

# 資料倉儲導論-第1章 資料倉儲簡介

## 大數據分析

- R/Python/Julia/SQL程式設計與應用  
(R/Python/Julia/SQL Programming and Application)
- 資料視覺化 (Data Visualization)
- 機器學習 (Machine Learning)
- 統計品管 (Statistical Quality Control)
- 最佳化 (Optimization)



李明昌博士

[alan9956@gmail.com](mailto:alan9956@gmail.com)

<http://rwepa.blogspot.com/>

# 大綱

- 1.1 教師簡介
- 1.2 資料倉儲簡介
- 1.3 資料倉儲的架構
- 1.4 線上分析處理 (OLAP)
- 1.5 商業智慧 (Business Intelligence, BI)
- 1.6 機器學習 (Machine Learning)

# 1.1 教師簡介

---

# 教師簡介 <http://rwepa.blogspot.com/>

- 姓名：李明昌 (ALAN LEE)
- 現職：中華R軟體學會 常務理事  
臺灣資料科學與商業應用協會 常務理事
- 學歷：中原大學 工業與系统工程所 博士
- 經歷：
  - 淡江大學 兼任教師
  - 佛光大學 兼任教師
  - 國立台北商業大學 兼任教師
  - 育達科技大學 資訊管理系(所) 專任助理教授
  - 東吳大學 兼任教師
  - 崇友實業 行銷企劃專員
  - 國航船務代理股份有限公司 海運市場運籌管理員
- 大專院校、資策會、工業技術研究院、國家發展委員會、中央氣象局、公平交易委員會、各縣市政府與日本名古屋產業大學等公民營單位演講達300餘場, 2800小時以上.
- 連絡資訊：[alan9956@gmail.com](mailto:alan9956@gmail.com)



- iPAS 巨量資料分析師 證照推廣
- iPAS 營運智慧分析師 證照推廣

## 1.2 資料倉儲簡介

---

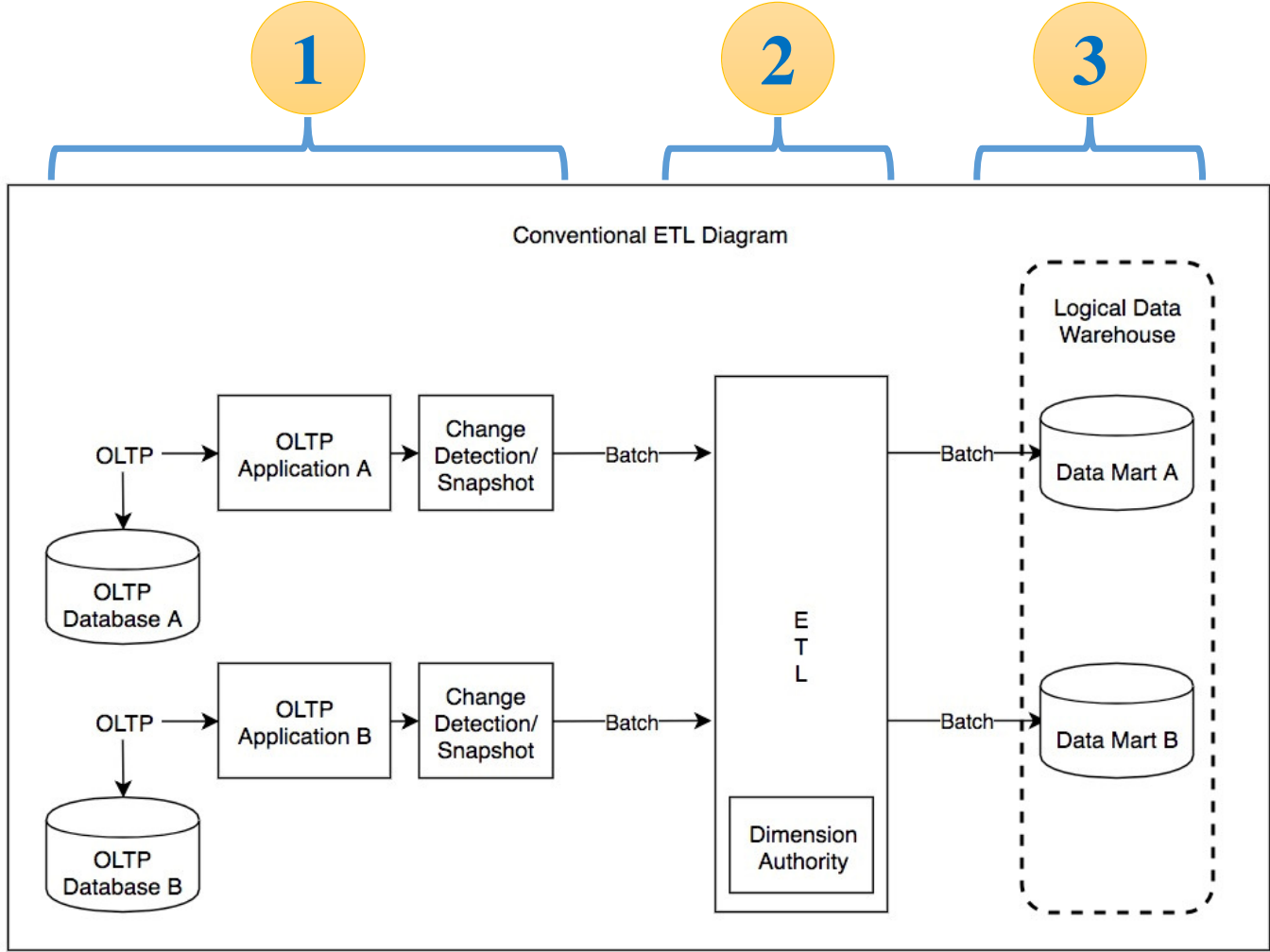
## 資料倉儲 (Data Warehouse)

- 在電腦領域, 資料倉儲, 也稱為企業資料倉儲, 一般用於報告和資料分析的系統, 被認為是商業智慧型的核心組件.
- 資料倉儲是來自一個或多個不同來源整合資料的中央儲存庫.
- 資料倉儲將當前和歷史資料儲存在一起, 可提供整個企業的員工建立分析報告.

# 資料倉儲建構方式

- 使用提取、轉換、載入 (Extract, Transform, Load, 簡稱ETL)或提取、載入、轉換 (Extract, Load, Transform , 簡稱ELT ) 的方式建立資料倉儲系統.
- 在資料倉儲中, 資料被區分為維度, 並區分為事實資料表和維度資料表 .
- 事實和維度的組合有時被稱為星狀綱要.

