

研究生：陳詩涵

指導教授：梁高榮博士

國立交通大學工業工程與管理學系

摘要

漁業資訊分享熱線是以網際網路為基礎的資訊系統，其建構的目的是為了探索重要的漁業知識。漁業資訊分享熱線的建構流程，由本質面來看依序可分為三個階段：資料倉儲，線上分析處理與資料探勘。在第一個階段中，定期地將異質性來源的漁業資料，匯集並存入資料倉儲。接下來在線上分析處理的階段，透過表格與圖表的呈現，依需求瀏覽與篩選出特定的訊息。第三個階段則是利用資料探勘技術，將前個階段所保留的訊息作進一步分析處理。在本論文中，共進行了三種資料探勘的案例。第一個案例使用序列模式分析方法，追蹤高漁獲量漁船的航行軌跡。第二個案例則在給定的準則下，利用二階段分群分析方法找出優良的漁船單位。第三個案則是藉由多變量技術進行漁區與月份對於漁獲情形關係的假設檢定與後續檢定。



關鍵字

資料倉儲(Data Warehouse, DW)

漁業業務情報網(Fishery Business Intelligence, FBI)

漁業資訊分享熱線(Fishery Information Sharing Hotline, FISH)

線上分析處理(On Line Analytical Processing, OLAP)

序列模式(Sequential Patterns)

分群分析(Cluster Analysis)

多變量變異數分析(Multivariate Analysis of Variance, MANOVA)

Design and Implementation of Fishery Information Sharing Hotline System
Using Data Warehouse and Data Mining Techniques

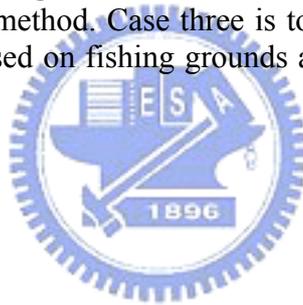
Student : Shih-Han, Chen

Advisor : Dr.Gau-Rong Liang

Department of Institute of Industrial Engineering & Management
National Chiao Tung University

Abstract

An Internet-based software system entitled Fishery Information Sharing Hotline (FISH) is constructed for exploring important fishery knowledge. Essentially the FISH system consists of three sequential stages: data warehouse, On-Line Analytical Processing (OLAP), and data mining. At the first stage, the fishery data from various sources are collected into the data warehouse periodically. Then selected tabular messages or visual diagrams are scanned and screened as required at the OLAP stage. Also the qualified messages will be exported for more advanced processing using data mining techniques. In this thesis, three data mining cases are studied at the final stage. Case one is to trace the voyaging trails of interested vessels using a mining method for searching sequential patterns. For the given criteria, case two is to filter out outstanding vessel groups using a two-stage clustering method. Case three is to test some well-known hypotheses for fisheries production relationships based on fishing grounds and seasons. The research results show the feasibility of FISH approach.



Keywords :

Data Warehouse, DW
Fishery Business Intelligence, FBI
Fishery Information Sharing Hotline, FISH
On Line Analytical Processing, OLAP
Sequential Patterns
Cluster Analysis
Multivariate Analysis of Variance, MANOVA

誌謝

本論文至今得以完成，心裡由衷地感謝身邊的人在過程中給予鼓勵、支持，沒有大家就沒有今天的詩涵。首先要感謝恩師梁高榮教授給予的指導。在這段求學的期間，老師培養了我做研究正確的態度和方法，更多方面地帶領我觸探其他領域的知識，帶給我最深的影響是讓我親眼見到一個成功人士背後的努力與自我砥礪的態度。希望老師在百忙之中仍要身體健康，少喝點可樂與少吃點漢堡喔。另外也感謝唐麗英教授以及李慶恩教授對本論文所提供的寶貴意見，使論文能更臻於完備。

此外本論文之進行，是伴隨著參與農委會的計畫進行。在此致謝「農委會漁業署-漁船資訊傳送系統整合計畫-92 農科-7.1.4-漁-F1」計畫，使我在論文進行中，除了理論的研究，更能與實務接軌。而在論文進行期間，曾多次拜訪漁業署與台北花卉產銷股份有限公司等單位實地拜訪，承蒙各領域之專家與學者協助提供資料與技術說明，特此致謝。另外在研究過程中曾參與漢康科技公司之專業課程，協助學習許多商業智慧之技術，亦在此致謝。

在學習與論文寫作的過程中，在此誠摯地感謝實驗室的同窗淙亮、士凱、英泰、楓凱等人的陪伴，無論是在課業上或是生活上，他們都是我好的夥伴，不僅是一起完成課業、計畫，也使我從中學到了許多做人處事、調劑身心的方法。另外也感謝學長哲煜、漢清、軒豪、雍仁、均豪以及世民在課業上樹立良好的楷模，以及學弟、學妹們的參與，為研究室帶來了歡樂，使整個團體更為融洽。在魚貝苗資料倉儲以及遠洋漁業資料倉儲的建置，分別感謝徐英泰以及溫師瀚的大力協助。此外感謝大個、Lucky、意琳、佳宜、齡玉、建秀、玉鳳、小潘、櫻溶、思明、小炳、Brian、Jason、佩欣、菁華、玉華、美玲、美良、碧姝、淑雍等太多好友們，從 92 年暑假開始陪我渡過最辛苦的日子。最後更要感謝我最親愛的家人們，是他們的鼓勵讓我能夠安心地求學，也是我心靈最大的依靠。

謹誌

中華民國九十三年六月
于風城交大

目錄

摘要.....	I
關鍵字.....	I
ABSTRACT.....	II
KEYWORDS :	II
誌謝.....	III
目錄.....	IV
圖目錄.....	VI
表目錄.....	X
第一章 緒論.....	- 1 -
1.1 研究動機.....	- 1 -
1.2 問題界定.....	- 3 -
1.3 研究目的.....	- 4 -
1.4 研究方法.....	- 5 -
1.5 論文架構.....	- 6 -
第二章 文獻回顧.....	- 7 -
2.1 漁業供應鏈.....	- 7 -
2.2 資料倉儲技術.....	- 9 -
2.3 資料擷轉.....	- 15 -
2.4 企業智慧系統架構.....	- 17 -
2.5 線上分析挖掘.....	- 19 -
2.6 序列模式挖掘.....	- 26 -
第三章 漁業資訊分享熱線的資料倉儲設計與實作.....	- 35 -
3.1 來源資料的性質與問題描述.....	- 35 -
3.2 遠洋魚貨銷售與漁船軌跡資料倉儲的設計與實作.....	- 40 -
3.3 魚貝苗養殖資料倉儲的設計與實作.....	- 60 -
3.4 水產加工製品資料倉儲的設計與實作.....	- 63 -
3.5 遠洋漁獲資料倉儲的設計與實作.....	- 67 -
3.6 各資料倉儲的研究限制.....	- 70 -
第四章 漁業資訊分享熱線的線上分析處理.....	- 71 -
4.1 漁業資訊分享熱線的系統架構.....	- 71 -
4.2 超方體建構與線上分析處理之設計.....	- 73 -
4.3 漁業資訊分享熱線之實作成果.....	- 75 -
第五章 資料探勘(I)－建構高漁獲量漁船的軌跡序列模式.....	- 85 -

5.1 序列模式挖掘的系統流程架構分析.....	- 85 -
5.2 結合資料轉換服務技術.....	- 87 -
5.3 WUM 系統的應用.....	- 91 -
第六章 資料探勘(II)－尋找優良漁船單位.....	- 99 -
6.1 優良漁船單位特性分析之架構.....	- 99 -
6.2 研究變數的定義與衡量值.....	- 101 -
6.3 分群分析與結果.....	- 103 -
6.4 透過資料轉換服務技術以連結漁船屬性.....	- 107 -
6.5 多變量變異數分析與結果.....	- 109 -
第七章 資料探勘(III)－各魚種在不同漁區、月份下與漁獲情形關係的假設檢定.....	- 111 -
7.1 探討漁區與月份對漁獲量影響之研究結果應用.....	- 111 -
7.2 多變量變異數分析的研究架構.....	- 112 -
7.3 資料分析方法與研究限制.....	- 113 -
7.4 資料分析結果與結論.....	- 117 -
第八章 結論與未來研究方向.....	- 140 -
8.1 結論.....	- 140 -
8.2 未來研究方向.....	- 141 -
參考文獻.....	- 144 -
附錄一 建構資料倉儲的各來源資料表內容與資料格式.....	- 146 -
1.1 「遠洋魚貨銷售與漁船軌跡資料倉儲」的來源資料表.....	- 146 -
1.2 「魚貝苗養殖資料倉儲」的來源資料表.....	- 150 -
1.3 「水產加工製品資料倉儲」的來源資料表.....	- 152 -
1.4 「遠洋漁獲資料倉儲」的來源資料表.....	- 154 -
附錄二 遠洋魚貨銷售與漁船軌跡資料倉儲的資料表.....	- 156 -
附錄三 使用「多變量常態分配檢定」.....	- 159 -

