一種基於 Spark 處理即時串流資料分析的編輯系統
Real-Time Stream Data Analysis Processing Editor Based on Spark

王蓁蒂, 陳鏡宇, 鄒育庭, 黎曲峯、王秉豐
財團法人資訊工業策進會
智慧網通系統研究所
Chen-Ti Wang, Ching-Yu Chen, Yu-Ting Tzou, Chu-Feng Li, Ping-Feng Wang
Institute for Information Industry Smart Network System Institute
Email: {ctwang, chingyuchen, tzyuting, chufengli, pfwang}@iii.org.tw

摘　要

目前智慧系統整合領域之各廠商已發展許多感測器技術，並受到感測串流的資料快速增加及複雜性提升，單一類型的感測器的資訊，已經無法滿足及正確辨識使用者所需的事件。故本技術開發複雜事件即時決策平台系統，平台以 Spark 為核心處理引擎，打造便於資料分析人員使用的完整資料分析流程。

希望提供給非專業的程式開發人員使用，讓一般的資料分析人員使用，在此平台上能夠建構即時串流資料分析的程式，透過圖形拖拉的方式建構決策分析規則，以更快速的提供企業所需的診斷、偵測及分析應用。

關鍵字：即時串流資料分析、複雜事件處理引擎、機器學習

一、前言

因應智慧聯網環境多元化趨勢，感測資訊多樣複雜的情形下，產生大量即時資訊串流匯入儲存空間，在尚未有完整決策分析的引擎與平台下，且缺乏專家知識的整體方案，導致於廠商接收許多感測資料並不知如何使用，需要仰賴具有程式專業能力的資料分析人員，批次性進行資料篩選、處理及分析，難以即時處理與分析大量且多元資料。

現階段資料分析產業應用發展情形，著重於智慧聯網的即時串流資訊分析，建構共通化引擎技術的能量，將智慧聯網從感測資料的偵測與蒐集，建立專家知識建立規則，發展可提供非程式開發者的專業領域人員容易操作與使用分析的軟體工具，提供資料分析與決策建議。

有鑑於傳統資料分析，必須由專業程式開發者才能進行分析的情形，本技術利用 Apache Spark[1] & Hadoop 等開放式原始碼專案，建立物聯網整合服務。導入 Spark Streaming，提供程式開發人員 Open API 的設計工具，並採用圖形化拖曳編輯工具，讓一般的資料分析人員也能透過圖形化拖曳編輯工具，建構即時串流資料分析的程式，編輯使用者所需之各種複雜處理的邏輯規則，並進一步利用事件分析偵測技術，創造各種應用情境。再透過核心引擎之 Spark Streaming，結合資料流處理和機器學習演算法，完成機器模擬與運算，進行大量即時串流資料分析。

二、技術應用情境

本平台以 Spark 為核心處理引擎，打造便於資料分析人員使用的完整資料分析流程。平台主要分為 5 大部分，分別是提供給使用者操作的[複雜事件決策規則編輯工具和處理模組]、Spark 核心引擎[Spark Cluster]、存放原始資料的[Data Warehouse 與 Message Queue]、存放便於查詢資料的[NoSQL Database] 以及與外部元件界接的[External Component Interface]。

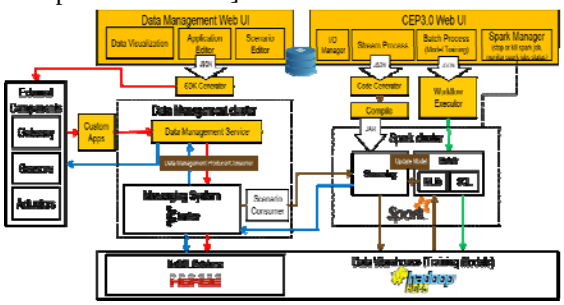

圖 1、複雜事件即時資料分析平台架構

複雜事件即時決策平台系統，希望提供給非專業的程式開發人員使用，讓一般的資料分析人員使用，在此平台上能夠建構即時串流資料分析的程式，並且透過圖形拖拉的方式建構決策分析規則，以更快速的提供企業所需的診斷、偵測及分析應用。本平台四大特色說明如下：

- 提供串流事件分析及圖形化規則編輯工具：傳統 Spark ecosystem 的分析工具，都以批次資料(batch data)分析為主，提供在 Spark Engine 上進行 Spark 的資料分析，本平台提供在串流資料的即時環境中進行即時資料的分析與診斷，提供更為即時的資料處理應用。

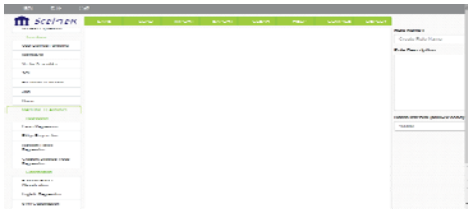

圖 2、串流事件分析及圖形化規則編輯工具

- 提供匯入自訂的分析模型演算法：藉由 import Lib 管理畫面來進行使用者自訂系統的資料匯入介面。當使用者將所開發的分析模型演算法註冊到平台中時，平台會檢查所提供的函式名稱、所屬參數介面及資料型態等內容是否一致，確保後續演算法執行時，能正確的運作。
- 提供使用者分析模型的儲存庫：使用者可以將以訓練學習完成的分析模型，註冊到平台中，系統會依選擇的 Model Type 進行相關分類，並在對應的規則元件中選擇已有的分析模型進行串流資料的分析偵測。
- 減少資料的執行時間：在資料接收的平行化方面，藉由 Message Cluster (如 Kafka[2]等) 考慮平行地接收網路資料，以解決資料接收成為系統的瓶頸。資料處理的平行化方面，如果運行在計算 stage 上的並發任務數不足夠大，就不會充分利用集群的資料來源。故藉由參數傳遞平行度，或者修改參數預設值，提供對於分散式 reduce 操作，並藉由配置屬性控制任務數以提供最佳的平行化運作。

