東海大學

資訊工程研究所

碩士論文

指導教授: 楊朝棟博士

運用 ELK Stack 於電子商務賣方價格分析與資料視覺化

Price Analysis and Data Visualization for E-Commerce Seller Using ELK Stack

研究生: 羅意智

中華民國一〇七年六月

東海大學碩士學位論文考試審定書

東海大學資訊工程學系 研究所

研究生 羅 意 智 所提之論文

運用 ELK Stack 於電子商務賣方價格分析與資料視覺化

經本委員會審查，符合碩士學位論文標準。

學位考試委員會

召 集 人 許慶賢 簽章

委 員 劉榮泰

黃國展

詹毓偉

指 導 教 授 楊朝棟 簽章

中華民國 107 年 6 月 8 日

摘要

雲端運算是近年來很熱門的領域，借由硬體效能的提升與軟體應用的密切結合，在產官學各界的帶動下，使得雲端服務越完整也越多元，已經到無所不在的境界了。電腦系統的運作主要為程式邏輯與資料儲存，雲端運算的架構改變了程式設計與執行的模式，也改變了資料存取的檔案系統架構，在 Amazon.com 平台上在網絡星期一收到的訂單超過 2800 萬件商品，Ebay.com 平台每年則有 660 萬戶的賣家在使用，如果沒有使用分散式運算與分散式儲存系統，將無法滿足使用者的需求。本研究是利用 ELK Stack 巨量資料分析軟體，透過 Logstash 讀取清洗後的資料並對資料類別進行過濾，輸出至 Elasticsearch 資料庫並建立索引，最終於 Kibana 顯示、分析結果。以 Ebay.com 購物網站為目標，蒐集使用者國籍資料並進行分析，以視覺化的方式呈現結果，讓管理人員能夠輕易解讀大量數據並對未來商務方向進行規劃。另外在簡易的查詢測試當中也發現 Elasticsearch 資料庫比一般的傳統資料庫快上非常的多。

關鍵字：雲端運算、分散式檔案系統、ELK Stack、電子商務、數據分析。

Abstract

Cloud computing is a very popular area in recent years, by enhancing the performance of the hardware and software applications closely, in industry, government and academic communities, making cloud services more complete and the more diverse, has come to the omnipresent realm. Mainly for the operation of the computer system program logic and data storage, cloud computing architecture has changed the programming and execution model, also changed the data access file system architecture. For example, on the Amazon.com more than 28 million items were received on the Cyber Monday, and there is about 6.6 million sellers in the Ebay.com, examples of these cloud services. Without the use of distributed computing and distributed storage system, unable to meet the needs of users. This study uses ELK Stack, a big data analysis software, read and basic filter the data by Logstash, than output to Elasticsearch to made index for the data, finial display on Kibana, let user can read data by visualize. We use Ebay.com as target, to analysis costumer' s nationality information to set managers easily interpret large of data and plan the future business direction.

Keyword：Cloud Computing, Distributed File System, ELK Stack, E-commerce, Data analysis.

致謝辭

本言就得以順利完成，承蒙恩師楊朝棟博士悉心指導，始能順利完成。老師盡心於對學生課業上的指導，在文獻蒐集與問題分析中給了學生許多概念及觀念挹注，使學生在各方面都有長足的進步、豐富的收穫，學生銘感五內，在此致上最高的謝意。也感謝台灣新蛋股份有限公司的協助，讓我有此機會研究爬蟲及大數據相關方面的題目。

同時感謝資工所每過老師的指導與鞭策及實驗室的舜文、智凱、泳盛、廣欽，與資工所的同學宇倫，各位的協助與相伴，使我研究所的生活更加豐富與充實，也感謝一眾好友家瑞、韋呈、鐘南、士益、景義的支持與鼓勵。

最後非常感謝親愛的父母、家人，當初支持我重回學校念研究所，並在這期間不斷付出愛與關懷，才能讓我得到想要的知識與能力，在此感謝大家並分享我的成果與喜悅，謝謝你們！

Table of Contents

List of Figures

List of Tables

Chapter 1

簡介

1.1 研究動機

網路的出現促使分散式運算的迅速發展，從 1990 年代為了解決大量計算問題而採用的網格計算開始，到現今的雲端運算，而海量資料需要的大量儲存空間是雲端運算平台中的一個重要議題，分散式檔案系統是儲存與分析海量資料的當然選擇，不但 Amazon 及 eBuy 的大數據 [1]，甚至根據 IDC(International Data Corporation) 的研究報告，2011 年全球數位資料的使用量約為 1.8 ZB，並預測 2020 年的總量為 35.2 ZB 之多 [2]。

在雲端運算及巨量資料蓬勃發展的時代，巨量資料的定義談的不僅僅是資料量 (Volume)，還包含了時效性 (Velocity)、多樣性 (Variety) 及可疑性 (Veracity) [3]。運用於巨量資料的技術成熟，有 Hadoop ecosystem [4] 與 Spark ecosystem [5] 這些巨量資料處理工具，不只提供分散式儲存空間更提供高效能處理技術，使得巨量資料的應用更為便利。

分散式檔案系統除了將資料分散的儲存在資料節點外，每個節點同時也具備分散式計算的能力，所以原本只能在單一主機處理的程式邏輯與資料處哩，可以分散到數以萬計的大量計算節點上運行處哩，等每個節點計算出結果後，再將處理完成的個別結果結合起來，產生最終的結果，這種運作方式有效率的

節省了資料存取與計算的排隊時間，提升了處理效能，例如 Google 搜尋在關聯式資料庫中無法達到這樣的效能與精準度。

分散式檔案系統另一個優點是高度的可擴充性，可以在任何時候增加節點，插上網路線，設定加入整個檔案系統，就能夠很容易的加入並且提供服務。如今的電子商務範圍非常的廣泛且從事相關行業的企業非常多，如果我們不善用分散式系統的可擴充性來動態增加我們的爬蟲，很有可能造成因為爬蟲數量的不足而導致漏掉重要的價格更新或是爬蟲本身的隱密性。

1.2 論文目標與貢獻

本論文運用 ELK Stack 也就是 Elasticsearch、Logstash、Kibana 這三個開源軟體組合而成的分析系統，對於電子商務的購物商場進行即時的分析與統計，提供給需要加入電子商務的族群，或電子商務購物網站的管理者，對競爭對手電子購物商場進行即時的分析與統計，可以將當天內價格更新頻率較快的商品即時呈現並依據商品的類別、排名、國別及關鍵字篩選後的結果自動繪製圖表，以一種簡單直覺的方式讓需要的人了解資訊，實驗的最後，會以實際的例子，來將使用者想要購買的商品進行分析，並且推薦給使用者，比較良好的賣家。

此系統架構更有別於過往應用於 syslog 收集與處理，不但可同時處理多種不同格式的開放性資料的收集與處理，更可以透過資料類別進行資料篩選過濾。本研究演示運用標準 Java 網絡爬蟲程序進行巨量資料收集，後利用 ELK Stack 這三個開源軟體組合而成的分析系統，並以分散式檔案系統架構運行。以 eBay.com 上的使用者國籍作為目標進行分析歸類，發現主要的使用者來自於美國，而使用者最少的國家為比利時，對於管理人員可以更針對的規劃未來計畫推進方向。

1.3 論文架構

本論文主要利用 ELK Stack 環境建置一個分析系統，運用視覺化套件將電子商務賣方資料的資料進行分析與呈現，其呈現結果可作為決策人員決策時參考依據。第一章簡述相關研究過程與結果，再於視覺化呈現的結果從中延伸出本論文的研究目標及方向；第二章介紹相關研究背景及其各項套件技術概念、本文中將使用之運用工具進行說明簡述；第三章節中將描述論文中使用的系統架構概觀與測試機器的資源配置規格；第四章實驗環境配置與實驗數據處理與結果呈現；最後於，第五章將對於本論文之實驗過程、結果，經彙整後作出最終結論，並且從中延伸出未來的研究方向。

Chapter 2

研究背景與相關研究

2.1 研究背景

2.1.1 巨量資料

大數據或稱為巨量資料 (Big Data) 就是是超過傳統數據庫系統處理能力的數據 [6]，大數據是一個抽象的概念。除了大量的數據外，還有一些其他的特徵決定了它與“海量數據”或“非常大的數據”的區別，大數據是一個龐大而且復雜的數據集合，使用一般的數據庫管理工具或傳統數據處理應用程序難以處理。這些挑戰包括資料的擷取，管理，儲存，搜尋，共享，傳輸，分析和視覺化。Big Data 4V 模型談的就是資料量 (Volume)、時效性 (Velocity)、多樣性 (Variety)、可疑性 (Veracity) 如 Figure 2.1，這 4V 就是巨量資料的四的特點: 大、快、雜、疑，而對大數據本質的觀察分析便從大型數據集中發現大量隱藏的價值，這些數據集是多樣的，複雜的，規模巨大的。而 4V 定義如下：[7] [8]。

- Volume：從不同來源產生的所有類型的數據量，並繼續擴大。收集大量數據的好處包括通過數據分析創建隱藏的信息和模式。如線上交易、網路搜尋等都會不斷產生資料。

- Variety：通過感測器，智慧型手機或社群網絡等收集的不同類型的數據，以結構化或非結構化格式收集的數據類型包括影片，圖片，文字，聲音和數據日誌，這些都很難以傳統關聯式的固定資料欄位架構來解決。[9]

- Velocity：數據傳輸頻率及處理效率。數據內容在不斷變化，為了收集及完善完整的數據，有時會導入以前存檔的數據或存放很久的數據，以及來自多個來源的即時數據等，例如用於市場預測，決策分析，那處理的時效如果太長就失去了預測的意義了，所以處理的時效對 Big Data 來說也是非常關鍵的。[10]

- Veracity：資料的可疑性是指我們如何從這麼大量的資料進行分析，如果資料的真實性本身就有問題，那分析出來的結果肯定也是有問題的，所以對於資料進行過濾，以去除異常或者偽造的資料。

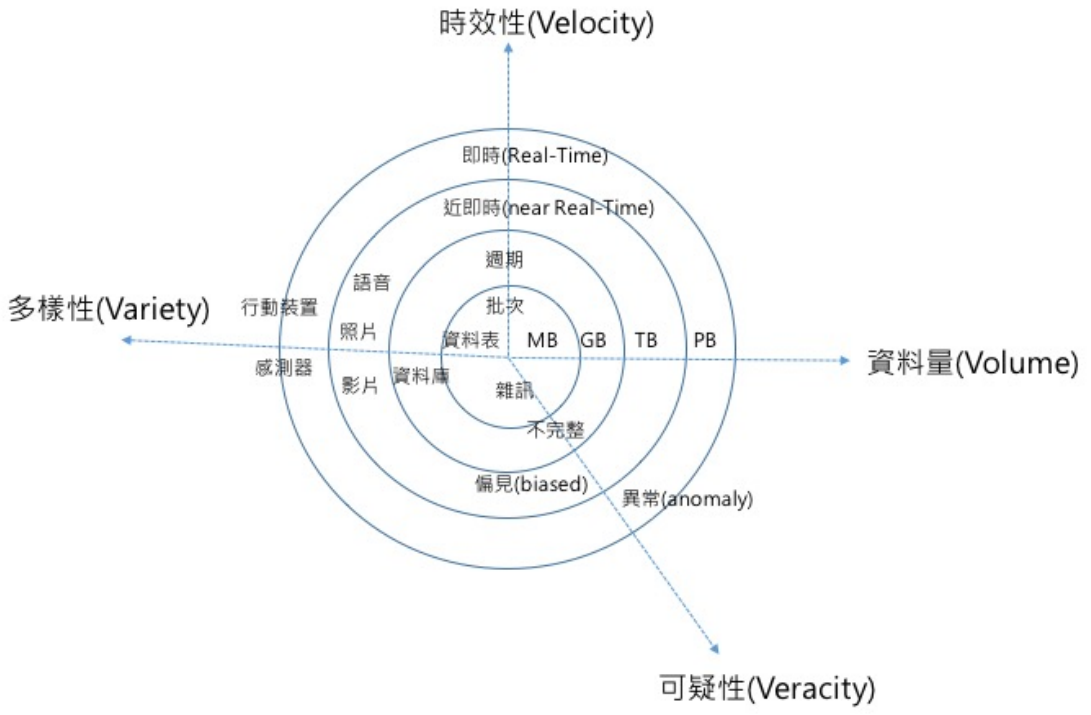


FIGURE 2.1: 巨量資料的 4V 定義

2.1.2 分散式檔案系統

分散式檔案系統 [11] 需由三台以上的電腦組成，其中至少有一台管理節點與兩台資料節點，各節點間透過網路連接，使用軟體管理各節點，讓多台用戶端的使用者可以共用檔案和儲存空間，某些大型的分散式檔案系統甚至多達上萬個資料節點。

分散式檔案系統的用戶端並不是直接存取實體硬碟所分割的資料區塊，而是透過網路，使用該分散式檔案系統提供的通訊協定來存取資料。分散式檔案系統具備資料複製與容錯的功能，即使再多個節點中有某一小部份的節點失效而離線，整個檔案系統仍然可以持續運作而不會造成資料遺失，用戶端也幾乎不會感覺任何異狀。

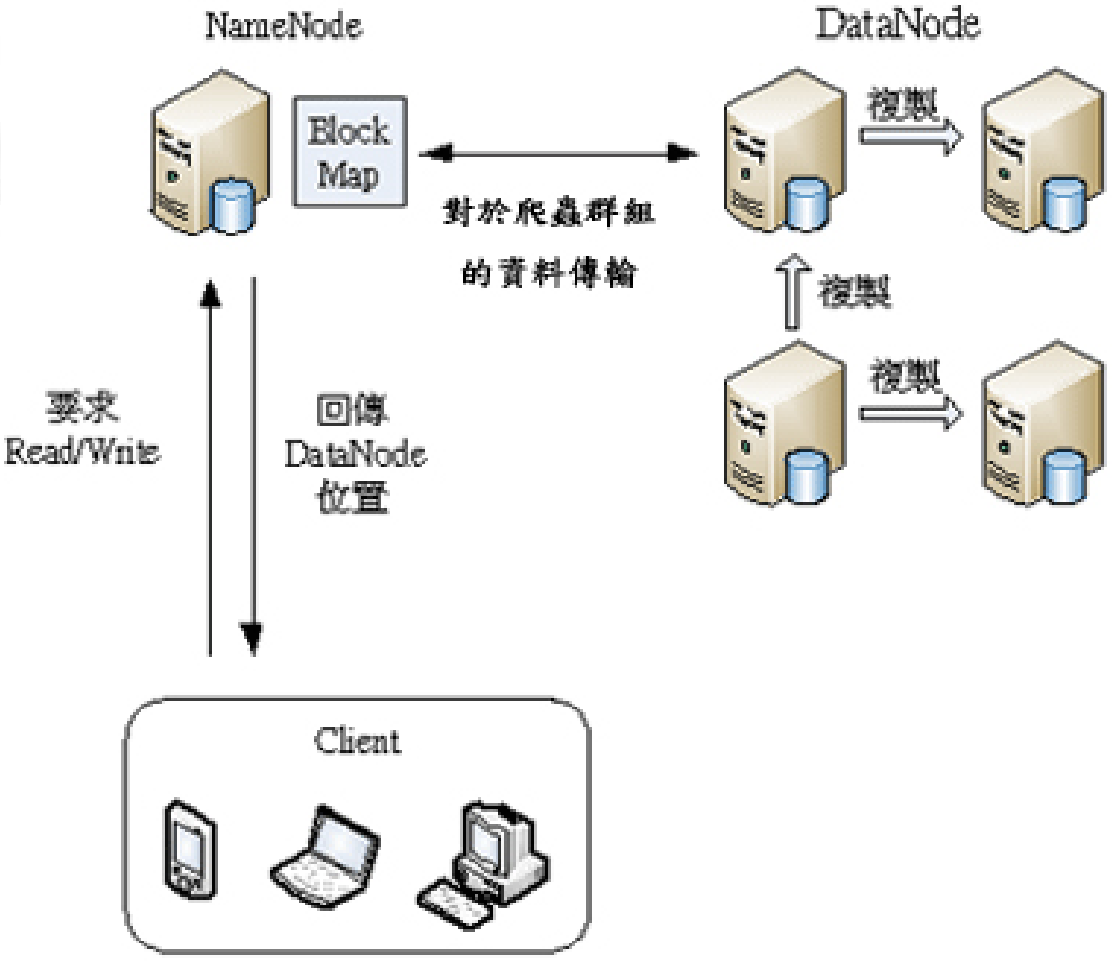


FIGURE 2.2: 分散式檔案系統

2.1.3 網路爬蟲

網路爬蟲（英語：Web Crawler），也叫網路蜘蛛（Spider），是一種用來自動瀏覽全球資訊網的網路機器人。其目的一般為編纂網路索引。網路搜尋引擎等站點通過爬蟲軟體更新自身的網站內容或其對其他網站的索引。網路爬蟲可以將自己所存取的頁面儲存下來，以便搜尋引擎事後生成索引供用戶搜尋。[12]

簡易的爬蟲大致是長的如下圖 4 ，由開發者進行調度，讓爬蟲能夠知道甚麼時候需要啟動，爬蟲起動之後內部則是三大步驟，第一是 URL 管理器，在此會將關注範圍的商品起點 URL 輸入，第二將網頁資訊放入下載器，也就是所謂將整個網站的資訊爬回來，第三就是網頁分析器，將網頁上的資料，配合需求，製做成資訊，最後傳到爬蟲使用者的資料庫。

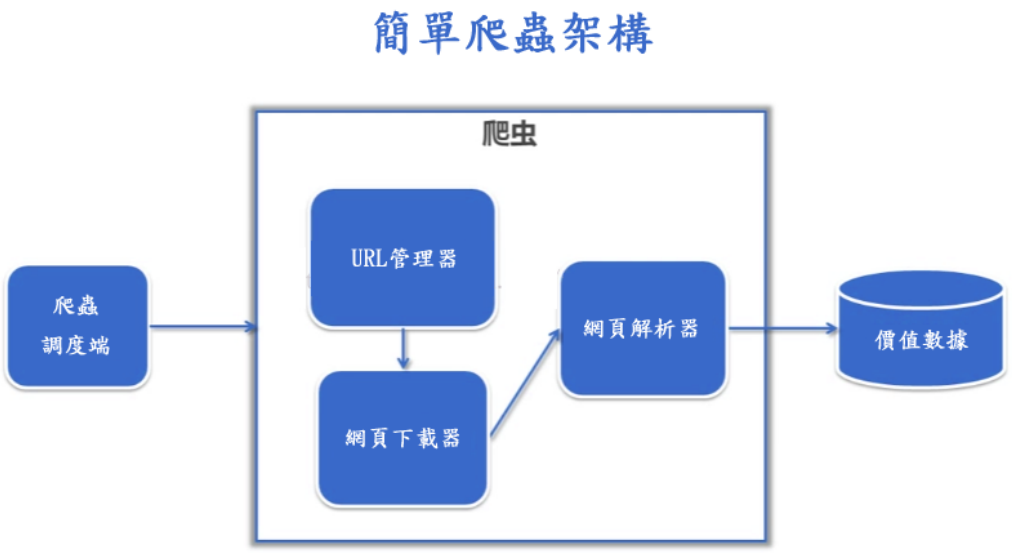


FIGURE 2.3: 簡易爬蟲架構

擁有爬蟲技術並不能成為競爭門檻，不論在日本或中國都有專門寫爬蟲的公司，台灣目前較傾向於地下化，但已經不在少數。對資料擁有者來說，想必是對於資料庫的價值利用不足，才會有人想來爬你的資料回去，所以資料擁有者該想的是要如何才能和這些來爬你資料的人共創價值，遠比費盡心機阻擋這些爬蟲來的重要多了。

Open Innovation 的概念告訴我們，研發不一定是自己。倘若我們反其道而行，將資料庫開放，告訴爬蟲們: ”別再爬了!! 這裡有 API 給你直接抓回去啦”。原本地下化的爬蟲為了獲得更大的利益，也許因而選擇公開合作，如此全世界就都是網站資料庫應用的開發者了，只要有一個可以像開心農場一樣的成功，獲益最大的相信會是網站本身。[13]