圖目錄

圖 1.1.1 中華民國台閩地區漁業統計.....	- 2 -
圖 1.1.2 研究背景與動機脈絡圖.....	- 2 -
圖 1.3.1 漁業資訊分享熱線的現階段架構與研究主題.....	- 4 -
圖 1.4.1 研究方法與步驟.....	- 5 -
圖 1.5.1 論文架構.....	- 6 -
圖 2.1.1 漁業產品運銷體系的關係圖.....	- 7 -
圖 2.2.1 資料倉儲系統架構與其相關應用.....	- 10 -
圖 2.2.2 星狀綱要.....	- 11 -
圖 2.2.3 雪花綱要(SNOWFLAKE SCHEMA).....	- 11 -
圖 2.2.4 星座綱要(FACT CONSTELLATION SCHEMA).....	- 12 -
圖 2.2.5 漁船資料倉儲的三階正規化關聯網要.....	- 13 -
圖 2.2.6 漁船資料倉儲的雪花綱要.....	- 14 -
圖 2.3.1 資料轉換服務連接異質來源資料的方式.....	- 15 -
圖 2.3.2 資料轉換服務功能及資料擷轉流程圖.....	- 16 -
圖 2.4.1 COGNOS 業務情報系統架構.....	- 18 -
圖 2.5.1 名目變數相似度的計算.....	- 20 -
圖 2.5.2 順序變數相似度的計算.....	- 20 -
圖 2.5.3 區間變數相似度衡量指標的定義.....	- 21 -
圖 2.5.4 階層式分群的過程.....	- 22 -
圖 2.5.5 非階層式分群的過程.....	- 22 -
圖 2.5.6 歸類分析的架構.....	- 23 -
圖 2.5.7 決策樹之架構.....	- 24 -
圖 2.5.8 決策樹形成之步驟.....	- 25 -
圖 2.6.1 序列模式與關聯性法則的比較.....	- 30 -
圖 2.6.2 尋找最大序列的方法.....	- 30 -
圖 2.6.3 WUM 開發環境以及架構.....	- 31 -
圖 2.6.4 AGGREGATED LOG 所包含的架構.....	- 32 -
圖 2.6.5 G-SEQUENCE $B^*[1; 5]D$ 的瀏覽模式.....	- 32 -
圖 2.6.6 A MINT QUERY(1).....	- 33 -
圖 2.6.7 A MINT QUERY(2).....	- 33 -
圖 2.6.8 A MINT QUERY(3).....	- 33 -
圖 2.6.9 A MINT QUERY(4).....	- 34 -
圖 2.6.10 序列的觀察.....	- 34 -
圖 3.1.1 漁業統計資料庫之資料來源.....	- 35 -
圖 3.1.2 各縣市政府傳送的文字檔.....	- 36 -
圖 3.1.3 VDW 建構流程.....	- 38 -
圖 3.2.1 遠洋魚貨銷售與漁船軌跡資料倉儲的架構與資料來源.....	- 40 -
圖 3.2.2 遠洋魚貨銷售與漁船軌跡資料倉儲的雪花綱要.....	- 41 -
圖 3.2.3 資料轉換服務.....	- 45 -

圖 3.2.4 利用資料轉換服務建立漁船相關屬性維度的流程	- 46 -
圖 3.2.5 SQL 更新	- 46 -
圖 3.2.6 經濟部商業司查詢系統	- 47 -
圖 3.2.7 處理主鍵重複的步驟	- 48 -
圖 3.2.8 更新維度表的步驟	- 49 -
圖 3.2.9 維度代表的地理意義	- 50 -
圖 3.2.10 利用資料轉換服務及預存程序建立漁船航行軌跡相關屬性維度的流程	- 50 -
圖 3.2.11 經緯度範圍的劃分以及編碼	- 51 -
圖 3.2.12 建立作業海域維度的語法	- 51 -
圖 3.2.13 建立漁獲區維度的語法	- 52 -
圖 3.2.14 利用資料轉換服務建立事實資料表的流程	- 53 -
圖 3.2.15 中文船名格式的統一	- 54 -
圖 3.2.16 時間格式的統一	- 55 -
圖 3.2.17 利用中文名稱尋查漁船代碼	- 56 -
圖 3.2.18 整合每一筆交易資料	- 57 -
圖 3.2.19 擴增維度所使用的方法	- 57 -
圖 3.2.20 以地圖建立銷售國別的維度資料表	- 58 -
圖 3.2.21 將漁船航海軌跡記錄存入事實資料表	- 59 -
圖 3.3.1 魚貝苗資料倉儲的雪花綱要	- 60 -
圖 3.4.1 水產加工製品資料倉儲的星座綱要	- 64 -
圖 3.5.1 遠洋漁獲資料倉儲的星座綱要	- 67 -
圖 4.1.1 目前建構於 FISH 的資料倉儲系統與線上分析處理系統	- 71 -
圖 4.1.2 目前 FISH 建構的系統目標圖	- 72 -
圖 4.2.1 以 MOLAP 的方式建構超方體	- 73 -
圖 4.2.2 線上分析處理對維度的操作	- 74 -
圖 4.3.1 紙編本漁業統計年報設計上的缺點	- 75 -
圖 4.3.2 FISH 與原系統的入口網頁	- 76 -
圖 4.3.3 原系統資料瀏覽的選擇頁面	- 77 -
圖 4.3.4 原系統單獨年度與遠洋漁業相關的 2 個統計主題	- 77 -
圖 4.3.5 漁業生產量值維度選單	- 78 -
圖 4.3.6 漁業年報中，漁業種類及魚類維度組合	- 78 -
圖 4.3.7 與遠洋漁獲相關主題的線上分析處理系統	- 79 -
圖 4.3.8 線上分析處理系統的轉軸操作	- 79 -
圖 4.3.9 FISH 中，漁業種類及魚類維度組合	- 79 -
圖 4.3.10 查詢魚種的季節性產量變化	- 81 -
圖 4.3.11 查詢魚種的季節性產量變化	- 81 -
圖 4.3.12 查詢魚種的季節性產量變化	- 82 -
圖 4.3.13 原系統的下載服務	- 84 -
圖 4.3.14 FISH 提供的下載服務	- 84 -
圖 5.1.1 挖掘航海軌跡序列模式之作業	- 85 -

圖 5.1.2 定義序列模式準則作業之流程.....	- 86 -
圖 5.2.1 利用資料轉換服務建立序列模式挖掘所需文字檔的流程.....	- 87 -
圖 5.2.2 序列模式分析的資料轉換服務.....	- 87 -
圖 5.2.3 序列模式分析的分群分析結果.....	- 88 -
圖 5.2.4 篩選出序列模式分析資料的 SQL 語法.....	- 89 -
圖 5.2.5 將高漁獲量的漁船航行資料轉換成網頁記錄檔案格式.....	- 90 -
圖 5.3.1 設定議期時間長度.....	- 91 -
圖 5.3.2 建立資料庫內的階層序列.....	- 91 -
圖 5.3.3 樹狀階層架構的序列.....	- 92 -
圖 6.1.1 挖掘優良漁船單位之流程.....	- 99 -
圖 6.1.2 漁船單位特性分析之程序圖.....	- 100 -
圖 6.1.3 資料分析流程之方法與目的圖.....	- 100 -
圖 6.3.1 各漁船單位之魚貨銷售情形之分群數判斷準則 1.....	- 103 -
圖 6.3.2 各漁船單位之魚貨銷售情形之分群數判斷準則 2.....	- 103 -
圖 6.3.3 分群效應圖.....	- 105 -
圖 6.4.1 資料轉換流程圖.....	- 107 -
圖 6.4.2 資料格式轉換語法.....	- 108 -
圖 7.2.1 操作性研究架構.....	- 112 -
圖 7.3.1 資料分析流程.....	- 113 -
圖 7.3.2 利用線上分析處理，挑選欲分析的魚種.....	- 114 -
圖 7.3.3 抽樣前由資料倉儲中建立某一魚種的資料.....	- 116 -
圖 7.4.1 正鯧各月份及各漁區之衡量值檢視圖.....	- 120 -
圖 7.4.2 正鯧各群體間「漁獲情形之 2 個衡量值」的交互作用檢視圖.....	- 120 -
圖 7.4.3 正鯧較佳漁獲量的海圖.....	- 122 -
圖 7.4.4 黃鰭鮪各月份及各漁區之衡量值檢視圖.....	- 125 -
圖 7.4.5 黃鰭鮪各群體間「漁獲情形之 2 個衡量值」的交互作用檢視圖.....	- 125 -
圖 7.4.6 黃鰭鮪較佳漁獲量的海圖.....	- 126 -
圖 7.4.7 長鰭鮪各月份及各魚區之衡量值檢視圖.....	- 129 -
圖 7.4.8 長鰭鮪各群體間「漁獲情形之 2 個衡量值」的交互作用檢視圖.....	- 129 -
圖 7.4.9 長鰭鮪較佳漁獲量的海圖.....	- 130 -
圖 7.4.10 大目鮪各月份及各漁區之衡量值檢視圖.....	- 133 -
圖 7.4.11 大目鮪各群體間「漁獲情形之 2 個衡量值」的交互作用檢視圖.....	- 133 -
圖 7.4.12 大目鮪較佳漁獲量的海圖.....	- 135 -
圖 7.4.13 黑鮪各月份及各漁區之衡量值檢視圖.....	- 138 -
圖 7.4.14 黑鮪各群體間「漁獲情形之 2 個衡量值」的交互作用檢視圖.....	- 138 -
圖 7.4.15 黑鮪較佳漁獲量的海圖.....	- 139 -
圖 8.2.1 本系統未來可能發生的人為失誤.....	- 141 -
附錄圖 A.1 「正鯧」轉換前原始數據之多變量常態分配檢定.....	- 159 -
附錄圖 A.2 「正鯧」轉換後的數據之多變量常態分配檢定.....	- 159 -
附錄圖 A.3 「黃鰭鮪」轉換前原始數據之多變量常態分配檢定.....	- 159 -

附錄圖 A.4 「黃鰭鮪」轉換後的數據之多變量常態分配檢定- 160 -
附錄圖 A.5 「長鰭鮪」轉換前原始數據之多變量常態分配檢定- 160 -
附錄圖 A.7 「大目鮪」轉換前原始數據之多變量常態分配檢定- 161 -
附錄圖 A.8 「大目鮪」轉換後的數據之多變量常態分配檢定- 161 -
附錄圖 A.9 「黑鮪」轉換前原始數據之多變量常態分配檢定- 161 -
附錄圖 A.10 「黑鮪」轉換後的數據之多變量常態分配檢定- 162 -



