Data Warehouse（資料倉儲）

> 主要的資料來源：
>
> Essentials of Database Management Jeffrey A. Hoffer, Heikki Topi, V. Ramesh
>
> 資料倉儲(Data Warehouse, DW) Dr. Tun-Wen Pai

關於資料倉儲

資料倉儲具有主題式、經過整合、分析具有時變性、資料不會流失的優勢，用來支援商業智慧，例如分析。

關於資料超市

由資料倉儲的資料，分離出一個更具主題式的資料集，稱為資料超市。

資料倉儲的需求

從各方蒐集出來的資料通常來說都是支離破碎的（例如：包含編碼、形式不同、格式不一等等）。

所以通常一個企業需要的是從這堆支離破碎的資料中經過整合，提供出一個企業想要的簡潔有力的資訊。

因此資料倉儲主要就是蒐集這些破碎的資料，然後提供給企業一個高質量的資料。

資料倉儲的開發考慮

1. 主題式：僅對特定主題感興趣
2. 整合性：各式各樣的資料進入資料倉儲都會經過整合成統一格式的資料
3. 時變性：分析具有時變性，根據不同時間具有不同的分析結果
4. 穩定性：進入資料倉儲的資料並不會變動。
5. 摘要整理：進入的資料會盡可能地最佳化資料欄位與維度。
6. 非標準化：Data Warehouse 使用星狀綱要，因此不是標準化成一個表格做查詢。
7. 後設資料：可以衍伸出其他的資料變數。
8. 即時／合適的時間

Operational System 與 Informational System

Operational System 主要負責即時的商業活動運行，例如：訂單整合、預訂系統...

Operational System 會將這個活動的資料完整的保存下來，但這些資料都是支離破碎的，會需要經過整合。

Informational System 主要被設計來從資料倉儲的資料提供出對於特定時間的快照資料，以利於支援資料探勘、分析等等。

總歸來說，Operational System 與 Informational System 可以利用以下的表格來比較。

	Operational System	Informational System
主要用途	即時的商業活動運行	支援決策、分析等等
資料型態	進行商業活動得到的數據（客戶、銷售商品. . . ）	特定時間的快照資料與預測
給哪些人使用	銷售者	管理層級的人員、商業分析、客戶...
使用範圍	狹窄、易於更新與查詢	非常廣、特設、在詢問與分析上非常複雜
設計目標	能夠被存取、能夠完整	易於使用與存取
資料大小	在更新與存取時，得到的只有帶有幾列的表格資料	定期更新、詢問，通常需要非常大量的資料