2.1.4 ELK Stack

ELK Stack [14] 是 Elasticsearch、Logstash、Kibana 這三個開源軟體的搭配如 Figure 2.4，各取頭一個英文字母的組合。在即時資料搜尋、檢索、分析和呈現場合，三者通常是相互搭配使用，而且目前都屬於 Elasic.co 公司所維護的開源軟體，故有此簡稱。ELK Stack 需要收集 Log 的服務伺服器上部署 Logstash，利用 Logstash agent 監控並過濾收集服務產生的 Log，將過濾後的內容發送到全文檢索搜尋服務 Elasticsearch，可以使用 ElasticSearch 進行自訂搜索通過 Kibana 來結合自訂搜索進行頁面展示。

ELK Stack 具有下列幾個優點 [15]：

- 處理及分析方式簡單靈活。
- 配置簡易上手，Elasticsearch 是採用 JSON 介面，Logstash 是採用 Ruby DSL 設計，都是目前最常見的配置語法設計。
- 檢索性能高效，每次查詢都是即時運算，基本上可以達到資料查詢的快速響應。
- 橫向擴展，Elasticsearch 可以叢集方式橫向擴展其儲存空間，並進達到資料複寫功能。
- 視覺化操作簡單，在 Kibana 界面上，只需要點擊滑鼠，就可以完成搜索、聚合功能，並產生淺顯易懂的儀表板。

FIGURE 2.4: ELK Stack 架構圖

Elasticsearch 分散式搜尋系統，提供搜集、分析、儲存數據等功能，具備 REST [28] 和 JAVA API 等傳輸架構提供高效率搜索功能，建構於 Apache Lucene 搜尋引擎庫上，也在 Apache 許可條款下以開放源始碼方式發佈，應用於雲端計算中，可達到即時搜索、穩定、可靠快速與安裝使用方便。 並且是一個用 Java 編寫的開源全文搜索引擎，它具有分佈式配置的能力。Elasticsearch 服務器易於安裝 [17]，預設的設定值就足夠獨立使用而不需要調整，但大多數用戶最終都希望調整一些參數。運行 Elasticsearch 服務的機器稱為節點，兩個或多個節點可以形成 Elasticsearch 叢集。要設置 Elasticsearch 叢集，需要在配置文件中設置的唯一值是叢集的名稱，Elasticsearch 會自動發現網絡上的節點並將其綁定到一個叢集中。[18]

Logstash 是一個用來蒐集、過濾、分析日誌及數據的開源工具，提供了各式各樣的日誌收集及各樣輸入、過濾及輸出的外掛模組 (Plugin)，任何類型的日誌幾乎都有支援 [19]，隨著版本的更新支援的檔案格式也更多樣化，並且能夠以多種方式輸出數據。Logstash 接收不同類型的日誌，例如系統日誌，Web 服務器日誌，錯誤日誌和應用程序日誌。這些通常分佈在使用不同格式的不同系統中。在儲存到分析數據資料庫之前，Logstash 可幫助用戶將數據解析成一種通用格式，如 Figure 2.5。此外，Logstash 提供了一種通過提供自己定義邏輯來解析日誌。[20]。Logstash 服務會依設定值，過濾日誌的雜訊並搜尋出符合的條件。當找到符合的條件時，則啟動輸出將訊息送到另一個 Logstash agent 或輸出到資料庫 [21]。

FIGURE 2.5: Logstash 輸入輸出示意圖

Log Analysis 日誌分析是理解日誌和提取有用信息的過程。其中一種開源日誌框架是由 Apache Software Foundation 在 Java 中開發的 Log4j [22]。它已經被用於記錄諸如時間戳，日誌級別，錯誤消息以及 IP 地址，用戶名稱和所請求的 URL 之類的資訊。一旦這些日誌被捕獲，然後進行日誌分析，就可以進一步對網站用戶上網行為進行分析與調查。

Log Collector 負責收集日誌，然後解析日誌，然後記錄 Windows 和 Linux 機器的事件。記錄、登錄、取消和安全相關事件的記錄。在安全日誌中，有助於為操作系統活動安全配置。從多個節點收集日誌後，對日誌進行格式正規化並發佈日誌，並且日誌收集器必須能夠處理大量，高速率和各種日誌 [23]。

Kibana 是一個開源的數據視覺化界面，用於即時彙總和繪製資料數據，起初 Kibana 有另一個名字叫 Elasticsearch Dashboard，隨著功能演進逐漸成為一個基於 Web 化圖形管理及資料呈現介面，可以用於搜索、分析及視覺化呈現存放於 Elasticsearch Index(索引) 中的數據。利用 Elasticsearch REST API 來檢索數據，使用者可以依據自己的需求創造個人化的視覺儀表板，通過提供不同的視覺化效果，如長條圖，原餅圖，線條點和地圖如 Figure 2.6，幫助使用者方便理解大量的數據集。它還提供了可以自行定義儀表板中的呈現不同可視化 [24]。

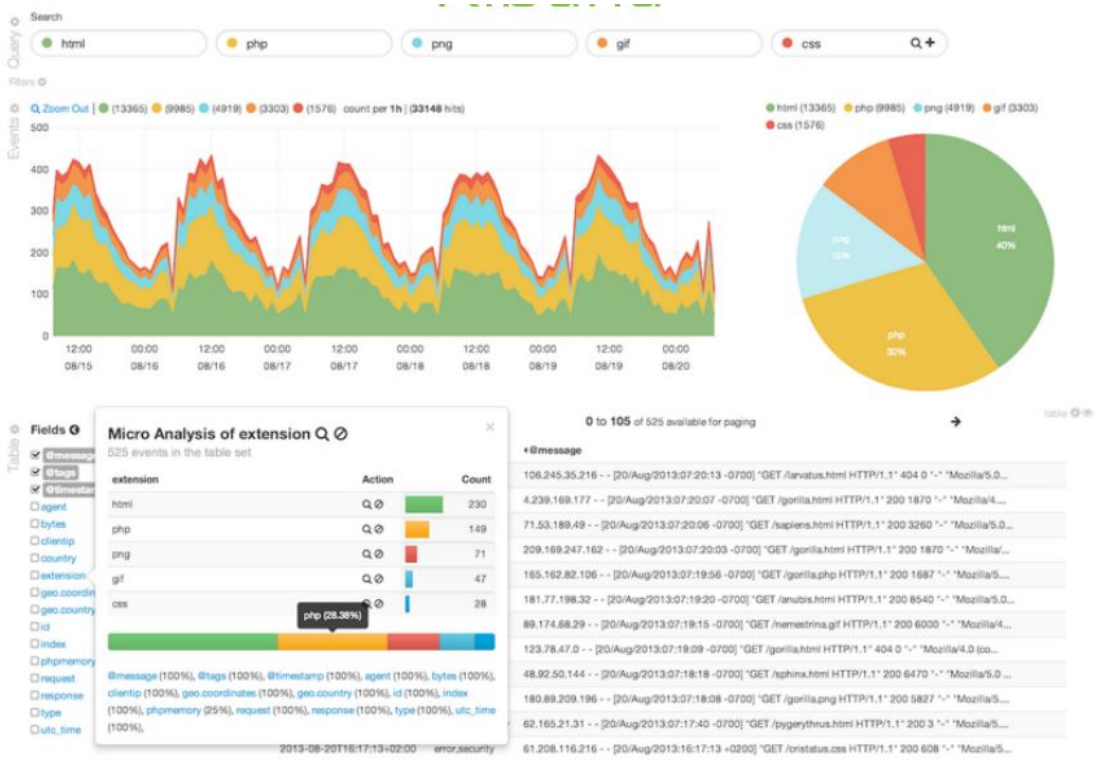


FIGURE 2.6: Kibana Dashboard 示意圖

2.1.5 商品資料蒐集爬蟲系統

資料主要來源為 Ebay.com 的官方網站，使用資料流程如下如 Figure 2.7：

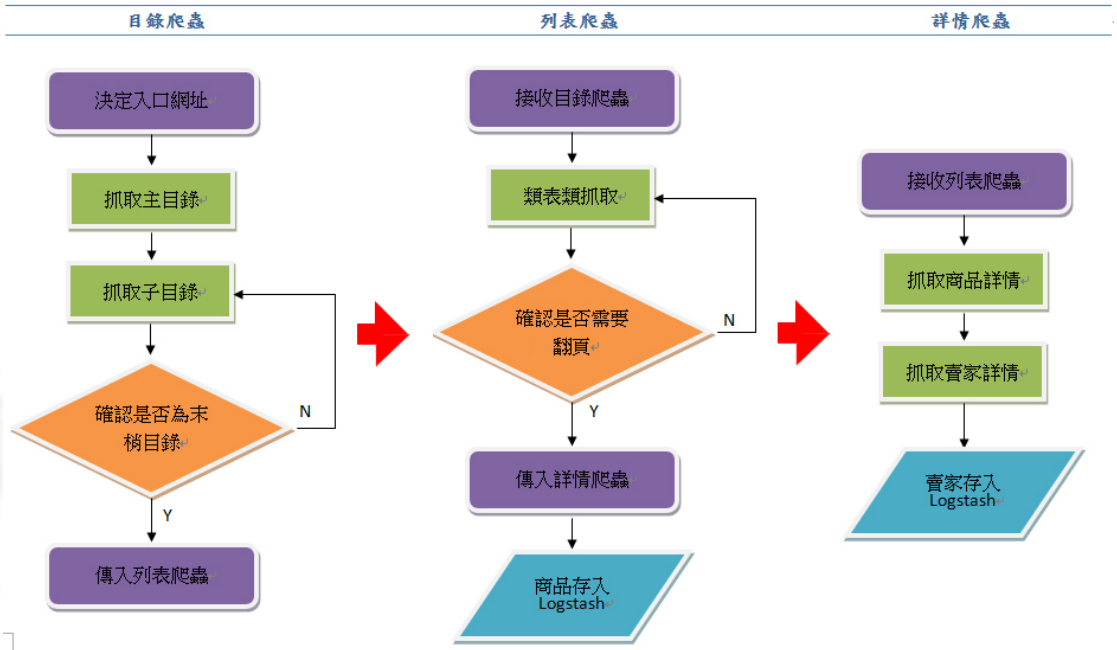


FIGURE 2.7: 爬蟲的資料流程判斷

因為每個站點的 site map 都不盡相同，所以通常必須要先探勘該網站，來知道該網站本身的架構，分析出，最適合爬蟲的移動路徑，例如某些網站 Sear.com [25]，或是 Amazon.com [26]，他們都會在列表頁面，設置一些翻頁上的陷阱，明明一頁有 40 個商品，總商品有 400000 個，理論上應該要有 10000 頁供使用者閱覽，但是實際上，能夠翻的頁數只有 400 頁而已，如果無法找出這些陷阱，很有可能浪費整整 9600 的流量，不但抓取速度會大幅下降，還有可能因為錯誤流覽量過大，而留下把柄在目標網站，進而遭到封鎖爬蟲的一些對策。並且因為網站的商品數量不見得一定大同小異的，所以也必須設定關注類別，已達到爬蟲的最大效益，因為一般在編寫爬蟲時，最在意的就是價格的更新速度，價格更新越是即時，對於自家商品的調價，就越有效益。所以如果將爬蟲資源浪費在不關注的類別上面，是非常不划算的。

所以我們會先讓目錄爬蟲進入對方的網站入口，然後目錄爬蟲就會將我們需求的資訊抓回給下一隻列表爬蟲，列表爬蟲會先將商品存入資料庫，並且再

將相對應的資料傳給詳情爬蟲，最後詳情爬蟲就會補齊商品及賣家的詳細資訊。

至於爬蟲的啟動，則是在各爬蟲機上安排排程系統，也就是所謂的爬蟲調度器，以此讓每台節點的 crawler 能夠按時啟動，去依照使用者需求，爬取相對應的任務資料。如 Figure 2.8

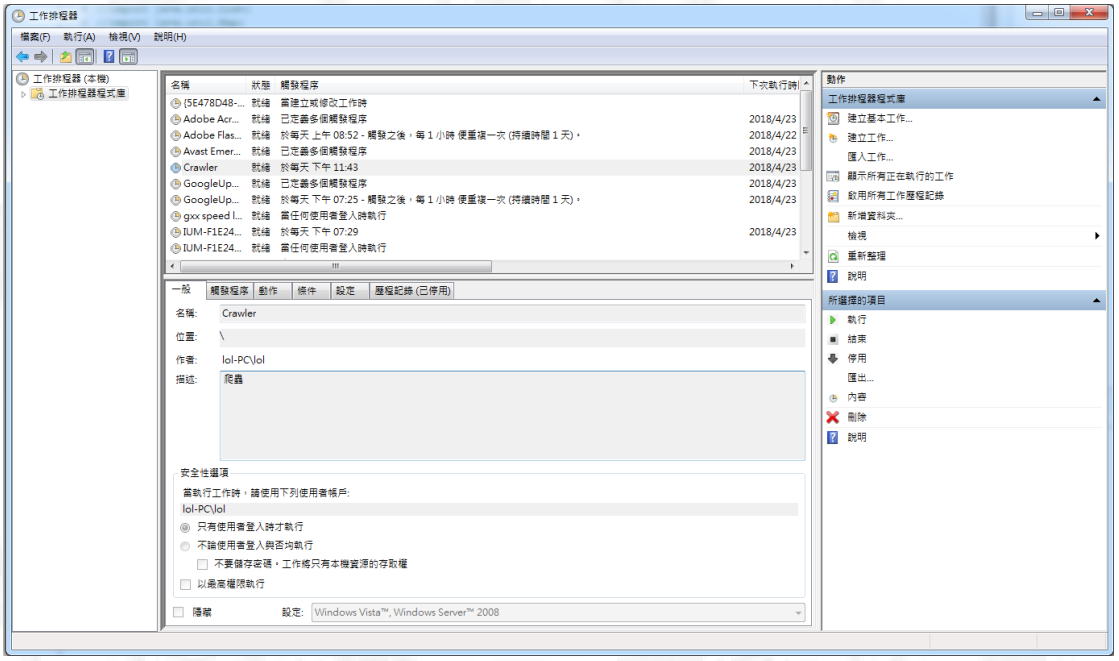


FIGURE 2.8: 爬蟲排程執行

啟動的各爬蟲軟體，會將頁面上的資訊進行資料化，並且藉由頁面上的一些資訊，如：產品名稱、賣家地點、價格、排名、類別.....，依照 CSV 的格式將其存入 Logstash 指定的空間位置，然後在藉由，Logstash 本身的功能，然後再依照分類、分別成有效的資料，來進行分析，如 Figure 2.9。

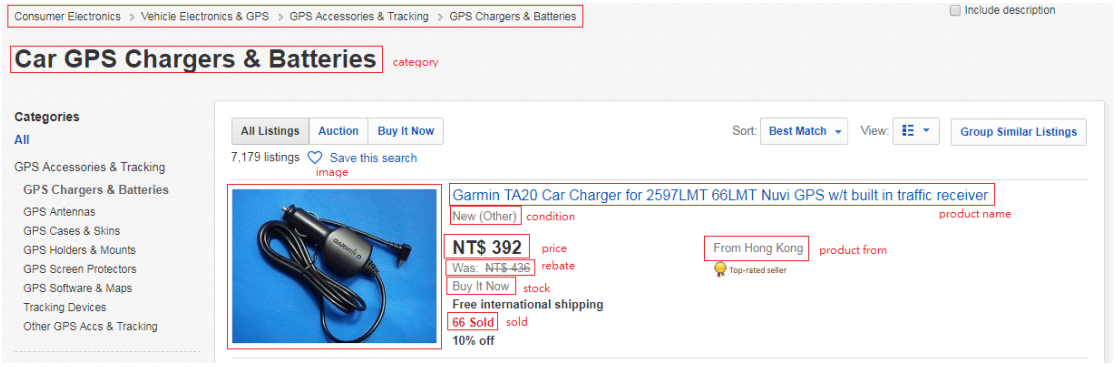


FIGURE 2.9: 商品資料化

2.1.6 電子商務

電子商務（Electronic Commerce, E-Commerce），指的是透過電子化技術（通常是網路）來進行交易、訊息傳遞的商業模式。

電子商務最早可以追溯至 1960 年代，企業間透過電子數據交換（Electronic Data Interchange, EDI, 又被稱為：無紙交易）的交易初步成形。這時，電子商務被定義為：「透過電子化簡化商業活動」。企業間用電子資料傳輸訂單、發票，或使用電子貨幣轉帳。1979 年，美國國家標準協會（American National Standards Institute）公認標準委員會（ASC）使用 X12 為 B2B 開發電子文件交換的統一標準。到了 1990 年代，eBay、Amazon 開啟了電子商務全新的時代。隨著信用卡及線上支付平台的普及，電子商務不再侷限於公司行號間的訊息傳送，傳統產業（Brick and Mortar）架設網路銷售平台，消費者則利用線上型錄及虛擬購物車進行採購。

EDI、文件傳輸協議（File Transfer Protocol, FTP）的訂定、網路服務及應用程式的發展，進一步推動了電子商務的普及。在網路平台上，消費者可以在網路上直接訂購、消費，企業也能利用用戶註冊的 E-Mai 信箱，將資訊直接發送至使用者信箱，或利用電子折價券、社群網站直接推播廣告訊息。同時，企業也能在平台上採集客戶的意見，更能透過客戶資料及交易內容，分析客群、消費習慣，並據此訂出行銷策略。

如今，電子商務的應用不再侷限於買／賣的貿易行為，包含電子銀行、資訊化的物流、倉儲管理等。電子商務的蓬勃發展與成長，也促使許多傳統的零售大廠（如：Costco、Safeway）也踏入了電子商務領域，電商領域進入了「磚塊加滑鼠」（Bricks and Clicks）的新時代。[27]

2.1.7 數據分析的實例

看看沃爾瑪 (Walmart)【啤酒 + 尿片】

TABLE 2.1: B2C 與 C2C 比較

購面	B2C	C2C
商品來源	商家提供	所有網友、小商家
產品線	特定或有限	多樣化
商品線深度	不錯	非常多樣化
欺偽	不多	有時發生
商品運送	不錯	品質不一
品質保證	不錯	需要規則提升
服務品質	依商家性質	由賣方提供

在 70 年代，美國的沃爾瑪超市有一個專賣啤酒的專櫃。門店經理從報表中發現：每個星期五晚上七八點的時候，啤酒的銷量非常好。而在這個特定的時間內，嬰兒尿片的銷量也非常好。經過抽驗調查後，這位經理髮現：因為第二次世界大戰剛結束，男人都出去工作，而女人則需帶小孩。週五晚上，男人便負責帶小孩去買尿布，然而這時間段有男人鍾愛的 NBA，因此他們便會順手帶回幾瓶啤酒。經過分析，沃爾瑪便將二者放在一塊，產生一加一大於二的效果。

目前不只沃爾瑪在使用爬蟲，還有非常多的商務網站及其他網站都友在使用爬蟲，如 Figure 2.10。

FIGURE 2.10: 誰在使用爬蟲技術

2.2 相關研究

隨著網路普及率的上升，網站瀏覽 log 交易量巨大，其 log 包含了巨大商業價值的隱藏信息，利用 ELK stack 開源軟體平台，可以有效地通過 log 來識別

網站用戶流量，對這些 log 的分析不僅有助於公司的決策，也有助於改善他們的產品和服務。然而，大多數小規模公司都無力負擔昂貴的日誌分析管理系統，ELK stack 提供了一個開源的日誌分析管理系統平台 [28]。

除了使用 Elasticsearch、Logstash、Kibana 這三個開源軟體的搭配外，也可以 Apache Flume 來代理收集日誌事件，存放於 Hadoop ecosystem 中的 Hbase，然後 ElasticSearch 根據搜索條件獲取 Hbase 數據並呈現於 Kibana 視覺儀表板 [29]。

ELK stack 也應用於系統監視，透過收集系統中的資源信息，例如：CPU 使用率、存儲器使用量、磁盤 I/O 使用等)，以便用於故障識別。從科學的分佈式系統收集信息，透過 SVM 分類演算法來做為故障檢測方法，準確率高達 90% [30]。

網絡服務已經在社會，商業，政府和學術界等各個領域成為不可或缺的一部分。雲端技術的出現，使得放在雲端的應用程序也可通過 Web 界面進行訪問和控制。因此，網絡安全是非常重要且具有挑戰性的。例如常見的安全問題就是 SQL-injection。現有用於檢測這些攻擊的大多數解決方案，大多都使用日誌分析，利用模式匹配方法來檢測這些攻擊類型，而這個日誌分析系統就是由 ELK Stack 組合而成 [20]。