表目錄

表 2.2.1 事實資料表和維度表的特性比較.....	- 11 -
表 2.2.2 關聯式資料庫系統和資料倉儲系統的比較.....	- 12 -
表 2.4.1 IMPROMPTU、POWERPLAY、VISUALIZER 的版本與功能比較[17].....	- 17 -
表 2.5.1 分群分析的步驟以及其方法.....	- 19 -
表 2.5.2 階層式分群分析的決定分群數之評量指標.....	- 22 -
表 2.5.3 線上學習挖掘的學習方法.....	- 24 -
表 2.6.1 挖掘序列模式的方法與步驟.....	- 26 -
表 2.6.2 顧客交易資料庫原始資料.....	- 27 -
表 2.6.3 顧客交易資料庫排序後的資料.....	- 27 -
表 2.6.4 由顧客交易資料庫中建立「大項目集合」以及其對應代碼.....	- 28 -
表 2.6.5 顧客交易資料庫的兩階段轉換.....	- 28 -
表 3.1.1 資料完整性.....	- 37 -
表 3.2.1 資料倉儲參數.....	- 42 -
表 3.2.2(A) 漁獲區重疊的狀況統計.....	- 53 -
表 3.2.2(B) 漁獲區分散的狀況統計.....	- 53 -
表 3.2.3 不同格式的中文船名.....	- 54 -
表 3.2.4 不同格式的時間欄位.....	- 54 -
表 3.2.5 建構事實資料表的步驟.....	- 55 -
表 3.3.1 魚貝苗資料倉儲各維度的資料階層關係.....	- 61 -
表 3.3.2 事實資料表.....	- 61 -
表 3.3.3 作業方式維度.....	- 61 -
表 3.3.4 魚貝苗種類維度.....	- 61 -
表 3.3.5 產地維度.....	- 62 -
表 3.3.6 時間維度.....	- 62 -
表 3.4.1 來源資料表的差異.....	- 63 -
表 3.4.2 水產加工製品資料倉儲之各維度的資料階層關係.....	- 64 -
表 3.4.3 事實資料表 1.....	- 65 -
表 3.4.4 事實資料表 2.....	- 65 -
表 3.4.5 時間維度 1.....	- 65 -
表 3.4.6 時間維度 2.....	- 65 -
表 3.4.7 水產生物種類維度.....	- 65 -
表 3.4.8 水產製品種類維度.....	- 66 -
表 3.4.9 地區維度.....	- 66 -
表 3.5.1 遠洋漁獲資料倉儲之各維度的資料階層關係.....	- 68 -
表 3.5.2 事實資料表.....	- 68 -
表 3.5.3 銷售地維度.....	- 68 -
表 3.5.4 漁獲生物種類維度.....	- 69 -
表 3.5.5 漁業作業種類維度.....	- 69 -
表 3.5.6 作業海域維度.....	- 69 -

表 3.5.7 時間維度.....	- 69 -
表 3.6.1 各資料倉儲的研究限制	- 70 -
表 4.3.1 「漁業資訊分享熱線」(FISH)與漁業資訊服務網漁業統計系統的比較.....	- 83 -
表 5.1.1 漁船和漁區所扮演的角色.....	- 86 -
表 5.2.1 漁獲量各分群組的上限值	- 89 -
表 5.2.2 資料轉換的型態.....	- 89 -
表 5.3.1 漁船軌跡 1 的「序列模式」結果.....	- 92 -
表 5.3.2 漁船軌跡 2 的「序列模式」結果.....	- 94 -
表 5.3.3 漁船軌跡 3 的「序列模式」結果.....	- 96 -
表 6.2.1 漁船噸位種類.....	- 101 -
表 6.2.2 漁船法長種類.....	- 101 -
表 6.2.3 漁船寬度種類.....	- 102 -
表 6.2.4 漁船深度種類.....	- 102 -
表 6.3.1 魚貨銷售情形分群之變異數分析表.....	- 104 -
表 6.3.2 魚貨銷售情形分群之 3 項衡量變數之平均值與命名	- 105 -
表 6.5.1 漁船屬性相關變數在魚貨銷售情形的 MANOVA 摘要表	- 109 -
表 6.5.2 產生魚貨銷售情形最佳的變因	- 110 -
表 7.3.1 分層隨機抽樣[14].....	- 115 -
表 7.3.2 長鰭鮪的樣本結構以及分層隨機抽樣後的結構.....	- 116 -
表 7.4.1 多變量常態分配檢定.....	- 117 -
表 7.4.2 正鰲月份因子和漁區因子組合下的此 354 個群體之「捕撈情形」差異分析表	- 118 -
表 7.4.3 正鰲月份和漁區兩因子組合下的此 354 個群體之「捕撈情形的 2 個衡量值」的檢定表.....	- 119 -
表 7.4.4(A) 正鰲捕獲情形具有顯著高值的漁區及月份	- 119 -
表 7.4.4(B) 正鰲之漁獲量具有顯著高值的「漁區-月份」組合	- 121 -
表 7.4.4(C) 正鰲之單位重量的平均價格具有顯著高值的「漁區-月份」組合.....	- 122 -
表 7.4.5 黃鰭鮪月份因子和漁區因子組合下的此 396 個群體之「捕撈情形」差異分析表	- 123 -
表 7.4.6 黃鰭鮪月份因子下的此 12 個群體之「捕撈情形的 2 個衡量值」的檢定表.....	- 123 -
表 7.4.7(A) 黃鰭鮪捕獲情形具有顯著高值的漁區及月份	- 124 -
表 7.4.7(B) 黃鰭鮪之漁獲量具有顯著高值的「漁區-月份」組合	- 126 -
表 7.4.7(C) 黃鰭鮪之單位重量的平均價格具有顯著高值的「漁區-月份」組合.....	- 126 -
表 7.4.8 長鰭鮪月份因子和漁區因子組合下的此 522 個群體之「捕撈情形」差異分析表	- 127 -
表 7.4.9 長鰭鮪月份和漁區兩因子組合下的此 522 個群體之「捕撈情形的 2 個衡量值」的檢定表.....	- 127 -
表 7.4.10(A) 長鰭鮪捕獲情形具有顯著高值的漁區及月份	- 128 -
表 7.4.10(B) 長鰭鮪之漁獲量具有顯著高值的「漁區-月份」組合	- 130 -
表 7.4.10(C) 長鰭鮪之漁獲量較佳的「漁區-月份」組合	- 130 -
表 7.4.11 大目鮪月份因子和漁區因子組合下的此 477 個群體之「捕撈情形」差異分析表.....	- 131 -
表 7.4.12 大目鮪月份因子下的此 12 個群體之「捕撈情形的 2 個衡量值」的檢定表.....	- 131 -
表 7.4.13(A) 大目鮪捕獲情形具有顯著高值的漁區及月份	- 132 -
表 7.4.13(B) 大目鮪之漁獲量具有顯著高值的「漁區-月份」組合	- 134 -
表 7.4.13(C) 大目鮪之單位重量的平均價格具有顯著高值的「漁區-月份」組合.....	- 135 -

表 7.4.14 黑鮪月份因子和漁區因子組合下的此 27 個群體之「捕撈情形」差異分析表.....	- 136 -
表 7.4.15 黑鮪月份和漁區兩因子組合下的此 27 個群體之「捕撈情形」的檢定表.....	- 137 -
表 7.4.16(A) 黑鮪捕獲情形具有顯著高值的月份及較佳情形的漁區	- 139 -
表 7.4.16(B) 黑鮪之漁獲量較佳的「漁區-月份」組合	- 139 -
表 7.4.16(C) 黑鮪之單位重量的平均價格較佳的「漁區-月份」組合	- 139 -
表 8.2.1 就組織面探討風險事件發生的錯誤鏈	- 142 -
表 8.2.2 針對可能發生的人為失誤之可行策略	- 143 -
附錄表 A.1.1.1 漁船軌跡事實表	- 146 -
附錄表 A.1.1.2 魚貨銷售紀錄表	- 146 -
附錄表 A.1.1.3 魚貨銷售紀錄表	- 146 -
附錄表 A.1.1.4 船公司或代理商資料表	- 147 -
附錄表 A.1.1.5 國外基地資料表	- 147 -
附錄表 A.1.1.6 漁船資料資料表	- 147 -
附錄表 A.1.1.7 銷售國別資料表	- 148 -
附錄表 A.1.1.8 作業海域資料表	- 148 -
附錄表 A.1.1.9 魚種資料表	- 148 -
附錄表 A.1.1.10 漁業作業種類資料表	- 149 -
附錄表 A.1.1.11 漁船明細資料表	- 149 -
附錄表 A.1.2.1 魚貝苗產量及價值資料表	- 150 -
附錄表 A.1.2.2 行政地區資料表	- 150 -
附錄表 A.1.2.3 魚貝苗種類資料表	- 151 -
附錄表 A.1.3.1 水產加工製品產量及價值資料表 1(魚產品別).....	- 152 -
附錄表 A.1.3.2 水產加工製品產量與產值資料表 2(月別).....	- 152 -
附錄表 A.1.3.3 水產生物種類資料表	- 152 -
附錄表 A.1.3.4 水產製品種類資料表	- 153 -
附錄表 A.1.3.5 行政地區資料表	- 153 -
附錄表 A.1.4.1 遠洋漁業生產量調查資料表	- 154 -
附錄表 A.1.4.2 區漁會資料表	- 154 -
附錄表 A.1.4.3 魚市場資料表	- 154 -
附錄表 A.1.4.4 漁獲生物種類資料表	- 155 -
附錄表 A.1.4.5 漁業作業種類資料表	- 155 -
附錄表 A.1.4.6 作業海域資料表	- 155 -
附錄表 A.2.1 事實資料表	- 156 -
附錄表 A.2.2 船主維度	- 156 -
附錄表 A.2.3 漁船公司維度	- 156 -
附錄表 A.2.4 國外基地維度	- 157 -
附錄表 A.2.5 作業海域維度	- 157 -
附錄表 A.2.6 漁船噸位維度	- 157 -
附錄表 A.2.7 漁船所屬地區維度	- 157 -
附錄表 A.2.8 漁業別維度	- 157 -

附錄表 A.2.9 漁獲區維度- 158 -
附錄表 A.2.10 魚種類維度- 158 -
附錄表 A.2.11 銷售國別維度- 158 -





第一章 緒論

近來，由於資料庫與資料倉儲技術的成熟與價格下降，這使得透過網路將各地交易資料、行政管理資料匯集於特定的資料庫與資料倉儲裡進行增值應用，已普遍成為各行業整合業務情報的必經之路。如何將儲存許久的歷史資料，經過加值的處理，成為決策的輔助工具，已成為必須探討的課題。本章主要的目的是要說明論文的著眼點。本章共分四節，第 1.1 節說明研究動機，第 1.2 節說明問題界定，第 1.3 節說明研究目的，第 1.4 節研究方法，第 1.5 節論文架構。

