【附件三】教育部教學實踐研究計畫成果報告格式(系統端上傳 PDF 檔)

教育部教學實踐研究計畫成果報告
Project Report for MOE Teaching Practice Research Program

計畫編號/Project Number：PEE107073
學門分類/Division：工程
執行期間/Funding Period：107 年 8 月 1 日起至 108 年 7 月 31 日

基於深度學習與大數據分析技術之空汙預測永續管理平台
雲端計算、大數據技術平台與應用

計畫主持人(Principal Investigator)：楊朝棟

共同主持人(Co-Principal Investigator)：

執行機構及系所(Institution/Department/Program)：東海大學資訊工程學系

繳交報告日期(Report Submission Date)：108 年 9 月 10 日

基於深度學習與大數據分析技術之空汙預測永續管理平台

一. 報告內文(Content)(請繳交 3 至 10 頁成果報告，不含封面、參考文獻、相關佐證附件與連結，檔案大小以 20mb 為限。)

1. 研究動機與目的(Research Motive and Purpose)

請描述所選擇研究議題的問題挑戰與背景、教學實務現場遇到之挑戰以及該議題的重要性與影響力。

(1)教學實踐研究計畫動機。(如：研究的發想背景、問題意識、問題重要性、影響及應用層面等)

近年來環境受到PM2.5的影響十分嚴重,許多地區都被列為空污高危害區域,且 PM2.5 更是影響人體健康的風險，使得民眾對於環境中的空氣品質感到憂慮。PM2.5 是大氣中的超細懸浮顆粒物，也稱為可入肺顆粒物。事實上，我們平常呼吸的空氣中充斥著 PM2.5，因為它們可以在空氣中懸浮或隨著氣流四處漂浮，也易吸附著有毒物質如二 s 出針對管制、監測 PM2.5 的標準，主要是為了更有效地監測因工業化而產生對人體有害的細小顆粒物（細懸浮微粒）。PM2.5 已經成為一項重要的監測空氣污染程度的指數。另外許多流行病理學研究，已證實 PM2.5 對健康的危害，包括早逝、支氣管炎、氣喘、心血管疾病、肺癌等相關疾病，PM2.5 的來源包含境外傳入、工業排放物、交通汙染等，在政府還沒有辦法解決空氣汙染的根本問題時，民眾應適當保護自己免於 PM2.5 的危害。

(2)教學實踐研究計畫主題及研究目的。(如：既有課程突破、新設跨領域課程規劃等)

目前校內現有課程面相大都屬於介紹及入門，並尚未有課程針對單一議題設計出一套專業且深入的教學方法，故本計畫將針對空汙議題開創跨領域課程，培養具備資訊，環境工程，社會科學領域之大數據人才。資訊課程設計方面為大數據相關應用技術，如：政府開放資料介接與爬蟲程式撰寫，分散式儲存技術與叢集運算分析等課程；增加自架感測器硬體設計課程，如 Arduino 開發加裝感測器接收 PM2.5，溫溼度等資料，環境空污監測由自身做起，分析周邊數據，讓這些空污開放數據不再只是遙不可及的數字，而是與自己息息相關。環境科學課程設計方面為環境開發相關評估，透過設置自架感測器於各地點，監測周遭環境，檢視居住環境是否健康，適宜外出活動時間等等。透過環境監測，進而擴展至人文科學上，空氣品質不僅與氣候，風向等因素影響，與人的習性也是有著密不可分的關聯性存在。透過社會科學的幫助，建立一套解決方案的模型。

本計畫也結合本校高等教育深耕計劃，本校所提出的高教深耕計畫，是一個以校務發展計畫為基礎，銜接現有學生學習問題與未來社會與產業需求的創新計畫。除了專業能力，我們更希望能培養學生的博雅通識素養與跨域學習能力，讓學生有能力認知、面對及解決未來社會各項關鍵議題。空氣汙染的議題應用可以結合環工系、生科系、景觀系等相關專業科系之能量，與資工系的資訊能量來為

資料實踐更多潛藏的價值。

希望整合深耕計劃及此教學實踐研究計畫，由專業到跨域的各項學習，我們希望能夠打造一個創新學習的生態系統，包括了從數位生活到行動學習、共學共創與跨域師資培育、產業鏈結與職涯發展、創業札根到新創加速等四個構面。本校電算中心、圖書館以及教學資源中心等單位搭配合作建立數位生活與行動學習平台、數位學習中心，以「東海學習的一天」串連學生一天的生活與學習。在創新學習模式中，問題導向式學習（Problem Based-Learning, PBL）已大量被教師在課堂教學活動式使用，藉由有趣以及與生活真實聯連結的問題做學習媒介，希望能有效引起學生的高度學習動機，並強化現有課程的深度與廣度。

(3)教學實踐研究計畫研究目的及目標。

培養學生具備實作一套具跨領域的巨量資料數據處理平台之能力，系統服務的面向可以更加多元，不再只是侷限於資訊應用層面，服務特定族群，而是將難以親近的數據透過系統的建置將數據提供給使用者，讓數據更加平易近人更容易取用，將應用層面拓展至一般社會大眾，創造各式各樣的生活應用。以提供社會大眾之所用。課程上不再只是接收既有知識，而是將課程所學應用於生活場域，利用跨領域的教學激盪出創新思維以建置一套跨領域應用的資訊數據處理平台模組。

2. 文獻探討(Literature Review)

請針對本教學實踐研究計畫主題進行國內外相關文獻、研究情況與發展或實作案例等之評析。

台北科技大學環境環境工程與管理研究所張哲鈞碩士，發表之「建立空氣污染地圖之分析、預測與預警機制以臺北市為例」研究，此究探討週遭空氣品質濃度時，因監測站彼此間相距數公里之遙，測站濃度完全無法描述其他任何位置受到局部固定源及移動源影響所造成之真實濃度。因此本研究利用大氣擴散模式 AERMOD 結合 2015 年氣象資料模擬點源及線源排放情況，以此為底圖，並利用 GIS 地理資訊系統調整為實際空品濃度分佈狀況，以達每小時或每日濃度預警。在探討週空品預測模式上則利用本研究週空品預測模式整合 AERMOD 底圖，並達到預測與預警未來週空品濃度以及週空品濃度分佈情形。

針對 104 年臺北市懸浮微粒(PM10)、細懸浮微粒(PM2.5)、二氧化硫(SO2)、二氧化氮(NOX)、一氧化碳(CO)及臭氧(O3)六大主要空氣污染物濃度變化，對臺北市民健康損失的影響，研究結果顯示健康損失區域與人口密度呈現顯著正相關，並針對人口密度集中的臺北市五區(大安區、信義區、松山區、中正區和中山區) ，與週空品預警地圖結合，與空品標準比較進行提前一週警報。