# ETL流程圖



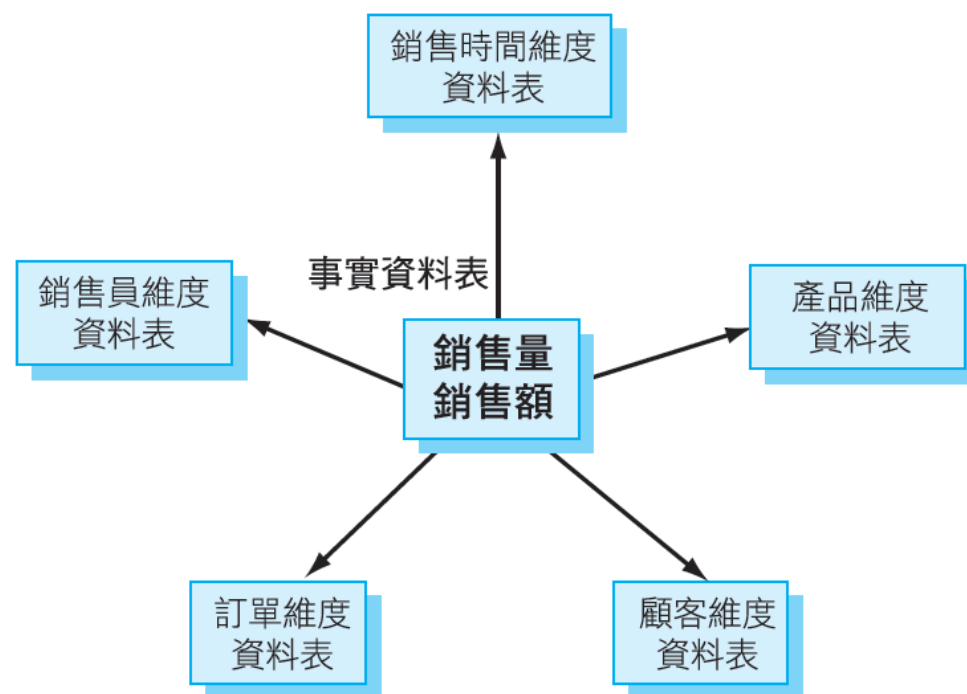


# 資料倉儲-維度模型 (Dimensional Model)

- 別稱—**星狀綱要(Star Schema)**
- 模型結構
  - 中央核心：**事實資料表 (Fact table)**→ 加總計算
  - 環繞核心的光芒：**維度資料表 (Dimension table)** — 群組計算→Excel樞紐分析
  - 此模型不建議以**實體關係體圖**(Entity Relationship Diagrams, ERD)來呈現
- **事實資料表(Fact table)**
  - 只有一個事實資料表
  - 可衡量(Measurement)數值績效統稱事實(Facts)
- 環繞核心的光芒：**維度資料表(Dimension table)**
  - 允許數個維度資料表
  - 每一道光芒代表決策者觀察績效的角度
  - 因此表格內所儲存的資料就是主管查看企業營運績效時所要的觀查的角度特色

# 維度模型範例

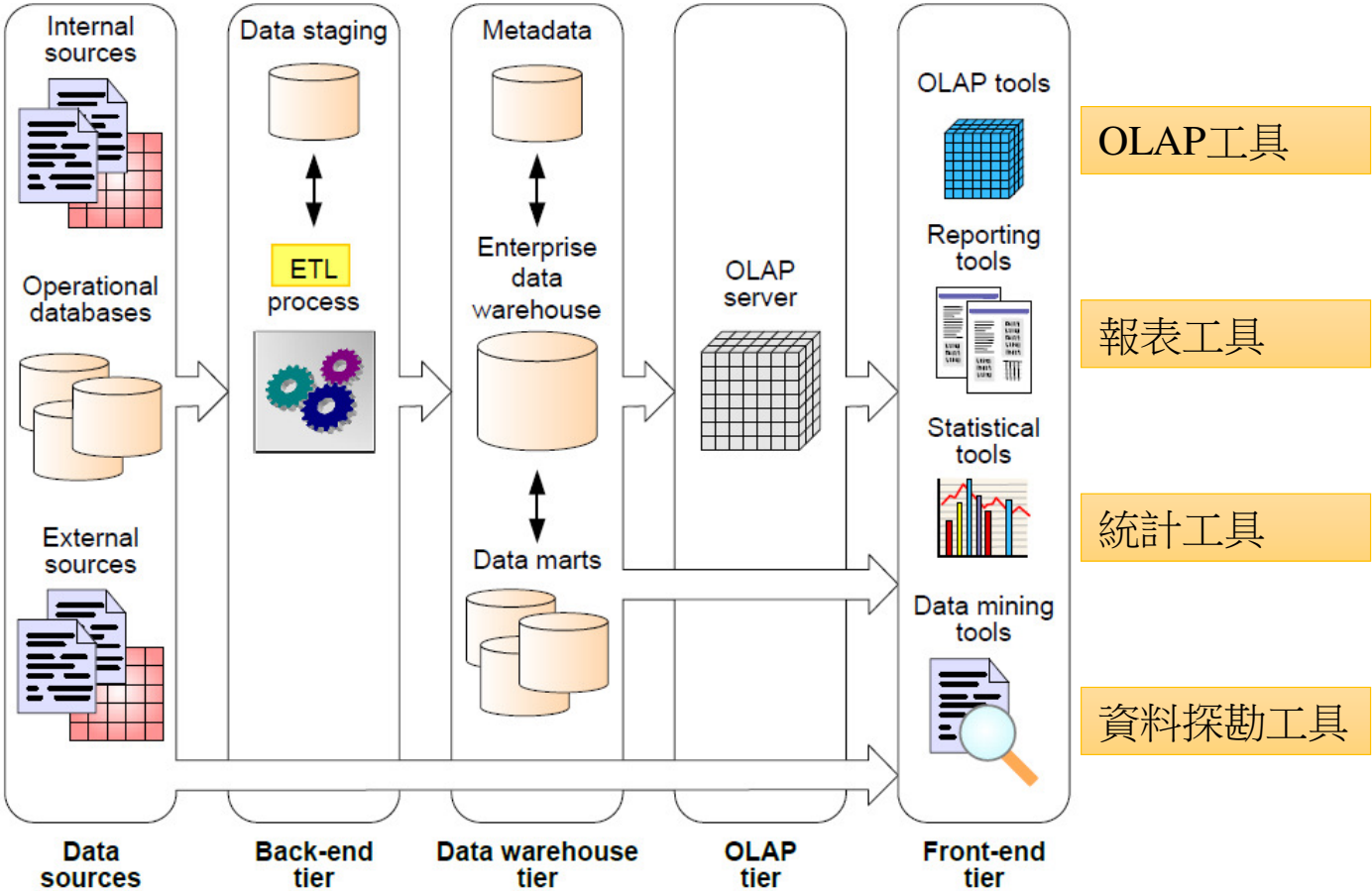
- 某企業經營績效的資料有銷售量與銷售額為二個事實資料表。
- 企業主有興趣查看經營績效的角度有五個,分別為銷售時間、產品、客戶、訂單以及銷售員等五個維度資料表。