1.1 研究動機

漁業產業一向是人類生活之基礎產業，以我國 91 年漁業生產為例，總產量於近年來已超過 140 萬公噸，總產值約為新台幣 928 億元。為了掌握這重要的漁業產銷情報，漁業管理單位除了每個月提供國內魚市場魚價行情的檔案下載服務，每年年底針對漁業生產與現況的年度統計結果，製作圖 1.1.1 的中華民國台閩地區漁業統計年報[3]以提供各決策單位參考。單行本發行的漁業統計年報有三大缺點，即不符合成本與環保觀念、缺乏彈性以及缺乏時效性。從成本與環保觀點而言，統計年報的大量印製以及運送各單位，消耗許多資金以及大量紙張。從彈性需求的觀點而言，年報內容無法以多維度的觀點呈現使用者所需要的資訊。從時效性需求的觀點而言，每一年的漁業生產訊息，如遠洋漁獲量、遠洋漁獲價格，必須等到年底才可得知，無法從該年度立即獲得即時統計的數據。

對遠洋漁業的現況來說，我國遠洋漁業近年來漁獲量已超過 82 萬公噸，佔我國漁業總產量之 49% 以上，作業漁場遍布三大洋之公海漁場，為我國十分重要的產業。由於世界各沿海國實施 200 浬經濟海域及公海資源共管之趨勢，再加上漁業捕獲技術的進步卻缺乏有效的管理規範，出現海洋資源逐漸衰竭的危機，遠洋漁業因此遭受重大衝擊。聯合國糧農組織 (Food and Agriculture Organization of the United Nations, FAO) 意識到目前漁業資源的危機，於是相繼制訂如「責任漁業行為準則」、「養護與管理跨界和高迴游性魚種」、「履行 1982 年聯合國海洋法公約有關養護管理跨界魚種及高度迴游魚種條款協定」(簡稱 UNIA) 等各種管理規範，要求作業漁船之船籍國，負起在公海漁業作業活動之責任[1][13]。

我國政府為了因應國際「責任制漁業」之相關漁業管理規範，積極推動民間魷釣漁船和鱒釣漁船加入漁船監控系統 (Vessel Monitoring System, VMS)，以即時掌控漁船動態的船位資料、漁獲量回報資訊[11]。並期望進一步運用漁船監控系統中各民間漁船回報的漁獲量，從事公海漁業資源挖掘，拓展遠洋漁場。由於漁業監控組織的建立，需要靠制度來維護；而制度本身就是一套協調人或物的賽局規則 (Rules) 或協定 (Protocol)，執行時再以契約方式進行。就本制度的設計而言，屬於「對人」的「交換組織」裡。也因此應用於漁業監控系統的制度之設計重點在於解決「委託人」與「代理人」之間的「非對稱資訊」以便管理並且能滿足組織成員需求以及管理者需求的分散式管理[13][7]。獲取漁船定位資訊以及漁獲訊息的漁業管理單位，屬於委託人；而提供訊息的各民間業者則屬於代理人。在非對稱資訊下可能有隱藏性資訊 (Hidden Information) 或是隱藏性動作 (Hidden Action)。例如代理人可能因隱藏漁獲量或是航速的資訊而使得委託人做出錯誤的判斷，這導致逆選擇 (Adverse Selection) 問題的發生。雖然漁業管理單位有制訂獎勵標準[4]以及限制制度[8]，使得參與此組織的所有利益團體都願意在規則下樂於加入或至少不反對此合約。但是缺少了「誘因匹配機制」，無法使得所有參與此組織 (即此賽局) 的利益團體都會誠實的依規則來回報漁獲資訊。

因此根據上述問題與發展背景，本研究建議運用線上分析處理 (On Line Analytical Processing, OLAP) 技術，建構漁業資訊分享熱線，可解決上述的問題，如圖 1.1.2 所示本研究動機的脈絡圖。另外在建立資料倉儲後，歷史資料的分析與增值應用更形重要，如何進一步透過分析資料倉儲的歷史資料，針對遠洋漁業的民間業者或是行政管理單位來進行各項重

要生產、銷售決策，更是本研究的重點。例如，本研究將會運用魚貨銷售記錄資料表記載該漁船在國外魚貨銷售的數量，來推估該漁船在當時航行軌跡的所捕獲得漁獲量；進一步利用該結果進行高漁獲量漁船的航海軌跡模式挖掘。



圖 1.1.1 中華民國台閩地區漁業統計

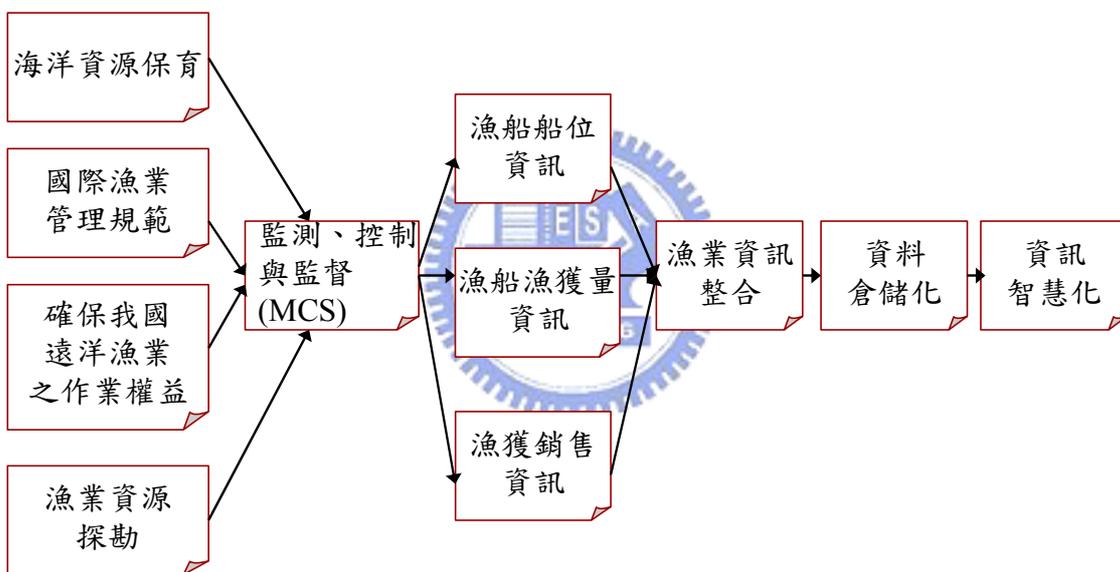


圖 1.1.2 研究背景與動機脈絡圖

1.2 問題界定

本研究之問題界定在解決漁業管理單位與民間漁業生產單位的資訊需求以及面臨的問題，分別是政治上的資訊管理需求和經濟上的資訊管理需求。就政治上的需求而言，針對漁船的船位資料進行管理，目的是因應國際「責任制漁業」之相關漁業管理規範和掌控漁船在公海上的各作業活動，例如禁漁區的管制、漁船海上救難、走私防範等管理。

就經濟上的資訊管理需求而言，針對遠洋漁獲量及銷售狀況、水產加工製品的數量及生產方式、魚貝苗養殖作管理。原因有下列三點：第一、藉由正確的漁獲量回報進行海洋資源管制，遵守國際漁業管理規範以確保我國遠洋漁船的作業權益。第二、提供即時線上漁業產銷情報，資訊分享更有效率。各單位的使用者透過網際網路經過認證與不同等級權限的授與，可利用七種常見的線上分析運算方式進行多維度分析(Multidimensional Analysis)，即上捲(Roll-up)、下挖(Drill-down)、切片/切丁(Slice-Dice)、轉軸(Pivoting)、排序(Sorting)、篩選(Selecting)及計算(Computing)，以獲得本身所需要的資訊。第三、運用資料探勘技術及多變量分析技術，歸納出知識以提供政府和民間單位。

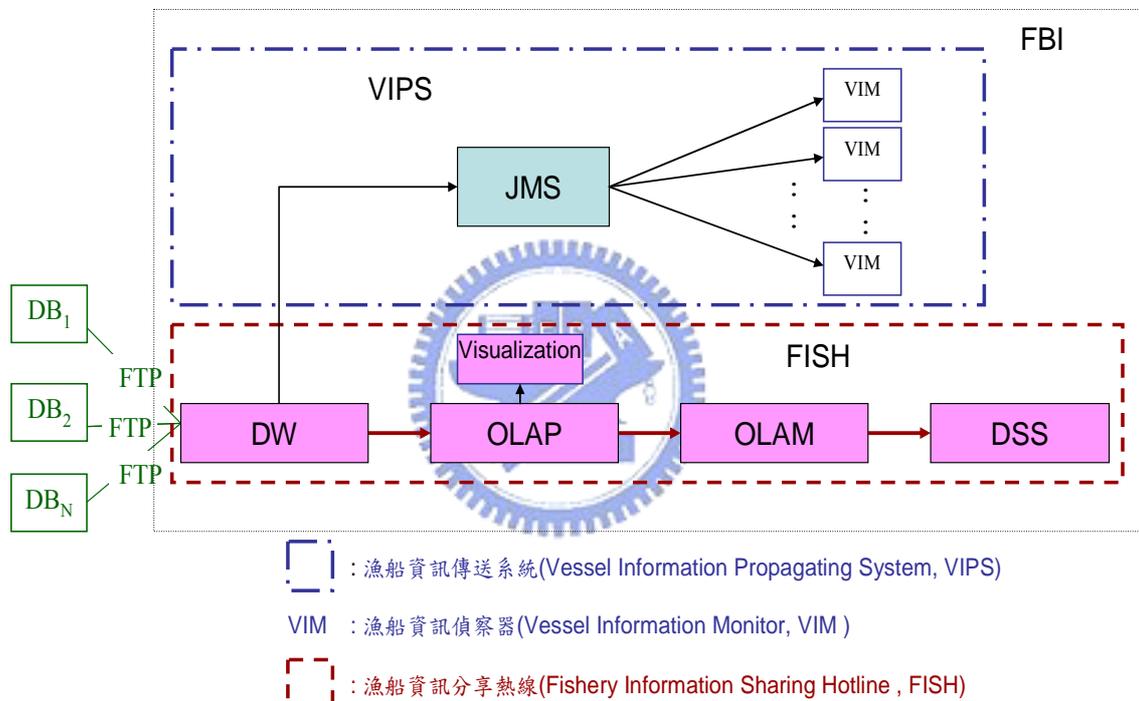


圖 1.2.1 漁業業務情報網 FBI 的架構圖

本研究欲建構「漁業資訊分享熱線」(Fishery Information Sharing Hotline, 以下簡稱 FISH)。由圖 1.2.1 可知 FISH 是為「漁業業務情報網」(Fishery Business Intelligence, 以下簡稱 FBI)的成員之一。FBI 是一個提供使用者透過網際網路，獲取漁業情報的資訊平台。在 FBI 裡包含兩大系統，為「漁業資訊分享熱線」系統與「漁船資訊傳送系統」(Vessel Information Propagating System, 以下簡稱 VIPS) 系統。就功能面來看，FISH 系統是負責進行異質來源資訊的整合與呈現，包括利用超方體提供線上分析處理與數值資訊的視覺化；VIPS 系統是透過以爪挖訊息服務(Java Message Service, JMS)技術，傳送 FISH 中資料倉儲的各船位資訊給漁船監控系統的各成員，並藉由漁船資訊偵察器(Vessel Information Monitor, VIM)將漁船航行軌跡呈現在電子海圖中。因此 FISH 是 VIPS 的資訊源，而 VIPS 是 FISH 的加值應用，兩項系統在 FBI 裡扮演的角色彼此相輔相成。本研究現階段建立四個主題的漁業資料倉儲，以及四個主題的線上分析處理，並且以遠洋漁業在國外銷售情況以及漁船航行軌跡為研究對象，進一步進行海洋資源的挖掘，也就是包含的範圍如圖 1.2.1 DW, OLAP, 和 OLAM 的 3 個建置階段。

