基於密文計算的即時串流紀錄資料分析方法

陳志華
中華電信研究院
副研究員
國立交通大學
兼任助理教授
chihua0826@gmail.com

楊雅婷
中華電信研究院
研究員
yatin@cht.com.tw

謝欣翰
中華電信研究院
副研究員
hsinhan@cht.com.tw

林佳宏
中華電信研究院
研究員
arphen@cht.com.tw

謝佳珉
中華電信研究院
研究員
charming@cht.com.tw

官大勝
中華電信研究院
高級研究員
ditto@cht.com.tw

摘要

有鑑於對即時且大量資料運算和分析的需求，本研究提出一個即時串流紀錄資料分析系統與方法，此系統架構由使用者設備、線上網頁伺服器設備、線上資料庫伺服器設備、紀錄資料處理設備、分散式資料庫裝置、管理者設備、紀錄資料分析設備、資料探勘模組裝置、分散式運算裝置、快取資料庫裝置、以及組合節點設備所組成。方法步驟主要包含 8 個步驟：(1) 紀錄線上資料、(2) 資料加密、(3) 存入分散式資料庫、(4) 選擇資料探勘模組、(5) 指派工作予分散式運算裝置，並進行密文計算、(6) 暫存運算結果至快取資料庫裝置、(7) 回傳和解密、以及(8) 通知、顯示結果。將可即時收集日誌紀錄，並結合快取機制，進行即時串流分析，並將同形加密方法和資料探勘方法整合，以支援在密文形式下進行資料分析。

關鍵詞：江河運算、同形加密、線性迴歸、紀錄資料、Splunk。

Abstract

For the requirements of real-time big data computation and analysis, this study proposes a real-time streaming log data analysis system and method. The system includes mobile devices, online servers, online database servers, log data processing device, distributed database devices, administrator device, log data analysis device, data mining module device, distributed computation devices, cache devices, and combination node. The method includes (1) recording online data, (2) data encryption, (3) storing data into distributed database devices, (4) selecting a data mining module, (5) assigning jobs to distributed computation devices and performing homomorphic encryption, (6) storing temporal results into cache devices, (7) returning and decryption, and (8) alerting and notification. This system and method can record log data immediately and provide streaming computing based on cache mechanisms. Then homomorphic encryption methods and data mining methods are integrated to support ciphertext analyses.

Keywords: Stream computing, homomorphic encryption, linear regression, log data, Splunk.

1. 前言

近年來，隨著物聯網和行動通訊的發展，設備與設備之間開始自動經由網路互連和交換訊息，進而產生出許多巨量資料分析的議題。在物與物交換訊息的過程中，存在著許多的商機與應用，例如：透過追蹤設備的異常資料以進行即時或預測維修，減少停機成本。然而，卻也因為物與物的連網間所傳輸的訊息過於大量，以至於現行系統和方法較無法有效收集這些日誌紀錄並進行即時分析。

在過去的文獻中，有許多方法被提出和實作以進行日誌紀錄資料存取、日誌紀錄資料分析、日誌紀錄資料探勘分析、以及日誌紀錄異常資料和事件追蹤等。但在目前發展的技術中，卻多為採用門檻值進行判斷，而無法對紀錄進行深度分析。而現行的日誌紀錄資料探勘方法也較不適用於處理巨量資料，且在事件追蹤需耗費大量的運算資源。

有鑑於此，本研究提出一個即時串流紀錄資料分析系統與方法。當客戶端設備向線上網頁伺服器設備發出要求時，將由線上網頁伺服器設備向線上資料庫伺服器設備存取相關資訊，並分別得到網路服務要求紀錄資料和資料庫操作紀錄資料。再將紀錄資料分別傳送至紀錄資料處理設備進行依各個格式解析，解析完成後再儲存至分散式資料庫裝置。管理者設備可向紀錄資料分析設備發出紀錄資料分析要求，並經由紀錄資料分析設備連結至資料探勘模組裝置，且可選擇合適的資料探勘模組。選定資料探勘模組後可指派予分散式運算裝置進行運算，並且在運算過程中可將運算結果和相關參數暫存於快取資料庫裝置，以提升後續即時串流紀錄資料分析的速度。之後，再由組合節點設備整合分散式運算裝置運算之結果，並且產製最後分析結果予紀錄資料分析設備，供管理者決策參考。

此論文以下分為六個章節，在第二節中將探討日誌紀錄資料存取與分析相關的技術背景。第三節說明基於密文計算的即時串流紀錄資料分析系統的系統架構和各個元件內容。第四節說明基於密文計算的即時串流紀錄資料分析方法，各個步驟的作法和設計原理。第五節則針對本研究提出之系統與方法進行實作和分析。最後一節則說明此論文之結論與未來研究方向。

2. 文獻探討

在日誌紀錄資料存取上，文獻「Distributed usage metering of multiple networked devices」主要可以側錄封包，並針對不同的封包進行紀錄[4]。另一文獻「Log data recording device, log data recording method and storage medium storing program」主要包括壓縮單元及記錄單元，可即時針對指定影像進行壓縮和儲存日誌資料[7]。然而，這些方法雖然可以即時側錄和紀錄封包，並對資料進行壓縮處理，但卻未對紀錄進行分析，以及產製分析結果供管理者參考。