義大利國家核物理研究所 (INFN) 杜林分部計算中心的私有雲，為各種不同的科學計算應用提供 IaaS 服務。基礎服務架構是由 OpenNebula 所構成的 [31]，除了追蹤資源使用情況之外，也需要為使用者動態分配資源，並對資源使用情況進行詳細的監控和計算。所以管理者建立了一個監控系統來檢查各節點活動，無論是 IaaS 還是託管虛擬機上運行的應用程序，這一個監控系統使用了 Elasticsearch，Logstash 和 Kibana 三個開源軟體。將不同的系統運作資訊送到不同的 MySQL 數據庫，並通過 Logstash 插件將資料送到 Elasticsearch，最後設置了一些預先定義查詢條件在 Kibana 儀表板上呈現，以便在每種情況下監控相關資訊。[32]。

Twitter 作為世界上用戶數最多的社交媒體之一，Twitter 提供了一個 API，使我們能夠即時觀看和使用 Twitter 數據。Elasticsearch 是一個有能力分析大數

據的工具。有兩種方法可以將 Twitter 數據輸入到 Elasticsearch。第一個是通過 Twitter River，第二個是通過 Logstash 這個元件。輸入數據的準確性和效率以及數據的存儲方式對於支持大數據系統是非常重要的。在這篇文章中，介紹了從 Twitter API 中輸入 Twitter 數據的 Twitter River 和 Logstash 性能評估。這項研究監測兩個 HPC 服務器上的 Elasticsearch 叢集，它們同時從 Twitter API 抓取數據。比較參數是 CPU 進程，RAM 使用率，空間使用率，Twitter 輸入數據和輸入字段數量。這項研究的結果顯示，Twitter River 平均每天的 CPU 進程為 33.96%，Logstash 為 34.95%。Twitter River 平均每天的內存使用率為 32.7%，而 Logstash 則為 39.9%。此外，Twitter River 平均每天的空間使用量為 431 MB，Logstash 為 544 MB。對於 Twitter 輸入數據，Twitter River 在一周內輸入比 Logstash 多 191 條推文。結果表明，Logstash 在很多數值上比 Twitter 原生的 API 程式 Twitter River 不理想。[33]

隨著物聯網普及率的提高，大量的日誌文件持續在成長，其中包含具有巨大商業價值的隱藏信息。為了挖掘這些隱藏的價值，日誌管理系統有助於做出業務決策。現在有許多的日誌管理系統的解決方案，但是他們有些不能橫向擴展或成本高昂。ELK 生態系統，即 Elasticsearch，Logstash 和 Kibana 聚集在一起，有效地分析日誌文件，並提供交互式且易於理解的見解。建立在 ELK Stack 上的日誌管理系統需要分析大型日誌數據集，同時通過交互式界面使整個計算過程易於監控。從開源社區開始，ELK Stack 有許多有用的功能用於日誌分析。Elasticsearch 被用作索引，儲存和檢索引擎。當 Kibana 使用儀表板執行數據可視化時，Logstash 充當日誌輸入切片器和切片機和輸出寫入器。通過實施 ELK 生態系統，我們有效地使用日誌對網站用戶流量進行地理標識 [34]。

Chapter 3

實驗環境與建置

3.1 實驗環境

實驗中環境以實體桌上型主機電腦建置，並使用虛擬化軟體 VMware Workstation，將 ELK Stack 中的三套分散式系統，Elasticsearch、Kibana 與 Logstash 分別建置在虛擬主機上，以節省實體主機成本且方便即時管理。

Table 3.1: 實體主機規格

CPU：Intel Core i5 3.10 GHz(64 位元)

硬體名稱	硬體規格
CPU	Intel Core i7-7700
Memory	8GB
Storage	256GB
System	Windows 7

ELK 是架構在 Lucene 是一套用於全文檢索和搜尋的開放原始碼程式庫。Lucene 提供了一個簡單卻強大的應用程式介面，能夠做全文索引和搜尋，在 Java 開發環境裡 Lucene 是一個成熟的免費開放原始碼工具；就其本身而論，Lucene 是現在並且是這幾年，最受歡迎的免費 Java 資訊檢索程式庫。而 Lucene 本身因為是全文檢索，所以他再輸入的時候，速度會非常不夠，甚至很多人，會因為將資料的輸入速度過慢，而不使用他，但是此篇論文的系統是讓爬蟲直接在網路上爬取資訊，所以系統本身輸入的速度是完全看網路速度，甚

TABLE 3.2: 虛擬化主機規格

虛擬硬體名稱	虛擬規格
Virtual machine ELK CPU core	1
Virtual machine ELK system	Ubuntu server 16.04.2
Virtual machine ELK memory	4GB
Virtual machine ELK DISK	100GB
Virtual machine Crawler CPU core	1
Virtual machine Crawler system	Ubuntu server 16.04.2
Virtual machine Crawler memory	2GB
Virtual machine Crawler DISK	20GB

至，在排名上的評價，被當成搜尋引擎。為此，此套電子商務賣方資料爬蟲相較於其他的分散式系統，更適合使用 ELK Stack 巨量資料分析軟體。

Rank			DBMS	Database Model
Sep 2017	Aug 2017	Sep 2016		
1.	1.	1.	Oracle	Relational DBMS
2.	2.	2.	MySQL	Relational DBMS
3.	3.	3.	Microsoft SQL Server	Relational DBMS
4.	4.	4.	PostgreSQL	Relational DBMS
5.	5.	5.	MongoDB	Document store
6.	6.	6.	DB2	Relational DBMS
7.	7.	↑ 8.	Microsoft Access	Relational DBMS
8.	8.	↓ 7.	Cassandra	Wide column store
9.	9.	↑ 10.	Redis	Key-value store
10.	10.	↑ 11.	Elasticsearch	Search engine

FIGURE 3.1: DBRank

實驗目的首要需求，是要將資料格式轉換為 Logstash 可讀資料前，要先將原始資料進行整理規劃及儲存，資料的前置作業往往是最費時的一項任務，所以為了使資料能快速的規劃與整理，結合了目前最為普遍的 Json 格式跟 CSV 檔的備用，以防萬一，如果資料沒有順利到主要的機器上面，可以在爬蟲機上面做一些基本上的備份，也可以讓爬蟲彼此的資料容易事後查詢除錯。

3.2 ELK Stack 與 Log 資料檔結合

Log 是記錄 HTTP Request 的重要記錄檔，我們可以從 Log 中得知爬蟲端對 Server 端的所有 HTTP 要求，Log 一般以檔案的型式儲存在磁碟中如 Figure 3.2，但是，我們不可能將這些 Log 一個一個點開來看，因為有些的連線失敗，或許只是對方站點的轉址，甚至是維修，因此要收 Log 就用監看檔案的格式，只要檔案有異動就把異動的部分往 Elasticsearch 傳輸，這樣不僅每次資料量不大，而且快又即時。[35] [36] [37]

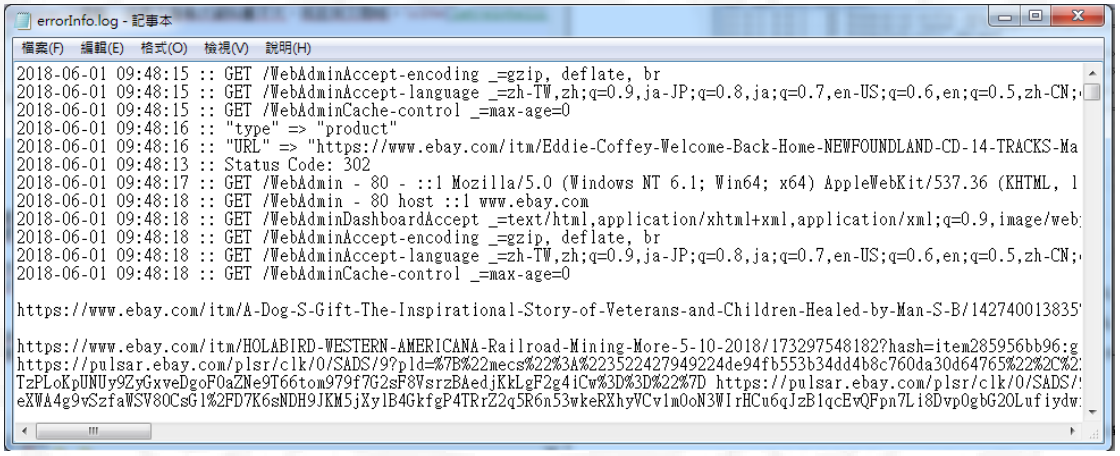


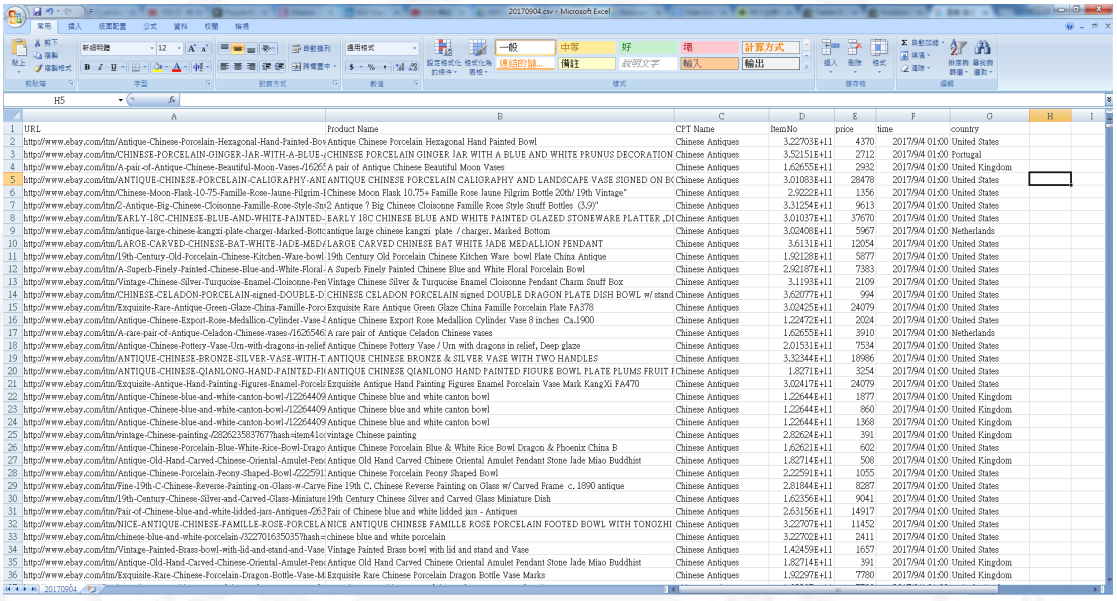
FIGURE 3.2: ERROR LOG

3.3 準備 CSV 或 JSON 資料

產生符合 JSON 標準字串，並將其存於 String 或 byte 類型的變數中，在每個不同類型的資料字串中，要將原始資料轉換為 JSON 檔案才可導入 Elasticsearch 中進行索引，當對產生的 JSON 格式文件進行索引時，可採用 Elasticsearch 索引資訊的方法。[38]

Elasticsearch 的索引、檢索、統計，在資訊處理的大致步驟是：以 Analyzer 分析結果為基礎建置 Query、給後台的 Elasticsearch 叢集發送、完成檢索並回傳結果，Elasticsearch 不僅可以透過 RESTful 等方式操作，透過不同程式語言的用戶端也可以使用。[20]

前端後端要進行資料交換是系統串階的重要環節，JSON 已成為顯學，不管是前端應用程式還是後端伺服器應用程式都已大大支持 JSON 格式。在開發階段要準備 JSON 資料，可以利用 JSON GENERATOR 極速來產生所需要的 JSON 測試資料。[39]



	URL	Product Name	CPT Name	ItemNo	price	time	country
2	http://www.ebay.com/itm/Antique-Chinese-Porcelain-Hexagonal-Hand-Painted-Bov	Antique Chinese Porcelain Hexagonal Hand Painted Bowl	Chinese Antiques	3.22703E+11	4370	2017/9/4 01:00	United States
3	http://www.ebay.com/itm/CHINESE-PORCELAIN-GINGER-JAR-WITH-A-BLUE-/	CHINESE PORCELAIN GINGER JAR WITH A BLUE AND WHITE PRUNUS DECORATION	Chinese Antiques	3.52151E+11	2712	2017/9/4 01:00	Portugal
4	http://www.ebay.com/itm/A-pair-of-Antique-Chinese-Beautiful-Moon-Vases-/16265	A pair of Antique Chinese Beautiful Moon Vases	Chinese Antiques	1.626555E+11	2932	2017/9/4 01:00	United Kingdom
5	http://www.ebay.com/itm/ANTIQUE-CHINESE-PORCELAIN-CALIGRAPHY-ANI	ANTIQUE CHINESE PORCELAIN CALIGRAPHY AND LANDSCAPE VASE SIGNED ON BC	Chinese Antiques	3.01983E+11	28478	2017/9/4 01:00	United States
6	http://www.ebay.com/itm/Chinese-Moon-Flask-10-75-Famille-Rose-Jaune-Pilgrim-I	Chinese Moon Flask 10.75+ Famille Rose Jaune Pilgrim Bottle 20th/ 19th Vintage"	Chinese Antiques	2.9222E+11	1356	2017/9/4 01:00	United States
7	http://www.ebay.com/itm/2-Antique-Big-Chinese-Cloisonne-Famille-Rose-Style-Snu	2 Antique ? Big Chinese Cloisonne Famille Rose Style Snuff Bottles (3.9)"	Chinese Antiques	3.31254E+11	9613	2017/9/4 01:00	United States
8	http://www.ebay.com/itm/EARLY-18C-CHINESE-BLUE-AND-WHITE-PAINTED-	EARLY 18C CHINESE BLUE AND WHITE PAINTED GLAZED STONEWARE PLATTER ,DI	Chinese Antiques	3.01037E+11	37670	2017/9/4 01:00	United States
9	http://www.ebay.com/itm/antique-large-chinese-kangxi-plate-charger-Marked-Bott	antique large chinese kangxi plate / charger. Marked Bottom	Chinese Antiques	3.02408E+11	5967	2017/9/4 01:00	Netherlands
10	http://www.ebay.com/itm/LARGE-CARVED-CHINESE-BAT-WHITE-JADE-MED/	LARGE CARVED CHINESE BAT WHITE JADE MEDALLION PENDANT	Chinese Antiques	3.6131E+11	12054	2017/9/4 01:00	United States
11	http://www.ebay.com/itm/19th-Century-Old-Porcelain-Chinese-Kitchen-Ware-bowl	19th Century Old Porcelain Chinese Kitchen Ware bowl Plate China Antique	Chinese Antiques	1.92128E+11	5877	2017/9/4 01:00	United States
12	http://www.ebay.com/itm/A-Superb-Finely-Painted-Chinese-Blue-and-White-Floral-	A Superb Finely Painted Chinese Blue and White Floral Porcelain Bowl	Chinese Antiques	2.92187E+11	7383	2017/9/4 01:00	United States
13	http://www.ebay.com/itm/Vintage-Chinese-Silver-Turquoise-Enamel-Cloisonne-Pen	Vintage Chinese Silver & Turquoise Enamel Cloisonne Pendant Charm Snuff Box	Chinese Antiques	3.1193E+11	2109	2017/9/4 01:00	United States
14	http://www.ebay.com/itm/CHINESE-CELADON-PORCELAIN-signed-DOUBLE-D	CHINESE CELADON PORCELAIN signed DOUBLE DRAGON PLATE DISH BOWL w/ stand	Chinese Antiques	3.62077E+11	994	2017/9/4 01:00	United States
15	http://www.ebay.com/itm/Exquisite-Rare-Antique-Green-Glaze-China-Famille-Roco	Exquisite Rare Antique Green Glaze China Famille Porcelain Plate FA578	Chinese Antiques	3.02425E+11	24079	2017/9/4 01:00	United States
16	http://www.ebay.com/itm/Antique-Chinese-Export-Rose-Medallion-Cylinder-Vase-/	Antique Chinese Export Rose Medallion Cylinder Vase 8 inches Ca.1900	Chinese Antiques	1.22472E+11	2024	2017/9/4 01:00	United States
17	http://www.ebay.com/itm/A-rare-pair-of-Antique-Celadon-Chinese-vases-/16265461	A rare pair of Antique Celadon Chinese vases	Chinese Antiques	1.626555E+11	3910	2017/9/4 01:00	Netherlands
18	http://www.ebay.com/itm/Antique-Chinese-Pottery-Vase-Urn-with-dragons-in-relief	Antique Chinese Pottery Vase / Urn with dragons in relief, Deep glaze	Chinese Antiques	2.01531E+11	7534	2017/9/4 01:00	United States
19	http://www.ebay.com/itm/ANTIQUE-CHINESE-BRONZE-SILVER-VASE-WITH-T	ANTIQUE CHINESE BRONZE & SILVER VASE WITH TWO HANDLES	Chinese Antiques	3.32344E+11	18986	2017/9/4 01:00	United States
20	http://www.ebay.com/itm/ANTIQUE-CHINESE-QIANLONG-HAND-PAINTED-FI	ANTIQUE CHINESE QIANLONG HAND PAINTED FIGURE BOWL PLATE PLUMS FRUIT I	Chinese Antiques	1.82711E+11	3254	2017/9/4 01:00	United States
21	http://www.ebay.com/itm/Exquisite-Antique-Hand-Painting-Figures-Enamel-Porcel	Exquisite Antique Hand Painting Figures Enamel Porcelain Vase Mark KangXi FA470	Chinese Antiques	3.02417E+11	24079	2017/9/4 01:00	United States
22	http://www.ebay.com/itm/Antique-Chinese-blue-and-white-canton-bowl-/12264409	Antique Chinese blue and white canton bowl	Chinese Antiques	1.22644E+11	1877	2017/9/4 01:00	United Kingdom
23	http://www.ebay.com/itm/Antique-Chinese-blue-and-white-canton-bowl-/12264409	Antique Chinese blue and white canton bowl	Chinese Antiques	1.22644E+11	860	2017/9/4 01:00	United Kingdom
24	http://www.ebay.com/itm/Antique-Chinese-blue-and-white-canton-bowl-/12264409	Antique Chinese blue and white canton bowl	Chinese Antiques	1.22644E+11	1368	2017/9/4 01:00	United Kingdom
25	http://www.ebay.com/itm/vintage-Chinese-painting-/252623583767?hash=item41c	vintage Chinese painting	Chinese Antiques	2.82624E+11	391	2017/9/4 01:00	United Kingdom
26	http://www.ebay.com/itm/Antique-Chinese-Porcelain-Blue-White-Rice-Bowl-Drago	Antique Chinese Porcelain Blue & White Rice Bowl Dragon & Phoenix China B	Chinese Antiques	1.62621E+11	602	2017/9/4 01:00	United States
27	http://www.ebay.com/itm/Antique-Old-Hand-Carved-Chinese-Oriental-Amulet-Pen	Antique Old Hand Carved Chinese Oriental Amulet Pendant Stone Jade Miao Buddhist	Chinese Antiques	1.82714E+11	508	2017/9/4 01:00	United Kingdom
28	http://www.ebay.com/itm/Antique-Chinese-Porcelain-Peony-Shaped-Bowl-/222591	Antique Chinese Porcelain Peony Shaped Bowl	Chinese Antiques	2.225911E+11	1055	2017/9/4 01:00	United States
29	http://www.ebay.com/itm/Fine-19th-C-Chinese-Reverse-Painting-on-Glass-w-Carve	Fine 19th C. Chinese Reverse Painting on Glass w/ Carved Frame c. 1890 antique	Chinese Antiques	2.81844E+11	8287	2017/9/4 01:00	United States
30	http://www.ebay.com/itm/19th-Century-Chinese-Silver-and-Carved-Glass-Miniatur	19th Century Chinese Silver and Carved Glass Miniature Dish	Chinese Antiques	1.62356E+11	9041	2017/9/4 01:00	United States
31	http://www.ebay.com/itm/Pair-of-Chinese-blue-and-white-lidded-jars-Antiques-/263	Pair of Chinese blue and white lidded jars - Antiques	Chinese Antiques	2.63156E+11	14917	2017/9/4 01:00	United States
32	http://www.ebay.com/itm/NICE-ANTIQUE-CHINESE-FAMILLE-ROSE-PORCELA	NICE ANTIQUE CHINESE FAMILLE ROSE PORCELAIN FOOTED BOWL WITH TONGZHI	Chinese Antiques	3.22707E+11	11452	2017/9/4 01:00	United States
33	http://www.ebay.com/itm/chinese-blue-and-white-porcelain-/322701635035?hash=	chinese blue and white porcelain	Chinese Antiques	3.22702E+11	2411	2017/9/4 01:00	United States
34	http://www.ebay.com/itm/Vintage-Painted-Brass-bowl-with-lid-and-stand-and-Vase	Vintage Painted Brass bowl with lid and stand and Vase	Chinese Antiques	1.42459E+11	1657	2017/9/4 01:00	United States
35	http://www.ebay.com/itm/Antique-Old-Hand-Carved-Chinese-Oriental-Amulet-Pen	Antique Old Hand Carved Chinese Oriental Amulet Pendant Stone Jade Miao Buddhist	Chinese Antiques	1.82714E+11	391	2017/9/4 01:00	United Kingdom
36	http://www.ebay.com/itm/Exquisite-Rare-Chinese-Porcelain-Dragon-Bottle-Vase-M	Exquisite Rare Chinese Porcelain Dragon Bottle Vase Marks	Chinese Antiques	1.92297E+11	7780	2017/9/4 01:00	United Kingdom