1.3 研究目的

根據前述問題界定以及經濟及政治管理需求之分析，本研究提出建構「漁業資訊分享熱線」(以下簡稱 FISH)的想法。圖 1.3.1 為漁業資訊分享熱線的現階段架構與研究主題，包含整合異質性來源資料庫、資料倉儲(Data Warehouse,DW)、網際網路、線上分析處理、線上分析挖掘(On-Line Analysis Mining, OLAM)及商業智慧(Business Intelligence)推導之六大資訊技術的整合。主要可達到下列目的：

1. 將異質性來源資料予以淨化與格式轉換處理，得到整合性的資訊。
2. 作為「漁船資訊傳送系統」的資料傳輸端。
3. 提供使用者透過網際網路，即可獲得個人所需的各種資訊。
4. 進一步將資料透過各種資料探勘的技巧，將訊息進一步處理為有用的知識。

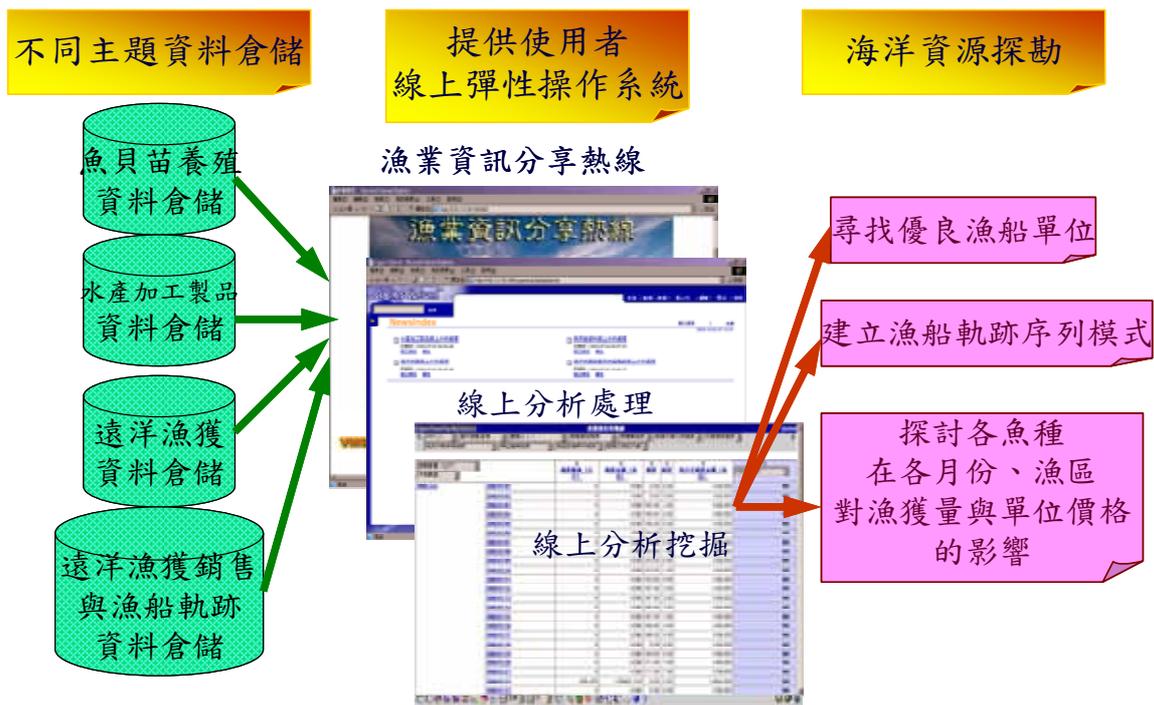


圖 1.3.1 漁業資訊分享熱線的現階段架構與研究主題

對第一個目標「將異質性來源資料予以淨化與格式轉換處理，得到整合性的資訊」而言，透過資料淨化與轉換技術以整合異質性來源的資料，建構四大主題的資料倉儲，將大量零散的資訊作有利的利用與整合。

對第二個目標「作為『漁船資訊傳送系統』的資料傳輸端」而言，以資料倉儲為資料傳輸端，透過爪挖訊息服務將資料傳送至各漁船資訊偵察器，使漁船船位訊息與漁獲量訊息進一步透過視覺化技術，有效呈現給使用者所需的訊息。

對第三個目標「提供使用者透過網際網路，即可獲得個人所需的各種資訊」而言，屬於 FISH 建構的「線上分析處理」階段。亦是以資料倉儲為資料來源，提供漁業管理單位或是民間業者透過網際網路進入 FISH，並依照個人需求下載或查詢所需的漁業資訊。

對第四個目標「進一步將資料透過各種資料探勘的技巧，將訊息進一步處理為有用的知識」而言，分析者經過線上分析處理的過程，得到初步直覺上的判斷後，在經過各種資料探勘的技巧，如關聯性法則、分群分析、歸類分析、序列模式分析、多變量分析、時間序列分析、回歸分析等其他統計方法，能夠由大量的訊息裡得到潛藏的漁業知識，也就是漁業業務情報。

本研究將於第四章第一節對漁業資訊分享熱線的系統架構與建構程序作進一步說明。

1.4 研究方法

建構漁業資訊分享熱線的研究方法一共分成六個步驟，如圖 1.4.1 所示，以下是針對六個步驟以及在各步驟中所需介入之相關技術與研究作說明：

1. 問題分析與界定：此階段首先需藉由訪談漁業管理相關單位以瞭解國內漁業產業的背景、現狀與面臨的問題。由於實作系統的使用者包含政府單位以及民間各級漁業生產者，因此經過數次的訪談和溝通，釐清資料的意義和資料庫的架構，涵蓋的資料分佈於各大系統中包括「漁船監控系統」、「漁船資料倉儲」、「漁船管理系統」、「漁業統計資料」，目的是：
 - A. 找出不同使用者對於資料倉儲的功能需求。
 - B. 瞭解各大系統的實際運作及功能以進行探討。
 - C. 由各大系統中挑選所需的資料作收集的工作，以利於後續研究工作的進行。
2. 系統架構之設計：此階段工作主要的目標是了解漁業相關資訊以及探討建構漁業資訊分享熱線的相關領域，於是探討下列領域之文獻：
 - A. 漁業供應鏈。
 - B. 資料倉儲技術。
 - C. 資料擷轉。
 - D. 企業智慧系統架構。
 - E. 線上分析處理。
 - F. 線上分析挖掘。
 - G. 序列模式挖掘。
3. 資料倉儲實作：找出各來源資料，並根據各來源資料的性質、綱要以及三階正規化等技術來建構資料倉儲系統。
4. 線上分析處理實作：依據資料流、資料儲存的模式、及預定解決的方案來設計與實作。此為漁業資訊分享熱線的基礎建設環境。
5. 線上挖掘技術實作：以資料倉儲為基礎發展線上分析挖掘的技術以導入漁業產銷情報。
6. 多變量技術實作：以資料倉儲為基礎發展多變量技術以導入漁業產銷情報。