IBM 中國研究院於 104 年提出之「基於深度學習的大數據空氣污染預報」，該研究為了更好地反映環境污染變化趨勢，為環境管理決策提供及時、全面的環境指標數據資料。為預防嚴重污染事件發生，開展城市空氣質量預報研究是十分必

要的。本研究針對環境大數據時代下的城市空氣品質提出預報，提出了一種基於深度學習的新方法。該方法通過模擬人類大腦的神經連接結構，將數據在原空間的特徵表示轉換到具有語義特徵的新特徵空間，自動地學習得到層次化的特徵表示，從而提高預報性能。得益於這種方式，新方法與傳統方法相比，不僅可以利用空氣質量監測，氣象監測及預報等環境大數據，充分考慮污染物的時空變化，空間分佈，得到語義性的污染物變化規律，還可以基於其他空氣污染預測方法的結果（如數值預報模式），自動分析其適用範圍，以及其優劣。因此，新方法通過模擬人腦思考過程實現更充分的大數據科學應用，一定程度上克服了現有方法的缺陷，應用上更加具有靈活性和可操作性。最後，通過實驗證明新方法可以提高空氣污染預報性能。

3. 研究方法(Research Methodology)

可包含實驗場域、研究對象、研究架構、資料蒐集方法與工具與分析方法等項目，但不限於列舉內容。

(1)研究說明。(請具體說明教學或課程設計，如：係針對整體課程、單元主題、教學方法、作業設計或評量策略...等不同研究主題所進行的具體設計。)

主題順序	主題	學習目標	教學單元影片	線上教學活動規畫
1	R 語言介紹	說明本課程授課大綱，並介紹課程開設目的以及目前空汙議題，如何透過本課程解決或改善空汙問題，安裝 R 語言編譯環境，介紹 R 語言統計相關語法與應用、政府公開資料抓取方法	單元 1：介紹課程架構 單元 2：R 語言環境建置與介紹統計語法 單元 3：R 語言套件抓取公開資料	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
2	佈署巨量資料系統平台	介紹巨量資料服務及相關技術如 Hadoop、Spark 等，並學習如何結合以上技術建置巨量資料平台，瞭解其運作原理並透過實際架設加深學習效果。	單元 1：Hadoop 服務介紹、架設與實例操作 單元 2：Spark 服務介紹、架設與實例操作	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
3	加入政府開放式資料並建置地理資訊系統(GIS)	加入環境汙染相關之政府開放式資料，。建置地理資訊系統，統整地理資訊與汙染資料，學習如何匯入開放式資料及地理資訊系統之架設與使用。	單元 1：政府開放式資料介紹與匯入 單元 2： 地理資訊系統建置與使用 單元 3: 開放式資料與 GIS 系統結合與進階應用	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業

主題順序	主題	學習目標	教學單元影片	線上教學活動規畫
4	建立污染熱區之空間插值基本模型(IDW)	介紹空間插值基本模型(IDW)之原理，並運用該技術來解決資料缺值之困境，藉由此過程使學員學習 IDW 空間插值方法之使用。	單元 1：空間插值基本模型原理介紹 單元 2：利用 IDW 推定目標數據實作與解析	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
5	建立深度學習模組進行資料預測	介紹深度學習理論以及 TensorFlow 深度學習架構，實作類神經網絡架構。了解其運作原理及如何利用深度學習技術訓練模組，以完成短期預估之目標。	單元 1：深度學習原理介紹 單元 2：TensorFlow 深度學習架構介紹、安裝與實例操作 單元 3：針對本系統之深度學習模組實作	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
6	定義預警機制並加入預警系統	依該空氣品質對人體造成影響之嚴重程度制定一空氣品質等級指標。並透過結合巨量資料運算技術及預警推播系統建置一預警系統。以此介紹預警推播機制之運作原理、了解預警指標是如何確立，以及熟悉預警系統的建置與使用。	單元 1：預警機制介紹與預警指標的確立 單元 2：預警系統之建置 單元 3：預警推播機制介紹，與預警系統之結合	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業

主題順序	主題	學習目標	教學單元影片	線上教學活動規畫
7	透過 ARIMA 改善深度學習的預測模組	介紹自回歸移動平均模型(ARIMA)時間序列方法之原理，並利用該方法優化深度學習模組。使學員理解 ARIMA 之運作原理並學習應用於深度學習模組上之方法。	單元 1：ARIMA 原理介紹與實例操作 單元 2：利用 ARIMA 改善深度學習模組	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
8	優化巨量資料系統平台並改進深度學習及預測模型	針對系統平台運行之流暢性、操作直覺性、圖形化顯示得清楚與否等角度做系統的修正與改進。認識、建立高斯模式擴散模型，加入汙染物擴散之評估方法與思維，並訓練學員思考可能改進方式的能力	單元 1：高斯模式擴散模型原理介紹 單元 2：系統平台之優化與內涵 單元 3：優化深度學習模組	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業
9	TensorFlow 預測模型精準度最佳化	學習透過 TensorBoard 輔助工具對深度學習之預測成效做評估，學習提升精準度的方法，並訓練學員自主思考解決問題的能力。	單元 1：TensorBoard 介紹與使用 單元 2： 預測模組之效能評估與改進	☒ 討論活動 ☐ 自動評分測驗 ☒ 同儕互評 ☐ 作業

(2)研究步驟說明。

A.研究架構

研究架構主要分為三階段，基本系統建置與資料蒐集、系統優化及資料分析、整合深度學習用於資料預測已達到本計畫所需之模擬與環境推算效果，使此計畫在往後的環境政策上提供有效的資訊與評估建議。