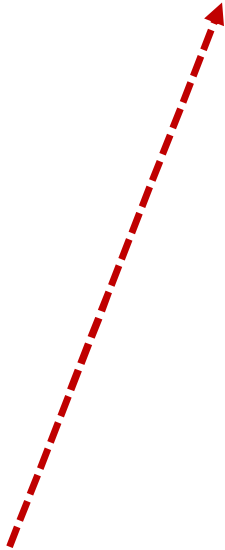
## 1.3 資料倉儲的架構

---

# 資料倉儲的架構



R 語言



# 資料倉儲之資料集結 (Data Staging)

- 資料倉儲架構包括前端 (Front-end tier)與後端 (Back-end tier)：
  - 後端負責準備資料, 亦稱為資料管理(Data management).
  - 前端負責應用資料, 亦稱為資料存取(Data access).
- 資料集結：
  - 後端會進行 ETL 的各項步驟 (抽取、清理、一致化與交付) 將所產生的資料進行儲存.
  - 一般每一個 ETL 步驟都需要進行資料的集結.
- 資料集結區(Data staging area)包括：
  - 持久集結區 (Persistent staging area)：維護歷史資料而使用的集結區.
  - 臨時集結區(Temporary staging area)：資料在每次載入過程後即被刪除.

## 1.4 線上分析處理 (OLAP)

---

# 線上分析處理 OLAP

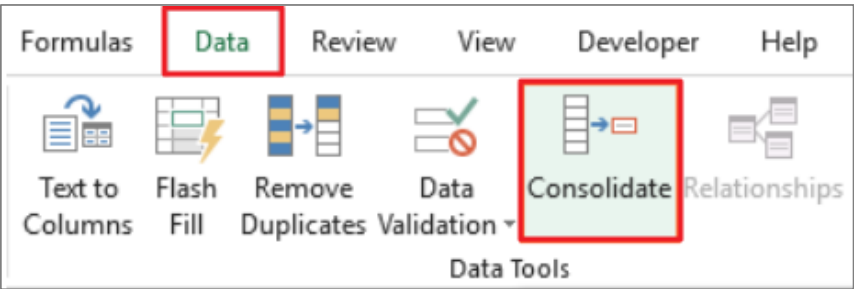
- 線上分析處理(On-Line Analytical Processing), 簡稱OLAP.
- OLAP是電腦技術中快速解決多維度分析問題 (Multi-Dimensional Analytical, MDA)的一種方法.
- OLAP更廣泛運用於商業智慧系統的一部分, 它還包括關聯式資料庫、報告編寫和資料探勘.
- OLAP的典型應用包括銷售業務報告、市場行銷、管理報告、業務流程管理(BPM)、預算和預測、財務報表.
- OLAP 與線上交易處理(Online transaction processing, OLTP)不相同.
  - OLTP 是指透過資訊系統、電腦網路及資料庫, 以線上交易的方式處理一般即時性的作業資料, 和更早期傳統資料庫系統大量批次的作業方式並不相同.

# OLAP 報表常用功能

- **Roll-up 向上匯總**: Roll-up 動作會讓資料根據所規範的屬性及其階層做彙匯總.
- **Drill-down 向下鑽取**: Drill-down 則是針對資料做細部展開, 以得到更詳細的資訊.



# excel-consolidate-answer.xlsx

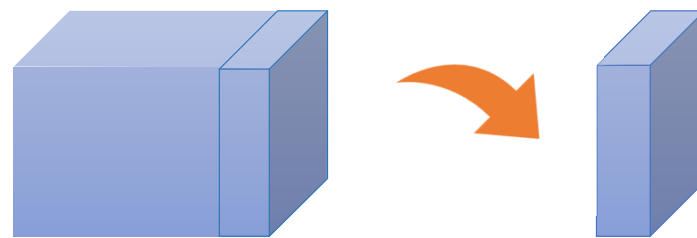


	A	B	C	D	E	F	G	H	I	J	K	L	M
1	一月	二月	三月	四月	五月	六月	七月	八月	九月	十月	十一月	十二月	
6	銷貨成本	1,100,000	1,300,000	1,500,000	1,023,000	1,209,000	1,395,000	1,142,000	1,349,000	1,557,000	1,205,000	1,424,000	1,643,000
7		815,000	845,000	870,000									
8					758,000	786,000	809,000						
9								846,000	877,000	903,000			
10											893,000	926,000	953,000
11	薪水	815,000	845,000	870,000	758,000	786,000	809,000	846,000	877,000	903,000	893,000	926,000	953,000
16	房租	185,000	185,000	185,000	172,000	172,000	172,000	192,000	192,000	192,000	202,000	202,000	202,000
21	折舊	100,000	100,000	100,000	93,000	93,000	93,000	103,000	103,000	103,000	109,000	109,000	109,000
26	出差費	60,000	80,000	100,000						103,000	65,000	87,000	109,000
31	其它	59,000	55,000	60,000						62,000	64,000	60,000	65,000
36	維護費用	40,000	50,000	60,000						62,000	43,000	54,000	65,000
41	辦公室用品	20,000	22,000	25,000						23,000	19,000	43,000	36,000
46	郵費	5,000	4,000	6,000	4,000	3,000	5,000	5,000	7,000	6,000	5,000	7,000	6,000
51	總計項目	\$ 2,384,000	\$ 2,641,000	\$ 2,906,000	\$ 2,214,000	\$ 2,458,000	\$ 2,695,000	\$ 2,479,000	\$ 2,752,000	\$ 3,011,000	\$ 2,605,000	\$ 2,912,000	\$ 3,188,000