FIGURE 3.3: 資料來源

本系統，不管是 JSON 還是 CSV 都是可以使用的，所以不管是即時的從指定資料夾中讀取 CSV 或是 JSON 檔案，又或是直接從 Redis 底下直接抓取，都是沒有問題，所以我們在爬蟲邊爬取的時候，就會邊將資料一筆一筆的順便讀入我們的 ELK 系統底下。

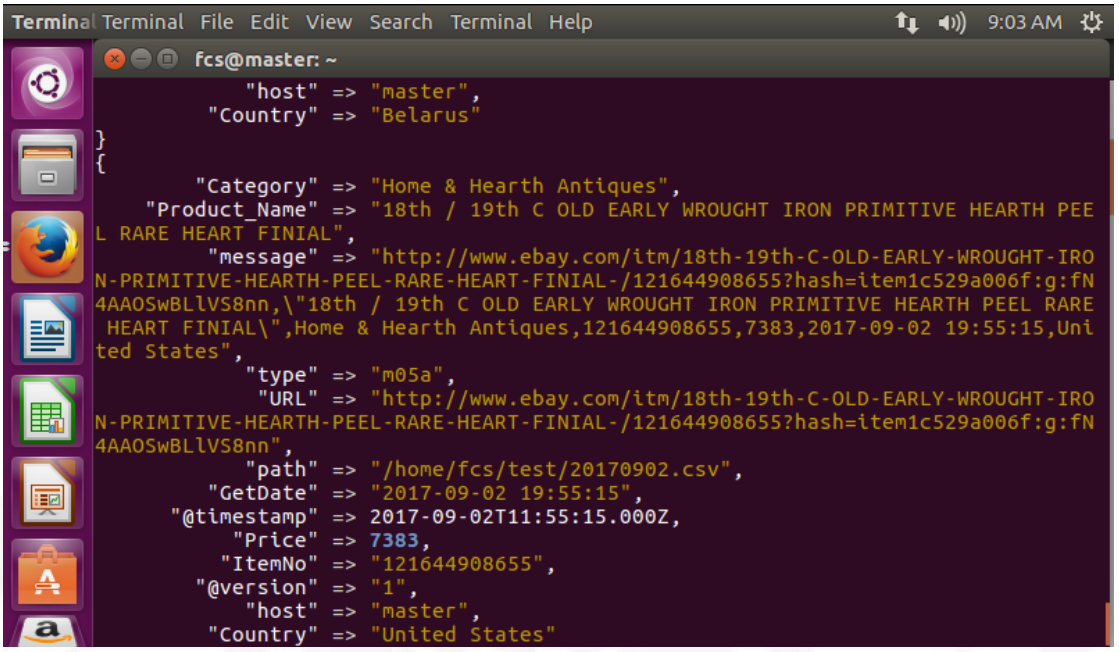


FIGURE 3.4: 資料灌入

3.4 系統架構圖

本系統整合架構如 Figure 3.5，透過排程來定期定時啟動爬蟲軟體，抓取 eBay.com 資料，經過資料清洗及整理，由 Logstash 讀取後再經過資料類型轉換和欄位處理，將資料輸出並儲存於 Elasticsearch，最後由 Kibana 分析過濾使用者需求的資料，即時以圖表呈現。

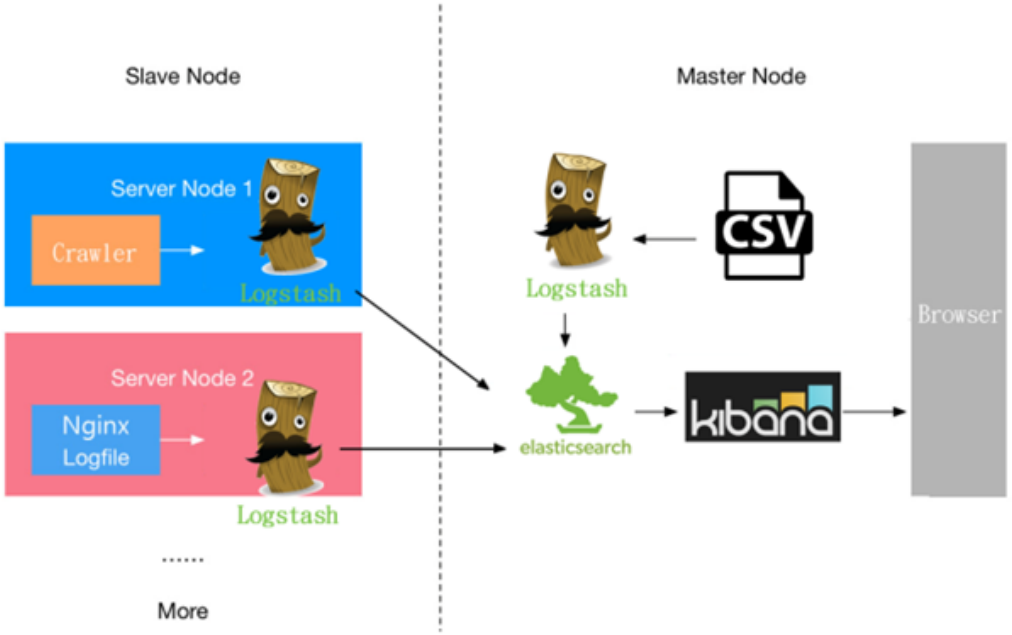


FIGURE 3.5: System

3.5 任務順序

我們的爬蟲在爬取一個敵對站點的時候，都會如前面所說，先決定網站的進入點（site map），而我們這次的 eBay.com 則是直接由下方的 site map 進入，如 Figure 3.6

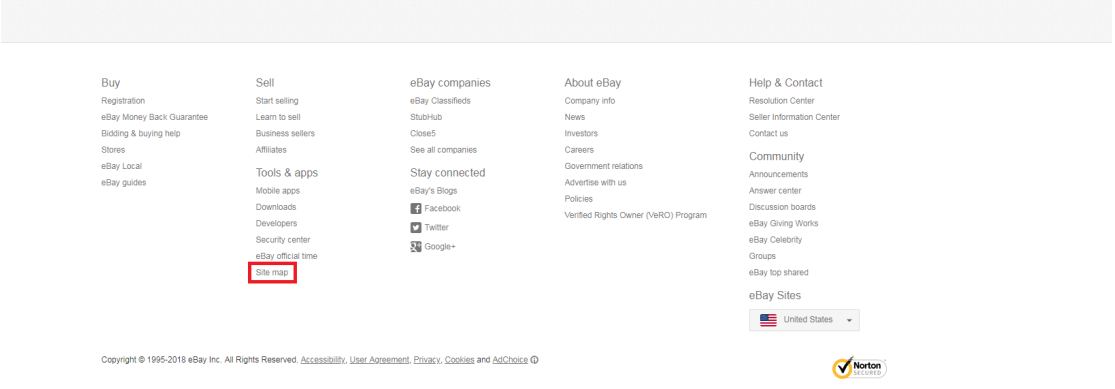


FIGURE 3.6: site map

接著在進入 site map 之後，這邊大多會有這個站點的所有類別分類，（All category）在這下方紅框框起來的，皆為爬蟲的一段目標，爬蟲會將這些類別

的 URL 以及相關的資訊，全部抓下來，並且將這些資訊傳遞給後方的任務。如 Figure 3.7

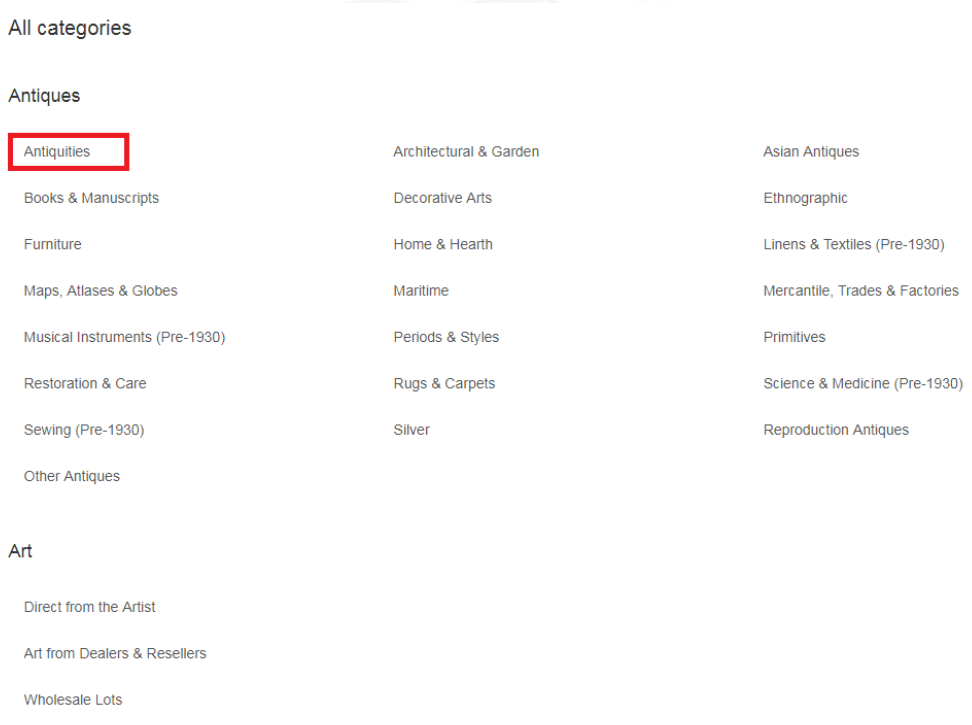


FIGURE 3.7: end category

在 category 爬蟲確認是末梢目錄，也就是所謂的對下端的網頁的時候，此爬蟲的任務就算是結束了，他會將右方商品列表的部分的資料傳遞給 List 爬蟲，這隻爬蟲則會負責將列表的商品全部翻完，並且對其資訊做更新。如 Figure 3.8

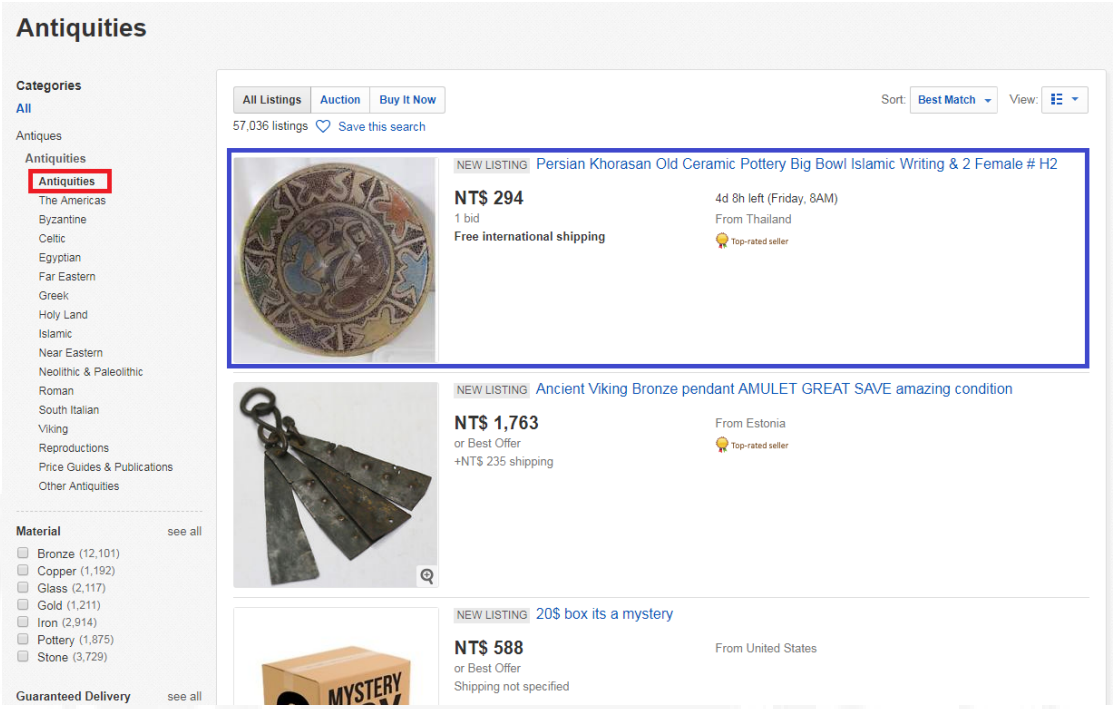


FIGURE 3.8: product list

接著，當 List 爬蟲將所有商品的 URL 抓取回來之後，他會將這些資料傳遞給 Info 爬蟲，這隻爬蟲則是負責爬取一個商品的詳情資料，他主要是將價格以及運費甚至是商品型號傳回，這些東西都是非常重要而且即時的，而且因為每個商品的價格更新速度不一，所以唯獨這隻爬蟲會在較短的時間內，重複一直的執行，根據過往經驗，在硬體允許的情形下，最好是一個商品能夠擁有一天更新 20 次的頻率是最好的。如 Figure 3.9

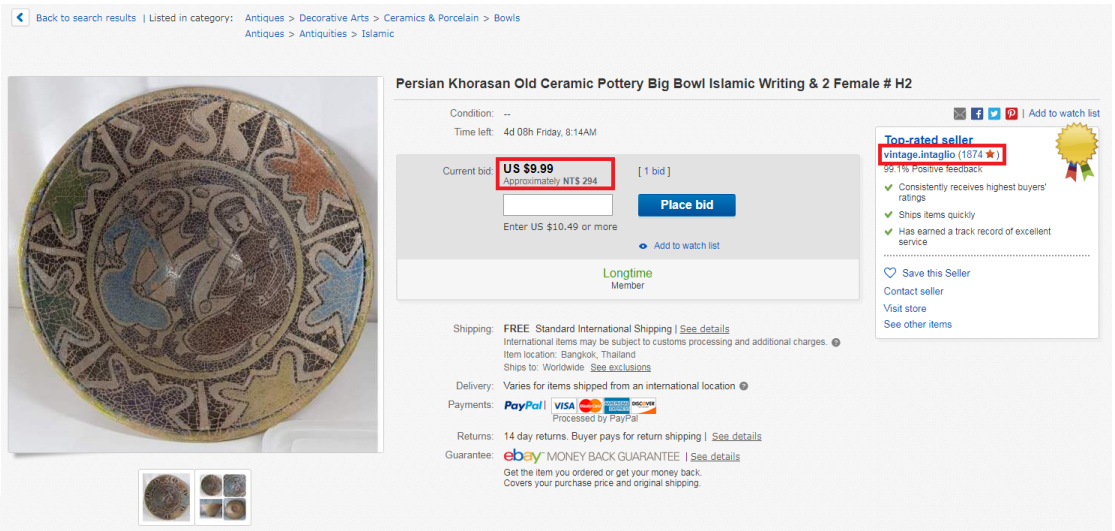


FIGURE 3.9: product Info

最後，也是這個任務比較在意的地方，就是怎樣的賣家，才能夠得到比較多的買家青睞，所以我針對此點，而近來這的賣家相關資訊的網頁。如 Figure 3.10

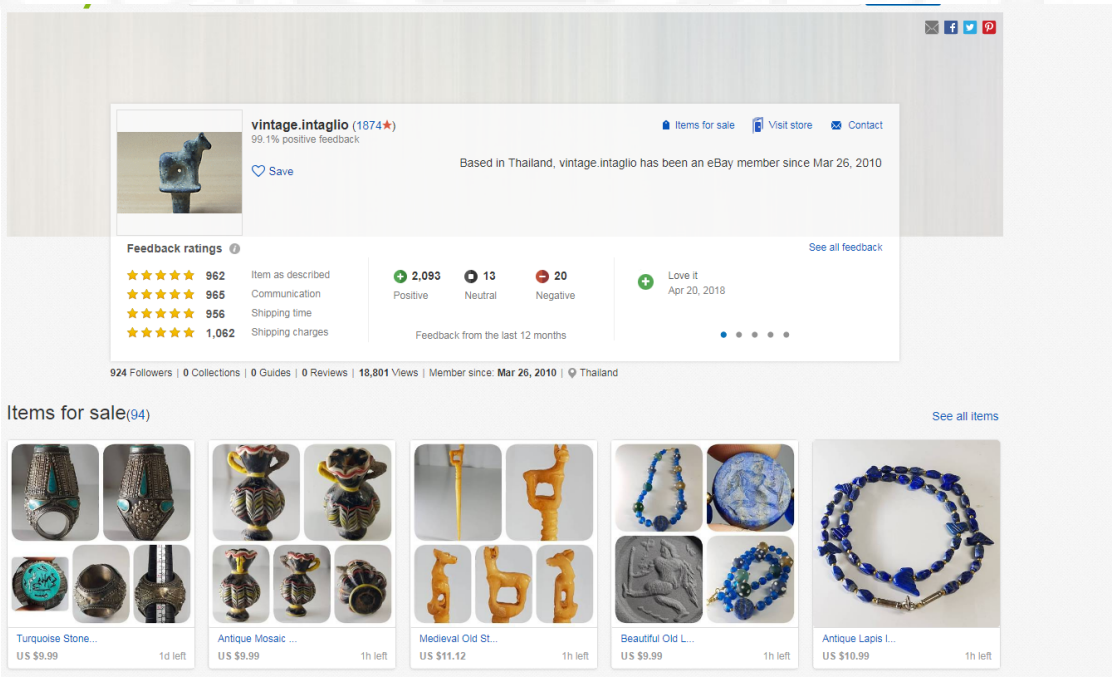


FIGURE 3.10: seller Info

3.6 文件索引

在資訊檢索過程中，文字資訊首先要經過加工處理後才能被檢索到，而這個加工處理過程就是建立索引檔案，一般的資訊檢索過程是使用者透過介面傳送查詢請求，在索引檔案中檢索出相關結果，並按相關度排序之後傳回給使用者，建立文件索引是最基礎的加工過程。[37]。由於檢索通常要針對大量使用者的查詢，因此索引檔案的設計要儘量高效，以便由索引項目快速找出到對應文件。[38]

TABLE 3.3: 關連式資料庫比較

相關概念在關聯式資料庫和 Elasticsearch 對應關係	
關聯式資料庫 (如 SQL 如 SQL Server)	Elasticsearch
資料庫 Database	索引 Index，支援全文檢索
表 Table	類別 Type
資料行 Row	文件 Document，但不需要固定結構
資料列 Column	欄位 Field
模式 Schema	映射 Mapping

3.7 安裝 Nginx

Nginx 是開源的、有效率的 Web 伺服器，支援 HTTP、HTTPS、SMTP、POP3、IMAP 等協定，使用 Logstash 來監聽 Nginx 紀錄檔，建立一個模式 (pattern) 來比對 Nginx 的 access.log 的模式檔案，啟動 Elasticsearch，使用指令./logstash -f conf.conf 啟動 Logstash，然後就可以利用 Elasticsearch 的 Head 工具看到 Nginx 的紀錄檔資訊已經進入到 Elasticsearch 的索引中 [20][21]。

```
Unpacking nginx (1.4.6-1ubuntu3.4) ...
Processing triggers for man-db (2.6.7.1-1ubuntu1) ...
Processing triggers for ureadahead (0.100.0-16) ...
Processing triggers for ufw (0.34~rc-0ubuntu2) ...
Setting up libapr1:amd64 (1.5.0-1) ...
Setting up libaprutil1:amd64 (1.5.3-1) ...
Setting up fontconfig-config (2.11.0-0ubuntu4.1) ...
Setting up libfontconfig1:amd64 (2.11.0-0ubuntu4.1) ...
Setting up libjpeg-turbo8:amd64 (1.3.0-0ubuntu2) ...
Setting up libjpeg8:amd64 (8c-2ubuntu8) ...
Setting up libjbig0:amd64 (2.0-2ubuntu4.1) ...
Setting up libtiff5:amd64 (4.0.3-7ubuntu0.4) ...
Setting up libvpx1:amd64 (1.3.0-2) ...
Setting up libxpm4:amd64 (1:3.5.10-1) ...
Setting up libgd3:amd64 (2.1.0-3) ...
Setting up libxslt1.1:amd64 (1.1.28-2build1) ...
Setting up apache2-utils (2.4.7-1ubuntu4.9) ...
Setting up nginx-common (1.4.6-1ubuntu3.4) ...
```