資料倉儲的資料處理

一般遵循 `ETL` 的方式處理，也就是 Extract（萃取）、Transform（轉換）、Load（讀入）。

通常來說，會從各式資料源萃取資料，將資料進行整合（轉換）、接著讀入 Data Mart 或 Data Warehouse。

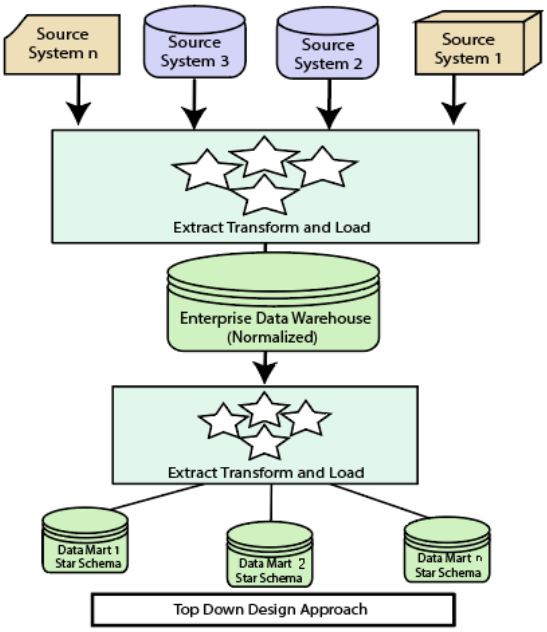
資料倉儲的開發方法

主要分成兩種：Top-Down、Buttom-Up。

Bill Inmon – Top-down

先從各方提取資料，接著經過 `ETL` 後，放入 Data Warehouse。

接著從 Data Warehouse 再經過 `ETL` 後，轉成各式各樣的 Data Mart。

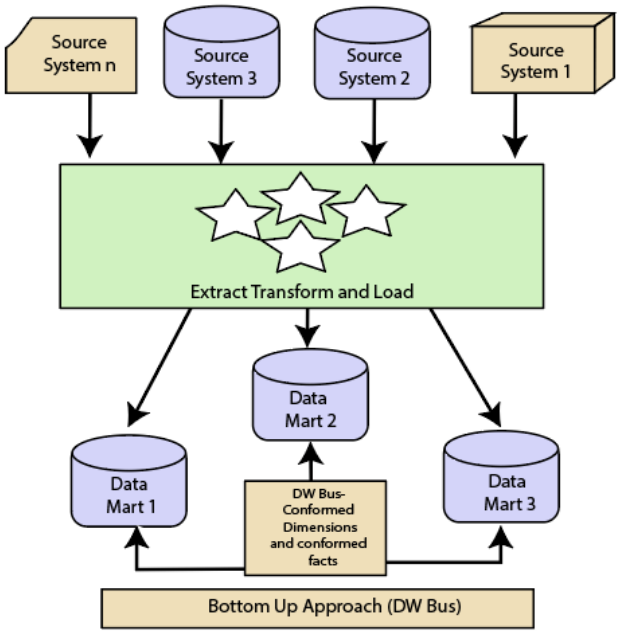


Top Down Design Approach

Ralph Kimball – Bottom-up

先從各方提取資料，接著經過 `ETL` 後，轉置成 Data Mart 的形式。

再經由 `ETL` 的處理後，將這些 Data Mart 合併成一個大的 Data Warehouse。



Bottom Up Design Approach

兩個開發方式的優缺點分析

分類	Top-Down	Buttom-up
問題分解	將大問題分解成小問題	將小問題解決，並且推廣至大問題
處理資料	處理資料的時候，架構是固定的 因此資料缺乏彈性	由於先建立 Data Mart 所以只需要增加 Data Mart 即可增加資料的彈性
管理	集中管理	去中心化管理
冗餘問題	可能包含冗餘的訊息	冗餘的訊息可以被移除
運行速度	由於單一窗口，所以運行較快	由許多 Data Mart 當作窗口，所以運行較慢

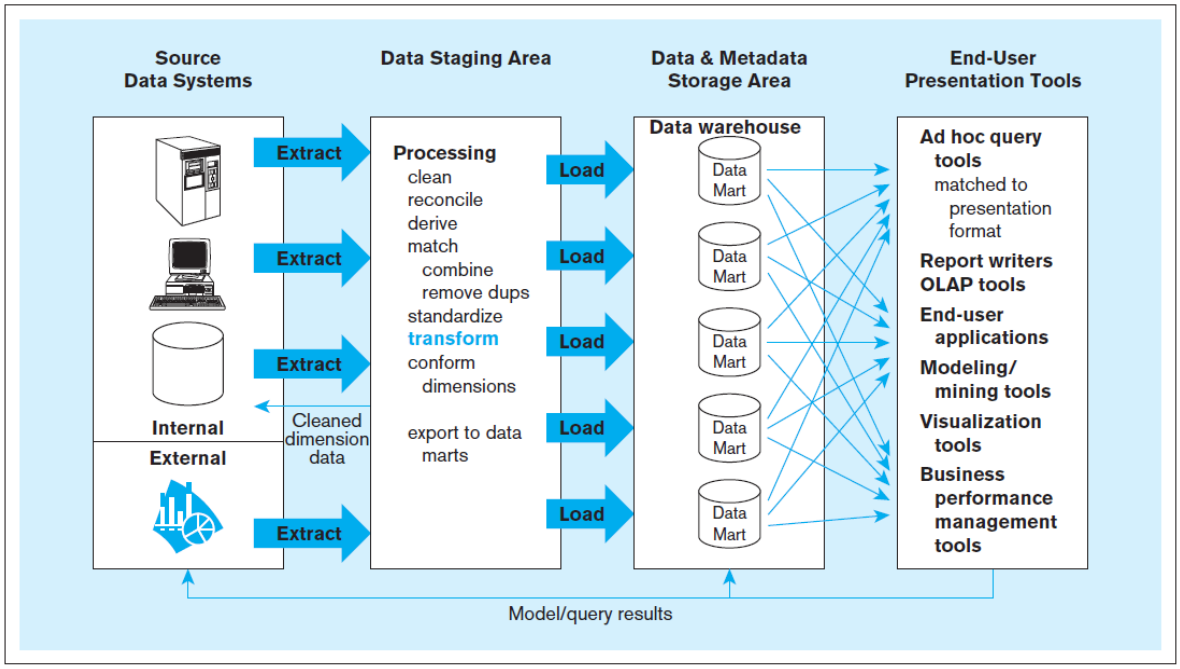
Reference Website:

1. S, V. (2018, February 28). *Various data warehouse design approaches:top-down and bottom-up*. DWgeek.com. Retrieved April 4, 2022, from https://dwgeek.com/various-data-warehouse-design-approaches.html/
2. *Data Warehouse Design - javatpoint*. www.javatpoint.com. (n.d.). Retrieved April 4, 2022, from https://www.javatpoint.com/data-warehouse-design

資料倉儲的設計架構

以 top-down 為例，主要分成四大子架構：資料來源、資料轉換、資料超市或資料倉儲、應用。

基本按照 `ETL` 架構來進行資料處理（見 資料倉儲的資料處理），轉換成 Data Warehouse 或 Data Mart。



資料倉儲的資料模型

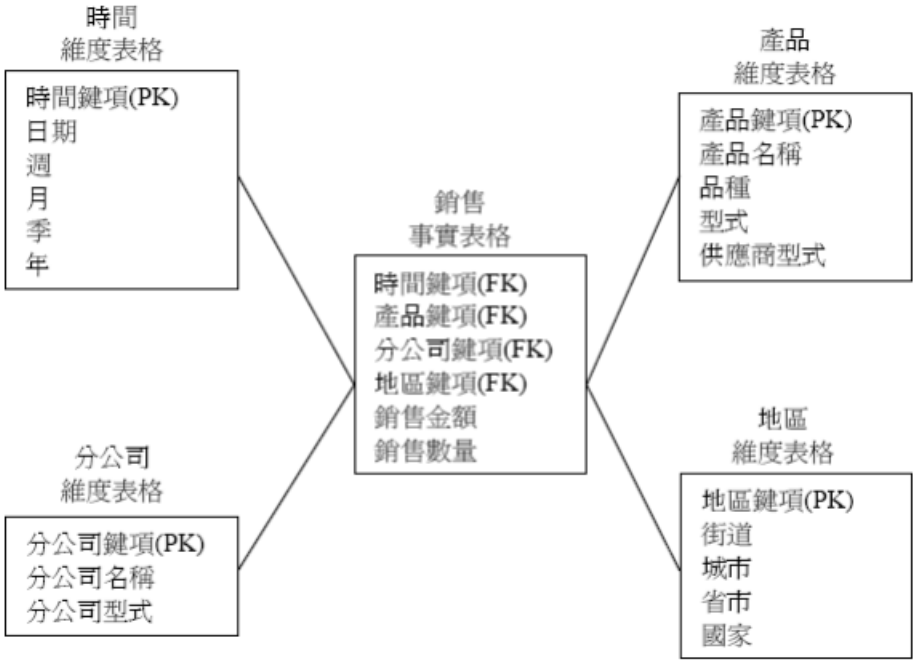
在資料倉儲中，資料模型都是多維度的，包含一個以上的事實表格與多個維度表格。

以下介紹三種不同的資料模型：星狀綱要、雪花狀綱要、事實星座綱要。

星狀綱要

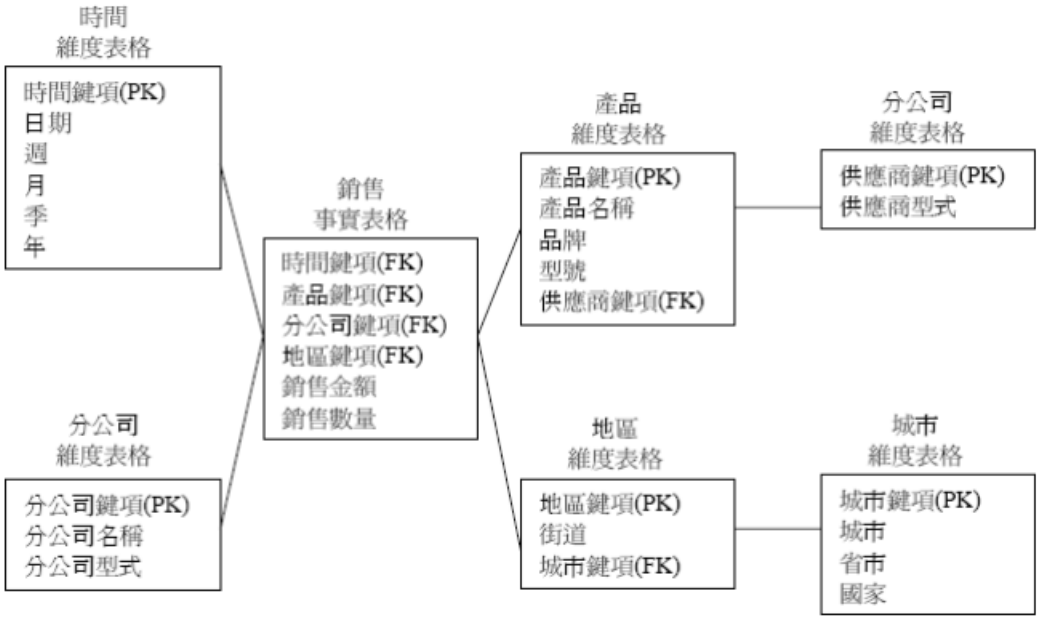
一個事實表格包含大量不重複的資料，與多個維度表格組成的綱要。

是目前最常用的綱要。



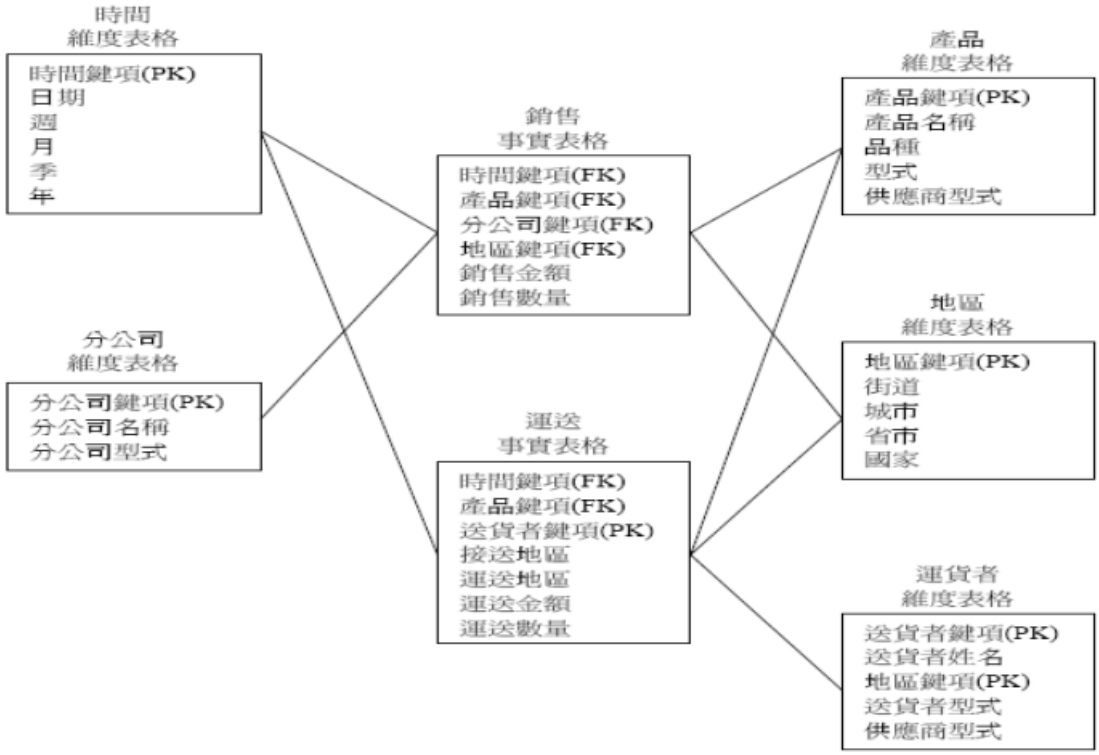
雪花狀綱要

一個事實表格包含大量不重複的資料，與多個維度表格組成的綱要，多個維度可再經過細分分成更多個維度的表格。



事實星座綱要

由多個事實表格共用維度表格。



資料倉儲的儲存架構

實際上的儲存可能是關聯式資料庫或資料立方體。

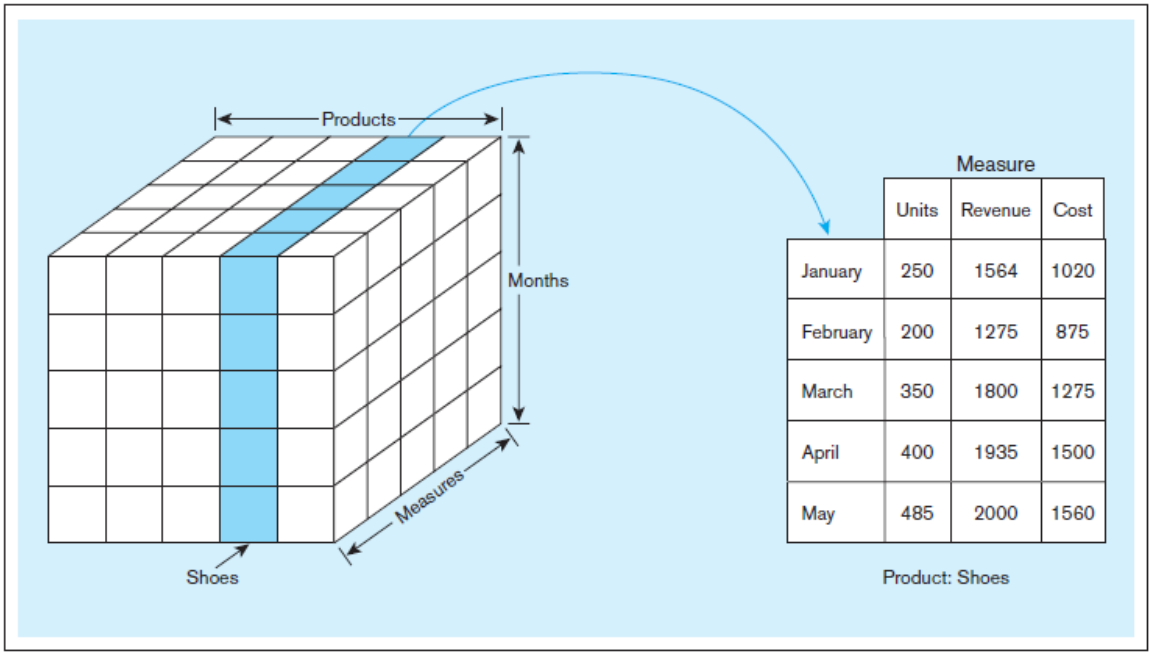
資料立方體 - 切塊與切片

由維度與事實組成，具有上捲、下鑽、切塊、切片、轉軸等操作，見以下例子。

操作方式 - 切片