在日誌紀錄資料分析上，文獻「Storing log data efficiently while supporting querying to assist in computer network security」主要包括事件接收器和儲存管理器，並由事件接收器接收日誌資料、處理日誌資料，並針對特定欄位設定最大值和最小值之門檻，並超出門檻之資料將進行儲存，並可作為安全資訊/事件管理系統之應用[5]。另一文獻「Storing log data efficiently while supporting querying」主要包括事件接收器和儲存管理器，並由事件接收器接收日誌資料、處理日誌資料，並針對特定欄位設定最大值和最小值之門檻，並超出門檻之資料將進行儲存，並可作為安全資訊/事件管理系統之應用[6]。然而，這些方法雖然可以即時紀錄和處理日誌資料，但卻只能做門檻值判斷，無法對紀錄進行深度分析。

在日誌紀錄資料探勘分析上，文獻「事件良率關聯分析系統及方法以及電腦可讀取儲存媒體」主要包括資料庫與紀錄分析單元，可儲存事件紀錄、影響紀錄以及良率紀錄，並依據良率紀錄建立關聯分析[2]。然而，此方法雖然可以紀錄和處理日誌資料，但卻只能做關聯分析，無法對紀錄進行即時深度分析，也無法進行即時追蹤和處理。

在日誌紀錄異常資料和事件追蹤上，文獻「基於跨層日誌記錄的資料軌跡追蹤系統與方法」主要可以取得不同的日誌資料來源，紀錄更多的日誌資料和存取軌跡，並且作為犯罪追查之用途[1]。另一文獻「以目錄服務存取日誌為媒介同步異動資料之系統」提出以目錄服務(Directory Service)存取日誌(Access Log)為媒介同步異動資料之系統，運用日誌資料的可讀性、規則性及完整性之特徵，紀錄異動資料，後續並可依此追蹤系統異動狀況[3]。然而，這些方法雖然可以紀錄和追蹤異動資料和事件，但卻無法對紀錄進行深度分析，並且需要大量的運算資源來進行處理和分析。

3. 系統架構

有鑑於對即時且大量資料運算和分析的需求，本研究提出基於密文計算的即時串流紀錄資料分析系統與方法，系統架構中主要包含有(1) 使用者設備、(2) 線上網頁伺服器設備、(3) 線上資料庫伺服器設備、(4) 紀錄資料處理設備、(5) 分散式資料庫裝置、(6) 管理者設備、(7) 紀錄資料分析設備、(8) 資料探勘模組裝置、(9) 分散式運算裝置、(10) 快取資料庫裝置、以及(11) 組合節點設備，如圖 1 所示。其中，在紀錄資料處理設備的部分，本研究將運用 Splunk 或 Logstash 等工具進行實作開發，以收集線上網頁伺服器設備和線上資料庫伺服器設備之紀錄資料；在分散式資料庫裝置的部分，本研究將結合 NoSQL 基礎之 HBase 或 MongoDB 進行實作開發，以儲存相關的紀錄資料；在分散式運算裝置的部分，將運用 Hadoop 或 MongoDB 中的 MapReduce 開發模

型來將資料進行分割和合併，加快資料運算和分析速度。此外，本研究設計快取資料庫裝置，以因應即時大量資料運算的需求，可同時收集各個異質資料來源，即時進行分析與運算，並避免重覆運算來達到高效率的即時串流紀錄資料分析系統，即時將資料回饋予管理者，詳述如下。

