Big Data 分散式平台 -Hadoop 技術實務

一、課程緣起：

Big Data 含括三種層面： 巨量、即時性及多樣性。

1. 巨量（Volume） - 大數據的特色就在於： 龐大。 企業資料包羅萬端，很容易便達到數兆位元組，甚至千兆位元組之譜。(1)Facebook 一天有 32 億筆使用者所產生的 po、按讚、回覆…等資訊，故一個月將近 1200 億筆的資料，這是關聯式資料庫無法處理的，所以 NoSQL(Not Only SQL)因此誕生。(2)以 Youtube 一天的影片上傳量來說，若一個人要全部看完，所需花的時間會來到 80 年，故窮盡一生若你會長壽的最多看個兩天也很了不起了。

2. 即時性（Velocity） - 海量資料通常具有時效性，一旦串流至企業便須立即使用，方能發揮其最大價值。另外，要注意的是有些網站是 24 小時流量跟資料不斷湧入，面對這種情況，我們通常可以把他稱為『data stream』，此時 data stream type 的 data mining 將成為一個有趣的議題。因為在此環境下，資料永遠沒有穩態的一天，量隨時再增加，傳統的資料挖掘方式得被迫重新思考其方式。

3. 多樣性（Variety） - 海量資料的範疇不僅止於結構化資料，還包含各類非結構化的資料： 諸如文字、音訊、視訊、點擊串流（click stream）、日誌檔等等。如何去妥善描繪圖片、影音檔的內容也成為一大議題。資料的運用不在是下了幾行 sql 語法就可以全部撈出來，因為更多的資訊藏匿在圖片跟影片之中。常見的手法可分為：(1) Meta description mode，在此模式下會將這些影音檔，設定好的描述資料(亦即 metadata)來陳述之，之後用一個 XML 檔來對應一部影片檔，如此才可以從 XML 去瞭解哪寫圖片想關聯性？哪些影片該如何推薦。(2) Behavior analysis mode，在此模式下會去記錄 user behavior 的關聯，從使用者行為去歸納，下一個使用者點擊某一內容時，跟它的行為模式最相近的群體其接下來最感興趣的會是什麼，就由大眾的力量來勾勒出群體模式。

這 3 個資料特性，已經是現在式，而不是未來式。然而該如何解決日漸緊迫的巨量資料處理問題呢？像 Facebook、Twitter 這樣面臨資料量大爆炸的網路公司，開始用 Hadoop、NoSQL 等新興技術來解決問題。

''挑戰還是機會？''

Big Data 不只是一項挑戰， 更是絕佳的機會，讓您能夠洞悉新興的資料類型、使企業運作更加靈敏並為過往所無法企及的問題提供解答。 但在此之前，這種機會並無實際方法可以掌握。 今天，Big Data 平台採用 Hadoop 等技術，能為充滿各種可能性的世界開啟一扇大門。

二、課程目標：

本課程旨在建立雲端運算之大量資料(Big Data)處理、分析、應用的根基，讓參訓學員瞭解正確的觀念與方法，重點並不在講解程式設計的細節，而在於透過體驗式教學方式的實作，經由指令剪貼方式來體驗實際的操作方式，以從體驗中驗證課程所學。

由於 Hadoop 是採用 Java 語言撰寫，對於許多不熟悉 Java 語言的學員來說有相當大的入門障礙，因此本課程針對資料分析運算這部份,主要是以 Aapache 基金會所開發的 Hadoop 原生分析工具(Pig, Hive) 及 大數據資料庫 (HBase), 做為課程操作與實務研討. 期能讓學員學會如何將 Hadoop 這項技術與現存資訊架構進行整合, 進而達到企業期望的預測分析。

上課方式採用 “雲中櫃” 實作教學環境，每位學員可在各自的 VMware 虛擬系統中，啟動 Hadoop 資料作業系統，得以完全操作多節點 Hadoop 運算分析平台。

三、適合對象：

- IT 專案經理、系統架構師 或 技術決策人員
- 網路管理工程師 或 應用程式設計師
- 欲親身體驗 Hadoop 資料科技

四、先備知識

Windows 檔案及目錄管理

五、課程日期：

110 年 7/5-7/7，週三四五白天 9:30 ~12:30,13:30~16:30 ，共 3 天、計 18 小時。

六、上課地點：

舉辦地點：工研院產業學院 產業人才訓練一部(台北)，實際地點依上課通知為準!!!!