FIGURE 3.11: Nginx 安裝流程

140.128.197.54:9200


```
{
  "name" : "master",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "8KEHyNe5T-O6143q7BquYQ",
  "version" : {
    "number" : "5.5.2",
    "build_hash" : "b2f0c09",
    "build_date" : "2017-08-14T12:33:14.154Z",
    "build_snapshot" : false,
    "lucene_version" : "6.6.0"
  },
  "tagline" : "You Know, for Search"
}
```


FIGURE 3.12: 重新啟動 Logstash

3.8 Nginx 功用

Nginx 在此系統還有著輔助的功能，他能夠管理爬蟲機上的 Tor，主要是讓用戶通過 Tor 可以在網際網路上進行匿名交流，為此，可以增加爬蟲本身的隱密度，現在的電子商務，不只是要使用爬蟲，避免自己的資訊過於落後、價格過高，也要同時防止，自己的價格相關資訊，被他人盜取，有些甚至會放大量的爬蟲，來造成伺服器本身的反應過漫，讓真的使用者感到不好的購物體驗。

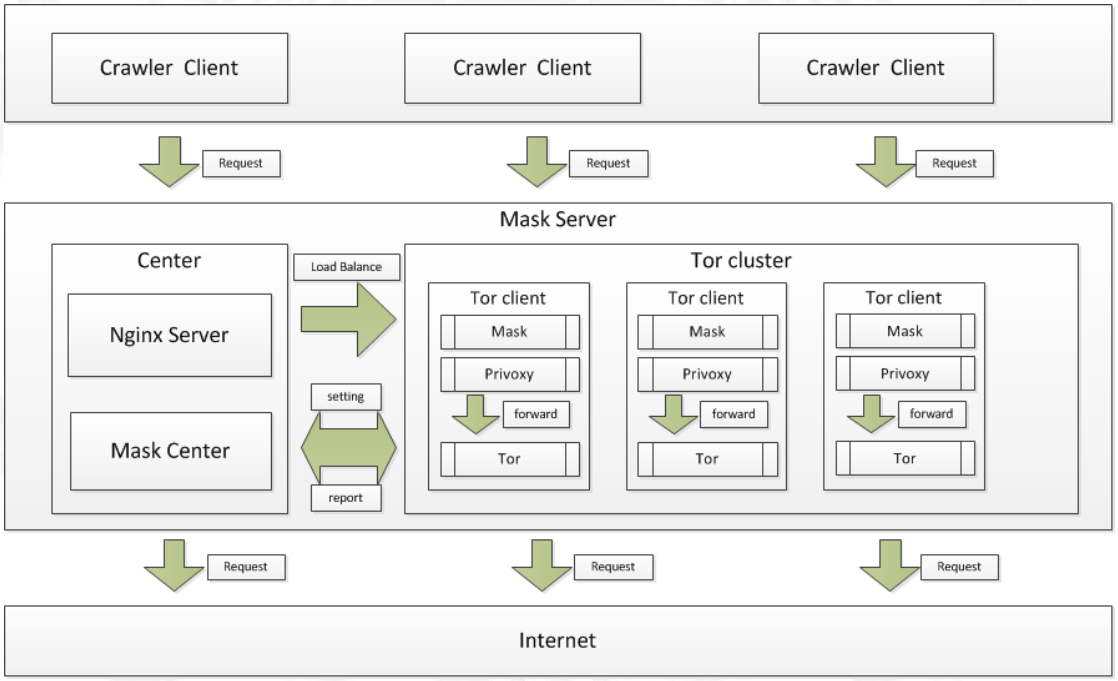


FIGURE 3.13: 含 TOR 架構圖

如何找出爬蟲，以下是可能的歸類：1. 抓取頻度過快 2. IP 變換頻繁 3. IP 不變且規律性訪問 4. IP 和登錄賬號不匹配 5. 國家限制

反爬蟲解決方案：1. 平台可配置訪問任務數以及線程數 2. IP 不變且規律性訪問 3. 爬蟲機設置代理進行抓取 4. IP 和登錄賬號不匹配 5. 賬號與爬蟲機 IP 綁定 6. 國家限制 7. 代理中心可針對代理設置國家 8. 圖形驗證 9. Auto Pricing 通過業務規則 + 爬蟲機快照分析

Chapter 4

實驗環境與結果

4.1 實驗介面與環境

Elasticsearch 的系統資源力道是較大的，許多 Java 開放原始碼系統，提供的豐富外掛程式可以有效地幫助 Elasticsearch 初學者和進階使用者有效使用 Elasticsearch。

4.2 Kibana 環境架構

實驗中，在 Kibana 中建置了許多欄位，進入 Kibana 的 Visualize 圖形呈現，建置了各種資訊。

在圖表數據中，X 軸屬於時間軸，代表所有資料的月份時間，Y 軸屬於每單日的資料量，以文具類別的資料為例，資料時間由 8 月份到 9 月份，共 1 個月的資料量進行分析，透過圖表，明顯得知 8 月底資料量將較其他月份資料為多，因此即可藉由圖表資訊瞭解，在八月份適合舉辦開學商品相關的促銷活動，以利於往後決策上的參考。

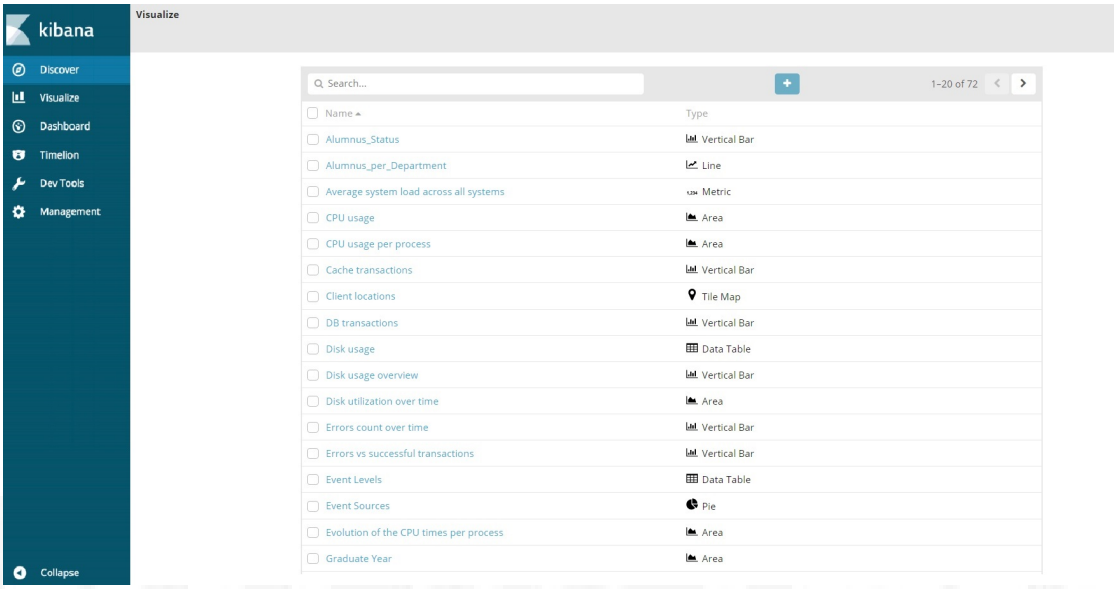


FIGURE 4.1: Kibana 圖表畫面

主功能選單，Kibana 的主功能列表包含了以下功能：

Discover：用以檢視各索引下的記錄內容及總記錄筆數。

Visualize：將搜尋出的數據以長條圖、圓餅圖、表格等方式呈現。

Dashboard：將以儲存的搜尋結果或已完成的圖表組合成一份快速報表。

Timelion：時序性的監看 query。

Dev Tools：提供一個在 Kibana 直接呼叫 Elasticsearch 的方式。

Managment：設定 Kibana 對應的 Elasticsearch index patterns，管理已經儲存好的搜尋結果物件、視覺化結果物件，及進階資料過濾設定。

Kibana 右上可以選擇要搜尋的時間區段，這一個區段主要是以 Elasticsearch 中的@ timestamp 為依據，而這個欄位預設是紀錄 Elasticsearch 寫入資料庫的時間，而在本文的分析系統中，特別將此欄位改成偵測站觸發的時間。Kibana 預設的時間工具列可以利用以現在時間為主往前計算，或著選擇特定的時間點如 Figure 4.2與 Figure 4.3。

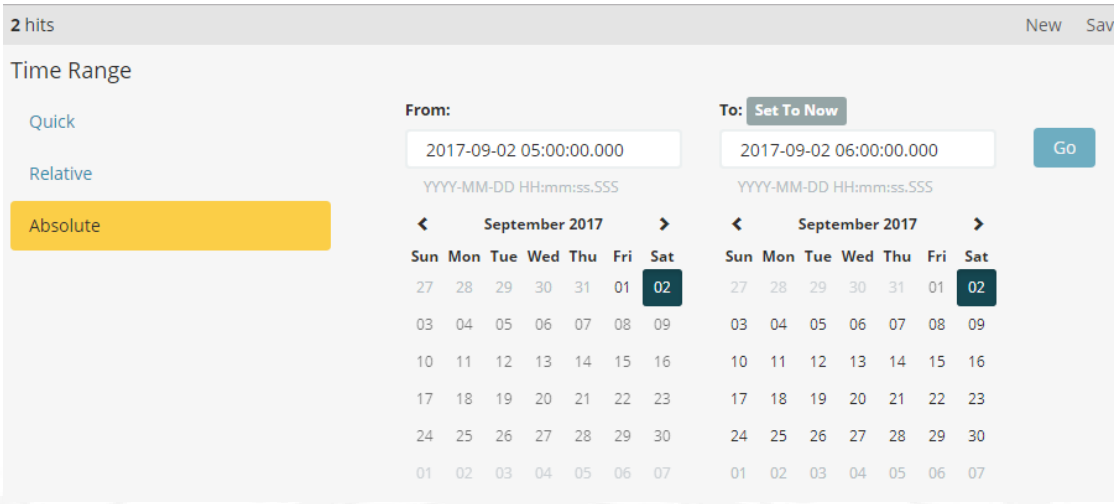


FIGURE 4.2: Absolute

FIGURE 4.3: Quick

Visualize 主要可以利用範本來建立視覺化圖表如 Figure 4.4，這些建立的圖表可以置放於 Dashboard 上呈現。預設 Visualize 包含了折線圖、直條圖、圓餅圖、資料表、數值百分比及地圖等，透過欄位的過濾、篩選及設定，就可以產生對應的圖表。

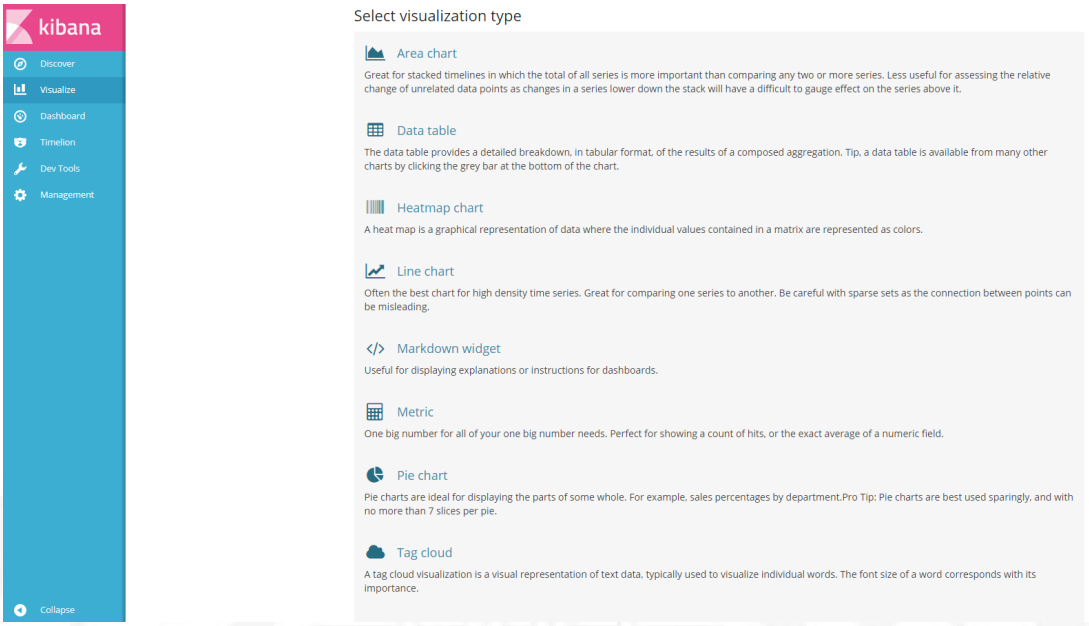


FIGURE 4.4: 視覺化圖表

雖然現今電腦的運算速度很快，但是網路連線的速度是有受限的，以本次目標站點 eBay.com 為例，如需要抓取全站商品，單一台爬蟲機，必須花費約 200 天左右，才有可能將商品詳情及使用者詳情全部抓取完畢。所以必須經由多爬蟲的架構方式，並減少關注類別來加速爬取的速度，才有辦法在比較合理的時間之內完成任務需求。

如 Figure 4.5 所示，平均每 30 分鐘可以取得 13 萬種商品類表，但是不包含商品詳情的部分。如果依照 eBay.com 的站點板型特性，需要抓取商品詳情資訊及賣家詳情資訊還需要再增加兩倍的連線數，為了考量時間上的問題，以及擔心可能會遭到 eBay.com 方面的封鎖，例如：圖型驗證，所以此實驗只有抓取商品的列表頁面，而未到商品詳情頁面及賣家詳細資訊。

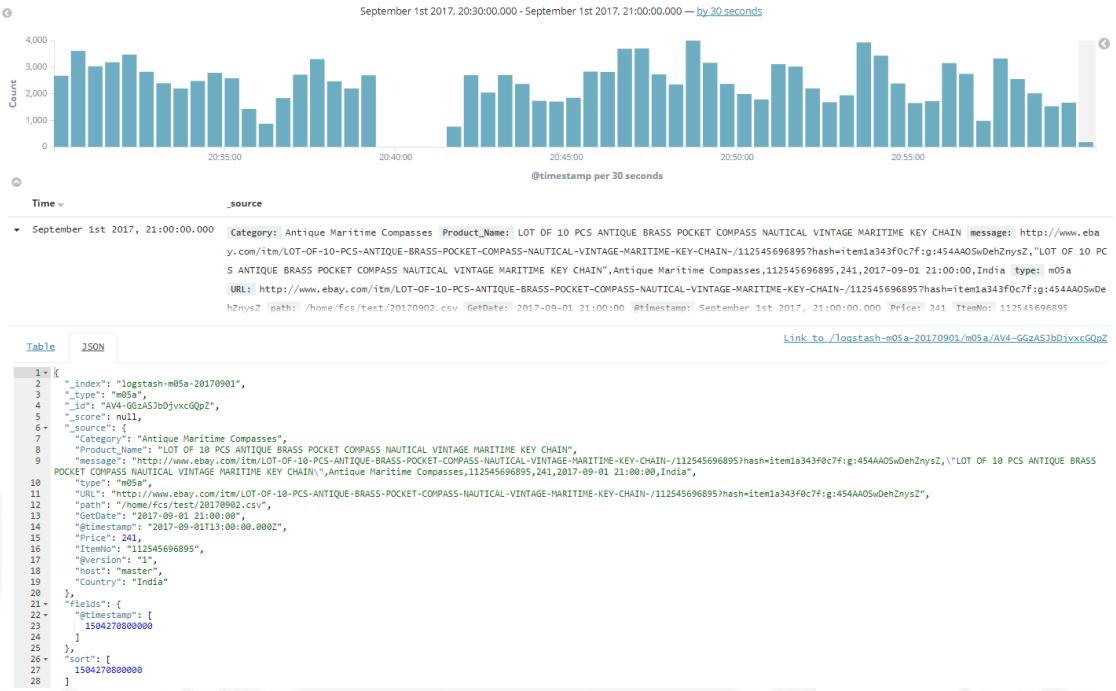


FIGURE 4.5: 商品爬取量

4.3 查詢執行時間

此系統架構最重要的分析功能，使用者可以自訂想要分析的資料區間及索引資料，更重要的是搜尋的反應時間如 Figure 4.6，五千萬筆資料中搜尋符合條件的資料顯示圖表，查詢及回應時間加總在 1 秒以下，凸顯由 ELK Stack 環境建置而成的系統擁有快速的分析能力。

FIGURE 4.6: 查詢執行時間

4.4 商品總數分類

而其中如 Figure 4.7，本文可以藉由取得的商品數量資料，來得知目前交易平台上商品數量是否有減少或是增加的趨勢。本文得到了文具類別的商品，在某一天內減少了一萬筆左右的商品，那麼本文可以推測有可能在這天有一萬筆左右的文具被買走，而這個時間是 8 月底、九月初，不但進而通知使用者注意當天有甚麼狀況發生，也有可能可以推算是，開學前夕，所以導致文具類的商品的被購買率上升。

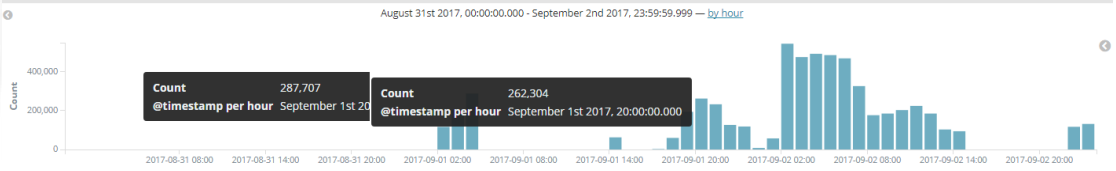


FIGURE 4.7: 商品爬取量

4.5 商品價格變化

而其中如 Figure 4.8，本文可以選取一個商品，來知道這個商品的價格變化，正常來說，一個電子商務網站，都會有自己養的爬蟲，然後利用爬蟲去同時爬取不同的敵對站點，然後參考敵對站點的價格，以及自己的進貨成本，來對自己的商品做自動調價的動作，這樣才可以避免自己的商品價格，遠高於其他站點，或是低於其他站點，讓自己的獲利下降。