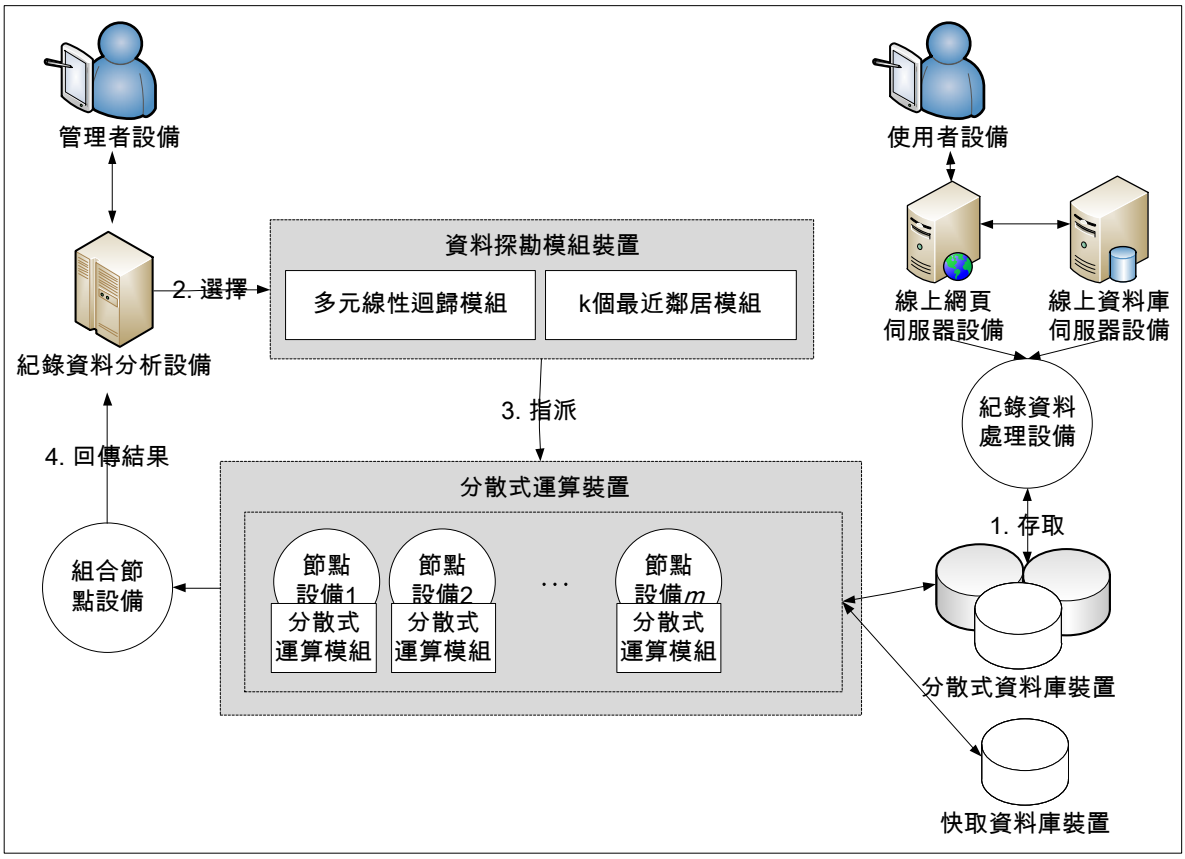

圖 1 系統架構圖

3.1 系統元件說明

在本節中主要將針對系統的各個元件分別進行描述。

3.1.1 使用者設備

使用者設備為個人電腦、平板、智慧型手機、個人數位助理、或車載設備等，並可運用設備中的瀏覽器元件(例如：Internet Explorer、Chrome、Firefox、Safari 等瀏覽器)或其他可連線之應用程式連線至線上網頁伺服器設備，並向線上網頁伺服器設備要求網路服務和相關資訊與應用。

3.1.2 線上網頁伺服器設備

線上網頁伺服器設備可運用微軟 Internet Information Services (IIS)、Apache 等網頁伺服器元件進行實作開發，並架設各種網路服務供使用者操作。可依使用者設備傳送的網路服務要求向線上資料庫伺服器存取相關資料後，提供網路服務和相關資訊給使用者設備。並且針對每個網路服務要求進行紀錄，可依據伺服器元件分別儲存網路服務使用紀錄資料，如：IIS Log 或 Apache Log。並且可將網路服務使用紀錄資料傳送至紀錄資料處理設備進行解析和儲存。

3.1.3 線上資料庫伺服器設備

線上資料庫伺服器設備可運用微軟 SQL Server、MySQL、甲骨文資料庫、IBM DB2、PostgreSQL 等資料庫伺服器元件進行實作開發，並提供各種資料庫操作方法(至少包含有新增、修改、刪除、查詢等)，供線上網頁伺服器設備存取。可接收線上網頁伺服器設備的資料庫操作要求，並依其要求回覆相關資訊。並且

針對每個資料庫操作要求進行紀錄，並可分別依不同的資料庫元件產生資料庫操作紀錄，且將資料庫操作紀錄傳送至紀錄資料處理設備進行解析和儲存。

3.1.4 紀錄資料處理設備

紀錄資料處理設備可運用 Splunk、Logstash 等紀錄處理和解析元件進行實作開發，並提供各種紀錄資料解析模組(至少包含有網路服務使用紀錄資料解析模組和資料庫操作紀錄資料解析模組)。其中，網路服務使用紀錄資料解析模組包含有 IIS Log 或 Apache Log 解析功能，可解析來自線上網頁伺服器設備的紀錄；並且，資料庫操作紀錄資料解析模組至少包含有微軟 SQL Server Log 等解析功能，可解析來自線上資料庫伺服器設備的紀錄。解析完成後再依其紀錄之格式進行解析後分別儲存至分散式資料庫裝置中。例如，表 1 為 IIS Log 紀錄資料，紀錄資料處理設備將可解析此資料，並至少分別得到該紀錄資料的日期為 2015-08-18、時間為 09:12:15、客戶端 IP 為 10.144.198.130、伺服器端 IP 為 10.144.192.1、連結埠號為 80、要求的網路服務為 /index.html、回應狀態碼為 200、客戶端瀏覽器為 Mozilla/4.0+ (compatible;MSIE+5.5; +Windows+2000+Server)。

表 2 IIS Log 資料案例

#Software: Microsoft Internet Information Services 6.0 #Version: 1.0 #Date: 2015-08-18 09:12:15 #Fields: date time c-ip cs-username s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status cs(User-Agent) 2015-08-18 09:12:15 10.144.198.130 - 10.144.192.1 80 GET /index.html - 200 Mozilla/4.0+(compatible;MSIE+5.5;+Windows+2000+Server)

3.1.5 分散式資料庫裝置

分散式資料庫裝置可運用 HBase、MongoDB 等分散式資料庫元作實作開發，以儲存和操作巨量資料，並可具備叢集互相備援，以支援紀錄資料分析和處理。分散式資料庫裝置主要可儲存經紀錄資料處理設備解析後的網路服務使用紀錄和資料庫操作紀錄。並且當得於分散式運算裝置進行分散式運算和紀錄分析時，提供紀錄資料供分散式運算裝置運算。

3.1.6 管理者設備

管理者設備為個人電腦、平板、智慧型手機、或個人數位助理等，並可運用設備中的瀏覽器元件(例如：Internet Explorer、Chrome、Firefox、Safari 等瀏覽器)或其他可連線之應用程式連線至紀錄資料分析設備，並經由紀錄資料分析設備連線至資料探勘模組裝置，以及選擇適合的資料探勘模組，再指派予分散式運算裝置進行運算，最後再由組合節點設備整合運算結果並回傳資料分析設備，以及由資料分析設備回覆予管理者設備。