複雜事件分析規則編輯工具 Web UI 操作，提供「資料的輸出入元件模組」、「資料處理元件模組」、「機器學習模型」選定資料來源提供相關規則元件。

資料分析人員透過[複雜事件分析規則編輯工具]之 Web UI 操作，在[Data Manager]選定資料來源後，可利用[Stream Process]編輯界面針對來源資料設計分析資料的規則，此編輯界面的特色在於分析人員不需要寫程式，即可利用圖形化工具迅速的建立、調整分析規則，如下圖所示。設計好的規則流程圖會轉換成 JSON 格式的文件或儲存於資料庫[sCEPter DB]中。

圖 3、複雜事件決策規則編輯工具

若[Stream Process]編輯界面提供的資料分析工具無法滿足資料分析人員分析資料的需求，則資料分析人員可透過[Model Manager]界面上傳自定義的資料分析模型函式庫插件，資料分析人員可自行利用 Java 撰寫所需的運算程式碼，再將編譯完成的程式碼自行打包成 Jar 檔，透過[Model Manager]界面上傳到[Model repository]，待[Stream Process]編輯界面採用到相關運算時，於 Compile 階段會自動帶入對應的 Jar 格式函式庫。

資料分析人員亦可在[Data Manager]選定歷史資料來源後，透過[Batch Process]編輯介面針對歷史資料設計歷史資料的資料探勘流程，設計好的探勘流程圖會轉換成 JSON 格式的文件，連同[Workflow Executor]一併部屬到[Spark Cluster]，即可反覆測試探勘參數，訓練出最佳的數值預測模型，將訓練好的模型儲存於[Model Repository]中，以便於在[Stream Process]階段可用來做 online prediction。

運行 Stream 或是 Batch 的過程中，資料分析人員若想監控 Spark 的運作狀態，可透過[Spark Manager]監看，並可針對特定的任務進行停止並刪除，以便符合最佳的資源分配狀況。

資料管理及規則處理流程方式方面，[Stream Process]的運作流程中，外部感測器資料首先推送到[Kafka MQ]上，Spark 上的[Stream Process Engine]再定期從[Kafka MQ]上抓取來源 topic 的資料，透過資料分析人員所撰寫的串流資料分析規則分析完成後，再將結果推送至目標 topic，外部的控制元件則可定期抓取此 topic 的資訊，以便進行特定動作。資料分析人員亦可針對不同的應用，將處理完的資訊存到 [Hadoop Database]。

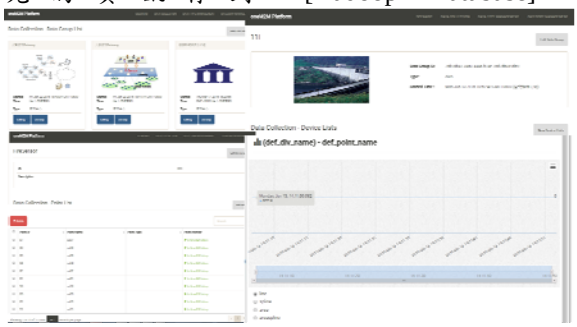

圖 4、複雜事件資料管理模組與分析工具

此外，資料分析人員所需之複雜事件即時決策規則模組技術，主要採用 Apache Hadoop 方法的架構在使用模型進行預測時，大多將中繼計算結果儲存於硬碟上，當資料量龐大時，容易導致系統花過多的時間於 I/O 上，進而降低系統的整體執行效能。

因此，開發複雜事件即時決策規則模組技術，依據 Apache Spark 之 In-memory Computing 架構下，從數個預測模型中篩選出適當的規則(Rule)與型樣(Pattern)放進 Memory 中，以分散式計算方式結合數個預測模型之預測結果，運用平行處理大量資料達到快速判斷與即時預測。

三、結論

本平台以 Spark 為核心處理引擎，設計資料分析流程，研發複雜事件決策開發平台，研發完成平台模組包含：複雜事件決策規則編輯工具和處理模組、Spark Cluster、Data Warehouse 和 Message Queue、NoSQL Database 以及與外部元件界接的 External Component Interface 等五大模組。

提供程式開發者 Open API 設計工具，利用程式擷取服務；與資料分析人員圖形化拖曳工具編輯

界面，提供包含串流事件偵測規則及機器學習資料分析編輯工具，可彈性、方便接取串流規則，簡易化資料處理模組與分析方式。

並開發複雜事件即時決策規則模組技術，依據 Apache Spark 架構下，從數個預測模型中篩選出適當的規則(Rule)與型樣(Pattern)放進 Memory 中，以分散式計算方式結合數個預測模型之預測結果，運用平行處理大量資料達到快速判斷與即時預測。

參考文獻

[1] http://spark.apache.org/
[2] http://kafka.apache.org/