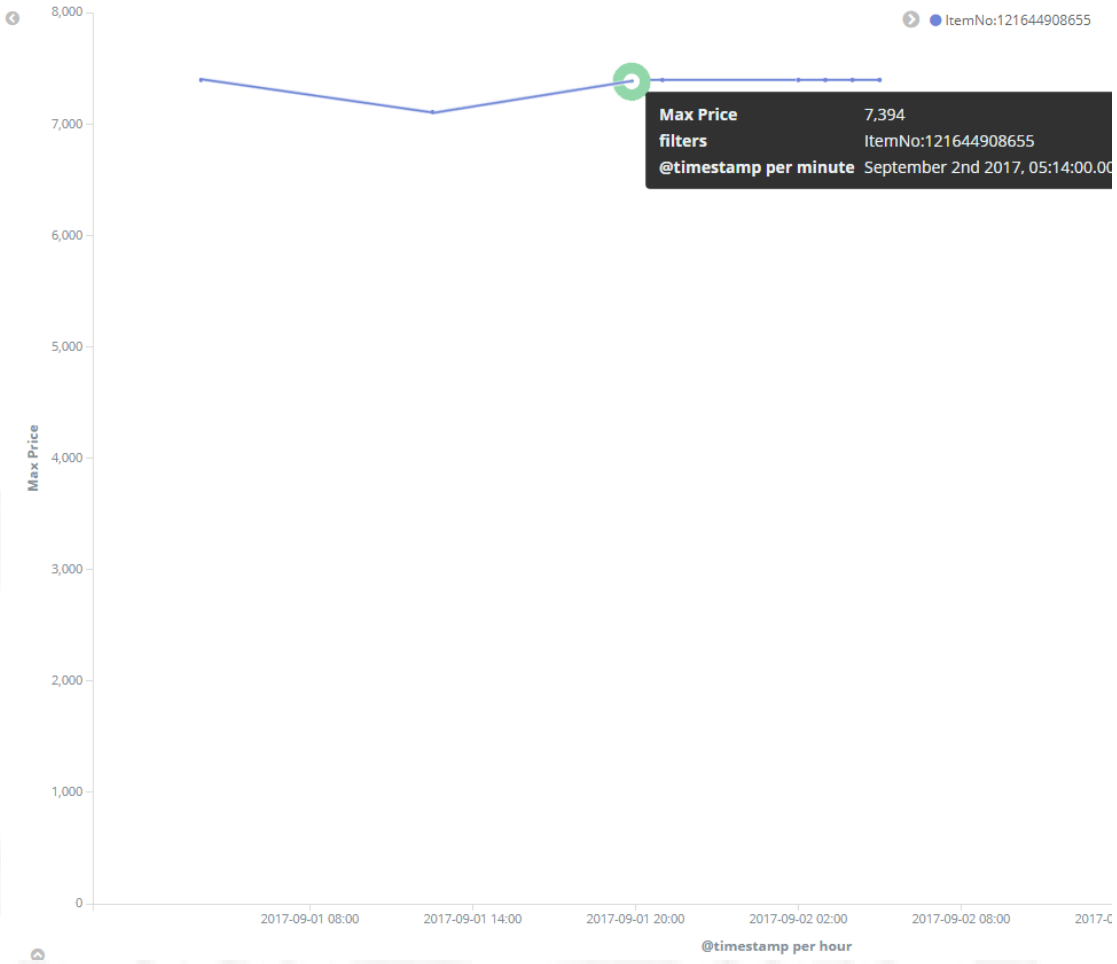


FIGURE 4.8: 商品價格變化

現今消費者重視的其實不是低價，而是高 C/P 低價跟高 C/P 之間的差異，是否分得出來？在低價之外，想找出能吸引消費者的誘因，關鍵就在於價格數字背後，可以提供的額外價值或服務。無論是吸引人的贈品或是加價購，還是對於消費者來說更有保障的商品附加服務，如何與其他的同類型商品或是提供更多產品外的價值，來凸顯價格以外的吸引力，其實才是中小企業面對電子商務該思考的網路行銷策略。相同的思考模式，在部分包含中小企業自有電商平台的多平台經營策略上，更可以加以提高自有平台的銷售利基。

4.6 賣家分析

對於大陸型國家的拍賣網站使用者來說，運費佔了整體購物費用的很大一部分，雖然同樣在美國，但是明顯東岸送到東岸的運費，跟東岸送到西岸的運費，會明顯有很大的差距，可能會因為昂貴的運費而放棄購買這件商品。同樣的狀況也可能發生在國際間的購物，如透過國際級的購物網站發現中意的商品，在折扣過後，雖然是比在本國購買價格較低廉，但是卻可能會因為運費的關係，而導致價格加上運費並沒有較為划算，甚至只是跟本地所得到的購買價格相同而已。

本實驗中以 eBay.com 來測試的結果有將近 70.68% 的商品來自於 United States，剩下 29.32% 則是其他國家，如 Figure 4.9。藉此可以將其站點的特性，盡量的推薦給美國的賣家，至於詳細地點，必須要進入商品詳情頁面才可以，礙於硬體設備的要求，暫時不對此做延展研究。

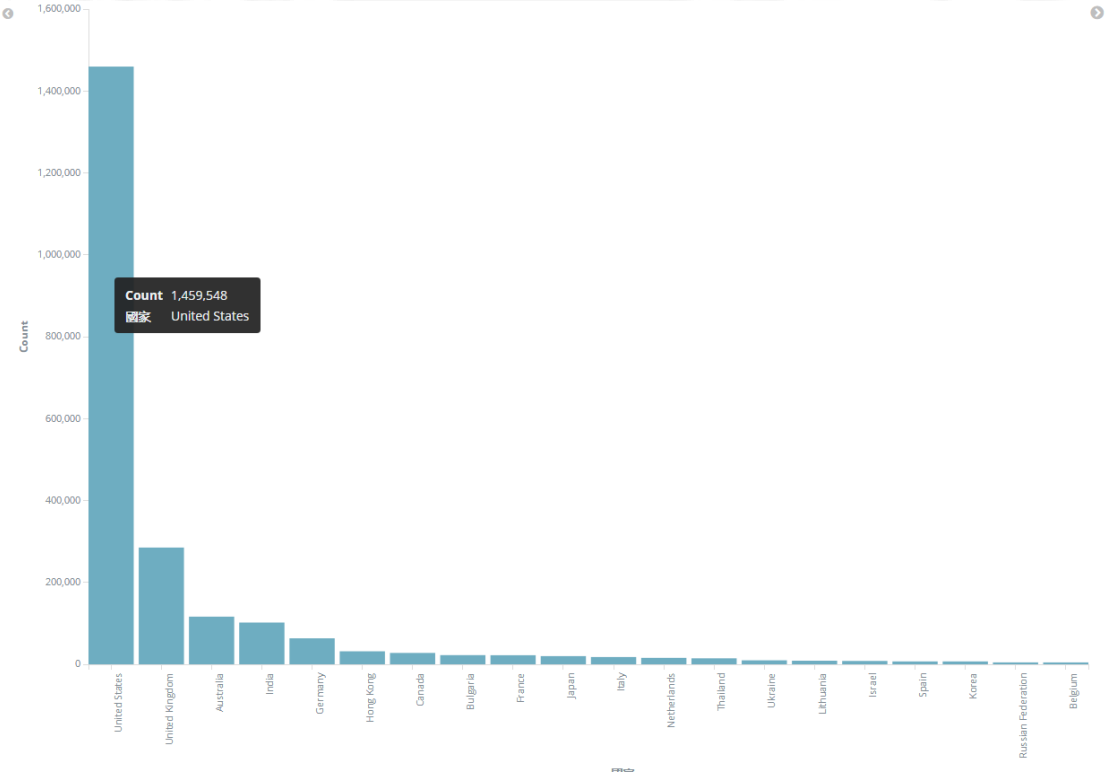


FIGURE 4.9: 賣家所屬國家

其中，本文隨機抽樣了幾個賣家，比較其賣出數量如 Figure 4.10

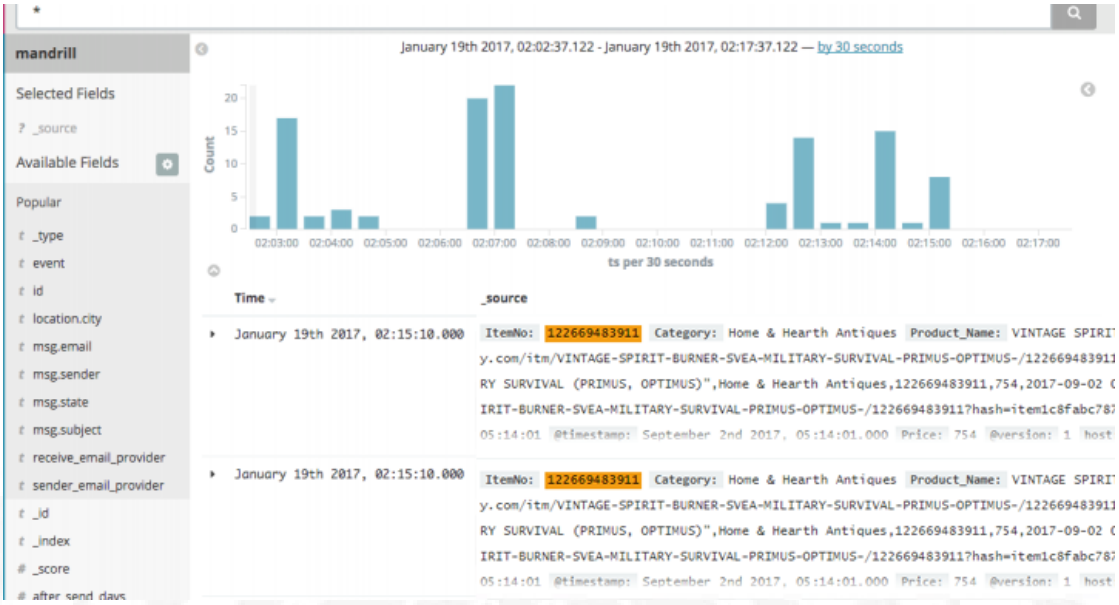


FIGURE 4.10: 賣家賣出數

而賣的比較好的賣家，幾乎都是在頁面上，具有最好賣家的標誌如 Figure 4.11。那麼本文就可以依據該站點的特性，來推薦給使用者，例如此站點就是：

TABLE 4.1: TOP 賣家

始終獲得最高買家評級
快速運送物品
在 eBay 上贏得了卓越服務的記錄

評分最高的賣家：

TABLE 4.2: 評分最高的賣家

必須持續獲得 eBay 買家的 4 或 5 星評級，才能提供優質的服務
快速發貨並提供準確的商品說明
每年必須出售 $ 1,000 或更多，並在 eBay 上完成至少 100 筆銷售交易
eBay 會定期審查是否符合這些標準

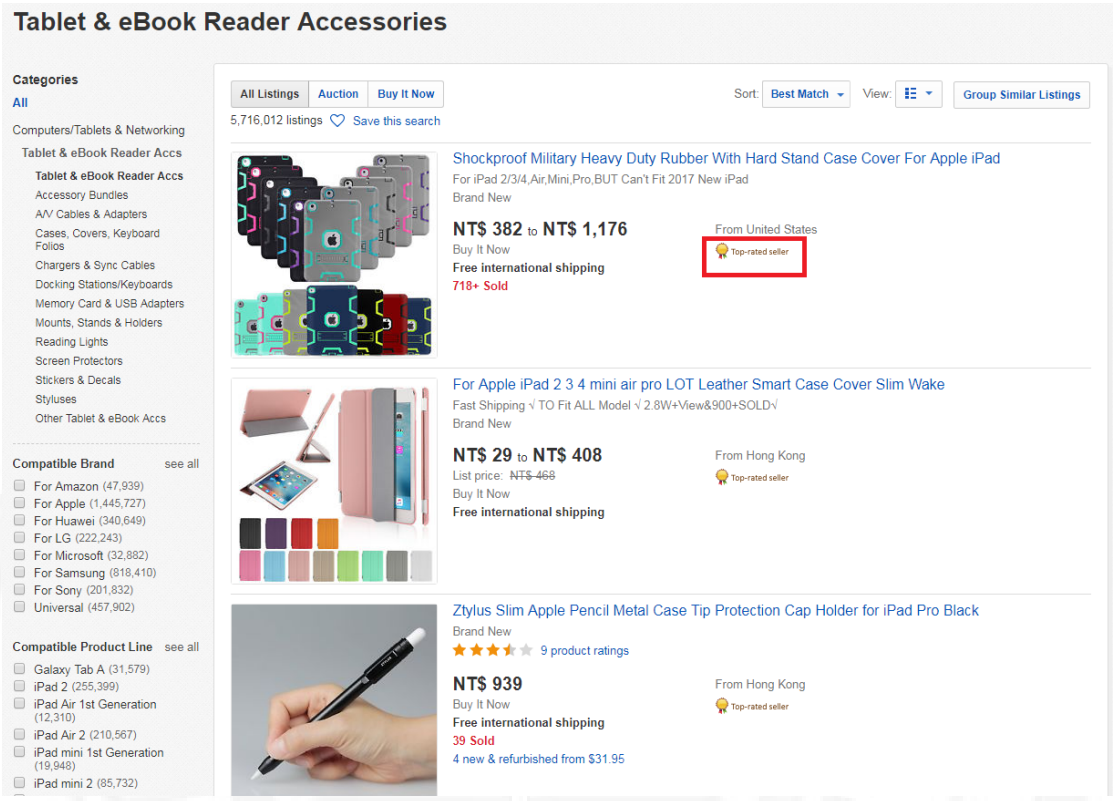


FIGURE 4.11: Top-rated seller

那本文則可以將這些資訊，推薦給使用者，讓使用者知道，怎樣的賣家，會比較容易得到買家的喜好，如同本文前面所說的，現在的買家已經不是光注重價格便宜這件事情了，他們同時會注重 C/P 值，他有可能會因為這個賣家是可信的，沒有欺騙過人，又或是他出貨的速度夠快，也或許單純的是因為該商品可以免運費之類的.....，這些都是買家非常注重的資訊，而本文就可以將這些東西傳遞給使用者。

4.7 價格分析

既然本文前面提到，現在的購物模式，其實是 B2B 經營方式的獲利會大於 C2B 經營模式，但是所謂的 B2B 經營模式，就不可能是一對一，而是多對一，因此本文就以圖 Figure 4.10商品來做分析。

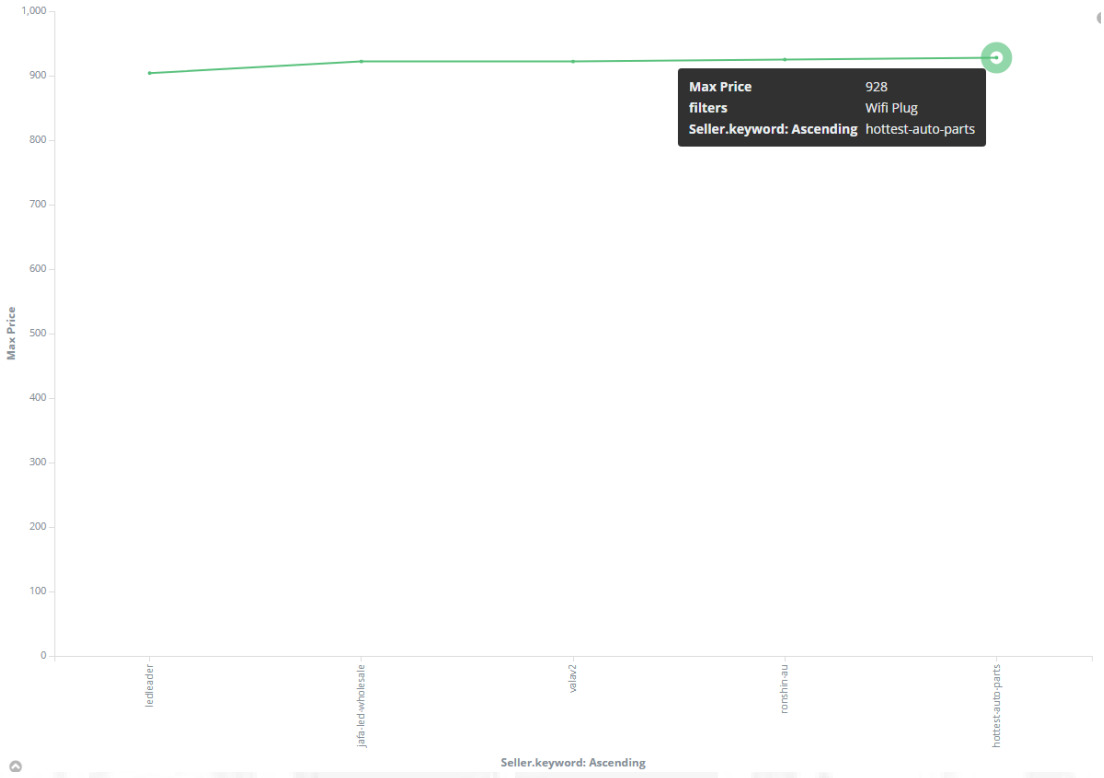
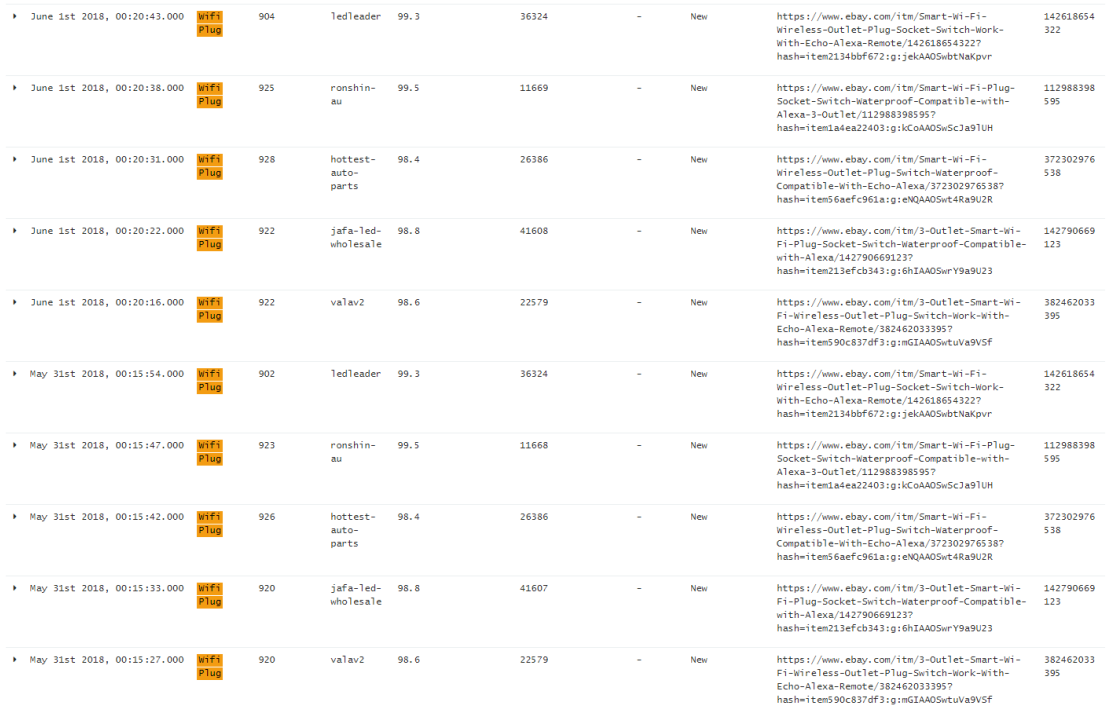


FIGURE 4.12: Smart Wi-Fi Wireless Outlet Plug 多方價格

在此實驗中，本文得知了，一個商品，會有複數的賣家，其中每個賣家，所賣的價格，並不會相等，我可以從詳細的圖表來看



June 1st 2018, 00:20:43.000	Wifi Plug	904	ledleader	99.3	36324	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Wireless-Outlet-Plug-Socket-Switch-Work-With-Echo-Alexa-Remote/142618654322?hash=item2134bbf672:g:jekAAOSwbtNaKpvr	142618654322
June 1st 2018, 00:20:38.000	Wifi Plug	925	ronshin-au	99.5	11669	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Plug-Socket-Switch-Waterproof-Compatible-with-Alexa-3-Outlet/112988398595?hash=item1a4ea22403:g:kCoAAOSwScJa9lUH	112988398595
June 1st 2018, 00:20:31.000	Wifi Plug	928	hottest-auto-parts	98.4	26386	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Wireless-Outlet-Plug-Switch-Waterproof-Compatible-With-Echo-Alexa/372302976538?hash=item56aefc961a:g:eNQAAOSwt4Ra9U2R	372302976538
June 1st 2018, 00:20:22.000	Wifi Plug	922	jafa-led-wholesale	98.8	41608	-	New	https://www.ebay.com/itm/3-Outlet-Smart-Wi-Fi-Plug-Socket-Switch-Waterproof-Compatible-with-Alexa/142790669123?hash=item213efcb343:g:6hIAAOSwrY9a9U23	142790669123
June 1st 2018, 00:20:16.000	Wifi Plug	922	valav2	98.6	22579	-	New	https://www.ebay.com/itm/3-Outlet-Smart-Wi-Fi-Wireless-Outlet-Plug-Switch-Work-With-Echo-Alexa-Remote/382462033395?hash=item590c837df3:g:mGIAAOSwtuVa9V5f	382462033395
May 31st 2018, 00:15:54.000	Wifi Plug	902	ledleader	99.3	36324	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Wireless-Outlet-Plug-Socket-Switch-Work-With-Echo-Alexa-Remote/142618654322?hash=item2134bbf672:g:jekAAOSwbtNaKpvr	142618654322
May 31st 2018, 00:15:47.000	Wifi Plug	923	ronshin-au	99.5	11668	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Plug-Socket-Switch-Waterproof-Compatible-with-Alexa-3-Outlet/112988398595?hash=item1a4ea22403:g:kCoAAOSwScJa9lUH	112988398595
May 31st 2018, 00:15:42.000	Wifi Plug	926	hottest-auto-parts	98.4	26386	-	New	https://www.ebay.com/itm/Smart-Wi-Fi-Wireless-Outlet-Plug-Switch-Waterproof-Compatible-With-Echo-Alexa/372302976538?hash=item56aefc961a:g:eNQAAOSwt4Ra9U2R	372302976538
May 31st 2018, 00:15:33.000	Wifi Plug	920	jafa-led-wholesale	98.8	41607	-	New	https://www.ebay.com/itm/3-Outlet-Smart-Wi-Fi-Plug-Socket-Switch-Waterproof-Compatible-with-Alexa/142790669123?hash=item213efcb343:g:6hIAAOSwrY9a9U23	142790669123
May 31st 2018, 00:15:27.000	Wifi Plug	920	valav2	98.6	22579	-	New	https://www.ebay.com/itm/3-Outlet-Smart-Wi-Fi-Wireless-Outlet-Plug-Switch-Work-With-Echo-Alexa-Remote/382462033395?hash=item590c837df3:g:mGIAAOSwtuVa9V5f	382462033395