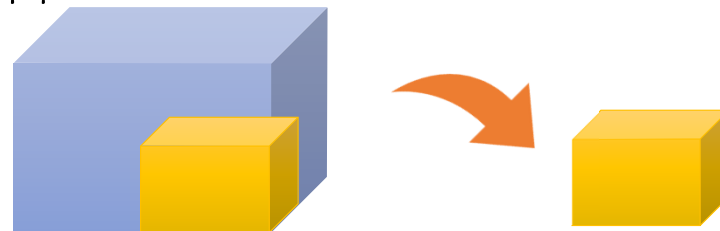
資料 \ 合併彙算

## OLAP 報表常用功能 (續)

- **Slice-and-dice 交叉剖析:** Slice 和 Dice 將多維度模型中的資料, 做進一步的**資料限制**, 以讓報表中只呈現出使用者所想要的資料.
  - 切片 (Slice) 的動作是指將某一個屬性的值規範在某一個值或某一個範圍, 以用來過濾多維度模型中的資料.



- 切丁 (Dice) 的動作是指我們透過一個可能包含多個屬性的條件(多個維度), 取出多維度模型中某一個區塊資料.



## OLAP 報表常用功能 (續)

- **Pivot 樞紐分析**: Pivot 是一個會改變報表排列的動作, 原本的主要分析維度和陪襯的維度會在 Pivot 動作中被對調, 整個報表的重點也可能會有所改變.
- **Drill-across**: Drill-across 動作指的是我們要將兩個或兩個以上的多維度資料模型建立關係, 以用來比較兩個不同模型裡面的資料.

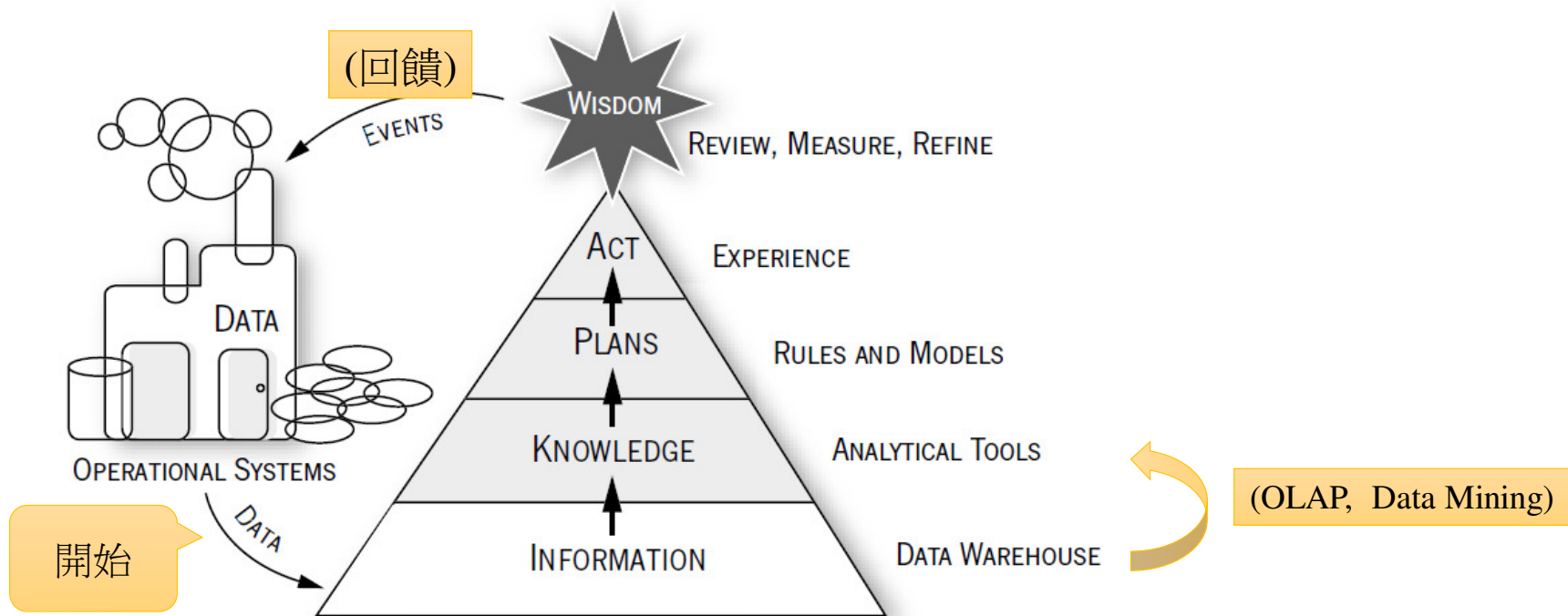
## 1.5 商業智慧 (Business Intelligence, BI)

---

# 商業智慧

## BI As a Data Refinery

資料煉油廠



參考: Wayne Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligence Solutions, 2013.  
[http://download.101com.com/tdwi/research\\_report/2003BIReport\\_v7.pdf](http://download.101com.com/tdwi/research_report/2003BIReport_v7.pdf)

# 商業智慧 – 步驟1

## 1. 從原始資料到資料倉儲的資訊

- 第一步是從企業間交易和企業內營運系統中萃取資料, 然後經過清理、轉換等處理步驟, 將定義清楚且一致的細節和彙總的資料, 載入資料倉儲的資料庫中, 從最底層的資料轉換成資料倉儲的資訊.
- 例如: 將分散在訂單、維修服務、銷售、出貨、和會員等系統中的顧客資料記錄整合成一個以顧客為主題的完整資料庫, 對於瞭解顧客及其需求產生有用而完整的資訊.

2. 從資訊到知識

3. 從知識到決策

4. 從決策到行動

5. 回饋迴圈

# 商業智慧 – 步驟2

1. 從原始資料到資料倉儲的資訊

## 2. 從資訊到知識

- 使用者可以運用各種報表和分析工具, 例如查詢、報表、線上分析處理(OLAP)、和資料探勘等, 存取並分析資料倉儲中的資訊.
- 這些分析可以找出資料中的趨勢(Trends)、型態(Patterns, 樣式)、和例外狀況等, 這些分析工具幫助使用者將資訊轉換成知識.
- 例如：零售通路從大量銷售資料中挖掘出特定類型的顧客會同時購買紙尿布和啤酒之間的關聯規則, 對於賣場而言, 這個發現對於商品陳列是有一定參考價值.

