高等資料庫 - 期中報告 Guide:

Version: 0911

請自行分組，組員人數 4 ~ 6人。

資料庫系統規劃 ( MSSQL 版本)

今有 政府公開之兩個資料來源如下，請進行資料庫系統規劃，以便後續查詢。

1. 違反勞動法令事業單位-勞工退休金條例 (<https://data.gov.tw/dataset/109898>)
2. 列管事業污染源裁處資料 (<https://data.moenv.gov.tw/dataset/detail/EMS_P_46>)

( 針對下列需求，應可進行 1. 單獨查詢某一資料源 2.單次查詢所有資料源)

1. 輸入公司名稱，查詢出該公司之違反資料
2. 輸入公司名稱、年度，可得知該年度公司罰款金額
3. 得知何公司違反最多事項
4. 得知何公司罰款金額最多

請依上列查詢需求進行相關資料表設計，資料表之欄位需依該欄位之可能值進行相關設計。於資料表結構設計完成後，將該公開資料下載至設計之系統。

下載過程中，可嘗試不同的資料量，包括 資料之數量為 (1) 1千筆、(2) 1萬筆 (3) 10 萬筆 (4) 50萬筆 (5)100萬筆 \*1，即進行相關查詢，甚至嘗試建立索引，以觀察不同數量的資料對查詢效率的影響，特別是在有無建立索引的情況下。

\*1: 當公開資料的資料量低於100萬筆時，我們會使用"資料混淆"（data obfuscation）的方式來擴充資料量。這種擴充方式是將原始資料複製，然後將資料來源1的欄位中的“事業單位名稱”的第一個字元以星號()取代，例如：將"金民股份有限公司"轉換為"民股份有限公司"。這樣一來，資料量就會變成原來的兩倍。接著，我們再次複製原始資料，但這次將第二個字元以星號()取代，這樣資料量就變成原來的三倍。如果資料量仍然不足，我們可以再次複製原始資料，依此類推，每次都將下一個字元以星號()取代。資料來源2的處理方式也類似，我們會以"FAC\_NAME（事業名稱）"這個欄位來進行資料混淆，同樣是將第一個字元和第二個字元以星號(\*)取代。這樣我們可以通過這種方式將資料量擴充至超過100萬筆。

報告之內容應包含:

Table schema、設計構想、index建立對應不同資料量之效率影響。

並針對 較特殊之欄位，補充說明為何需指定該資料格式。

此公開資料有中文資料，在選定之character-set 上，應該選哪一種？

是否可應用 store procedure or table view ? 若有使用 store procedure or table view請加以補充說明。

高等資料庫 - 期末報告 Guide:

已於期中完成 MSSQL 之資料庫，現欲將 MSSQL 資料庫裡面之資料轉移至 MariaDB上。請於期末報告說明以下事項：

1. 如何將 MSSQL 上之資料 轉移至 MariaDB
2. 轉移過程中如何進行資料一致性驗證
3. 轉移過程後如何進行資料一致性驗證
4. MariaDB上有 explain 指令[1]，請使用該指令進行優化動作，並說明優化後之成果。
5. 針對MSSQL 與 MariaDB 進行差異性說明

是否可以使用外部程式進行資料轉移？若使用外部程式，如何進行轉移過程中的資料一致性驗證？

(a) 老師同意可以使用外部程式進行資料轉移，但需要明確說明使用哪一個外部程式以及轉移的相關限制。由於一般外部程式轉移可能存在資料筆數限制，如果使用外部程式，請將資料擴增至1千萬筆以上，並確認外部程式是否有資料筆數的限制。如果在轉移過程中發現其他限制，請一併記錄並在期末報告中提出。

(b) 轉移過程中的資料一致性驗證，請透過截圖說明外部程式在轉移過程中是否進行了資料一致性驗證。同時，請記錄是否可以分批轉移，以及是否有轉移日誌等相關資訊。

(c) 轉移結束後，仍需進行兩邊資料的一致性驗證，有關資料一致性驗證的內容將在後續進行補充說明。

資料一致性驗證的內容可分為兩部分，一為資料表綱要(table schema)，另一為儲存的資料(data)。

資料表綱要的驗證包括確認在原先的資料表中所能接受的 "極大" 或 "極小" 值，在轉移後的資料表中是否仍能正確寫入，並確保資料的型態未發生問題。

轉移資料的過程可以想象成從資料表 A 複製到資料表 B，因為資料表內有多筆資料，在複製過程中可能有多種作法。一種相對較差的作法是每複製一筆資料後就立即確認這筆資料是否正確（即從來源和目的地資料表各抓出同一筆資料作比對）。這種方式在未建立索引或主鍵的情況下可能會導致效能低落的問題。因此，在轉移資料過程中，如果發現轉移錯誤，應立即記錄日誌以便後續處理。

[1] MariaDB explain, <https://mariadb.com/kb/en/explain/>