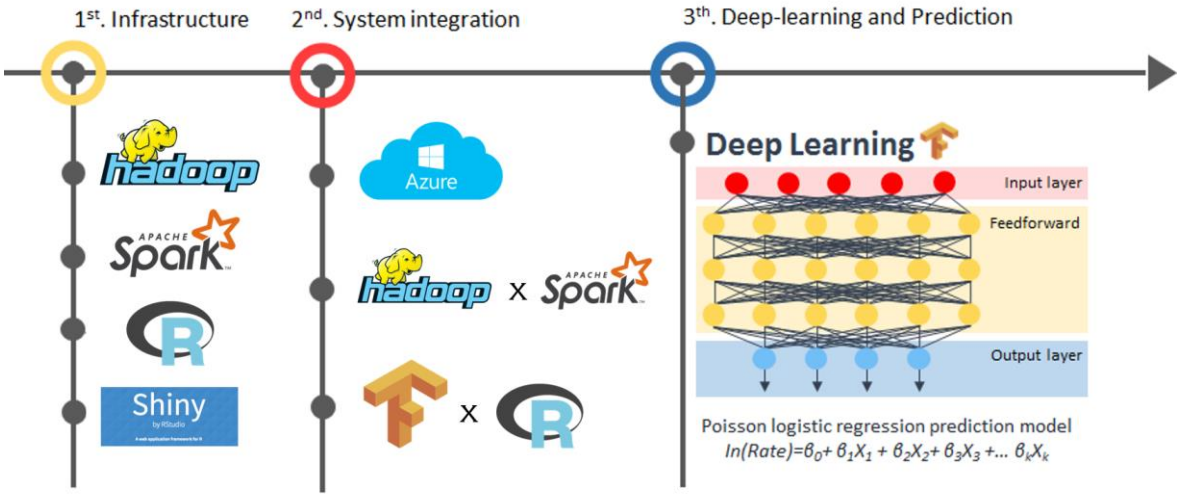

圖 1 計畫架構圖

B.研究假設

為了更好地反映環境污染變化趨勢，為環境管理決策提供及時、全面的環境指標數據資料。為預防嚴重污染事件發生，開發空氣品質預報研究是十分必要的。本研究假設針對環境大數據時代下的城市空氣品質提出預報，提出了一種基於深度學習的新方法，並透過 IDW 空間插值的方法改善感測器搜集資料密度不夠高和準確準確性的問題以及利用 ARMIMA 時間序列結合深度學習預測空氣品質的走向。該方法通過模擬人類大腦的神經連接結構，將數據在原空間的特徵表示轉換到具有語義特徵的新特徵空間，自動地學習得到層次化的特徵表示，從而提高預報性能。得益於這種方式，新方法與傳統方法相比，不僅可以利用空氣質量監測，氣象監測及預報等環境大數據，充分考慮污染物的時空變化，空間分佈，得到語義性的污染物變化規律，還可以基於其他空氣污染預測方法的結果（如數值預報模式），自動分析其適用範圍，以及其優劣。

C.研究範圍 (請說明教學投入及實施相關規劃，如：課程規劃為單一性或系列性、課程規劃相關說明、教學使用之相關資源、採用評量方式等，或社群教師課程設計與協作實踐方式等)

本課程透過一完整連貫性的規劃，期許學員在完成本課程後，可以在巨量資料、軟體工程、深度學習以及建置並維護一系統平台等領域皆能有所成長。藉由一連串扎實的實作過程，讓學員對系統之開發有更深刻的體驗，並能真實的參與開發過程中會遇到問題與困境，透過思考與不斷嘗試的過程，提供了自主學習的動力，也訓練了於此專業領域獨立思考之能力。教學將會使用政府提供之開放式資料，透過與政府機關的合作，節省了蒐集資料所花費的成本本，縮減研究所需的時間成本。本課程評量方式包含上課參與，上課進度的完整度、以及同儕互評等方式，重視團隊的討論表現以及心態的積極度。

授課教室使用本校教學卓越計畫與電算中心於今年 9 月 6 日在大智慧科技大

樓 ST039 會議室與 ST020 電腦教室建置 3D 軟體電腦教室，配有 76 台最新型 Dell 工作站，規格包含第 6 代 Intel® Core™ i7-6700 處理器、16GB DDR4 2133MHz 記憶體、1TB 7200rpm SATA3 介面硬碟、2GB DDR3 繪圖卡及 24”液晶螢幕…等，此教室電腦支援三種作業平台(Windows 10、Windows 7、CentOS)，並提供 SolidWorks、CADWork、ZWCAD、Unity、Photoshop、SAS、SPSS、MatLab、Illustrator、Dreamweaver…等軟體，支援高階繪圖、統計及多媒體設計軟體，讓本校高階資訊課程所需軟體都能執行。同時此教室配備更大的 24 吋螢幕及更清晰畫質，大幅提升教學彈性運用，讓師生使用更加便利。為了充分應用校園授權軟體於教學與研究，電算中心也建置 3D 軟體雲端電腦教室系統並搭配廣受師生好評的 3D 軟體雲服務，充分結合了實體電腦教室和雲端環境，讓學生的學習跳脫時間及空間的限制，使得學習更有效率。

D.研究對象 (請說明教學實踐研究對象特性及背景分析-如：學生先備特質或學習經驗的起始行為)

本課程之目標對象應具備基本資訊領域常識，基本虛擬機操作經驗，基礎程式語言能力，對數學不會感到抗拒，對巨量資料、深度學習及統計分析等領域具有濃厚興趣等特質。

E.研究方法及工具 (對於所提研究主題將採行何種方法及工具進行資料蒐集與分析)

針對本計畫研究方法及工具將部屬巨量資料系統平台(Hadoop+Spark)且搭配 R 語言進行資料分析與視覺化，並加入政府開放式資料並建置地理資訊系統(GIS)。

Apache Spark 是由美國柏克萊大學 AMPLab 所開發的一套開放原始碼的叢集計算框架，其目的是能與現有的 Hadoop 相容，但速度卻更快、更簡單且方便使用的系統，相對於 Hadoop 的 Disk-Based MapReduce，Spark 的 In-Memory Computing 設計對於某些應用能提供比 Hadoop 快 100 倍的速度，Spark 的 in-memory 允許使用者將資料載入到一個機器的記憶體中和進行多次查詢，所以 Spark 適合於機器學習的演算法，Spark 改善了 Hadoop 的硬碟存取時間過長的問題。Spark 有以下特點:

- Resilient Distributed Datasets (RDD):RDD 是整個 Spark 的核心，它提供了分散式的任務分配和排程及基礎的 I/O 功能，RDD 是 Spark 應用到 In-Memory 技術最主要的架構，它能控制切割資料的大小。
- Spark SQL:SparkSQL 介紹了一個新的資料抽象 SchemaRDD，它提供在 RDD 中查詢結構化或半結構化的資料，Spark SQL 提供了包含 Python、Scala 和 Java 的 API，SchemaRDDs 提供了單一介面達到有效地使用 Spark SQL 的數據資料，它還連接 JDBC/ODBC serve，透過 Shark 來提供了 SQL 語言的支援。

