）、周詩函（10030096）、王君瑋（10030114）

三、**系統環境：**

(一) 軟體：Ubuntu、Hadoop、PHP、Mysql、Simple HTML DOM Paeser

(二) 硬體：PC 3 台

(三) 通訊設備/協定：無

四、**系統功能與特色：**

(一) **功能**

1. 功能一:利用 Hadoop 透過 MapReduce 進行分散式爬取網路資料。
2. 功能二:可以將爬取的資料轉化為內部資源，並且以圖文並茂的方式呈現。
3. 功能三:針對專利名稱以及專利內容進行關鍵字搜尋。
4. 功能四:針對專利內容進行專利摘要比對，可以比對文章的相似度百分比。
5. 功能五:可統整資料庫的海量資料。

(二) **特色**

本專題以專利局網頁為執行網頁爬取,由於專利局網頁每一筆資料的網址每幾秒鐘就會變動一次，所以抓起來更具有挑戰性，透過將抓取的網頁資料儲存於 Mysql 中，將資料轉化為內部資源，並讓使用者可在海量資料庫中有效且快速的進行資料搜尋、處理及分析等應用。

針對使用者欲蒐集的資料，藉由雲端運算平台強大的運算能力，快速的抓取使用者所需的資料，確保資料來源的完整性、可靠性，及利用網路爬蟲的技術原理，將眾多資料中過濾出有用的資訊，減少使用者親自篩選的時間，進而大幅提升資訊的可靠性、可用性、存取效能及海量資料處理能力。