3.1.7 紀錄資料分析設備

紀錄資料分析設備為一個具有網路服務的伺服器，可經由網路服務介面與管理者設備、資料探勘模組裝置、組合節點設備介接和傳收資料。紀錄資料分析設備由管理者手動或自動連線至資料探勘模組裝置，並選擇適合的資料探勘模組，並指派予分散式運算裝置進行運算，以及向組合節點設備取得運算結果。

3.1.8 資料探勘模組裝置

資料探勘模組裝置為一個具有網路服務的伺服器，可經由網路服務介面與資料探勘模組裝置、分散式運算裝置介接和傳收資料，並且可包含多個資料探勘模組，可供模組予分散式運算裝置進行運算和分析。其中，資料探勘模組裝置包含有 k 個最近鄰居模組、多元線性迴歸模組等資料探勘模組；並且 k 個最近鄰居模組包含可運算 k 個最近鄰居方法之分散式運算模組，而多元線性迴歸模組包含可運算多元線性迴歸方法之分散式運算模組。將可依選定之資料探勘模組指派給分散式運算裝置進行運

算和分析。

3.1.9 分散式運算裝置

分散式運算裝置可運用 Hadoop、MongoDB 等分散式運算元作進行實作開發，並至少包含有多個節點設備、多個分散式運算模組以分析巨量資料。其中，節點設備可依紀錄資料分析設備選定之資料探勘模組產生多個分散式運算模組，並可向分散式資料庫裝置取得紀錄資料，指派予分散式運算模組進行分析；分散式運算模組得依選定的資料探勘模組分別進行運算和分析紀錄資料。例如，可運用 Hadoop 或 MongoDB 所提供之 MapReduce 分散式運算模組分別依指派之資料探勘模組任務執行分散式運算，再將運算結果整合傳送予組合節點設備。

3.1.10 快取資料庫裝置

快取資料庫裝置可運用關聯式資料庫或非關聯式資料庫元件進行實作開發，儲存分散式運算裝置暫存各個紀錄資料分析要求運算結果和相關參數，以作為加速運算使用。例如，在分散式運算裝置執行 k 個最近鄰居模組之分散式運算模組後，將可取得最相似的數筆紀錄資料，並可將此數筆紀錄儲存至快取資料庫裝置，後續在即時運算時可先取出快取資料庫裝置中最相似的數筆紀錄資料進行比對和分析。在分散式運算裝置執行多元線性迴歸模組之分散式運算模組後，將可運算得到之線性迴歸模型參數(包含有斜率和截距)儲存至快取資料庫裝置，後續在即時運算時可先取出快取資料庫裝置中線性迴歸模型參數，以及加入新的紀錄資料和刪除舊的紀錄資料，避免重覆計算，可大幅提升運算效率。

3.1.11 組合節點設備

組合節點設備為一個具有網路服務的伺服器，可經由網路服務介面與紀錄資料分析設備、分散式運算裝置介接和傳收資料，可擷取分散式運算裝置各個運算結果，並進行整合和分析，再將分析結果回傳予紀錄資料分析設備。

3.2 k 個最近鄰居之即時串流運算

在本節中將以分析網路紀錄資料來產製定位資訊為案例說明 k 個最近鄰居之即時串流運算。

3.2.1 紀錄資料處理設備

紀錄資料處理設備得收集智慧型手機回報之經緯度座標資料(即訓練位置)和基地台訊號強度集合資料。並且紀錄資料處理設備得解析上述資料，紀錄每個訓練位置 $L = \{l_1, l_2,..., l_m\}$ 及其對應的基地台訊號強度集合資料 $c_i = \{c_1^i, c_2^i,..., c_n^i\}$，並將其紀錄於分散式資料庫裝置中。其中，$c_j^i$ 代表第 j 個基地台之訊號強度，$j = 1,\ldots, n$ (在此案例共有 n 個基地台)。

後續當手機移動時，手機可以測量和回報其附近的基地訊號強度集合 $r = \{r_1, r_2,..., r_n\}$，並將由資料探勘模組裝置、分散式運算裝置、快取資料庫裝置以 k 個最近鄰居模組計算基地訊號強度集合 r 與分散式資料庫裝置中所有位置及其訊號強度集合進行比對，評估手機當時可能的位置。

3.2.2 分散式資料庫裝置

分散式資料庫裝置得運用 HBase、MongoDB 等分散式資料庫元作實作開發，得儲存每個訓練位置 $L = \{l_1, l_2,..., l_m\}$(在此案例共有 m 個位置)及其對應的基地台訊號強度集合資料 $c_i = \{c_1^i, c_2^i,..., c_n^i\}$。當得於分散式運算裝置進行分散式運算和紀錄分析時，分散式資料庫裝置得提供紀錄資料供分散式運算裝置運算。