透過 Hadoop 與 Spark 兩個不同的分散式系統結合運用，建立一個可儲存大量數據資料，並且能夠達到快速運算分析的系統環境，以利於未來在面對龐大政府

開放資料時，能夠將資料進行完整快速的處理與分析。

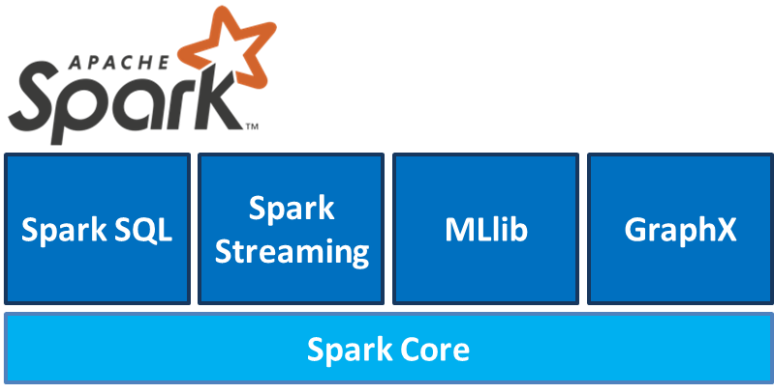

圖 2 Spark 架構

R 語言，一種自由軟體程式語言與操作環境，主要用於統計分析、繪圖、資料探勘。本計畫主要是使用其可以透過撰寫套件的方式加強資料探勘的功能。R 分散式環境架構會有兩個主要儲存的部份 HDFS 和 MySQL。這裡選擇了 HDFS 作為 R 分散式儲存的部分，其理由為一般用 R 語言進行資料分析時，最常取得資料的方式是從 localhost 的家目錄讀取，若資料較大或是要即時讀取較常用的方法是 R 讀取 MySQL。但資料量超出 single machine 能容納的 Big Data 量時，此時應該嘗試用 HDFS 的架構對資料進行讀和寫。

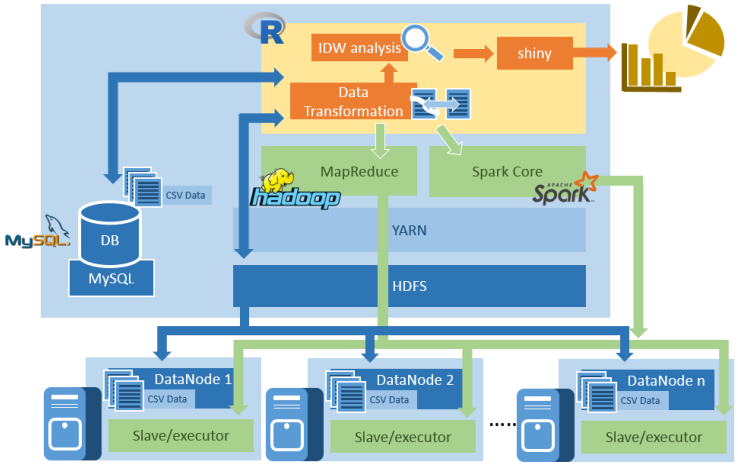

圖 3 R 分散式架構

開發介面則用到了 Rstudio，Rstudio 若在一般的 personal computer 就是一個對 R 語言開發的圖像介面，Rstudio 是一個友善方便操作的 R 語言分析環境介面。但是 Rstudio 不只是一個在 personal computer 進行分析的環境，也可以將該環境建置成在後端 server 上進行雲端計算的 Rstudio server。RStudio Server 是一个 IDE 集成環境並提供了 web 的功能，也就是說可以安裝遠程的計算機並通過 web 進行訪問後可以支持多個用户。

Rstudio server 不只是圖像介面方便操作這一項好處而已，更重要的是它可以讓 R 語言對一些後端的環境進行操作。除了可以直接用 R 的相關 packages 直接對 MySQL 進行遠端的操作外，開發者也可以用 Rstudio server 的環境讓 R 語言對

Hadoop 和 Spark 等分散式環境進行操作。對於用 R 開發分散式環境的使用者來說是一個相當重要的介面。

雖說資料讀取與儲存已有 HDFS 作為 R 分散式儲存的架構，但在 Data Transformation 方面也希望有個能利用 R 語言套件強化的方法去做處理，以利在資料重組上更有效率，這裡便選用了 SparkR 作為一個溝通的工具，SparkR 的環境依照 R 語言和 spark 系統之間的溝通方式可分成兩個部份，SparkR packages 和 JVM back-end。SparkR 是 R 環境的擴展 packages，SparkR 在 master machine 運行時提供了 RDD 和 DataFrame API。SparkR API 在 R Interpreter 中運行，而 Spark Core 則是在 JVM back-end 的環境中運行，需要有一種機制能讓 SparkR API 元件調用 Spark Core service。JVM back-end 是 Spark Core 中的其中一個組件，JVM back-end 提供了 R Interpreter 和 JVM 之間的橋接功能，JVM back-end 能夠讓 R Script 創建 Java class 的實例並且調用 Java 對象的實例方法

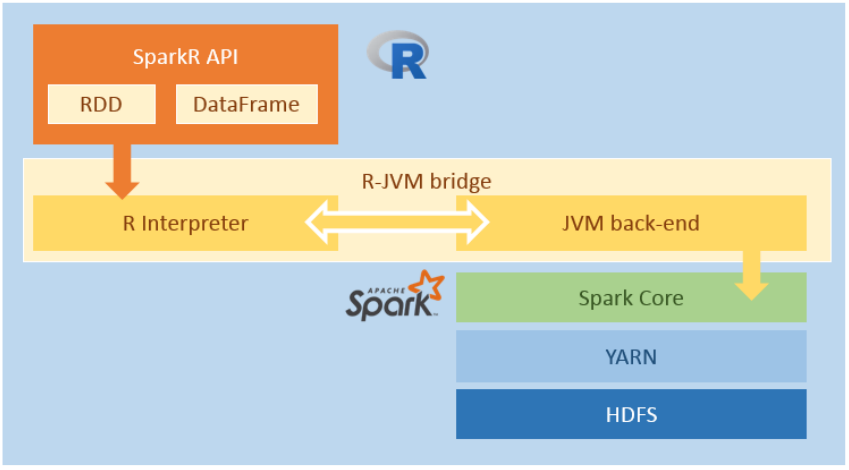

圖 4 SparkR 中與 JVM back-end 溝通

另外，在 Spark of distributed computing 這個環節中，Local host 的 master machine 主要扮演 Spark 的 Driver 角色(圖 8)：

1. 在 Driver 的 Spark Context 部份，將平行化的 R Script 傳遞至 R-JVM Bridge 中的 R Interpreter。
2. R Interpreter 和 JVM back-end 用 R-JVM Bridge 進行溝通，並將 R Script 重新編碼為 Java Class 的實例。
3. 轉換的 Java Class 可在 R-JVM Bridge 中呼叫其它 Spark Core service，並且它會呼叫 Java Spark Context 的過程中將工作分成多個 Tasks 分配至叢集中的 Worker Nodes。
4. 在叢集中的每個 Node 裡的 Executor 會將 Task 傳至 R 環境中的 worker 進行計算。

