商業智慧 第五章練習題解答

問答題：

1. Ans: (參考5.2節)
* 規劃與設計流程之步驟：參考圖5-1、
	+ 1. 需求與現況(Requirements and realities)
		2. 架構設計(Architecture)
		3. 系統建置(System implementation)
		4. 測試與發布(Test and release)
* 各步驟之任務：參考表5-1
1. Ans: (參考5.2節)
	* 資料處理流程之步驟：參考圖5-2、
		1. 萃取(Extract)
		2. 清理(Clean)
		3. 一致化(Conform)
		4. 交付(Delivery)
	* 各步驟之任務：參考表5-2
2. Ans: (參考5.3.1節)
* 業務需求 (Business needs)
* 法規需求 (Compliance requirements)
* 品質需求 (Quality requirements)
* 安全需求 (Security requirements)
* 整合需求 (Integration requirements)
* 時程需求 (Latency requirement)
* 備份需求 (Backup requirements)
* 交付需求 (Delivery requirements)
* 技能需求 (Skills requirements)
* 資源需求 (Resource requirements)
1. Ans: (參考5.3.2節)
* 購買 ETL 工具還是自行開發 (ETL tool versus hand coding)
* 批次還是串流的處理流程 (Batch versus streaming data flow)
* 水平還是垂直的任務依賴 (Horizontal versus vertical task dependency)
* 排程自動化 (Scheduler automation)
* 異常處理的機制 (Exception handling)
* 品質控制的機制 (Quality handling)
* 復原與重新啟動 (Recovery and restart)
* 詮釋資料的處理方式 (Metadata)
* 資料安全的控管方式 (Security)
1. Ans: (參考5.3.2.2節)
* 資料集結(Staging) 概念： 一詞意指後端所進行ETL 的各個步驟(萃取、清理、一致化與交付) 將所產生的資料進行儲存。
* 設計資料集結區工具：製作集結區規模估算表(Staging tables volumetric worksheet) 給資料庫管理員和系統管理員進行集結區的空間分配。
1. Ans: (參考5.3.2.3節)
* 平面文件 (Flat files)
* XML 資料集 (XML Data sets)
* 關聯資料表 (Relational tables)
* 獨立的 DBMS 工作表 (Independent DBMS working tables)
* 第三正規化 ER 模型 (Third normal form entity/relation models)
* 非關聯式資料來源 (Nonrelational data sources)
* 維度資料模型 (Dimensional data models)
1. Ans: (參考5.4.1節)
* 邏輯資料對照表(The logical data map) 的功能：為ETL 開發者提供一個藍圖，清晰地描述在轉換過程中來源系統的資料定義，目標資料倉儲的資料模型，以及從原始資料到最終目的資料轉換所需要的操作
* 邏輯資料對照表組成之內容包括：

(1) 目標表名稱 (Target table name)

(2) 目標欄位名稱(Target column name)

(3) 資料表類型 (Table type)

(4) 緩時變維度類型 (SCD type)

(5) 來源資料庫(Source database)

(6) 來源資料表名稱 (Source table name)

(7) 來源欄位名稱(Source column name)

(8) 轉換規則(Transformation)

1. Ans: (參考5.4.2節)
* 資料探索階段的步驟：

(1) 收集來源系統：包含來源系統的說明檔、資料字典、報告等。

(2) 追蹤來源系統：建置來源系統追縱報告(Source system tracking report)。

(3) 確定原始系統：確定資料發源地，解決相同資料不同版本之問題。

(4) 分析來源系統：如完整的瞭解來源系統之 ER 模型。

* 資料探索階段的工具：來源系統追蹤報告。
1. Ans: (參考5.4.3.1節)
* 使用審計欄
* 使用資料庫日誌
* 按時萃取
* 排除處理
* 初始和增量載入
1. Ans: (參考5.5.1節)

(1) 正確性檢查(Correct)

(2) 明確性檢查(Unambiguous)

(3) 一致性檢查(Consistent)

(4) 完全性檢查(Complete)。

1. Ans: (參考5.5.1節)
* 四種壓力

(1) 完備性 (Be thorough)：必須保證在資料檢測的全面性。

(2) 快速性 (Be fast)：必須快速的處理大量且不斷增長的資料。

(3) 正確性 (Be corrective)：必須保證資料的正確性。

(4) 透明性 (Be transparent)：必須誠實的揭露資料的錯誤與缺陷。

* 二種平衡
	+ 1. 衝突一：完備性與快速性

資料品質系統無法同時兼顧快速性和完備性。解決的方法就是思考資料品質的問題來找到平衡點，比如：(1) 時效性：什麼時間點資料會變得過時？及 (2)正確性：保證這些資料的正確性有多重要？等等。

* + 1. 衝突二：正確性與透明性

資料品質系統無法同時滿足正確性和透明性。解決的方法是對不同的錯誤建立一個可判斷的原則界限，能夠在清理過程進行糾正或者標識，並生成一個容易進行審計的工具，用來記錄那些錯誤與修正。

1. Ans: (參考5.5.2節)
* 清理前：資料概況分析報告 (Data-profiling analysis)
* 清理中：錯誤事件事實表 (Error event fact table)
* 清理後：審計維度表 (Audit dimension)
1. Ans: (參考5.5.3.1節)
* 資料計數(Data count) ：是對有問題的欄位進行分組時計算該表的資料筆數。透過簡單的查詢、分析值的分布，如發現奇異值，就可能發現損壞的資料。
* 資料採樣 (Data sampling)：對於資料量大的表，一般通過採樣技術來減少資料量，然後分析其分布的合理性。
1. Ans: (參考5.5.3.2節)
	* 欄位檢查 (Column property enforcement)
	* 結構檢查 (Structure enforcement)
	* 資料檢查 (Data enforcement)
	* 值檢查 (Value enforcement)
2. Ans: (參考5.5.3.2節)
* 處理流程(參考圖 5-8)