3.2.3 資料探勘模組裝置

資料探勘模組裝置得至少具備一 k 個最近鄰居模組，可用以評估每一個訊號強度集合 r 之位置 $loc(r)$。在本案例中運用 Euclidean 距離運算方法，採用公式(1)將訊號強度集合 $r = \{r_1, r_2,..., r_n\}$ 與資料庫中的每一個位置 l_i 及其訊號強度集合 $c_i = \{c_1^i, c_2^i,..., c_n^i\}$ 進行距離計算。再針對每一個訓練位置進行 Euclidean 距離運算，並使用公式(2)找出訊號強度最接近的位置 h_1 和最接近的 k 個位置(即 $\{h_1, h_2,..., h_k\}$)。資料探勘模組裝置得將 k 個最近鄰居模組指派予分散式運算裝置執行。

$$dist(r, c_i) = \sqrt{\sum_{j=1}^{n} (r_j - c_j^i)^2} \quad (1)$$

$$h_1 = \arg\min_{l_i \in L} dist(r, c_i) \quad (2)$$

3.2.4 分散式運算裝置

分散式運算裝置具備多個節點設備，而每個節點設備具備多個分散式運算模組，並且可依資料探勘模組裝置選定之資料探勘模組進行運算。在此案例中，於分散式資料庫裝置共有 m 個位置(即 m 筆資料需比對)，可將此 m 筆資料均勻分配於每個節點設備，再於每個節點設備中的分散式運算模組分別執行 k 個最近鄰居模組，並分別取得最接近的 k 個位置(即$\{h_1, h_2, ..., h_k\}$)。並將此最接近的 k 個位置傳送予組合節點設備，以供組合節點設備計算出最後的位置資訊。

3.2.5 組合節點設備

組合節點設備可接收來自分散式運算裝置運算所得到之資訊，並進行整合和產製分析結果。在此案例中，組合節點設備可接收多個節點設備分別計算所得到之 k 個位置(即$\{h_1, h_2, ..., h_k\}$)，再從此集合中進行比對取得 k 個絕對接近位置，運用公式(3)產製訊號強度集合 $r = \{r_1, r_2, ..., r_n\}$對應的位置資訊 $l(r)$。

$$l(r) = \frac{\sum_{i=1}^{k} \left[\frac{h_k - h_i}{h_k - h_1} \times z_i \right]}{\sum_{i=1}^{k} \frac{h_k - h_i}{h_k - h_1}} \quad (3)$$

3.2.6 快取資料庫裝置

快取資料庫裝置主要將可儲存由分散式運算裝置運算的結果和相關參數資訊，以供後續分析使用，加速分析效率。在此案例中，將可由每個節點設備計算取得最接近的 $q \times k$ 個位置資訊(其中 $q \times k$ 小於 m，並且 q 為正整數)及其對應的基地台訊號集合。後續分析同一智慧型手機回報之基地台訊號強度集合時，可對該最接近的 $q \times k$ 個位置資訊及其對應的基地台訊號集合進行分析，而不用再比對原始的 m 筆資料，以加速分析效率。此外，可分析該智慧型手機移動的速度，當智慧型手機移動速度緩慢或靜止時，q 值可設定為極小值(例如：1)；而當智慧型手機快速移動時，q 值可設定為較大值。

3.3 多元線性迴歸模組之即時串流運算

在本節中將以分析交通紀錄資料來產製交通預測資訊為案例說明多元線性迴歸模組之即時串流運算。

3.3.1 紀錄資料處理設備

紀錄資料處理設備得收集清潔車之車載設備回報到站時間資訊，並得由紀錄資料處理設備解析到站時間資訊，並產製站到站之間的旅行時間，如：第 r 筆資料的第 i-n-j 個清運點到第 i-n 個清運點間的旅行時間為 $t^r_{i-n-j,i-n}$。並且紀錄資料處理設備可將旅行時間集合儲存至分散式資料庫裝置，以供後續分析處理使用。

3.3.2 分散式資料庫裝置

在此案例中，分散式資料庫裝置得運用 HBase、MongoDB 等分散式資料庫元作實作開發，並得儲存每個站到站之間的旅行時間。

3.3.3 資料探勘模組裝置

資料探勘模組裝置得至少具備一多元線性迴歸模組，可用以產製站到站之間的旅行時間之關聯性(至少包含斜率和截距等)。在本案例中，將分析歷史資料中的 m 筆資料來產生 k 個加權線性迴歸模型 $T_{t_{i,i-n,i-n-j}}\left(t^r_{i-n-j,i-n}\right)$。第 i-n 個清運點到第 i 個清運點的預測旅行時間 $t^r_{i-n,i}{}'$ 可以運用多元加權線性迴歸模型(如公式(4)所示)進行計算取得。在執行階段中主要將依據第 i-n 個清運點的前 k 個清運點到達第 i-n 個清運點的旅行時間(即 $\left\{t_{i-n-1,i-n}, t_{i-n-2,i-n}, \cdots, t_{i-n-k,i-n}\right\}$)和訓練好之多元加權線性迴歸模型，以預測第 i-n 個清運點到第 i 個清運點的預測旅行時間(如公式(5)所示)。

$$t^r_{i-n,i}{}' = \frac{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}} \times T_{t_{i,i-n,i-n-j}}\left(t^r_{i-n-j,i-n}\right)}{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}}}$$
$$= \frac{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}} \times \left(a_{t_{i,i-n,i-n-j}} \times t^r_{i-n-j,i-n} + b_{t_{i,i-n,i-n-j}}\right)}{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}}} \quad (4)$$

where, $w_{t_{i,i-n,i-n-j}} = 1 - \frac{\sum_{r=1}^{m} \frac{\left| t^r_{i-n,i} - t^r_{i-n,i}{}' \right|}{t^r_{i-n,i}}}{m}$