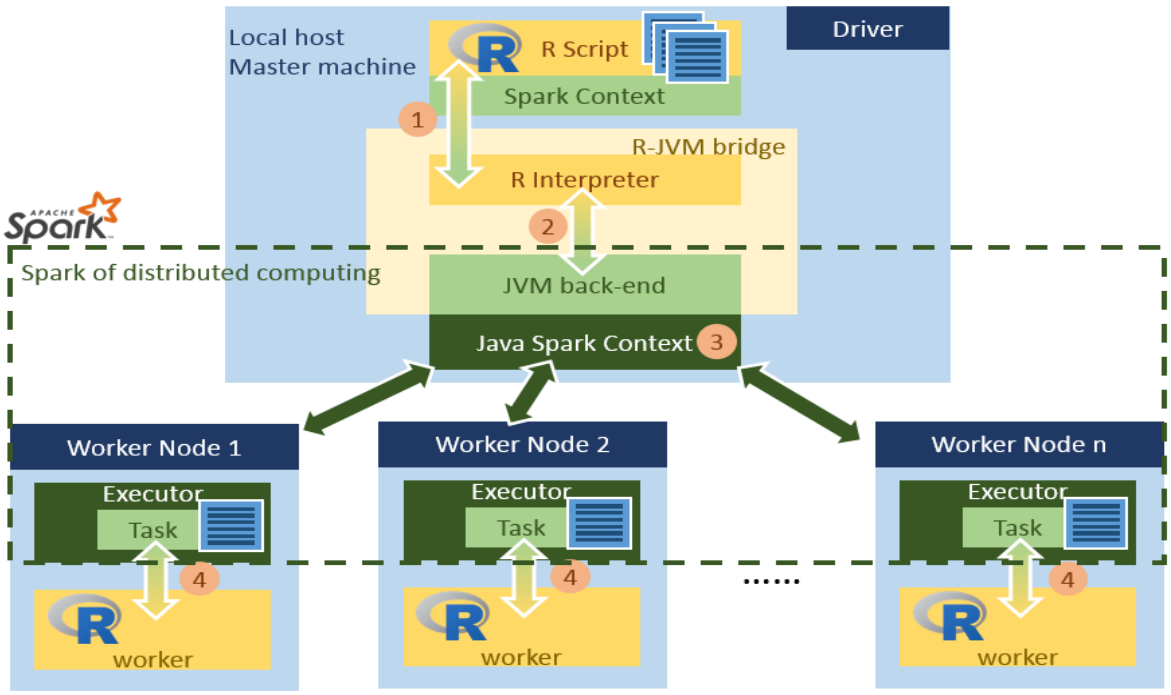

圖 5 SparkR 運作流程

反距離權重 (IDW) 空間插值法假設：彼此距離較近的事物之相似度會比彼此距離較遠的事物來的更高。當對於特定位置數值發生有缺漏的情況時，反距離權重法會採用預測位置周圍與距離預測位置較遠的測量值做比較，距離預測位置最近的測量值對預測值有更大的影響力。反距離權重法假定每個測量點都有一種局部影響，而這種影響與距離成反比的關係。由於這種方法為距離預測位置近的點擁有較大的權重，而權重卻作為距離的函數而減小，因此得名反距離權重法。如圖所示，距離紅區的點越遠其影響權重越小，而在紅區頂部的點影響權重最大。

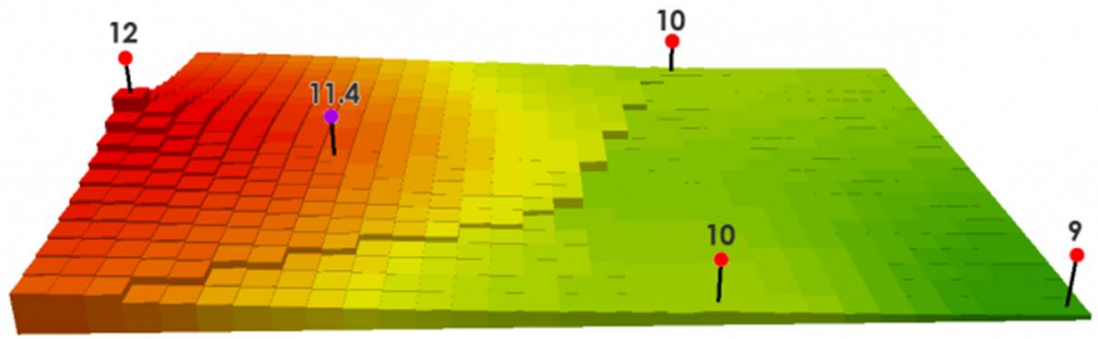

圖 6 IDW 模型範例

由於與目標位置距離較遠的事物的相關性會隨著距離越來越小，因此，當位置之間的距離增大，測量值與預測位置的值的關係將變得非常不密切時。為了節省計算所花費之時間，可以將距離遙遠到幾乎不會造成影響之數據點排除在外。因此，通過指定搜索鄰域來限制測量值的數量是一種非常常見的方法。鄰域的形狀將限制測量值的搜索距離和搜索位置。其他鄰域參數將限制在該鄰域中搜索的位置。

地理資訊系統是一種具有資訊系統空間專業形式的資料管理系統。在嚴格的意義上，這是一個具有集中、儲存、操作、和顯示地理參考資訊的電腦系統。例如，根據在資料庫中的位置對資料進行識別。其提供資料採集、操作、展現、座標與投影系統、空間分析等豐富的功能，使管理者在監控抑或是模擬某區域環境在某便因下可能產生的結果都能清楚呈現，透過建置 GIS 並結合 Shiny-Server 服務於客戶端呈現圖形化顯示時，可提供更豐富的地理屬性以及空間要素。

在此系統上匯入環境汙染相關之政府開放式資料，利用政府之空氣品質檢測站提供的資料，獲得更加豐富的資料，提升資料來源之可靠度，並節省架設與維護感測器材的成本。

透過上述圖像化技術並匯入政府的開放式環境資料後，即可在地圖上呈現各項環境數據，並透過分析與模擬即可清楚了解即時或未來可能發生的環境變化，並從區域思考相應環境對策或評估。