3. 從知識到決策

4. 從決策到行動

5. 回饋迴圈

# 商業智慧 – 步驟3

1. 從原始資料到資料倉儲的資訊

2. 從資訊到知識

**3. 從知識到決策**

- 使用者從分析所發現的趨勢和型態中可以**建立業務規則**, 也可以將知識作為建立**決策模型**的依據, 來規劃業務的進行並作為決策的參考.
- 例如: 庫存降至10單位時, 就要下單採購30個單位.
- 規則也可能是根據過去的趨勢所做的預測, 或是根據假設或估計所產生的情境 (what if) 分析.
- 統計分析和最佳化分析也可以產生比較複雜的規則, 例如機動的定價機制以回應變動的市場狀況, 其規則可以用統計方法產生, 也可以使用利潤最大化最佳化模型.

4. 從決策到行動

5. 回饋迴圈



## 商業智慧 – 步驟4

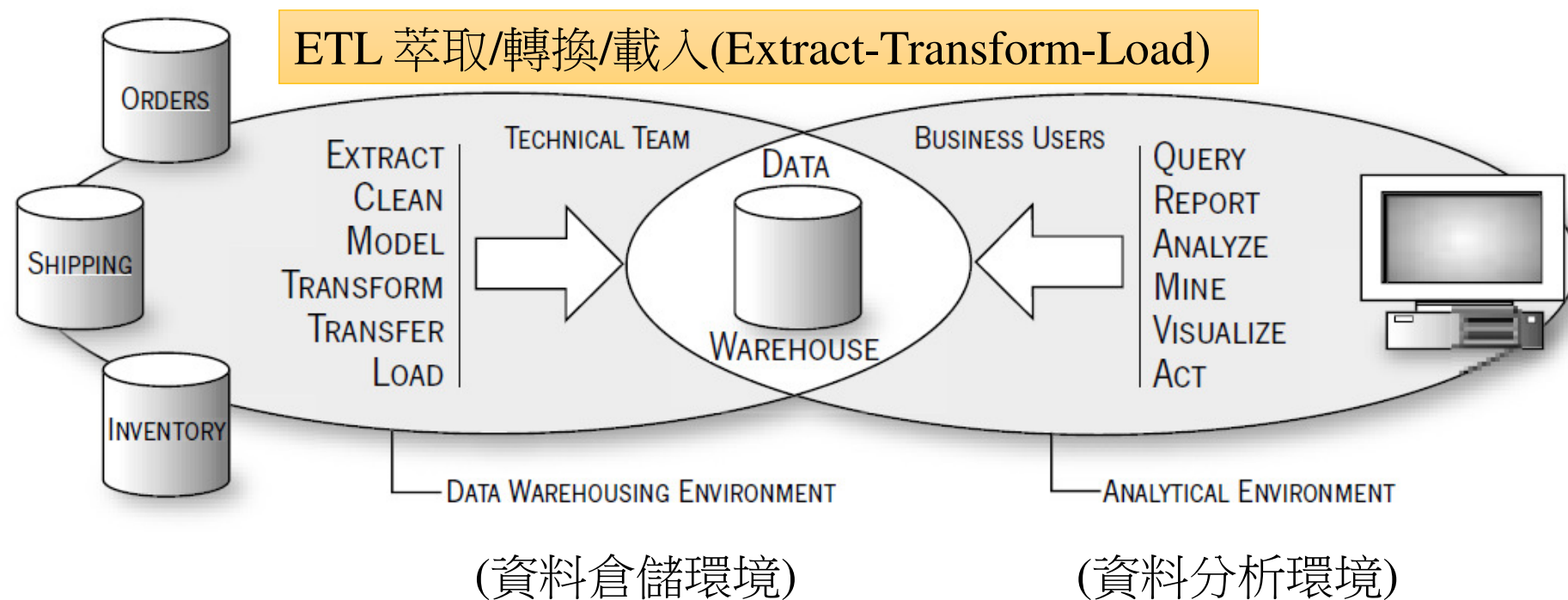
1. 從原始資料到資料倉儲的資訊
2. 從資訊到知識
3. 從知識到決策
- 4. 從決策到行動**
  - 根據前一步驟的業務規則或決策規劃, 使用者要產生執行計畫.
  - 例如: 行銷人員要根據顧客區隔的分析, 顧客回應特定優惠的預測模型, 以及過去的促銷活動經驗, 來規劃各種促銷活動.
  - 針對不同的客戶透過不同的通路所提供的優惠方案為何. 這些執行計畫將決策方案轉換成實際的行動.
5. 回饋迴圈

## 商業智慧 – 步驟5

1. 從原始資料到資料倉儲的資訊
2. 從資訊到知識
3. 從知識到決策
4. 從決策到行動
- 5. 回饋迴圈**
  1. 一旦計畫開始實施, 整個循環便會重複.
  2. 營運系統中會有顧客對於優惠的反應以及後續的交易.
  3. 這些資料會被萃取出來, 並和相關資料整合後載入資料倉儲.
  4. 行銷人員就可以進一步分析以評估其促銷活動的成效, 並據以修正其促銷活動的規劃.