七、報名方式：

線上報名：到工研院產業學院官網報名
課程洽詢：02-2370-1111 分機 306 黃小姐 Email：wenhsin.huang@itri.org.tw
報名確認：02-2370-1111 分機 304 黃小姐 Email：finn@itri.org.tw

九、課程大綱：

課程單元	課程內容
認識資料科技 (Data Technology)	● 資料科技的現在與未來 ● Hadoop 資料科技平台架構
Hadoop 分散系統基礎建置	● Hadoop 分散系統架構規劃與設定 ● 建置 Hadoop 分散運算主機 (Docker) ● 管理 Hadoop 分散運算主機 (Master, Worker) ● 登入 Hadoop 資料分析主機 (Hadoop Client)
建置 HDFS 分散檔案系	● 認識 HDFS 分散檔案系統運作架構

統	● 設定與啟動 HDFS 分散檔案系統 ● 管理 HDFS 分散檔案系統 ● 規劃與建置團隊分析目錄架構(家目錄, 資料集,…)
認識 MapReduce 開發模式（Program Model）	● 認識 MapReduce 開發模式運作架構 ● 撰寫與執行 MapReduce 開發模式程式
建置 YARN 分散運算系統	● 認識 YARN 分散運算系統運作架構 ● 設定與啟動 YARN 分散運算系統 ● 使用 YARN 分散運算系統 ● 規劃 YARN 分散運算系統資源（NodeManager, YARNChild）

＊ 課程執行單位保留調整課程內容、日程與講師之權利

十、課程費用與繳費：

1. 本課程費用含課程、講義、餐點。

方案	課程費用
課程原價（每人）	$15,000 元
14 天前報名 優惠價(每人)	$12,000 元
14 天前報名+兩人揪團同行 優惠價(每人)	$11,400 元
14 天前報名+四人(含)以上揪團同行/工研人 優惠價(每人)	$10,800 元

5. 課程若未如期開班，費用將全額退還。
6. 繳費方式
   - ATM 轉帳（線上報名）：繳費方式選擇「ATM 轉帳」者，系統將給您一組轉帳帳號「銀行代號、轉帳帳號」，但此帳號只提供本課程轉帳使用，各別學員轉帳請使用不同轉帳帳號！！轉帳後，寫上您的「公司全銜、課程名稱、姓名、聯絡電話」與「收據」傳真至 02-2381-1000 黃小姐 收。
   - 信用卡（線上報名）：繳費方式選「信用卡」，直到顯示「您已完成報名手續」為止，才確實完成繳費。
   - 銀行匯款(公司逕行電匯付款)：土地銀行　工研院分行，帳號 156-005-00002-5（土銀代碼：005）。戶名「財團法人工業技術研究院」，請填具「報名表」與「收據」回傳真至 02-2381-1000 黃小姐 收。
   - 即期支票或郵政匯票：抬頭「財團法人工業技術研究院」，郵寄至： 100 台北市中正區館前路 65 號 7 樓 704 室 黃小姐收。
   - 計畫代號扣款(工研院同仁）:請從產業學院學習網直接登入工研人報名；俾利計畫代號扣款。

十一、報名確認與取消：

1. 已完成報名與繳費之學員，課程主辦單位將於開課三天前以 E-mail 方式寄發上課通知函；

若課程因故取消或延期，亦將以 E-mail 方式通知，如未收到任何通知，敬請來電確認。

2. 已完成繳費之學員如欲取消報名，請於實際上課日前以書面通知業務承辦人，主辦單位將退還 80% 課程費用。

3. 學員於培訓期間如因個人因素無法繼續參與課程，將依課程退費規定辦理之：上課未逾總時數三分之一，欲辦理退費，退還所有上課費用之二分之一，上課逾總時數三分之一，則不退費。

4. 本單位保留是否接受報名之權利。

5. 如遇不可抗拒之因素，課程主辦單位保留修訂課程日期及取消課程的權利。