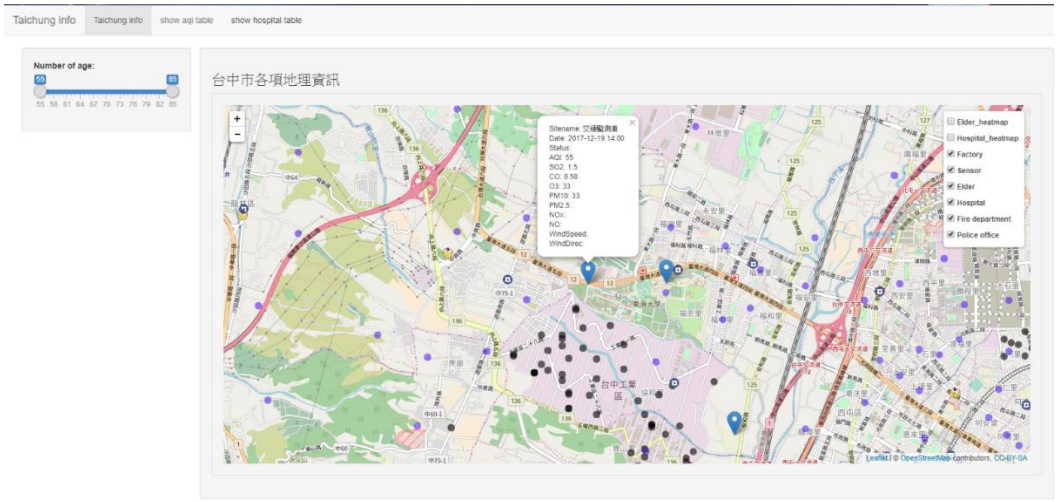

圖 7 GIS 示意圖

F.實施程序

1. 系統範疇界定
2. 佈署巨量資料系統平台
3. 加入政府開放式資料並建置地理資訊系統(GIS)
4. 建立污染熱區之空間插值基本模型(IDW)
5. 建立深度學習模組進行資料預測
6. 定義預警機制並加入預警系統
7. 透過 ARIMA 改善深度學習的預測模組
8. 優化巨量資料系統平台並改進深度學習及預測模型
9. TensorFlow 預測模型精準度最佳化

G.資料處理與分析

本計畫系統使用空氣品質指數（Air Quality Index AQI）做為空氣品質優劣的重要指標，AQI 值為一描述空氣品質狀況的數值指標，其數值大小可直接反映為空氣汙染之嚴重程度、與對人體之影響大小。AQI 分級計算參考的指標污染物為

以細顆粒物（PM2.5）、可吸入顆粒物（PM10）、二氧化硫（SO2）、二氧化氮（NO2）、臭氧（O3）、一氧化碳（CO）六項，以以上之指標求出空氣品質分指數（Individual Air Quality Index　IAQI）。

$$IAQI_P = \frac{IAQI_{Hi} - IAQI_{Lo}}{BP_{Hi} - BP_{Lo}} (C_P - BP_{Lo}) + IAQI_{Lo}$$

並利用求得各空氣污染指數之 IAQI 求得 AQI 值。

$$AQI = \max \{IAQI_1, IAQI_2, IAQI_3, \ldots, IAQI_n\}$$

藉由此指標獲得一量化的空氣品質優劣程度的判斷標準。然而欲使用該指標，若是該指數有數值缺漏之情事，由於該指數運算包含的數值眾多，補上缺漏的困難度較高，因此採用反距離權重法來解決該困境。

反距離權重 (IDW) 插值顯式假設：彼此距離較近的事物要比彼此距離較遠的事物更相似。當為任何未測量的位置預測值時，反距離權重法會採用預測位置周圍的測量值。與距離預測位置較遠的測量值相比，距離預測位置最近的測量值對預測值的影響更大。反距離權重法假定每個測量點都有一種局部影響，而這種影響會隨著距離的增大而減小。由於這種方法為距離預測位置最近的點分配的權重較大，而權重卻作為距離的函數而減小，因此稱之為反距離權重法。

由於彼此距離較近的事物比彼此距離較遠的事物更加相似，因此，隨著位置之間的距離增大，測量值與預測位置的值的關係將變得越來越不密切。為縮短計算時間，可以將幾乎不會對預測產生影響的較遠的數據點排除在外。因此，通過指定搜索鄰域來限制測量值的數量是一種常用方法。鄰域的形狀限制了要在預測中使用的測量值的搜索距離和搜索位置。其他鄰域參數限制了將在該形狀中使用的位置。

其中，我們透過計算 mape 值來進行數據準確度的驗證，以下是其計算方式：假設有n組驗證數據和預測值，驗證數據分別為：v1、v2、v3.....、v4，而預測數據則有：p1、p2、p3.....p4。先算出第 i 組的 percentage error，算出 percentage error 的方法就是 pi 和 vi 的相差絕對值，之後再除以第 i 組的驗證值 vi 即可。

$$error_1 = |\frac{p_1 - v_1}{v_1}|$$

求出每組的 percentage errors：error1、error 2、error3.....error i...error n。最後，將 n 組的誤差百分比進行平均後就求出 MAPE：

$$MAPE = \frac{\sum_{i=1}^{n} |\frac{p_1 - v_1}{v_1}|}{n}$$

由前文公式可以得知，MAPE 就是在計算 n 組 data training 資料的誤差呈度，或也可以說是誤差值與實際值的比值的平均。因此 MAPE 數值越小就越精準，相對越大其精準度就越低。若 MAPE 大於 50%則該組數據就沒有參考價值。

MAPE	Ability to predict
<10%	high accuracy
10% - 20%	good
20%-50%	reasonable
>50%	incorrect

利用 IDW 模組可針對某一區域內的汙染特區進行視覺化分析，以便於了解在特定的區域內，不同距離的污染熱源對觀測區域的影響大小。

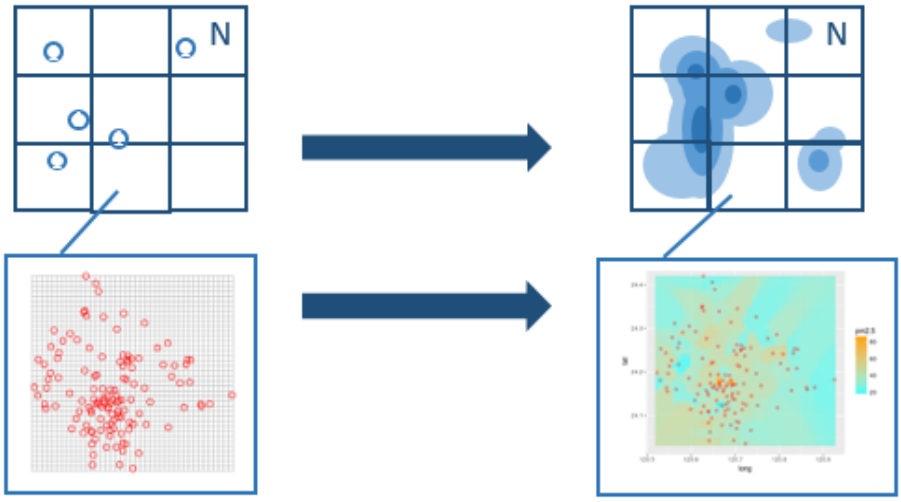

圖 8 IDW 示意圖

預測模組方面利用 ARIMA 演算法改善，ARIMA 模型全稱為自回歸移動平均模型(Autoregressive Integrated Moving Average Model，簡稱 ARIMA)。是由博克思(Box)和詹金斯(Jenkins)於 70 年代初提出的一著名時問序列預測方法，所以又稱為 box--jenkins 模型、博克思－詹金斯法。其中 ARIMA(p，d.q)稱為差分自回歸移動平均模型，AR 是自回歸，P 為自回歸項；MA 為移動平均，q 為移動平均項數，d 為時間序列成為平穩時所做的差分次數。ARIMA 模型可分為 3 種：(1)自回歸模型(簡稱 AR 模型)；(2)滑動平均模型(簡稱 MA 模型)；(3)自回歸滑動平均混合模型(簡稱 ARIMA 模型)。