# 商業智慧系統的組成架構

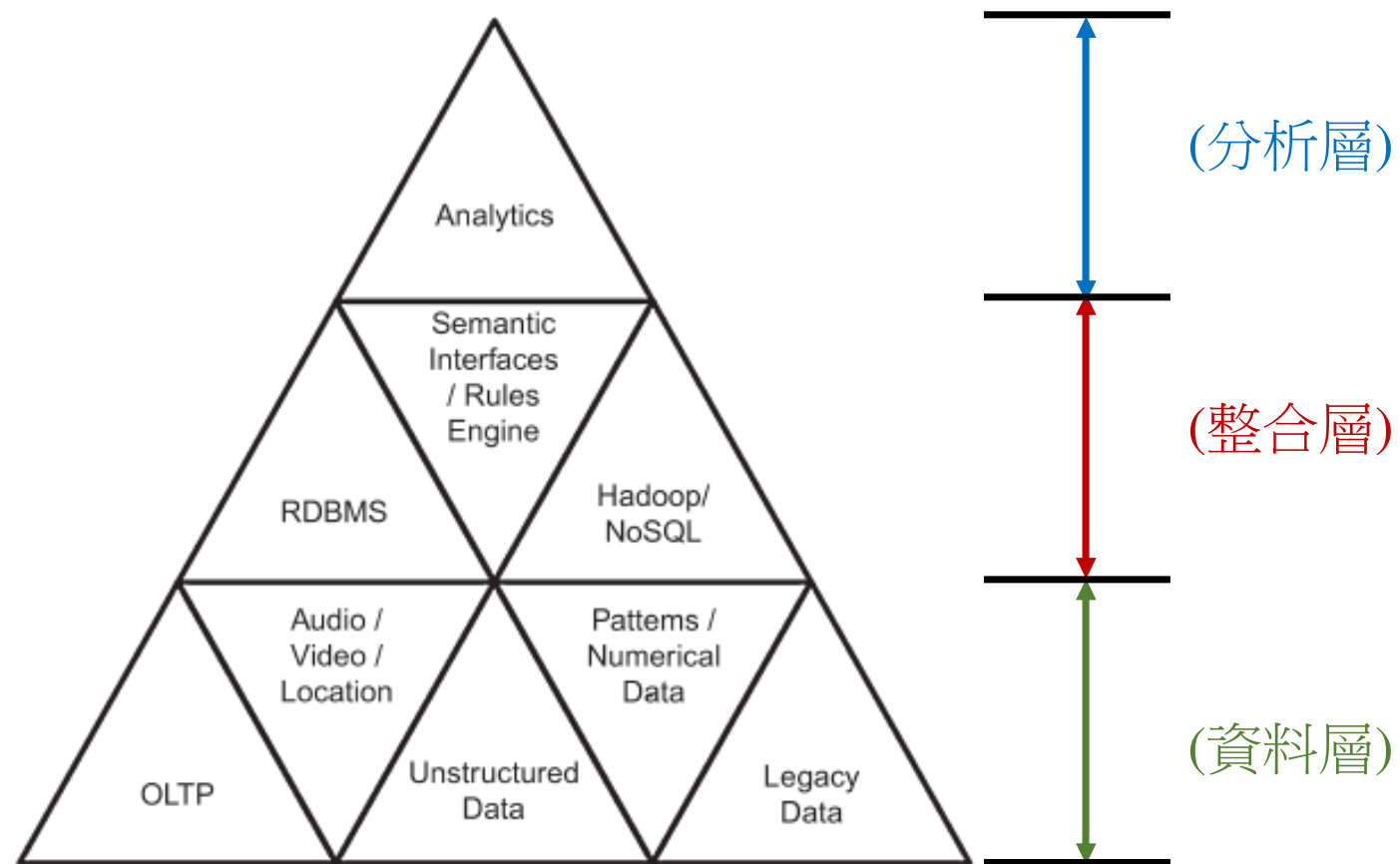
## BI Component Framework



參考: Wayne Eckerson, Smart Companies in the 21st Century: The Secrets of Creating Successful Business Intelligence Solutions, 2013. [http://download.101com.com/tdwi/research\\_report/2003BIReport\\_v7.pdf](http://download.101com.com/tdwi/research_report/2003BIReport_v7.pdf)

# 新世代資料倉儲架構

## (Components of the next-generation data warehouse)



# 商業智慧在企業應用

- 關鍵結果指標(key result indicators, **KRI**)：公司在某方面的表現, **長時間評估**, 例：每月, 每季報表交付董事會.
- 績效指標(performance indicators, PI)：公司該做的是什麼, 例：獲利前10%的顧客, 其主要產品的淨利.
- 關鍵績效指標(key performance indicators, **KPI**)：該做什麼可以**大幅改善**公司的績效, **短時間評估**, 例：每日, 每週. → 中高階主管
- 顧客滿意度、稅前淨利、和員工滿意度等, 是常被誤認為**KPI**的**KRI**(本質是**KRI**).
- 較佳數量：10 **KRI**(少), 80 **PI**(多), 10 **KPI**(少).
- 商業智慧中上列三大指標需要被檢視、衡量、和修正.

# BI專案生命週期



## 評估：

- 公司營運上的評估
- 成本效益分析
- 風險評估

## 規劃：

- 技術面的基礎設施-硬體平台, 中介軟體平台, 資料庫管理系統平台
- **非技術面**的基礎設施的評估-功能部門的運作, 營運活動的作業流程, 企業營運資料, 企業應用系統, 詮釋資料庫(Meta data repository)

## 專案分析：

- 需求分析
- 資料分析：(1).邏輯資料模型(Logical data modeling)法—確保資料整合與一致性  
(2).來源資料分析(Source data analysis)法—確保資料品質

## 1.6 機器學習 (Machine Learning)

---

# 深度學習發展史




資料探勘 (Data Mining)

- 1943年：美國數學家 Walter Pitts和心理學家 Warren McCulloch提出人工神經元。
- 1957年：美國心理學家 Frank Rosenblatt 提出了感知器(Perceptron)。
- 1980年：多層類神經網路失敗，淺層機器學習方法(Support Vector Machine, SVM等)興起。
- 2006年：Geoffrey Hinton 成功訓練多層神經網路(限制玻爾茲曼機, RBM)，命名為深度學習。
- 2012年：ImageNet 比賽讓深度學習重回學界視野，開啟 NVIDIA GPU 為重要運算硬體。



# 機器學習 Machine learning

- **監督式學習 (Supervised learning)**
    - **Telling the algorithm what to predict**
  - **非監督式學習 (Unsupervised learning)**
    - **No label or target value given for the data**
- 
- **半監督學習 (Semi-supervised learning)**
    - 具有少量標記資料
  - **強化學習 (Reinforcement learning)**
    - 為了達成目標，隨著環境的變動，而逐步調整其行為，並評估每一個行動之後所到的回饋是正向的或負向的。
  - **深度學習 (Deep learning)**

# 監督式學習 vs. 非監督式學習

- 監督式學習 Supervised learning - 執行  $X \rightarrow$  預測  $\rightarrow Y$ 
  - 迴歸分析 Regression analysis
  - 廣義線性模型 General linear model (GLM)
  - 天真貝氏法 Naïve-Bayes
  - K近鄰法 k-nearest neighbors (KNN)
  - 決策樹 Decision tree
  - 支持向量機 Support vector machine (SVM)
  - 類神經網路 Neural network (NN)
  - 集成學習 Ensemble learning: 使用多種學習算法來獲得比單獨使用演算法更好預測結果
- 非監督式學習 Unsupervised learning
  - 集群法 Clustering
  - 關聯規則 Association rule
  - 主成分分析 Principal Component Analysis