$$a_{t_{i,i-n,i-n-j}} = \frac{m\left(\sum_{r=1}^{m} t_{i-n-j,i-n}^{r} t_{i-n,i}^{r}\right) - \left(\sum_{r=1}^{m} t_{i-n-j,i-n}^{r}\right)\left(\sum_{r=1}^{m} t_{i-n,i}^{r}\right)}{m\left(\sum_{r=1}^{m} \left(t_{i-n-j,i-n}^{r}\right)^2\right) - \left(\sum_{r=1}^{m} t_{i-n-j,i-n}^{r}\right)^2}$$

and $b_{t_{i,i-n,i-n-j}} = \frac{1}{m}\left(\sum_{r=1}^{m} t_{i-n,i}^{r} - a_{t_{i,i-n,i-n-j}} \sum_{r=1}^{m} t_{i-n-j,i-n}^{r}\right)$

$$t_{i-n,i}' = \frac{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}} \times \left(a_{t_{i,i-n,i-n-j}} \times t_{i-n-j,i-n} + b_{t_{i,i-n,i-n-j}}\right)}{\sum_{j=1}^{k} w_{t_{i,i-n,i-n-j}}} \quad (5)$$

3.3.4 分散式運算裝置

分散式運算裝置具備多個節點設備，而每個節點設備具備多個分散式運算模組，並且可依資料探勘模組裝置選定之資料探勘模組進行運算。在此案例中，由於多元加權線性迴歸模型中多為加法和乘法運算，並具有結合律之特性，得依歷史資料 m 筆均勻分配於每個節點設備，再於每個節點設備中的分散式運算模組分別執行多元加權線性迴歸模型；亦或得依待產製之 k 個加權線性迴歸模型均勻分配於每個節點設備，再於每個節點設備中的分散式運算模組分別執行多元加權線性迴歸模組。並且，在運算完成後，得將每個多元加權線性迴歸模型之斜率(如 $a_{t_{i,i-n,i-n-j}}$)、截距(如 $b_{t_{i,i-n,i-n-j}}$)、以及權重(如 $w_{t_{i,i-n,i-n-j}}$)分別儲存於快取資料庫裝置，以供後續即時分析使用。

3.3.5 組合節點設備

組合節點設備可接收來自分散式運算裝置運算所得到之資訊，並進行整合和產製分析結果。在此案例中，組合節點設備可接收多個節點設備分別計算所得到之 k 個加權線性迴歸模型及其相關參數(即斜率、截距、以及權重)，運用公式(5)產製預測之第 i-n 個清運點到第 i 個清運點的旅行時間。

3.3.6 快取資料庫裝置

快取資料庫裝置主要將可儲存由分散式運算裝置運算的結果和相關參數資訊，以供後續分析使用，加速分析效率。在此案例中，將可由每個節點設備計算得到之每個多元加權線性迴歸模型之斜率、截距、以及權重，將分別儲存於快取資料庫裝置，以供後續即時分析使用。此外，後續資料異動時，由於多元加權線性迴歸模型中多為加法和乘法運算，並具有結合律之特性，故搭配快取資料庫裝置暫存之資料，只需加入新增資料或減去刪除資料即可調整每個多元加權線性迴歸模型之斜率、截距、以及權重，而不用再計算原始的 m 筆資料，以加速分析效率。

4. 系統方法

一種具備密文計算的即時串流紀錄資料分析方法，流程如圖二所示，此方法主要將包含 8 個步驟：(1) 紀錄線上資料；(2) 資料加密；(3) 存入分散式資料庫；(4) 選擇資料探勘模組；(5) 指派工作予分散式運算裝置，並進行密文計算；(6) 暫存運算結果至快取資料庫裝置；(7) 回傳和解密；以及(8) 通知、顯示結果，詳述如下。

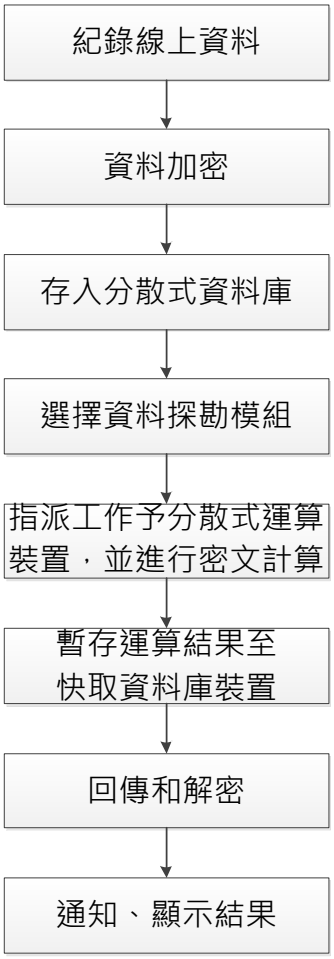

圖 2 方法流程圖

4.1 紀錄線上資料

紀錄資料處理設備將把線上網頁伺服器設

備和線上資料庫伺服器設備的運作紀錄進行收集和儲存。例如：車載機 1 (使用者設備)於 09:00:00、09:03:20、09:07:00 分別到站點 1、站點 2、站點 3；車載機 2 (使用者設備)於 10:00:00、10:04:00、10:08:10 分別到站點 1、站點 2、站點 3；車載機 3 (使用者設備)於 11:00:00、11:03:30、11:07:20 分別到站點 1、站點 2、站點 3；以及車載機 4 (使用者設備)於 12:00:00、12:03:40 分別到站點 1、站點 2，如表 2 所示。使用者設備到達各站點時，將經由中介軟體(如：RESTful API)回報其位置資訊和時間資訊至線上網頁伺服器設備和線上資料庫伺服器設備，而紀錄資料處理設備將可把此紀錄儲存和分析，並計算出站到站時間之間的旅行時間，如：車載機 1 從站點 1 到站點 2 的旅行時間($t_{1,2}$)為 200 秒、從站點 2 到站點 3 的旅行時間($t_{2,3}$)為 220 秒，如表 3 所示。