ARIMA 模型的基本思想是：將預測對象隨時問推移而形成的數據序列視為一個隨機序列，以時間序列的自相關分析為基礎，用一定的數學模型來近似描述這個序列。這個模型一旦被識別後就可以從時間序列的過去值及現在值來預測未來值。ARlMA 模型在經濟預測過程中既考慮了經濟現象在時間序列上的依存性，又考慮了隨機波動的干擾性，對於經濟運行短期趨勢的預測準確率較高，是近年應用比較廣泛的方法之一。

未破損時序預測

驗證值	預測值	誤差
29.90	27.20	2.70
32.20	28.98	3.22
32.00	30.95	1.05

原始資料時序分析MAPE誤差值為:0.0742852149886525

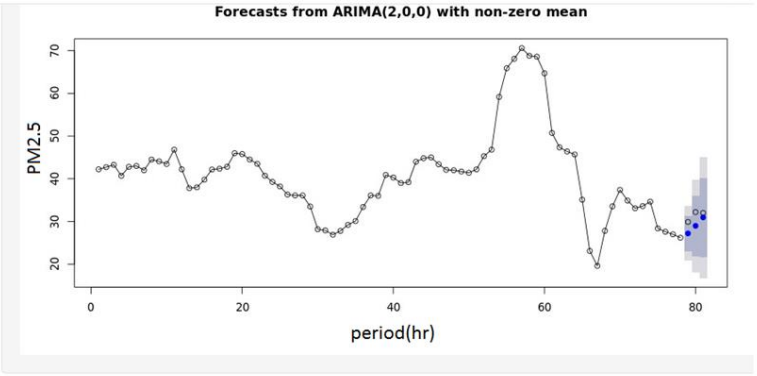

圖 9 ARIMA 示意圖

每年環境數據。隨著時間的推移而形成一個隨機時間序列，通過對該時間序列上空氣品質的隨機性、平穩性以及季節性等因素的分析，將這些環境數值之間所具有的相關性或依存關係用數學模型描述出來，從而達到利用過去及現在的數值信息來預測未來環境變化情況的目的。

短期空氣品質警報及應對方面，本計畫採用高斯煙流模式(閉合解析 analytical closed analytical closed-form)。

$$C(x,y,z)=\frac{Q}{2\pi u\sigma_y\sigma_z}\exp(-\frac{y^2}{2\sigma_y^2})\cdot\left\{\exp\left[-\frac{(Z+H_e)^2}{2\sigma_z^2}\right]+\exp\left[-\frac{(Z-H_e)^2}{2\sigma_z^2}\right]\right\}$$

C：污染物濃度，g/m^3或ug/m^3

Q：污染源強度（排放量），g/s

U：平均風速，m/s

σ_y、σ_z：y及z方向之擴散係數(diffusion coefficient)，m

H：污染源的有效高度，m

透過高斯模式可有效進行空氣污染物容許增量限值、健康危害風險評估、減量成效評估，其具體可模擬的汙染源模式有：點源（point sources）、體源（volume sources）、面源（area sources）、礦坑源（open pit sources），加上豐富的使用經驗使其易於使用並且適合做為長期評估，以及具有極大的彈性易於修改。

4. 教學暨研究成果(Teaching and Research Outcomes)

(1) 教學過程與成果

課程所規劃之教學過程主要為讓學生了解巨量資料分析技術與相關應用，課程上介紹巨量資料分析工具，以及發想創新校園應用。本課程規劃了整體符合業界所需的物聯網處理流程，並且配合創新校園應用議題，讓學生可以由此課程完成一整套的資料擷取、資料處理及資料分析流程。資料擷取由第一項主題開放資料使用介紹開始，本課程提供東海大學開放資料(http://opendata.thu.edu.tw)，讓學生於課堂上使用程式語言實作網路爬蟲抓取空氣品質資訊，此東海大學開放資料網站也提供了多種資料 API 供學生於課

後練習，最終也可以於期末發揮創造力實作專題。第二主題引導學生將資料整合至校園實際應用，讓學生實作資料視覺化，學習視覺化套件，如 ECharts 等，第三主題開始讓學生們實作資料分析流程，在大數據的應用上，資料遺失值是最常見的問題，於各場域數值中實作空間插值基本模型(IDW)。最終，第四主題帶領同學們為此專案做一個最終呈現，一項作品最重要的為完整性，加入預警機制及通報系統才能提升作品實用度，並結合智慧校園應用，來為師生提供參考依據。

其主要達成之目標如下：

- 針對巨量資料分析技術與應用課程模組開發課程教材與實驗教材。
- 讓學生了解與熟悉代表性的巨量資料分析技術之操作、應用與實現的方法。
- 讓學生了解與熟悉常見的巨量資料運算平台之原理、架構，並實際建置與操作。
- 讓學生了解與熟悉巨量資料之熱門議題，如：高效能運算(High Performance Computing)、機器學習(Machine Learning)、雲端計算(Cloud Computing)、資料探勘(Data Mining)。
- 提高學生對於巨量資料分析技術與相關應用的興趣，培植國內相關領域之可用人才。
- 利用實作達到理論與實務兼備的教學目的。

(2) 教師教學反思

課程內容包含了能處理大數據龐大資料的叢集運算技術 Hadoop、HDFS 以及 HBASE，能快速穩定執行的計算框架以及整合其他的資料管理工具的大數據運算平台 Cloudera，自網站上取得數據的 Python 爬蟲技術並結合 Python 對於資料前處理、分析以及視覺化的程式技巧，用於大量數據交換的雲端私有雲技術 NextCloud、Freenas，處理大數據資料更有效率的 in-memory computing 技術 Spark，易於使用的資料分析軟體 WEKA 以及 ELK 數據視覺化平台等，透過以上的課程安排為帶給學員自資料的蒐集及前處理、大數據巨量資料的叢集運算技術、滿足大數據資料交換的私有雲建置，到數據分析的統計分析與機器學習應用以及最後的視覺化呈現，讓學員可以完整的學習到一整個完整資料分析過程中需要的各種技術，期望培養全面的大數據人才。而修習完本課程之學員可以接續 LORA 無線傳輸技術以及邊緣運算相關課程，將大數據的議題應用範圍擴展到更大的應用層面，結合了無線傳輸距離大幅度增加的優點和邊緣計算相關的應用，可以將大數據技術不僅僅是侷限於在 Server 上面運算，更可以佈署到實際應用的場域，為即時反應或是早期預警等應用上提供所需要的關鍵技術硬實力。