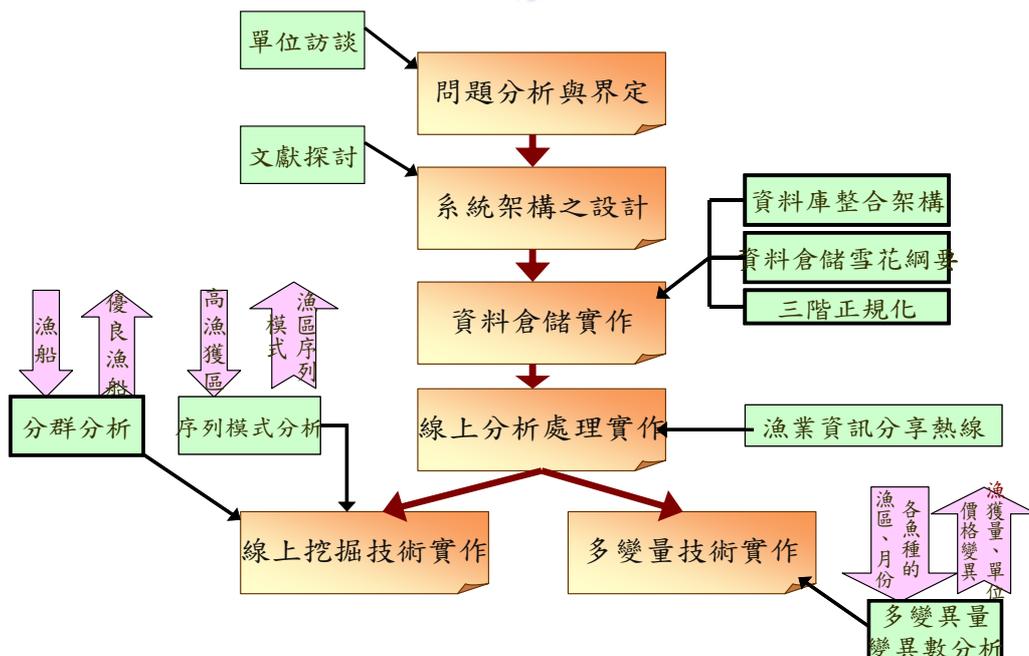


圖 1.4.1 研究方法與步驟

1.5 論文架構

本研究架構一共包含八個章節，如圖 1.5.1 所示：

- 第一章、『緒論』，說明本論文之研究動機與背景、研究主題、研究方法與步驟。
- 第二章、『文獻回顧』，探討本論文中所需參考的觀念。
- 第三章、『漁業資訊分享熱線的資料倉儲設計與實作』，說明「魚貝苗養殖資料倉儲」、「水產加工製品資料倉儲」、「遠洋漁獲資料倉儲」及「遠洋魚貨銷售與漁船軌跡資料倉儲」四個資料倉儲系統的原始交易資料，三階正規化的雪花綱要以及資料轉換服務的設計內容及操作方式。
- 第四章、『漁業資訊分享熱線的線上分析處理』，設計多維度之超方體架構，並使用軟體實作線上分析處理系統。
- 第五章、『漁業資訊分享熱線的資料探勘－建構高漁獲量漁船的軌跡序列模式』，藉由資料轉換技術及序列模式挖掘軟體，挖掘出漁船的軌跡序列模式。
- 第六章、『漁業資訊分享熱線的資料探勘－尋找優良漁船單位』，藉由資料轉換技術、線上挖掘分析及分群分析，尋找高漁獲量及高單位重量價格的優良漁船單位。
- 第七章、『漁業資訊分享熱線的資料探勘－探討各魚種在不同漁區、月份下與漁獲量以及單位重量價格的關係』，藉由事實資料表之查詢及多變量變異數分析，探討漁區、月份對於漁獲量及單位重量價格的影響。
- 第八章、『結論與未來研究方向』，結論、分析及未來研究方向。

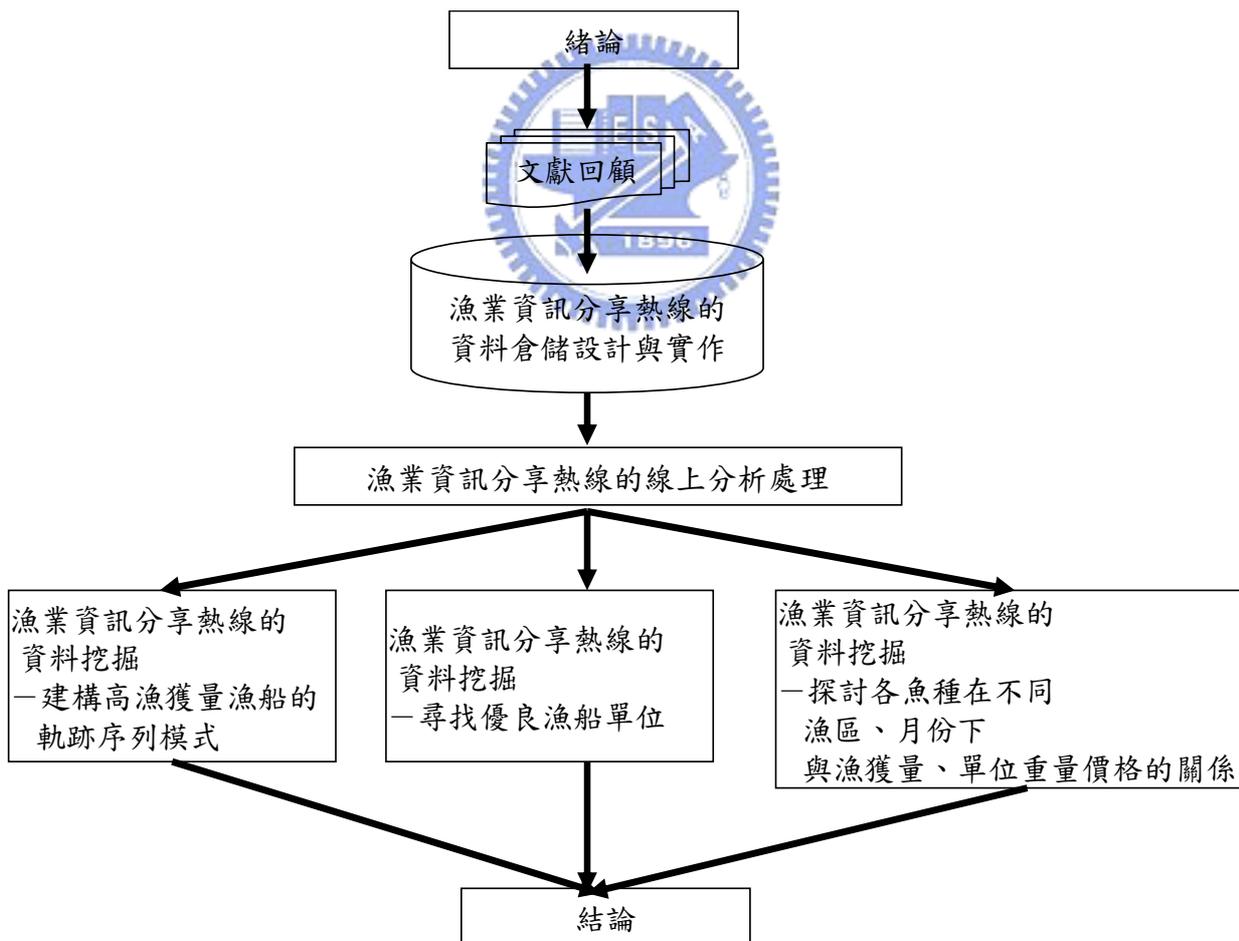


圖 1.5.1 論文架構

第二章 文獻回顧

本章主要目的是建立漁業資訊分享熱線的架構，因此必須了解漁業供應鏈的特性，探討資料倉儲技術及其相關技術、線上分析處理技術、線上分析挖掘技術、資料探勘技術等技術以利本研究之應用。本章共分六節，第 2.1 節介紹漁業供應鏈的體系，第 2.2 節主要說明資料倉儲技術的相關定義，第 2.3 節主要說明資料擷轉的運作與架構，第 2.4 節主要介紹 COGNOS 企業智慧系統架構，第 2.5 節主要說明線上分析理論，第 2.6 節說明序列模式挖掘。

2.1 漁業供應鏈

漁業產業一向是人類生活之基礎產業，台灣本身地理環境特性是四面環海，擁有豐富的海洋資源，漁業產業更行重要。以我國 91 年漁業生產為例，總產量於近年來已超過 140 萬公噸，總產值約為新台幣 928 億元，漁業生產者約 13 萬戶，從業人數約 34 萬人，占農業（包含農作物、林產、漁業、畜牧業）總值的 26.41%[2]。本節將對漁業供應鏈之架構做簡述，介紹魚貨產銷之流程。以供應鏈中生產者角色為劃分，本節將在第 2.6.1 小節介紹漁撈漁業，第 2.6.2 小節介紹養殖漁業。

漁產品透過漁業供應鏈進行產銷，整個漁業產品生產體系的關係圖如圖 2.1.1[5]。對漁業供應鏈而言，在政府單位的輔導、協助及管理下，分成四個主要單位：「漁業生產者」、「民間企業」、「經銷方式」及「消費者」。就「漁業生產者」而言，為漁民團體組成，以漁會為例，在縣市政府的監督下，結合其生產技術的優勢成立各種不同團體的產銷班，透過產銷班的经营來辦理該水產品加工、冷藏、調配、運銷。漁會也可以從事生產地魚市場與消費地魚市場的经营，以及協助漁用物資進口、加工、製造、配售、漁船修造及會員生活用品供銷。故漁會可以協助最基層漁民的生產品有組織地與消費者做接觸。漁船公司也類似同樣的情形，尤其在進行遠洋漁獲作業，根據其以往的經驗來指揮、協助出海的漁船，並且規劃其在國外市場進行魚貨銷售。就「民間企業」而言，將漁撈漁半產品以及養殖漁半產品經過加工、冷藏、分類、包裝等步驟，以進行下一步的運銷，屬於漁業生產者在原料上的供應商。就「經銷方式」而言，可以由生產地魚市場（如高雄、宜蘭）直接透過物流業者送到消費者手中，飯店或餐廳多屬於此種運銷方式。另外消費者也可以透過各縣市的農漁會、消費地魚市場（如台北、台南）、量販店業者、傳統市場零售商、農漁電子商務網站等不同的經銷據點及經銷方式，購買所需的魚貨。

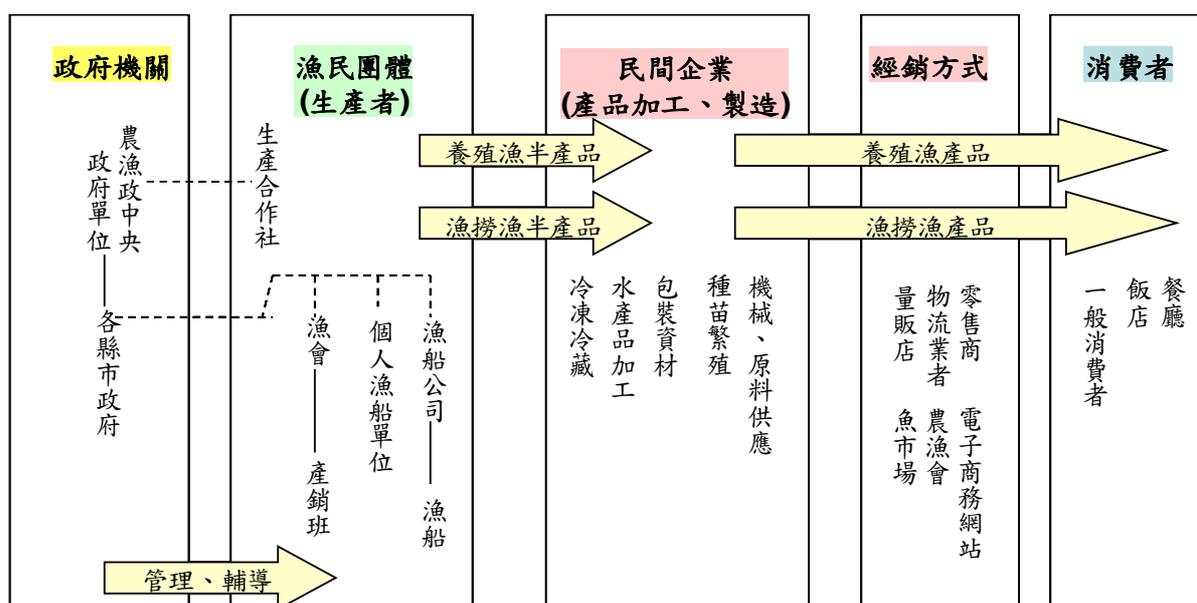


圖 2.1.1 漁業產品運銷體系的關係圖

2.1.1 漁撈漁業

漁撈漁業包括了遠洋漁業、近海漁業以及沿岸漁業三部分。對遠洋漁業而言，該作業活動屬於我國 200 浬經濟水域之外的公海，或是與其他漁業國家合作而得以在該合作國家的經濟海域內捕魚。遠洋漁業由於使用國際海洋資源，因此必須遵循各種國際規範。自民國 77 年起，遠洋漁業漁獲量以占總漁獲量已高達約 50%，其產值也超過總漁業價值的 50%，為我國最重要的漁業。遠洋漁業主要的作業方式，包括鮪魚延繩釣、鰹鮪圍網、拖網、魷釣及秋刀魚棒受網等。遠洋漁業作業漁場遍及三大洋，大部分之鮪釣及鰹鮪圍網漁業均在世界各地以國外港口設有基地進行補給、整修及卸魚轉載，目前經核准之國外基地港口有 67 個。