# CRISP-DM標準流程

---

# 資料探勘生命週期－CRISP-DM

- 跨產業資料探勘標準作業流程 (CRoss Industry Standard Process for Data Mining)
- 資料探勘方法論
- CRISP-DM是於1990年起，由SPSS以及NCR兩大廠商在合作戴姆克萊斯勒-賓士(Daimler Benz)的資料倉儲以及資料探勘過程中發展出來的。

## CRISP-DM 資料探勘流程(續)

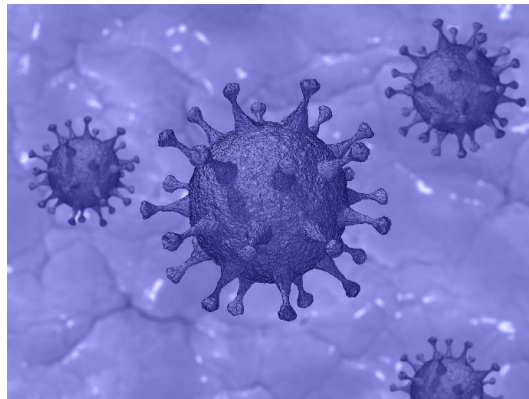
- 步驟 1：商業理解
- 步驟 2：資料理解
- 步驟 3：資料準備
- 步驟 4：模式建立
- 步驟 5：評估與測試
- 步驟 6：佈署應用

佔整專案時間的  
~80%

- 訓練資料70%
- 測試資料30%

## 步驟1.商業理解

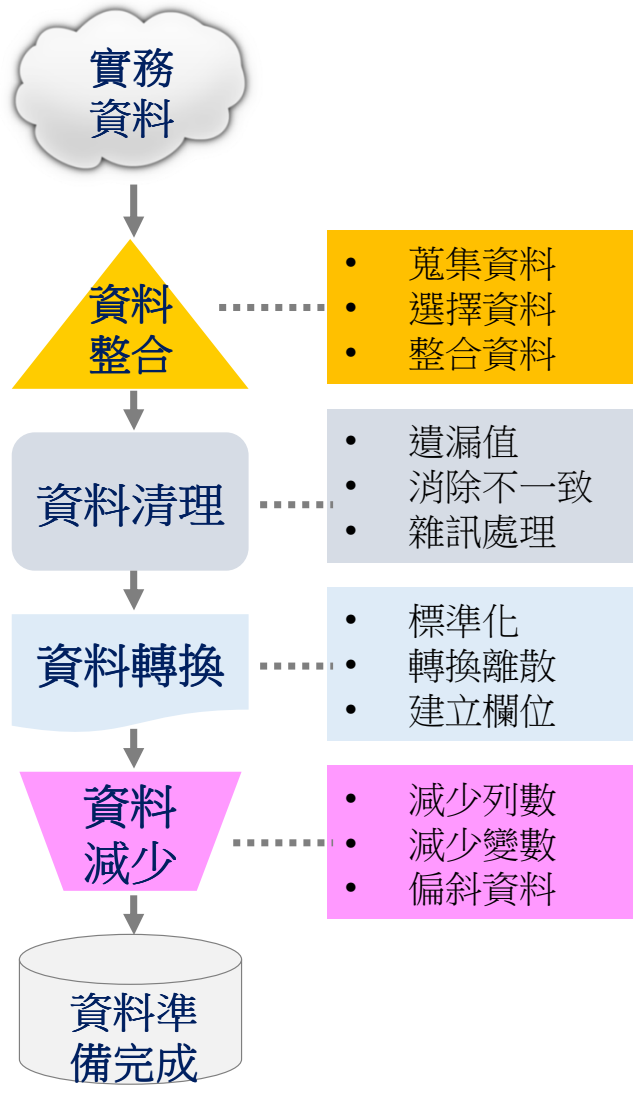
- 終極目標是要解決具體的產業問題，諸如提高購買率、找出詐欺交易、銷售預測與異常偵測等，因此以專業知識 (domain knowledge)進行商業理解是重要的第一步，處理重點：
  - 擬定商業目標
  - 進行當前處境評估
  - 決定資料探勘目標/成本
  - 產生專案計劃
  - 解決顧客問題



## 步驟2. 資料理解

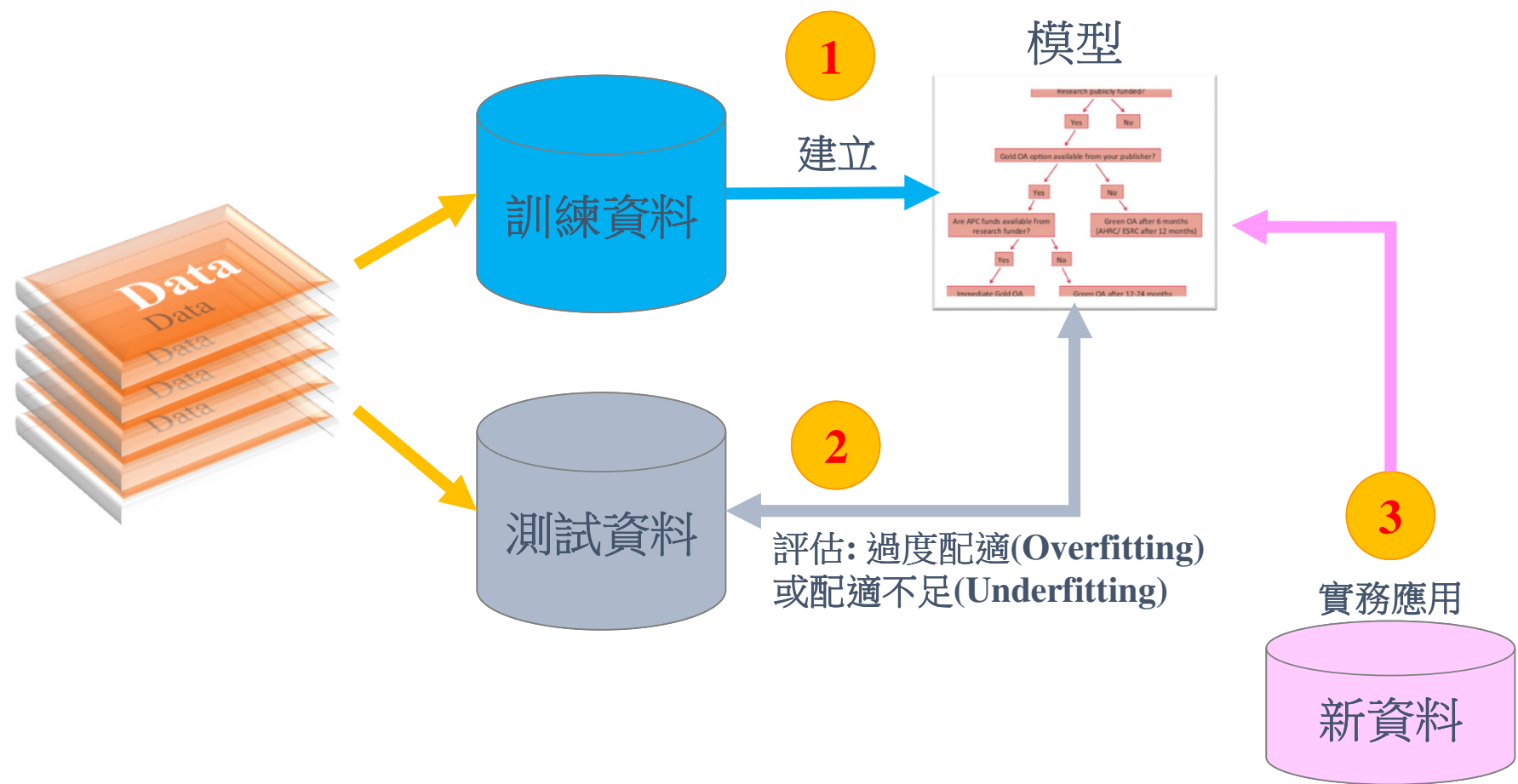
- 包括描述資料、探索資料、核驗資料品質
- 敘述統計分析
  - 六力分析(summary函數)
- 繪圖
  - 依群組特性
  - 依時間特性
  - 新增評估欄位
  - 趨勢
  - 離群值 (outlier)
  - 散佈圖、散佈圖矩陣
  - 盒鬚圖