例如下圖的資料方塊是由（產品類別、月份、數量）三種維度所組成。

若我們想要知道鞋子的資料，就相當於將產品類別固定為鞋子，得到一個降成二維的（月份、數量）資料。



操作方式 - 切塊

延續上面的例子，若我們想要知道鞋子與衣服的資料，相當於將鞋子、衣服的立方塊切下來，此時資料依然是三維的。

操作方式 - 下鑽

延續上面的例子，若我們將這個鞋子的資料，想要知道某個月份的上旬、中旬或下旬的銷售資料。

則我們可以針對於資料立方體的月份下鑽，得到上旬、中旬、下旬的資料，藉此得到更詳細的資料。

操作方式 - 上捲

延續上面的例子，若我們想要知道整年的銷售資料，則我們可以針對於資料立方體的月份上捲，得到更為概觀的資料。

操作方式 - 轉軸

我們可以將立方體做旋轉，得到不同視角的資料。

資料倉儲的查詢處理

- 若是為了快速查詢，在記憶空間充裕的情況下，可能會根據這些資料先建立出可能查詢的方式，可藉由目前有的維度欄位慢慢上推整合合併成一個包含所有維度的資料。
- 多維度的資料與一般的關聯式資料庫相比，一般的關聯式資料庫只能慢慢地計算累積，而多維度資料庫具有更好的效能。