對近海漁業而言，作業海域的範圍在 12 浬至 200 浬經濟海域內從事漁撈作業。主要漁場分佈在在東海、台灣海峽、南中國海、巴士海峽及台灣東部海域等。主要漁業為拖網、巾著網、鯖魚參圍網、刺網、延繩釣、焚寄網等。由於近年來近海水域漁業資源逐漸枯竭，加上漁船大型化，漁撈技術及漁航儀器大幅改進，近海鮪延繩釣漁船作業漁區已逐漸轉向遠洋。民國 80 年至 84 年間，政府為紓解勞力之不足，減輕近海漁業資源過漁之壓力，及防止漁船從事走私等不法行為，乃編列經費收購老舊漁船，進行減船計畫。

對沿岸漁業而言，作業海域的範圍在係指在我國領海（12 浬）內從事漁撈作業，近年來產量約維持在 4 萬公噸左右，主要作業方式包括刺網、定置網、地曳網、魚苗捕撈業及其他釣具漁業等。一般來說，沿岸海域屬於由於生態環境有較豐富的浮游植物，因此屬於海洋漁業生產力較高之區域。然而目前沿岸漁業面臨船筏數目過多，導致漁業資源被過度利用，以及沿岸海域污染日趨嚴重，資源遭受破壞的問題。政府一方面陸續於沿岸海域設置漁業資源保育區、人工魚礁區及保護礁區、從事魚、貝、介類種苗大量放流，以積極改善漁場、復育漁業資源。另一方面，輔導漁民從事娛樂漁業，如海釣、賞鯨豚等，以多元化經營漁業。

2.1.2 養殖漁業

我國養殖漁業主要分為淡水養殖、鹹水養殖及海面養殖三大類。由於水土資源有限及生態環境的改變，導致生產成本之提高，使我國養殖漁業逐漸喪失競爭力，然而水產養殖在台灣確有其存在條件。基於下列三點原因：第一點是水產養殖已有三百餘年之歷史技術十分純熟，第二點則由於我國地處亞熱帶，適合種苗培育，第三點是水產種苗生產技術之研發與改進，使得養殖技術之不斷提昇，使得可養殖種類已近 100 種。在淡水養殖方面主要種類包括鰻魚、吳郭魚、鯉魚、淡水長臂大蝦、蜆等；鹹水養殖則包括鯛類、虱目魚、草蝦、斑節蝦等；淺海養殖則以牡蠣、文蛤、九孔為主，而箱網養殖則以嘉臘、石斑、海鱺、紅甘魚參等為主。養殖漁業主要是以吳郭魚和虱目魚為大宗，總年產量約 34 萬噸。政府希望鼓勵漁民以生產、生活、生態三生兼顧的休閒漁業為主軸，並致力產品品質提昇捨棄追求高產量策略。執行方式是重視環境生態保育，規劃出觀光休閒與體驗漁業之親子教育遊憩地帶，並且建立有品牌之水產品以提供消費者高品質的養殖漁產品。

2.2 資料倉儲技術

資料倉儲系統為整合性資料的儲存體，是一種能適當的組合及管理不同來源的資料的技術，使這些資料能提供農產運銷情報中的統計分析及決策輔助，可供使用者查詢擷取、篩選、整合相關資訊。相對於傳統系統的被動式查詢--當需要查詢時才做運算；資料倉儲使用主動式查詢--當來源資料更動時，即做出相對應的反應。從資訊技術的角度看，資料倉儲的目標是在組織中、在正確的時間，將資料交給正確的人。

2.2.1 資料倉儲的特性與架構

在 90 年代初期，當主從架構技術發展之後，許多的大型企業中既有的資料庫科技架構在一時無法更新的情況下，這時出現了資料倉儲能夠兼顧主從式架構線上查詢的即時功能及現有舊系統資料設備的運作，資料倉儲便隨即受到重視，並進而由學術性的理論衍生到廣而被接受的技術。

資料倉儲之父之稱的 W.H. Inmon 在 1992 年首次對資料倉儲做了一個定義[34]：資料倉儲具有主題性(Subject-Oriented)、整合性(Integration)、時變性(Time Variation)與永存性(Non Volatilization) 四大特色。分述如下：

1. 主題導向：

資料倉儲的建立是為了分析特定的主題，而和主題不相關的資料就必須加以去除，例如：如果企業針對以花卉銷售為分析主題的資料倉儲，那麼承銷店家的營業登記證號，聯絡電話及家庭成員等等不相關的資料便要去除。

2. 整合性：

資料倉儲內的資料必須具有相當整合性，在資料的轉換過程中，要讓欄位名稱、測量變數、編碼方式、日期時間等等主題屬性具有一致性的格式。

3. 時間變化性：

日常的作業系統每天都有新資料增加，為維持資料倉儲的可用性，必須在某些特定的時間點到作業系統中擷取新資料，稱之為「快照 (Snapshot)」。以台北花卉市場拍賣資料為例，所儲存的資料是依時間序列累計個時期的拍賣資料而來，也因此資料倉儲中，可以有不同時間資料提供做比較分析。

4. 非揮發性：

資料倉儲僅允許資料最初的載入與讀取，資料一旦存入資料倉儲即被保留，不因新資料的加入而被取代或改變。

實際上的根據需求不同有兩種型態的資料倉儲系統[15]：分散式架構和集中式架構。以分散式架構來說，也就是大家耳熟能詳的資料超市(Data Marts)，資料超市是一個單一目標的資料倉儲系統，可以在短期間內建置而成。相反的集中式資料倉儲是最複雜的一種，它將各個部門的資料經由多重的轉換連結在一起，藉以建立一個具有一致性的加值性的系統。如果以資料的集合而言，資料超市為資料倉儲的一個子集，一般僅提供部份資訊，給某一群使用者或某一部門使用，以符合不同使用者的不同需求。一般來說資料倉儲的建置與管理往往非常昂貴且耗時；建立的方法通常是從上到下。而資料超市通常只為了特定的決策支援應用程式或使用群組而設計，採由下而上的建置方式。

由圖 2.2.1 說明資料倉儲架構以及其後續應用，一共分成四層，從圖中可以看出線上交易處理(On-Line Transaction Processing, OLTP)與框線內資料倉儲系統的互動關係，此兩者的關係 2.2.3 小節會有更詳細的說明。由下到上的開始為資料來源(Data Source)，這代表資料來自不同的地方，有 IBM DB2 資料庫，MS SQL sever 資料庫，文字檔等異質來源資料。進入底層前，必須先將資料過濾和轉換，此階段是屬於資料倉儲對於收集資料的前置作業，而根據經驗，這個步驟會花費建置整個資料倉儲 80%的時間。底層代表資料倉儲本身，也包括各部門或分析單一目標的資料超市(Data Market)。第二層是線上分析處理(On-line

Analytic Processing, OLAP)層，是資料初步加值的步驟，將建立的超方體透過線上分析處理伺服器，提供使用者自行對多維度的資料操作。第一層後端加值應用的步驟，所採取的方式可能為客戶關係管理(Customer Relation Management, CRM)，決策支援系統(Decision Support System, DSS)和線上分析挖掘 (On-Line Analytical Mining, OLAM)。