FIGURE 4.13: Smart Wi-Fi Wireless Outlet Plug 多方價格表

價格最低的是 ledleader 這位賣家，雖然他的評分數是 99.3 分，但是他的總評價數卻是最高的 36324，而最高的賣家則是 ronshin-au，雖然他的評價分數是最高的 99.5 分，但總評價數卻是最少的 11669，所以代表，很有可能，ledleader 這位賣家，他藉由壓低價格，來讓自己的賣出的總商品數來更高。而且，本文如果加入了時間參數，可以得知，其實這塊餅競爭很激烈。

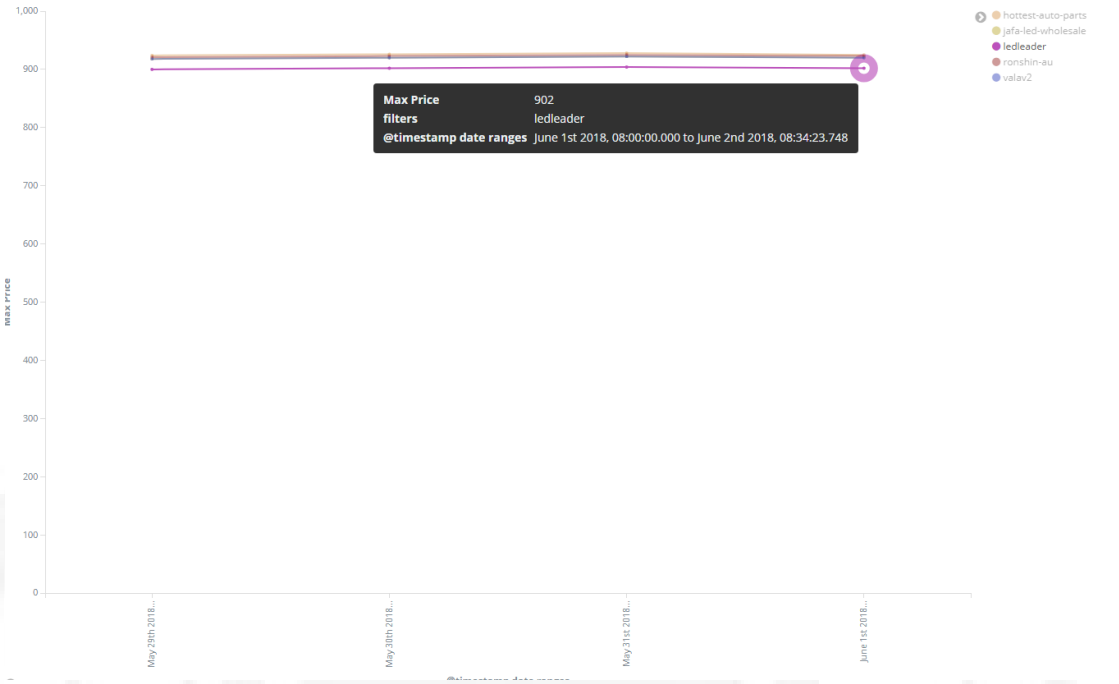


FIGURE 4.14: Smart Wi-Fi Wireless Outlet Plug 多方時間價格表

當其中一位進行調價的時候，其他的賣家居然也同時調價，很有可能代表，他們不是有自己的爬蟲，就是有在定時關注競爭對手的價格。

4.8 結合 ELK 的優勢速度

既然 ELK 能夠做這麼多方面的資訊顯示，來提供使用者分析，他們需要得知的相關訊息，例如價格漲幅、運費、賣出數... 等等，自然他的查詢時間，就不應該太長，而且本文當初也是看上，ELK 俱備自動索引的功能，因此我先對 ELK Stack 進行查詢時間的實驗，分別對 5 萬、20 萬、100 萬、200 萬、500 萬、900 萬的商品數，進行亂數取一的資料查詢，並且將同樣的資料數，輸入至 SQL 裡面，而在 ELK 中，他的 Query 及 Request 如同圖 Figure 4.15，雖然會隨著資料量的上升。

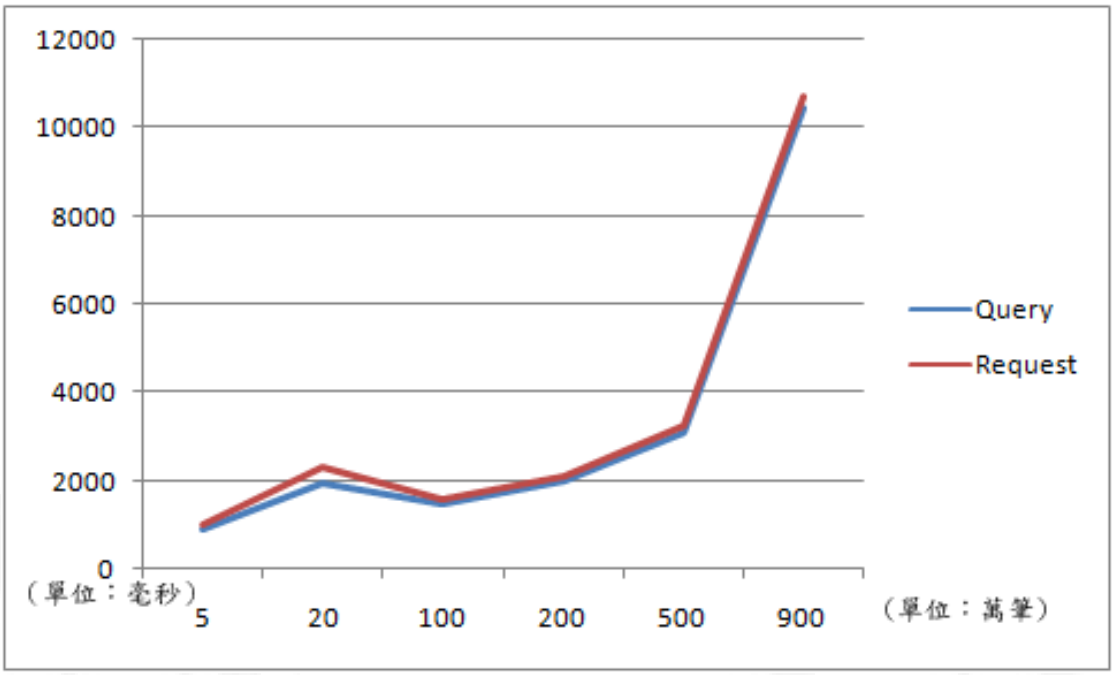


FIGURE 4.15: ELK 的查詢速度成長

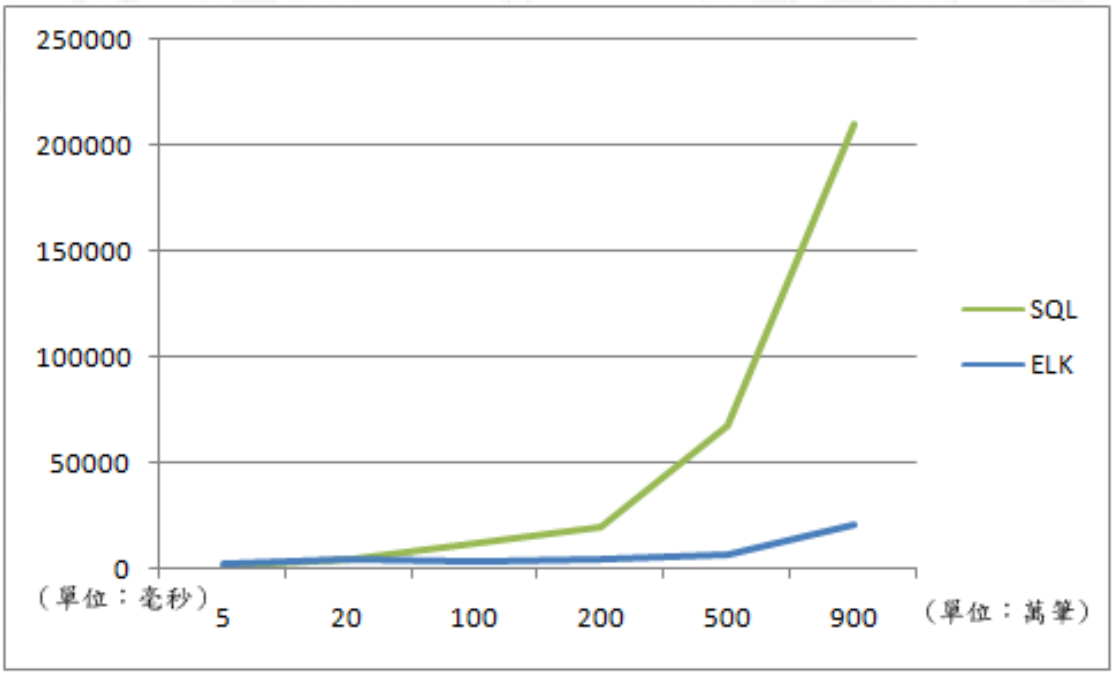


FIGURE 4.16: ELK 對比 SQL 的查詢速度成長

4.9 找出熱門賣家

既然要參考，就勢必要找出最好的賣家，這樣才會顯得有效益，如圖 Figure 4.17。

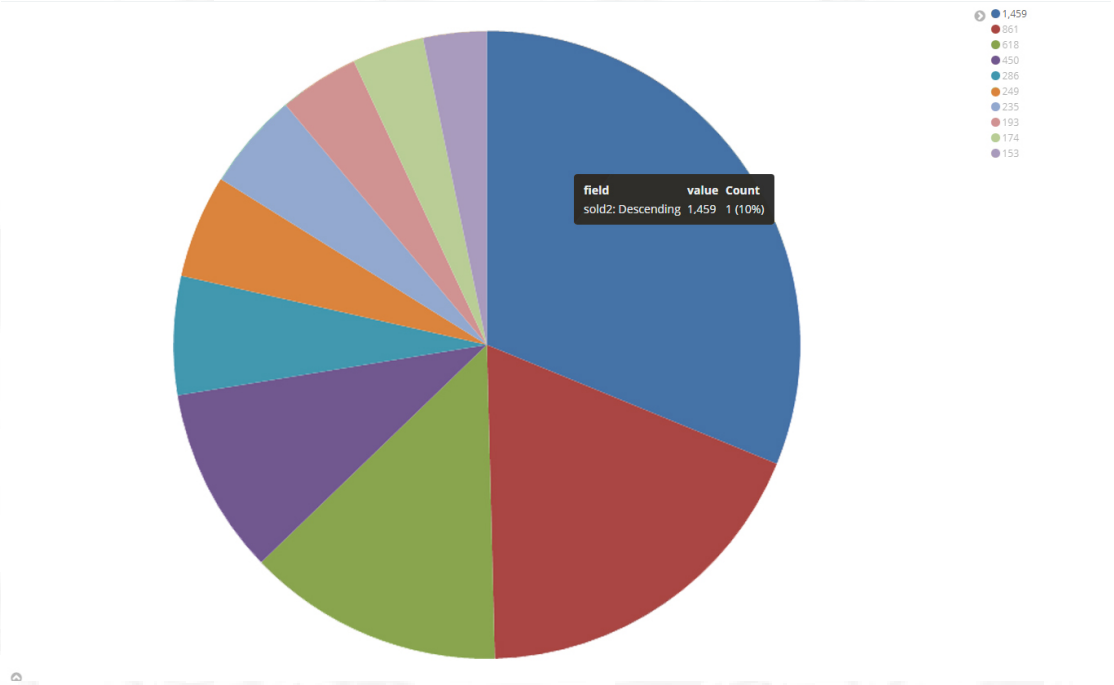


FIGURE 4.17: 熱門賣家圓餅圖

他能夠在眾多賣家中找出最賣最多的賣家，並且 Kibana 還具備有，像是以下的熱門字搜索的方式，來讓使用者能夠輕鬆一覽，到底是哪一些賣家，是賣的最好，最有參考價值的賣家，如圖 Figure 4.18。

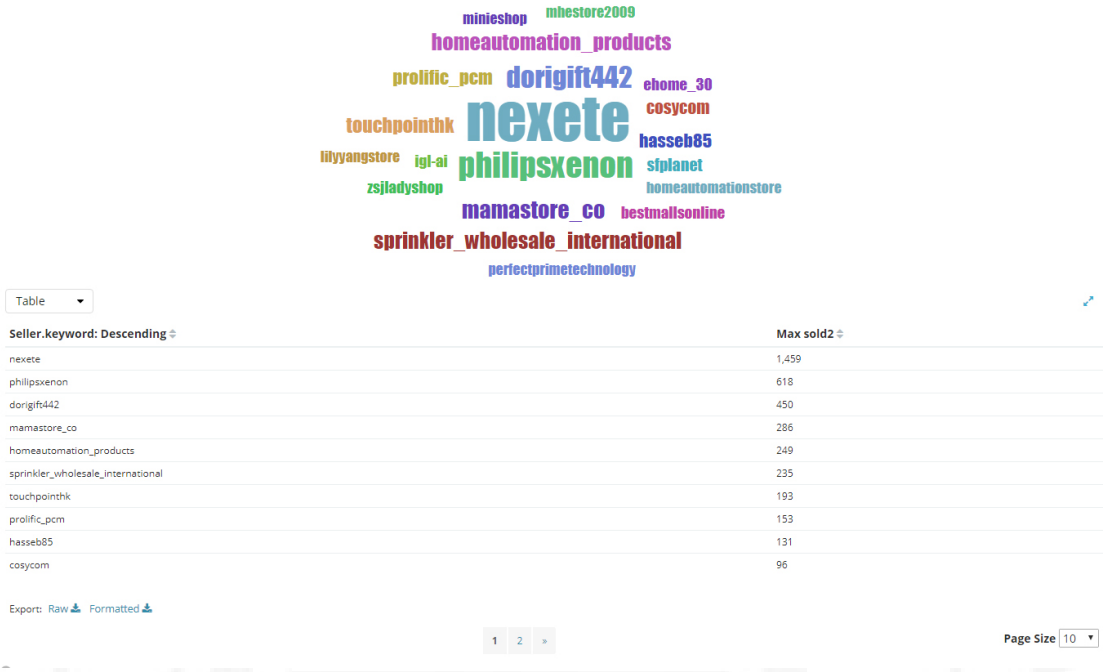


FIGURE 4.18: 熱門賣家文字雲

那同理，就可以拿來利用上面實驗的結果套進去，我今天想要買一個商品 WIFI Plug，但是賣家卻有五個，此套系統就能夠將比較適合的賣家推薦給使用者，如圖 Figure 4.19。

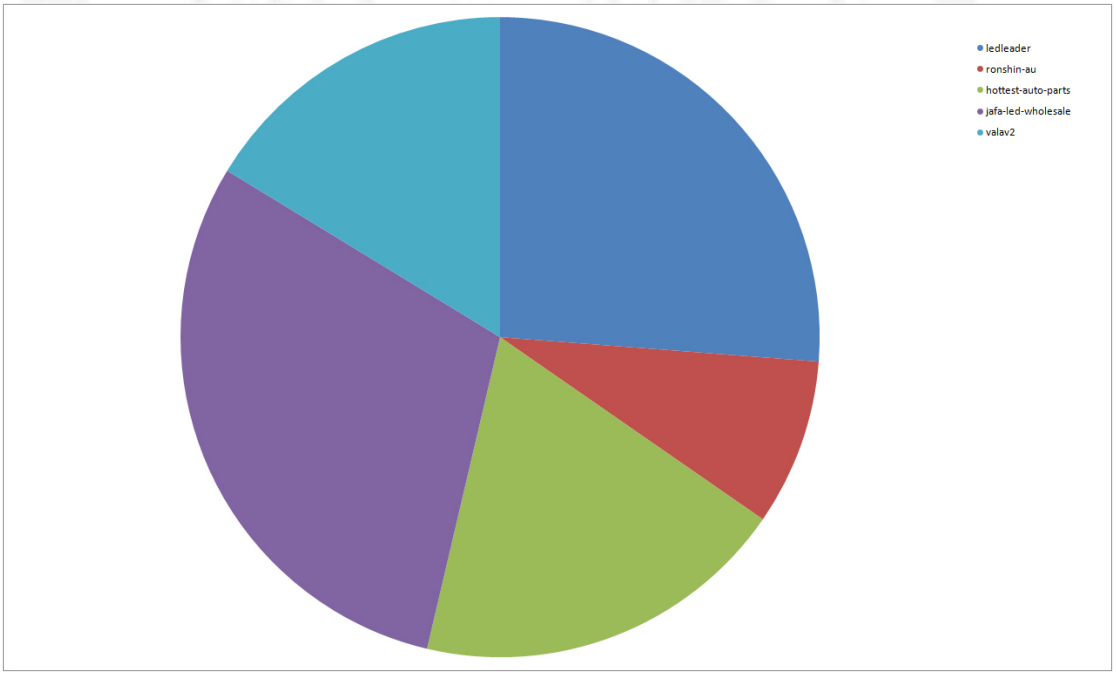


FIGURE 4.19: 推薦賣家

Chapter 5

結論與未來方向

5.1 結論

資料的應用已是現在主流，透過對資料的分析和應用，可以從中挖掘出未知的趨勢及有價值的資訊。本論文主要將爬蟲與 ELK Stack 結合，建置一個分析系統，這次的實驗發現，能夠使用於大數據的系統很多，但是唯獨 ELK Stack 卻特別的合適。並且從各方面的數據分析，得到類似大多使用 Ebay.com 的用戶都來自於 United States，那麼廠商就可以依照這個數據，來推薦給美國的買賣家，並搭配一些關於運費的活動折扣，來促進商品的交易，的建議結果。

在電子商務的世界中，要做到頂尖其實是需要非常多的資料量，而這些資料，並不會由敵對站點直接公佈在自己的網站上，例如銷售量、商品價格、熱賣商品，買方資訊.....，因此在做投入電子商務的同時，必須要能取得大量的資料來進行交叉比對，從中得知商品目前在那家是最為便宜，又或者是最為昂貴，而他最便宜的理由是甚麼，以及是否有 bundle 商品。

本文以 Ebay.com 為例，基於銷售商品大數據分析，採用數據挖掘技術，通過標準 Java 網絡爬蟲程序，來進行數據採集。為電子商務業在銷售方式提出一些建議，來減少企業的庫存和資源浪費。

5.2 未來方向

這套系統目前僅具備資料分析功能，未來預計應用一些簡單的 Data Mining 來將一些推算的資訊，推薦給使用者。例如，可以藉由商品的銷售量、或者是新品量，來得知目前比較熱門的商品類別，來告知其他使用者，目前哪些種類的商品比較熱門，是否可以多進一些庫存量，或是少進一些冷門的商品，藉此調整自己的進貨量，來達到最高的利潤。

參考文獻

[1] eBay 公布美國賣家數據. ebay 公布美國賣家數據：哪個州銷售額最高、賣家數量最多,http://www.gooread.com/article/20122787466/, 2017.

[2] TechOrange 科技報橘. 當你和 FB 一樣每天要處理 300TB 的資料，你就會知道虛擬化技術的重要！,http://techorange.com/2013/04/16/why-virtual-machine-is-so-attractive/, 2017.

[3] Rajkumar Buyya, Rodrigo N Calheiros, and Amir Vahid Dastjerdi. *Big Data: Principles and Paradigms*. Morgan Kaufmann, 2016.