(1) 執行：一系列的篩檢程式按照詮釋資料定義的順序執行。

(2) 記錄：篩檢程式所偵測到的每個錯誤記錄在錯誤事件事實表中。

(3) 判斷：當每一篩檢程式執行後，就會在錯誤事件事實表中查詢是否遇到致命的錯誤 (如整批資料遺失)。

(a) 如果沒有發現：持續運行 ETL 流程；

(b) 如果發現：停止 ETL 處理流程。

(4) 審計：當篩檢程式執行完畢後，根據錯誤事件事實表為審計維度表計算整個資料品質得分。

* 指導原則：持續運行原則
	+ 資料清理的指導原則是發現並且記錄存在的資料品質錯誤，而不是跳過資料或者是停止ETL 過程。
1. Ans: (參考5.6節)
* 一致化原因：

由於終端使用者會從多個分離的事實表組合所需的資料進行應用，所以在清理資料後，就需對資料進行一致化處理，解決不同來源資料間資料不一致的問題，使資料能夠在資料倉儲中一起使用。而ETL 小組必須為這些事實表提供統一的介面，這樣資料方能被整合，此統一的介面就是一致化維度(Conformed dimensions) 和一致化事實 (Conformed facts)。

* 一致化維度

一致化維度對於每個被關聯的事實表來說都是相同的，為所有事實表提供統一的資料內容和屬性描述，這樣才可以將不同事實表間的資料進行資料整合。一個好的一致化維度常需要融合多個舊系統的資源及外部資料來源。

* 一致化事實

在確定一致化維度的過程中同時也確定了標準的事實定義。建立一致化維度與一致化事實是一個合作的過程，因為每一張事實表的管理員都必須就使用一致化維度來達成一致化的事實。有了一致化的事實方可以直接進行比較，甚至可以直接進行數學計算，如計算總和、比率等。

1. Ans: (參考5.6.1節)
* 標準化 (Standardizing)

標準化的目的是使不同資料來源採用相同的資料編碼方式、資料格式等，為下一步比對(去重複) 步驟作準備。

* 比對 (去重複) (Matching drives deduplication)

比對，或叫做去重複，指刪除重複的標準化資料。

* 篩選 (Surviving)

篩選是指整合一組完成比對的資料作為一致化維度的主資料，成為最終交付的一致化維度，提供最豐富完整之維度資料。

1. Ans: (參考5.6.2節)
* 一致化維度表的交付步驟：

(1) 增加新的維度資料到一致化維度，產生新的代理鍵。

(2) 為第二型緩時變維度(保留歷史) 的變化，增加新維度記錄到現有一致化維度中，並生成新的代理鍵。

(3) 為第一型緩時變維度(覆寫) 和第三型緩時變維度(替換實體) 的變化，修改維度資料，但不改變代理鍵。

(4) 一旦完成修改，則更新維度的版本號。

(5) 為所有的事實表提供者同步發布修訂後的一致化維度。

* 一致化事實表的交付步驟：

(1) 接收或者下載更新一致化維度。

(2) 處理標記為新且當前在用的維度資料，改映射到代理鍵的當前鍵值。

(3) 處理標記為新並且為遲到的維度資料。

(4) 用正確的代理鍵替換自然鍵後，把所有新事實資料添加到事實表中。

(5) 修改所有的事實表。

(6) 刪除無效的彙總。

(7) 重新計算受影響的彙總。

(8) 確保所有基礎資料表和彙總事實表的資料品質，確保正確計算彙總事實表。

(9) 把更新完成的事實表和維度表上線。

(10) 通知終端使用者資料倉儲已被更新。

1. Ans: (參考5.7.2節)
* 彙總事實表的用途：

藉由控制維度表的細度(如產品項目(小) 改成產品類別(大)) ，來改變原事實表事實之細度，並對原事實表之量值進行彙總計算後，另產生一組細度較大之彙總事實表，可以大幅的減少原事實表的資料量，作為查詢的可用版本。藉此，就可以提供不同查詢細度之SQL 指令，選擇最適合之彙總版本進行查詢。透過建立彙總之技術可以顯著的提高性能。

* 彙總導航架構(參考圖 5-14)
* 彙總導航架構之設計要點，：

(1) 針對資料倉儲中重要維度常用的彙總層級，建立多組彙總事實表。

(2) 彙總導航器是中介軟體的一部分，位於用戶端和 DBMS 之間。

(3) 彙總導航器用於將查詢基礎事實表的SQL 轉化為查詢彙總事實表的SQL。

(4) 彙總導航器使用特殊的詮釋資料進行要點(3) 之轉化，該詮釋資料描述了資料倉儲之彙總方式。

1. Ans: (參考5.8節)

詮釋資料的架構與分類(參考圖 5-15)：

* 業務詮釋資料 (Business metadata)
* 技術詮釋資料 (Technical metadata)
* 處理過程詮釋資料(Process metadata)

二、選擇題：

1. (C) ;2. (A) ; 3. (B) ;4. (B) ;5. (D)