表 2 到站時間

	站點 1	站點 2	站點 3
車載機 1	09:00:00	09:03:20	09:07:00
車載機 2	10:00:00	10:04:00	10:08:10
車載機 3	11:00:00	11:03:30	11:07:20
車載機 4	12:00:00	12:03:40	

表 3 站到站之間的旅行時間(單位：秒)

	站點 1-站點 2	站點 2-站點 3
車載機 1	200	220
車載機 2	240	250
車載機 3	210	230
車載機 4	220	

4.2 資料加密

紀錄資料處理設備收集到線上網頁伺服器設備和線上資料庫伺服器設備的運作紀錄後，得運用加密演算法對紀錄資料進行加密。紀錄資料處理設備可計算所得之站到站之間的旅行時間，再分別計算出 $t_{1,2}$ 乘上 $t_{2,3}$ 的值和 $t_{1,2}$ 平方的值，得到相關參數值，如表 4 所示。

紀錄資料處理設備得運用一私密金鑰 p、一公開金鑰 q、一任意整數值 z，以公式(6)對相關參數值進行加密，在此案例中設定私密金鑰 p 為 39,916,801、公開金鑰 q 為 112,909、任意整數值 z 為 7。如：明文 44,000 經由加密後得到密文為 279,461,607，整理如表 5 所示。

$$f(x)=(x+p\times z)\bmod(p\times q) \quad (6)$$

表 4 相關參數值

	$t_{1,2} * t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} * t_{1,2}$
車載機 1	44000	200	220	40000
車載機 2	60000	240	250	57600
車載機 3	48300	210	230	44100

表 5 加密後相關參數值

	$t_{1,2} * t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} * t_{1,2}$
車載機 1	279,461,607	279,417,807	279,417,807	279,457,607
車載機 2	279,477,607	279,417,847	279,417,847	279,475,207
車載機 3	279,465,907	279,417,817	279,417,817	279,461,707

4.3 存入分散式資料庫

紀錄資料處理設備可將紀錄資料的明文或密文儲存至分散式資料庫裝置。在本案例中，紀錄資料處理設備可將表四之加密後相關參數值儲存至分散式資料庫裝置中，在資料庫中儲存密文，可防範資料庫被竊取時造成資料外洩的風險。

4.4 選擇資料探勘模組

管理者設備可連線上紀錄資料分析設備，經由紀錄資料分析設備存取資料探勘模組裝置，選擇其偏好的資料探勘模組。在本案例中，管理者可選擇多元線性迴歸模組，後續步驟將以多元線性迴歸模組進行分析和運算。

4.5 指派工作予分散式運算裝置，並進行密文計算

資料探勘模組裝置可依管理者選擇的資料探勘模組，指派給分散式運算裝置，並由分散式運算裝置以多個分散式運算模組進行計算，且得以密文計算方式對密文進行處理。分散式運算裝置將可依管理者所選定之多元線性迴歸模組，運用公式(4)和公式(5)的運算需求，運用多個分散式運算模組分別加總所需之參數值，加總後結果如表 6 所示。在本案例中，以計算一組迴歸模組參數 a 和 b 為例，但不以此為限；分散式運算裝置可同時運用多個分散式運算模組，進行多組迴歸模組參數計算。

表 6 加密後相關參數值之加總

	$t_{1,2} * t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} * t_{1,2}$
加總	838,405,121	838,253,471	838,253,471	838,394,521

4.6 暫存運算結果至快取資料庫裝置

分散式運算裝置運算完之結果得暫存運算結果至快取資料庫裝置，作為後續資料分析處理使用。在本案例中，已加總完車載機 1、車載機 2、車載機 3 的資料，將可把加總結果暫存至快取資料庫裝置，後續直接套用加總結果，無需再重新加總車載機 1、車載機 2、車載機 3 的資料。

4.7 回傳和解密

分散式運算裝置將把運算完結果回傳至組合節點設備，並由組合節點設備進行資料組合，以及得對密文進行解密。當組合節點設備收到分散式運算裝置運算結果後，得運用與紀錄資料處理設備相同之一私密金鑰 p、一公開金鑰 q、一任意整數值 z，運用公式(7)進行解密，在此案例中設定私密金鑰 p 為 39,916,801、公開金鑰 q 為 112,909、任意整數值 z 為 7。如：加總後結果之密文 838,405,121 經由解密後得到明文為 152,300，整理如表 7 所示。

$$g(x) = (x) \bmod (p) \tag{7}$$

表 7 解密後相關參數值之加總

	$t_{1,2} * t_{2,3}$	$t_{1,2}$	$t_{2,3}$	$t_{1,2} * t_{1,2}$
加總	152,300	650	700	141,700