[4] Apache. Apache hadoop,http://hadoop.apache.org, 2017.

[5] Apache. Apache spark,https://spark.apache.org, 2017.

[6] Andrej Trnka. Big data analysis. *European Journal of Science and Theology*, 10(1):143–148, 2014.

[7] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.

[8] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, 47:98–115, 2015.

[9] Daniel E O'Leary. Artificial intelligence and big data. *IEEE Intelligent Systems*, 28(2):96–99, 2013.

[10] Jules J Berman. *Introduction in Principles of big data: preparing, sharing, and analyzing complex information*. Newnes, 2013.

[11] Jin Xiong, Yiming Hu, Guojie Li, Rongfeng Tang, and Zhihua Fan. Metadata distribution and consistency techniques for large-scale cluster file systems. *IEEE Transactions on Parallel and Distributed Systems*, 22(5):803–816, 2011.

[12] 網路爬蟲. 網路爬蟲,https://zh.wikipedia.org/wiki/2015.

[13] marsz. 網路爬蟲扼殺了網站經營者!?,https://ppt.cc/fad9cx, 2017.

[14] Elastic. Elastic stack,https://www.elastic.co, 2017.

[15] Clinton Gormley and Zachary Tong. *Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine*. O'Reilly Media, Inc., 2015.

[16] Roy T Fielding and Richard N Taylor. *Architectural styles and the design of network-based software architectures*. University of California, Irvine Doctoral dissertation, 2000.

[17] Abdelkader Lahmadi and Frédéric Beck. Powering monitoring analytics with elk stack. In *9th International Conference on Autonomous Infrastructure, Management and Security (AIMS 2015)*, 2015.

[18] Oleksii Kononenko, Olga Baysal, Reid Holmes, and Michael W Godfrey. Mining modern repositories with elasticsearch. In *Proceedings of the 11th Working Conference on Mining Software Repositories*, pages 328–331. ACM, 2014.

[19] Alexander Reelsen. Using elasticsearch, logstash and kibana to create realtime dashboards. *Dostupné z: https://secure. trifork. com/dl/goto-berlin-2014/GOTO_Night/logstash-kibana-intro. pdf*, 2014.

[20] Melody Moh, Santhosh Pininti, Sindhusha Doddapaneni, and Teng-Sheng Moh. Detecting web attacks using multi-stage log analysis. In *Advanced*

Computing (IACC), 2016 IEEE 6th International Conference on, pages 733–738. IEEE, 2016.

[21] Xiwei Xu, Ingo Weber, Len Bass, Liming Zhu, Hiroshi Wada, and Fei Teng. Detecting cloud provisioning errors using an annotated process model. In *Proceedings of the 8th Workshop on Middleware for Next Generation Internet Computing*, page 5. ACM, 2013.

[22] Ceki Gülcü. *The complete log4j manual*. QOS. ch, 2003.

[23] Sushma Sanjappa and Muzameel Ahmed. Analysis of logs by using logstash. In *Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications*, pages 579–585. Springer, 2017.

[24] Elastic. Kibana: Explore, visualize, discover data,https://www.elastic.co/products/kibana, 2017.

[25] sears. sears.com,https://www.sears.com, 2018.

[26] Amazon. Amazon.com,https://www.amazon.com, 2018.

[27] Hsinlan Chen. 硬塞科技字典,https://www.inside.com.tw/2016/07/28/what-is-ecommerce, 2016.

[28] Tarun Prakash, Misha Kakkar, and Kritika Patel. Geo-identification of web users through logs using elk stack. In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, pages 606–610. IEEE, 2016.

[29] Jun Bai. Feasibility analysis of big log data real time search based on hbase and elasticsearch. In *Natural Computation (ICNC), 2013 Ninth International Conference on*, pages 1166–1170. IEEE, 2013.

[30] Khanasin Yamnual, Phond Phunchongharn, and Tiranee Achalakul. Failure detection through monitoring of the scientific distributed system. In *Applied System Innovation (ICASI), 2017 International Conference on*, pages 568–571. IEEE, 2017.

[31] S Bagnasco, D Berzano, S Lusso, M Masera, and S Vallero. Managing competing elastic grid and cloud scientific computing applications using opennebula. In *Journal of Physics: Conference Series*, volume 664, page 022004. IOP Publishing, 2015.

[32] S Bagnasco, D Berzano, A Guarise, S Lusso, M Masera, and S Vallero. Monitoring of iaas and scientific applications on the cloud using the elasticsearch ecosystem. In *Journal of Physics: Conference Series*, volume 608, page 012016. IOP Publishing, 2015.

[33] Pingkan PI Langi, Warsun Najib, Teguh Bharata Aji, et al. An evaluation of twitter river and logstash performances as elasticsearch inputs for social media analysis of twitter. In *Information & Communication Technology and Systems (ICTS), 2015 International Conference on*, pages 181–186. IEEE, 2015.

[34] Tarun Prakash, Misha Kakkar, and Kritika Patel. Geo-identification of web users through logs using elk stack. In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, pages 606–610. IEEE, 2016.

[35] Steven Latré, Marinos Charalambides, Jérôme François, Corinna Schmitt, and Burkhard Stiller. Intelligent mechanisms for network configuration and security.

[36] Ivan Ermilov, Axel-Cyrille Ngonga Ngomo, Aad Versteden, Hajira Jabeen, Gezim Sejdiu, Giorgos Argyriou, Luigi Selmi, Jürgen Jakobitsch, and Jens Lehmann. Managing lifecycle of big data applications. In *International Conference on Knowledge Engineering and the Semantic Web*, pages 263–276. Springer, 2017.

[37] Dong Nguyen Doan and Gabriel Iuhasz. Tuning logstash garbage collection for high throughput in a monitoring platform. In *Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), 2016 18th International Symposium on*, pages 359–365. IEEE, 2016.

[38] Damian Hermanowski. Open source security information management system supporting it security audit. In *Cybernetics (CYBCONF), 2015 IEEE 2nd International Conference on*, pages 336–341. IEEE, 2015.

[39] Ruben Tous, Jordi Torres, and Eduard Ayguadé. Multimedia big data computing for in-depth event analysis. In *Multimedia Big Data (BigMM), 2015 IEEE International Conference on*, pages 144–147. IEEE, 2015.

附錄 A

ELK Stack 安裝步驟

A.1 基本環境設定

步驟 1. 更新 vim 避免舊版 vim 上下左右鍵異常，為時間錯亂安裝 NTP。

```
sudo apt-get update
sudo apt-get install -y vim ntp curl ssh
```

步驟 2. 將所有主機加入本機 IP 對應表。

```
sudo vim /etc/hosts
#加入下列參數
192.168.56.134 master
```

步驟 3. 安裝 open JDK。

```
sudo apt-get install -y openjdk-8-jdk
sudo ln -s /usr/lib/jvm/java-8-openjdk-amd64 /usr/lib/jvm/jdk
```

步驟 4. 編輯環境參數。

```
vim .bashrc
#在最後面加入jdk.hadoop等相關路徑
export JAVA_HOME=/usr/lib/jvm/jdk/
#匯入環境變數
source .bashrc
```

A.2 Elasticsearch 安裝

步驟 1. 加入 Elastic 套件庫來源。

```
https://packages.elastic.co/GPG-KEY-elasticsearch | sudo apt-key add -
```

步驟 2. 建立 Elastic Stack 來源表。

```
sudo apt-get install -y apt-transport-https
echo "deb https://artifacts.elastic.co/packages/5.x/apt stable main" | sudo tee -a /
    etc/apt/sources.list.d/elastic-5.x.list
```

步驟 3. 更新套件庫列表及安裝 Elasticsearch。

```
sudo apt-get update
sudo apt-get install -y elasticsearch
```

步驟 4. 編輯 elasticsearch.yml 讓外部可以存取。

```
sudo vi /etc/elasticsearch/elasticsearch.yml
```

步驟 5. 更新參數，將下列幾行參數#移除並修改如下。

```
cluster.name    : hpc
network.host   : 192.168.56.134
http.port          : 9200
```

步驟 6. 重啟服務並將 elasticsearch 設定成開機就啟動。

```
sudo systemctl restart elasticsearch
sudo systemctl enable elasticsearch
```

A.3 Logstash 安裝

步驟 1. 下載安裝 logstash。

```
sudo apt-get install -y logstash
```

步驟 2. 簡易功能測試。

```
cd /usr/share/logstash
sudo bin/logstash -e 'input {stdin{ } } output {stdout{ } }'
輸入Hello
螢幕畫面會顯示master Hello則表示基本功能正常
```

步驟 3. 建立設定檔。

```
sudo vi /etc/logstash/conf.d/first-pipeline.conf
#加入下列參數進行測試，實際加入參數請參考附錄D “Logstash configuration file”

input {
                file {
                        path => "/var/log/*.log"
                        }
                }
output {
                elasticsearch {
                        hosts => " 192.168.56.134:9200"
                        manage_template => false
                        index => "%{[@metadata][beat]}-%{+YYYY.MM.dd}"
                        document_type => "%{[@metadata][type]}"
                        }
                }
```

步驟 4. 重啟服務並將 logstash 設定成開機就啟動。

```
sudo systemctl restart logstash
sudo systemctl enable logstash
```

A.4 Kibana 安裝

步驟 1. 下載安裝 Kibana。

```
sudo apt-get install -y kibana
```

步驟 2. 編輯 kibana.yml。

```
sudo vi /etc/kibana/kibana.yml
```

步驟 3. 更新參數，將下列幾行參數#移除並修改如下。

```
server.port: 5601
server.host: "192.168.56.134"
elasticsearch_url :http://192.168.56.134:9200
```

步驟 6. 重啟服務並將 kibana 設定成開機就啟動。

```
sudo systemctl restart kibana
sudo systemctl enable kibana
```

附錄 B

Crawler Script

B.1 目錄爬蟲

category.java

```
Date date = new Date();
SimpleDateFormat bartDateFormat = new SimpleDateFormat("yyyyMMdd");
String saveName = bartDateFormat.format(date);     //爬蟲區別用

HttpUtil httpUtil = new HttpUtil();
String content = httpUtil.getProxyContent(url, "").getUrlContent();
HashMap<String, String> subURL = new HashMap<String, String>();
String categoryContent = Util.subString(content, 0, "<div class=\"h3content\">{@{arg
    }</div>").GetItem("arg");
categoryContent = categoryContent.replaceAll("\" >", "\">");
String name = "";
String subUrl = "";
int i = 0;
int j = 0;
while(i > -1) {
        SubStrResult ssr = Util.subString(categoryContent, i, "<li><a href=\"{@{url
    }\">{@{name}</a></li>");
        i = ssr.CurrentPos;
        if(i > -1) {
                name = ssr.GetItem("name");
                subUrl = ssr.GetItem("url");
                subURL.put(name, subUrl);
        if(j > 2) {
```

```
                subCategory(name, subUrl, saveName);    //將任務內容傳遞給下一隻爬蟲，
    並且存到Logstash
        }
        j++;
        }
}
```

B.2 列表爬蟲

productlist.java

```
HttpUtil httpUtil = new HttpUtil();
String content = httpUtil.getProxyContent(url, "").getUrlContent();
int totel = 0;
String allProduct = Util.subString(content, 0, "listingscnt\"  >{@{arg} listings</").
    GetItem("arg");
if(allProduct != null && !allProduct.equals("")) {
        allProduct = allProduct.replace(",", "");
        allProduct = str2number(allProduct);
        totel = Integer.parseInt(allProduct);
}
int allPage = 0;
if(totel%200 == 0) {
        allPage = (totel / 200);
}else {
        allPage = (totel / 200)+1;
}
if(allPage > 50) {allPage = 50;}
parseList(content, url, saveName);
for (int i = 2; i < allPage; i++) {
        content = httpUtil.getProxyContent(url+"&_pgn="+i, "").getUrlContent();
        parseList(content, url+"&_pgn="+i, saveName);                    //將任務內容傳
    遞給下一隻爬蟲，並且存到Logstash
}
```

B.3 詳情爬蟲

productInfo.java

```
HttpUtil httpUtil = new HttpUtil();
String content = httpUtil.getProxyContent(url, "").getUrlContent();
parseInfo(content, url);                    //將商品詳情抓回，並且存到Logstash


productName = Util.subString(content, 0, "<title>{@{arg}</title>").GetItem("arg").trim
    ();
price = Util.subString(content, 0, "convbidPrice{@{*}>{@{arg}<").GetItem("arg").trim()
    ;
fromWhere = Util.subString(content, 0, "<div class=\"u-flL\">{@{arg}</div>").GetItem("
    arg").trim();
```

附錄 C

Logstash Script

C.1 去取得爬蟲抓回來的資料

logstash

```
input {
file {
path => "/home/fcs/test/*.csv"
start_position => beginning
type => "csv"
}
}
#進行資料格式轉換及欄位參數設定
filter{
#  if [type] == "csv"{
csv{
columns => ["URL", "Product_Name", "Category", "ItemNo", "Price", "GetDate",  "Country
    "]
}

date {
match => [ "GetDate", "yyyy-MM-dd HH:mm:ss" ]
timezone => "Asia/Taipei"
target => "@timestamp"
}

mutate{
convert => ["Price", "integer"]
```

```
}
#  }
}

output {
#  if [type] == "csv"{
elasticsearch {
hosts => "192.168.56.134:9200"
index => "logstash-%{type}-%{+YYYYMMdd}"
document_type => "%{type}"
}
stdout { codec => rubydebug}
#  }
}
```

附錄 D

Product Info file

productList.csv 是由產品目錄及產品列表清洗過後的資料，主要保留商品名稱、URL、位置、價格及商品唯一碼。

```
http://www.ebay.com/itm/MV01-Pre-Columbia-Inca-Antique-Nazca-Small-Polychrome-Pottery-
    pot-vase-/352153509349?hash=item51fdfc1de5:g:k4kAAOSwnJVZmj9b,"MV01 Pre-Columbia
    Inca Antique Nazca Small Polychrome Pottery pot vase",352153509349,45267,Japan
http://www.ebay.com/itm/Authentic-Mayan-Pre-Columbian-Chac-Bowl-/282627351482?hash=
    item41cde723ba:g:ebwAAOSwrRlZoz8k,"Authentic Mayan Pre Columbian Chac Bowl
    ",282627351482,39231,United States
http://www.ebay.com/itm/ANTIQUE-KACHINA-KATSINA-ZUNI-circa-1940-1950-/222627010061?
    hash=item33d59a960d:g:N54AAOSwXz9ZlfLs,"ANTIQUE KACHINA KATSINA ZUNI circa
    1940/1950",222627010061,39077,France
http://www.ebay.com/itm/Taino-Caribbean-Indians-Shaman-s-Vomit-Stick-circa
    -1300-1500-/152680511867?hash=item238c77b97b:g:5ZYAAOSwGJlZNirA,"Taino (Caribbean
    Indians) Shaman' s Vomit Stick…….circa (1300 - 1500)",152680511867,72427,Canada
http://www.ebay.com/itm/round-oak-A-18-finial-/253123017546?hash=item3aef4ebb4a:g:
    kZ4AAOSw3GFZn2SX,"round oak A 18 finial",253123017546,10592,United States
http://www.ebay.com/itm/Small-powder-Horn-With-3-Tacks-On-Bottom-/263175424961?hash=
    item3d467a4bc1:g:q80AAOSweQBZpe2e,"Small powder Horn With 3 Tacks On Bottom
    .",263175424961,1056,United States
http://www.ebay.com/itm/Pre-Columbian-COLIMA-COILED-SNAKE-VESSEL-EX-SOTHEBYS
    -79-/122668976668?hash=item1c8fa40a1c:g:MIcAAOSw1-RUZ7E6,"Pre-Columbian COLIMA
    COILED SNAKE VESSEL, EX: SOTHEBY&#039;S &#039;79",122668976668,60205,United States
http://www.ebay.com/itm/Awesome-Large-TAINO-Celt-MAKE-AN-OFFER-Precolumbian-Mayan-
    Aztec-/272825809931?hash=item3f85af7c0b:g:oJ0AAOSwYlBZpetM,"Awesome Large TAINO
    Celt, MAKE AN OFFER Precolumbian Mayan Aztec",272825809931,5281,Canada
```

```
http://www.ebay.com/itm/Rare-Precolumbian-Veracruz-Warrior-Figure-Attached-To-An-
    Offering-Vase-/122674715929?hash=item1c8ffb9d19:g:cREAAOSw539Zb-Cx,"Rare
    Precolumbian Veracruz Warrior Figure Attached To An Offering Vase
    ",122674715929,30027,United States
http://www.ebay.com/itm/Unidentified-lead-head-of-alien-being-sliced-in-half-f-m-South
    -America-70s-L204-/172829057152?hash=item283d6a2080:g:WRAAAOSwo4pYk3NW,"
    Unidentified lead head of alien being? sliced in half f/m South America 70s L204
    ",172829057152,781,United Kingdom
http://www.ebay.com/itm/Rare-Precolumbian-Veracruz-Figure-Attached-To-An-Offering-Bowl
    -/122674713481?hash=item1c8ffb9389:g:zsIAAOSwAAVZb-Al,"Rare Precolumbian Veracruz
    Figure Attached To An Offering Bowl",122674713481,30027,United States
http://www.ebay.com/itm/Round-oak-e-18-swing-cover-and-trademark-indian-and-finial
    -/263175467606?hash=item3d467af256:g:mhAAAOSwGNdZn163,"Round oak  e 18 swing cover
     and trademark indian and finial",263175467606,8238,United States
http://www.ebay.com/itm/AZTEC-ORIGINAL-TERRACOTA-FROG-NECKLACE-PRE-COLUMBIAN-INDIAN-
    RELIC-ARTIFACT-/142491152783?hash=item212d22718f:g:m08AAOSwkRpZphiz,"AZTEC
    ORIGINAL TERRACOTA FROG NECKLACE PRE-COLUMBIAN INDIAN RELIC, ARTIFACT
    ",142491152783,45267,United States
http://www.ebay.com/itm/PRE-COLUMBIAN-CHANCAY-MASK-HEADBAND-EX-SOTHEBYS
    -81-/122667684382?hash=item1c8f90521e:g:HaEAAOSwFqJWsiYl,"PRE-COLUMBIAN CHANCAY
    MASK & HEADBAND EX: SOTHEBYS &#039;81",122667684382,90382,United States
http://www.ebay.com/itm/Huge-Clava-stone-sceptre-Mapuche-culture-Pre-columbian-stone
    -/162643682178?hash=item25de51a382:g:~OQAAOSwFWVZnhxs,"Huge Clava stone sceptre.
    Mapuche culture, Pre-columbian stone.",162643682178,43758,United States
http://www.ebay.com/itm/Incredible-Mayan-Dagger-Spear-found-in-Lamanai-Belize-
    Precolumbian-/272820833190?hash=item3f85638ba6:g:1DgAAOSwzaJYA7rp,"Incredible
    Mayan Dagger/Spear found in Lamanai, Belize, Precolumbian",272820833190,33950,
    Canada
http://www.ebay.com/itm/Pre-Columbian-Mayan-9-Inch-Heavy-Stone-Shaman-Effigy-Statue
    -/172843515590?hash=item283e46bec6:g:NfIAAOSwRJ1ZpZ6g,"Pre-Columbian Mayan 9 Inch
    Heavy Stone Shaman Effigy Statue",172843515590,36455,United States
http://www.ebay.com/itm/ARTEMIS-GALLERY-Olmec-Green-Stone-Ceremonial-Hand-Axe-Celt
    -/381788292287?hash=item58e45b04bf:g:L24AAOSwCGVX6VLn,"ARTEMIS GALLERY Olmec Green
     Stone Ceremonial Hand Axe (Celt)",381788292287,36062,United States
http://www.ebay.com/itm/RARE-AUTHENTIC-PRE-COLUMBIAN-COCLE-EFFIGY-POTTERY-VESSEL-
    HUNCHBACK-PANAMA-/132311684269?hash=item1ece6414ad:g:wnUAAOSwSDZZpjAp,"RARE
    AUTHENTIC PRE COLUMBIAN COCLE EFFIGY POTTERY VESSEL HUNCHBACK PANAMA
    ",132311684269,75444,United States
http://www.ebay.com/itm/Rolex-Gold-Blue-Face-/172829359663?hash=item283d6ebe2f:g:7
    b8AAOSwKoRZYxq3,"Rolex Gold (Blue Face)",172829359663,60220,United States
```