# 步驟3.資料準備

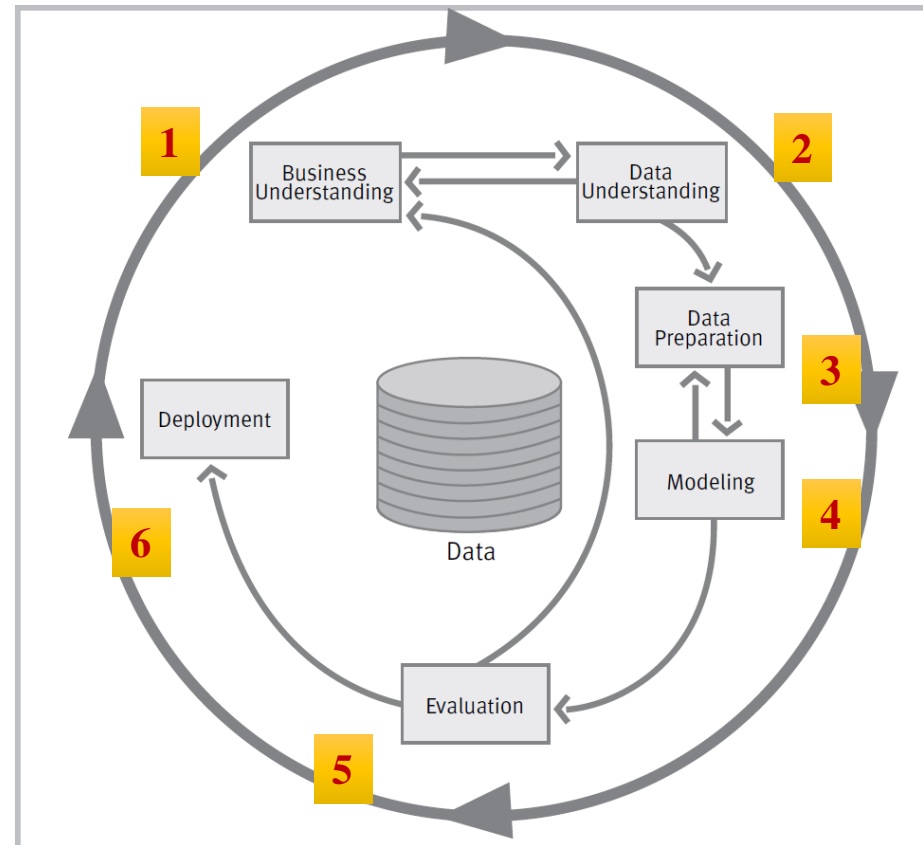




# 步驟4.模型建立、步驟5.評估與測試



# CRISP-DM 資料探勘流程(續)



參考 [https://en.wikipedia.org/wiki/Cross\\_Industry\\_Standard\\_Process\\_for\\_Data\\_Mining](https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

# 數值模型績效指標

- 不可直接使用誤差的算術平均!

$$Total\ error = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$$



- 均方誤差 (Mean Squared Error, MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- 均方根誤差 (Root Mean Squared Error, RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

- 平均絕對誤差 (Mean Absolute Error, MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

# 類別模型績效指標

- <http://rwepa.blogspot.com/2013/01/rocr-roc-curve.html>

```
#          | 真實P類別 真實N類別
# *****
# 預測P類別 | TP真陽數  FP假陽數
# 預測N類別 | FN假陰數  TN真陰數
# *****
#          | P      N
```

混淆矩陣  
(Confusion Matrix)

```
# 1.TPR(True positive rate) 真陽性率, 愈大愈好 -----
# =TP/(TP+FN)
# =TP/P
# =Sensitivity 靈敏度
# =Recall 召回率
# =Probability of detection
# =Power
# 實際為陽性的樣本中, 判斷為陽性的比例。
# 例如真正有生病的人中, 被醫院判斷為有生病者的比例。
```

## 參考資料

- Alejandro Vaisman and Esteban Zimányi, *Data Warehouse Systems Design and Implementation, Second Edition*, Springer, 2022.
- Data warehouse, [https://en.wikipedia.org/wiki/Data\\_warehouse](https://en.wikipedia.org/wiki/Data_warehouse)
- OLAP, [https://en.wikipedia.org/wiki/Online\\_analytical\\_processing](https://en.wikipedia.org/wiki/Online_analytical_processing)
- OLTP, [https://en.wikipedia.org/wiki/Online\\_transaction\\_processing](https://en.wikipedia.org/wiki/Online_transaction_processing)
- RWEPA, <http://rwepa.blogspot.com/>



# 謝謝您的聆聽

Q & A

李明昌

EMAIL: [alan9956@gmail.com](mailto:alan9956@gmail.com)

WEB: <http://rwepa.blogspot.com/>