可將解密後的加總資料，以及資料筆數 $m = 3$，運用公式(4)分別計算出 a 和 b 參數，如公式(8)所示。再運用公式(9)預測車載機 4 從站點 2 到站點 3 的旅行時間，可得估計約為 236 秒，故預測車載機 4 到達站點 3 的到站時間為 12:07:36。

$$a = \frac{3 \times (152300) - (650)(700)}{3 \times (141700) - (650)^2} = 0.730769$$

and

$$b = \frac{1}{3}(700 - 0.730769 \times 650) = 75 \tag{8}$$

$$t = 0.730769 \times 220 + 75 = 235.7692 \approx 236 \tag{9}$$

4.8 通知、顯示結果

組合節點設備可將運算後結果傳送至紀錄資料分析設備，由紀錄資料分析設備通知管理者設備，並於管理者設備呈現分析結果。當組合節點設備計算完預測結果時，可將預測結果傳送至紀錄資料分析設備，再由紀錄資料分析設備通知管理者設備，並於管理者設備上呈現車載機 4 到達站點 3 的預測到站時間為 12:07:36。

5. 系統實作

本研究目前以車訊快遞系統為例進行開發與實作，在管理者設備操作介面上，管理者可以先連結至紀錄資料分析設備，並選擇資料探勘模組裝置中所欲採用的資料探勘模組(如：k 個最近鄰居模型)，如圖 3 所示。當管理者選擇資料探勘方法完成後，將會進入到輸入資料來源和表單欄位的設定，再選擇其所欲分析的資料表及其欄位。以行動定位服務為例，將分析蜂巢網路的電信訊號強度，因此需選擇相關的資料表(如：Cell Information)，並設定輸入的資料欄位資訊(如：Location 和 RSSI)，如圖 4 所示。

圖 3 選擇資料探勘方法

圖 4 設定輸入資料

如此，將可以依輸入的位置資訊和網路訊號強度進行資料訓練和學習。在設定完成後，將進入輸出資料欄位設定，並由於行動定位服務主要將進行定位，所以在資料輸出的部分設定為 Location，如圖 5 所示。如此，在訓練和學習的過程中將會依歷史資料，訓練得到位置估計的模型，以利後續運用。

圖 5 設定輸出資料

最後，管理者將相關參數設定完成後，將由高效率資料分析平台分析出結果，並顯示其正確率等資訊予開發者瀏覽。如果管理者覺得目前的資料探勘模型和訓練結果合適，將可點擊部署，後續將以此模型持續運用，並進行即時串流紀錄資料分析。

6. 結論與未來發展

由於物聯網和行動通訊的發展，帶動起許多巨量資料分析的議題，並由於在目前發展的日誌紀錄資料分析技術中，卻多為採用門檻值進行判斷，而無法對紀錄進行深度分析。而現行的日誌紀錄資料探勘方法也較不適用於處理巨量資料，且在事件追蹤需耗費大量的運算資源。有鑑於對即時且大量資料運算和分析的需求，本研究提出一個即時串流紀錄資料分析系統與方法，此系統架構由使用者設備、線上網頁伺服器設備、線上資料庫伺服器設備、紀錄資料處理設備、分散式資料庫裝置、管理者設備、紀錄資料分析設備、資料探勘模組裝置、分散式運算裝置、快取資料庫裝置、以及組合節點設備所組成。方法步驟主要包含 8 個步驟：(1) 紀錄線上資料、(2) 資料加密、(3) 存入分散式資料庫、(4) 選擇資料探勘模組、(5) 指派工作予分散式運算裝置，並進行密文計算、(6) 暫存運算結果至快取資料庫裝置、(7) 回傳和解密、以及(8) 通知、顯示結果。將可即時收集日誌紀錄，並結合快取機制，進行即時串流分析，並將同形加密方法和資料探勘方法整合，以支援在密文形式下進行資料分析。

在本研究目前主要以 k 個最近鄰居之即時串流運算和多元線性迴歸模組之即時串流運算進行實作和案例研究。在未來研究，將可以考慮發展更多不同的資料探勘方法的即時串流運算，並將資料探勘方法結合同形加密技術，以強化在物聯網環境中的資訊安全。

參考文獻

[1] 李忠憲、劉奕賢、盧建同、張家瑋，“基於跨層日誌記錄的資料軌跡追蹤系統與方法”，***中華民國發明專利***，I484331，2015。

[2] 邱文仁、蕭維勤、藍玉潔、祁孝麟、陳信雄 、吳冠良、許嘉慧、游明蒼，“事件良率關聯分析系統及方法以及電腦可讀取儲存媒體”，***中華民國發明專利***，I251752，2006。

[3] 黃子峻、黃華泰、黃茁淳，“以目錄服務存

取日誌為媒介同步異動資料之系統”，***中華民國發明專利***，I396104，2013。

[4] Drucker, B. and Teng, A., “Distributed Usage Metering of Multiple Networked Devices,” ***US Patent***, US20030123442 A1, 2003.

[5] Huang, W., Tang, W., Beedgen, C.F., “Storing Log Data Efficiently While Supporting Querying to Assist in Computer Network Security,” ***US Patent***, US9031916 B2, 2015.

[6] Huang, W., Zhou, Y., Yu, B., Tang, W., Beedgen, C.F., “Storing Log Data Efficiently While Supporting Querying,” ***US Patent***, US9166989 B2, 2015.

[7] Komatsu, T., “Log Data Recording Device, Log Data Recording Method and Storage Medium Storing Program,” ***US Patent***, US7664331 B2, 2010.