(3) 學生學習回饋

東海大學107學年第2學期教學意見反映統計表

研究所工學院　資工系楊朝棟先生　　2019/08/20 21:02列印

			一學生學習自評	1 我出席(約)為(A)90以上(B)80(C)70(D)60(E)50以下。	2 每週用課自時(小約)為(A)6以上(B)4-6(C)2-4(D)1-2(E)0-1。	3 瞭這課課目與要養核能。	4 完門課，的心力所修這課後我核能有提升。	二教學準備	5 師供備授大(含學目、課教材、進度成評式)。	6 課容分準備，與學綱授內充相符。	三教學內容與方法	7 程材內充，合學目標提專知識。	8 師課法引學興，幫學有學。	四教學態度	9 師視生，反應師互良好。	10 師據學綱親授。	11 師遲不退上授時。	五教學評量	12 師量式反教重點。	13 體而，對課滿意。	應填寫/不可評量學生人數	問卷填寫率%	問卷未填寫數
5695	大數據技術平台與	A		19	16	20	21		24	23		22	22		23	22	23		23	21	40	72.50	11
		B		6	6	7	6		4	5		5	6		6	6	4		5	6	0		
		C		3	4	2	2		1	1		2	1			1	2		1	1			
		D		1	3															1			
		E																					
評量值	(A+B) / (A+B+C+D+E)								96.6	96.6		93.1	96.6		100.0	96.6	93.1		96.6	93.1	單科	A+B人數%	95.8
	(D+E) / (A+B+C+D+E)								0.0	0.0		0.0	0.0		0.0	0.0	0.0		0.0	3.4		D+E人數%	0.4
	單科各題分數								95.9	95.2		93.8	94.5		95.9	94.5	94.5		95.2	92.4		單科分數	94.6

東海大學107學年第2學期教學意見調查

研究所工學院　資工系楊朝棟先生　　2019/08/20 21:02　列印(A4直印)

◎對於本課程及授課教師你認為最值得肯定的是？(請盡量舉例說明)

5695-大數據技術平台與應用

1. the learning methods
2. 課堂錄影
3. 老師教學認真，學到很多工具，非常實用
4. 楊老師的課都很用心準備，上課都有錄製影片可以看VOD，可以複習不熟悉的上課內容、或是睡過頭了沒有上到課也有補救的方式。老師也時常關心學生、聊聊一些課程的延伸應用、談談人生經歷、或是教導我們如何找資料參考等等。上老師的課能學習到各式各樣的知識，對升學或是離開學校的未來都非常有幫助。
5. 實作項目明確易懂

◎對於本課程及授課教師你認為最需要改進的是？(請盡量舉例說明)

5695-大數據技術平台與應用

1. easy to follow
2. 作業需安裝太多軟體在個人電腦上，可以多宣導學校提供的線上虛擬環境請大家在上面作業
3. 0，太棒了怎麼辦
4. 無

◎對本課程的其他意見：

5695-大數據技術平台與應用

1. 作業需安裝太多軟體在個人電腦上，可以多宣導學校提供的線上虛擬環境請大家在上面作業
2. 我愛朝棟 最棒老師>
3. 無

◎這位老師如有明顯性別偏見之語言或行為，請舉例說明。

5695-大數據技術平台與應用

1. 無

二. 參考文獻(References)

[1] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," J. Biomed. Inform., vol. 46, no. 5, pp. 774–781, 2013.

[2] C. T. Yang, W. C. Shih, G. H. Chen, and S. C. Yu, "Implementation of a cloud computing environment for hiding huge amounts of data," in Proceedings - International Symposium on Parallel and Distributed Processing with Applications, ISPA 2010, 2010, pp. 1–7.

[3] D. Agrawal, A. El Abbadi, S. Das, and A. J. Elmore, "Database scalability, elasticity, and autonomy in the cloud (Extended abstract)," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6587 LNCS, no. PART 1, pp. 2–15, 2011.

[4] D. Agrawal, A. El Abbadi, S. Das, and A. J. Elmore, "Database scalability, elasticity, and autonomy in the cloud (Extended abstract)," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, vol. 6587 LNCS, no. PART 1, pp. 2–15

[5] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber, "Bigtable: A distributed storage system for structured data," in 7th Symposium on Operating Systems Design and Implementation (OSDI '06), November 6-8, Seattle, WA, USA, 2006, pp. 205–218.

[6] H. Demirkan and D. Delen, "Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud," Decis. Support Syst., vol. 55, no. 1, pp. 412–421, 2013.

[7] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," in Proc. of the OSDI - Symp. on Operating Systems Design and Implementation, 2004, pp. 137–149.

[8] J. Dittrich and J. Quian, "Efficient Big Data Processing in Hadoop MapReduce," Proc. VLDB Endow., vol. 5, no. 12, pp. 2014–2015, 2012.

[9] J. Korpela, "Lurching toward Babel: HTML, CSS and XML," Computer (Long. Beach. Calif)., vol. 31, no. 7, 1998.

[10] J. Spillner, J. Müller, and A. Schill, "Creating optimal cloud storage systems," Futur. Gener. Comput. Syst., vol. 29, no. 4, pp. 1062–1072, 2012.

[11] K. K. Y. Lee, W. C. Tang, and K. S. Choi, “Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage,” Comput. Methods Programs Biomed., vol. 110, no. 1, pp. 99–109, 2013.

[12] M. Chen, S. Mao, and Y. Liu, “Big data: A survey,” Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, 2014.

[13] M. Purvis, J. Sambells, and C. Turner, Google Maps Applications with PHP and Ajax From Novice to Professional. 2006.

[14] P. Atzeni, F. Bugiotti, and L. Rossi, “Uniform access to NoSQL systems,” Information Systems, 2013.

[15] Cristiane Silva da Silva, Juliana Marzari Rossato, Jocelita Aparecida VazRocha, and Vera Maria Ferrão Vargas. Characterization of an area of reference for inhalable particulate matter (pm2.5) associated with genetic biomonitoring in children. Mutation Research/Genetic Toxicology and Environmental Mutagenesis, 778:44 – 55, 2015.

[16] Takashi Yorifuji, Saori Kashima, Midory Higa Diez, Yoko Kado, Satoshi Sanada, and Hiroyuki Doi. Prenatal exposure to outdoor air pollution and child behavioral problems at school age in japan. Environment International, 99:192 – 198, 2017.

三. 附件(Appendix)

與本研究計畫相關之研究成果資料，可補充於附件，如學生評量工具、訪談問題等等。