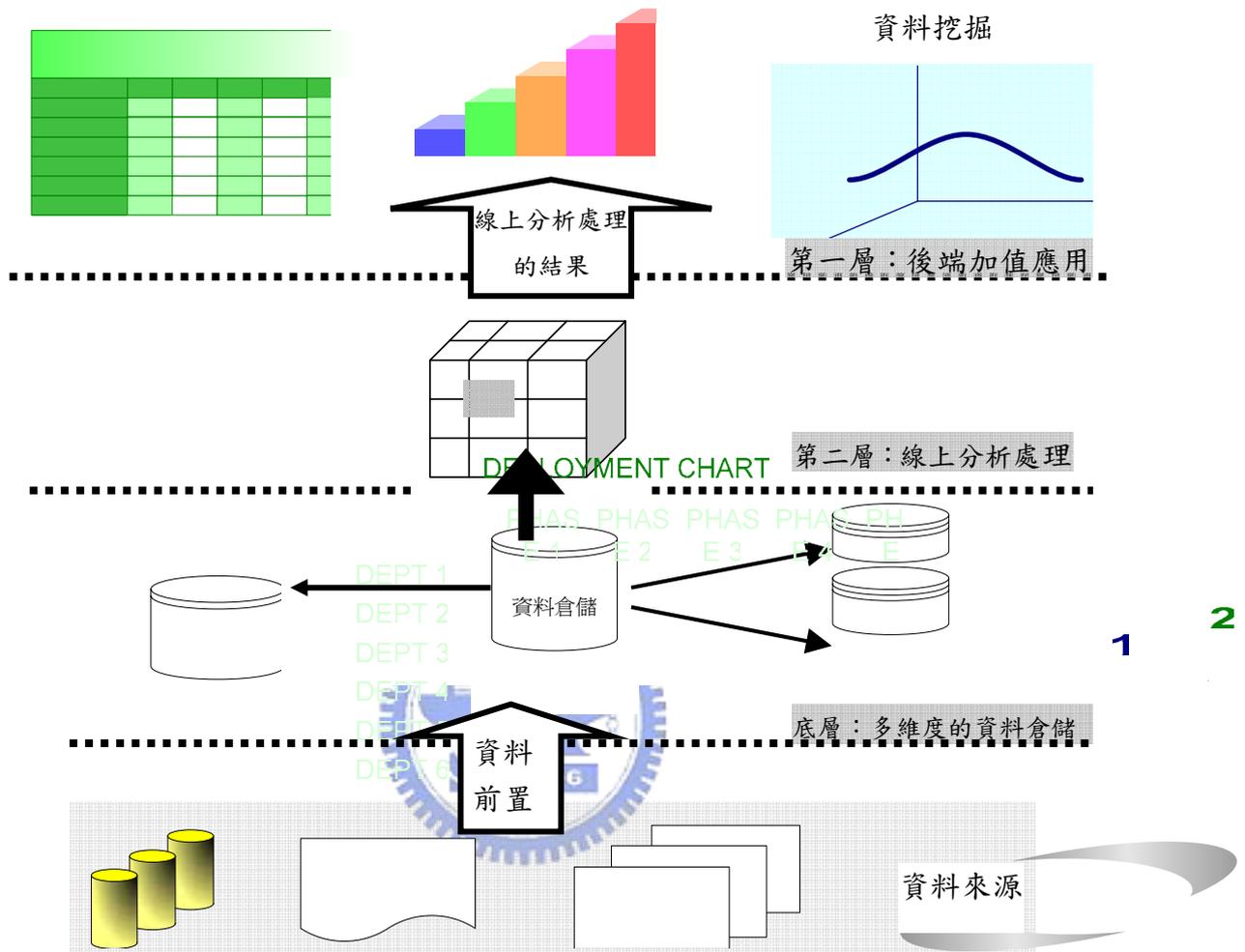


圖 2.2.1 資料倉儲系統架構與其相關應用

線上分析處理就是一種快速擷取多維資訊或超方體資料分析的工具，通常有多維度線上分析處理(Multi-dimensional OLAP, MOLAP)，關聯線上分析處理(Relational OLAP, ROLAP)及混合線上處理(Hybrid OLAP, HOLAP)三種方式[16][19]。決策支援系統必須先建立業務模式(Business Model)，配合使用者的 Know-How，在將資料倉儲的資料分析和預測作為未來決策的參考。線上分析挖掘主要是應用資料庫技術和數學統計工具，從大量的資料中歸納出事先未知的有用知識過程。

最後為中介資料(Meta Data)，是專門管理資料的資料，它記錄了資料倉儲中資料的原先存放的位置、資料格式和涵義等資訊。它所主要扮演的功用有三種：

- 一.引導分析工作找出資料倉儲的內容；
- 二.資料由作業環境轉換到資料倉儲整個資料的說明；
- 三.是作為分析資料間演算的引導。

2.2.2 資料倉儲的維度模型

資料倉儲可以藉由『維度模型』(Dimensions Model)來建立其綱要，以下將分別介紹三種不同的綱要架構：1.星狀綱要(Star Schema)、2.雪花綱要(Snowflake Schema)與 3.星座綱要

(Fact constellation Schema)[18][20]。

1. 星狀綱要：

此種架構是最常見的的模型，如圖 2.2.2 所示。中間的表為事實資料表（簡稱事實表，Fact Table），存放著我們所需要的真實資料，其內容包括我們所要分析的衡量值 (Measure) 以及連接各維度的外鍵屬性。例如某一產品的生產量及產品的銷售。Fact 表格中的資料內容是無法修改的，只能新增資料到表格中而已，是資料倉儲中存放資料最多的資料表。環繞其週遭儲存資料較少的資料表皆為維度表(Dimension Table)，此表格的資料是用來參考事實資料表格資料的，例如產品、客戶名稱、住址及供應商....等，透過某些事件的需求將一些煩雜敘述分割成幾個小部份，例如某個時點的銷售量等，應盡量藉由這些表格的設計減少對事實資料表格的讀取將查詢最佳化。事實資料表和維度表的特性比較如表 2.1 所示。

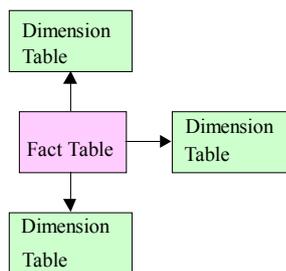


圖 2.2.2 星狀綱要

表 2.2.1 事實資料表和維度表的特性比較

事實資料表	<ol style="list-style-type: none"> 1. 大量的資料列，可能達上千萬、甚至上億筆歷史資料。 2. 要都是數值資料，只有少數的文字資料。 3. 多和維度表連接的外鍵 (Foreign Key)。 4. 靜態資料，不能修改。
維度表	<ol style="list-style-type: none"> 1. 其資料遠少於事實資料表，可能只有上百或上千個資料而已。 2. 大都為文字屬性資料。 3. 欄位會較多，其是用來管理維度的階層。 4. 只有一個主鍵 (Primary Key)。 5. 動態資料，資料可以修改。

2. 雪花綱要：

雪花綱要和星狀綱要的事實資料表和維度表觀念相同，主要差異在於對某些未做正規化的維度表，更進一步做三階正規化以減少重複性的資料儲存空間，並且在同一維度內建立階層性架構(Concept Hierarchy)。如圖 2.2.3 所示，本論文多是採用此架構去建構資料倉儲系統。

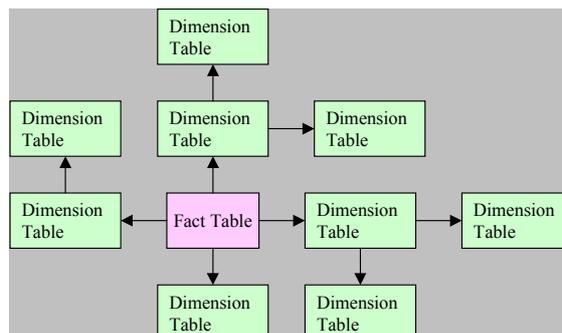


圖 2.2.3 雪花綱要(Snowflake schema)

3. 星座綱要：

又可稱為銀河綱要(Galaxy Schema)，是一種較為複雜的應用，由兩個以上的事實資料表共同去參考某些維度表，如圖 2.2.4 所示。本研究的也有應用此綱要建立資料倉儲系統。

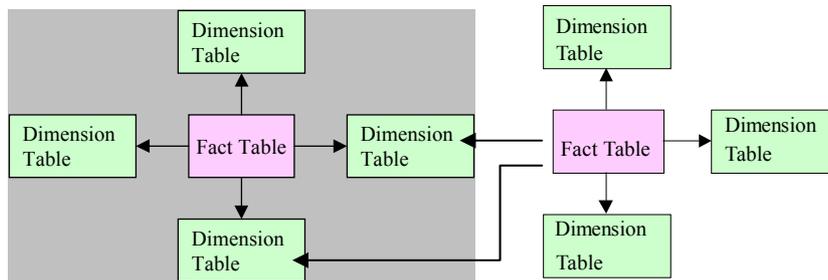


圖 2.2.4 星座綱要(Fact constellation schema)

2.2.3 資料倉儲系統與關聯式資料庫系統的比較

資料倉儲系統不同於關聯式資料庫系統(Relational Database Management System, RDBMS)、線上交易處理系統(以下簡稱 OLTP)。過去關聯式資料庫系統的組織架構設計是為了支援線上交易處理，因此系統設計最重要的目的在資料輸入和更新，例如訂單的輸入、工廠存貨的紀錄等作業。由此可知，線上交易處理系統，需具備較快速的交易處理需求。

資料倉儲通常則是為了分析資料和產生報表等目的，並且常被設計成唯讀式的系統。OLTP 系統所處理的資料類型通常為企業中以日為單位所處理的資料，但是資料倉儲所含的則是用來分析企業運作狀況的資料，其中可能包含了月的彙總資料或年的彙總資料；OLTP 系統所得到的資料，多具高揮發性和即時的特色，因為輸入的可能為不完整甚至是無法判讀的資料，而資料倉儲中的歷史資料則為已更正過的非揮發性資料。

再者這兩種系統的差異甚大，所以 OLTP 和資料倉儲系統使用的是完全不同的資料模型，運作速度與彈性是核心資料庫之設計關鍵。在 OLTP 系統中，由於複雜的資料更新作業無時不在進行，所以 OLTP 系統的資料庫多半已經高度正規化，而且通常使用實體關係圖(Entity-Relationship Model)資料模型來設計[29]，只是在蒐集資料後，無法將資料作多向廣度面的運用。資料倉儲系統為了簡化並提高使用者存取資料的效能，大大降低了資料表使用的數目，並且在其中充斥著多餘的資料。不像 OLTP 系統，僅專注於目前的瞬間時點，資料倉儲必須提供任何時刻，全面綜覽資料的能力，因此資料倉儲系統多半不使用實體關係資料模型，而改採多維資料庫的設計，故可以提供比傳統關聯式資料庫快的及時查詢效能。

基於上述幾點以及表 2.2.2 的比較所示，可看出關聯式資料庫系統和資料倉儲系統差異性，但是資料倉儲的出現並非取代傳統關聯式資料庫的地位，而是在其基礎上進行資料的加值處理及應用，可由圖 2.2.1 明白看出資料倉儲系統和關聯式資料庫系統的互動關係。透過資料倉儲將企業內部和外部的異質資料庫加以整合，便可以將企業間零散、無法串聯和分析的歷史資料，賦予再利用的價值，可利用線上分析處理加以分析查詢或製作報表，或是利用資料探勘進行分析和決策輔助，因此資料倉儲和關聯式資料庫系統並非對立，而是相輔相成的。

表 2.2.2 關聯式資料庫系統和資料倉儲系統的比較

	關聯式資料庫系統	資料倉儲系統
組織架構	依執行的交易而定(如訂單處理、進銷存貨處理)。	依主題而定(如產品、顧客)。
資料特性	及時更新、常更改，少歷史性及重複性質的資料。	為歷史性、不常更動且資料重複性高。

	關聯式資料庫系統	資料倉儲系統
功能特性	供每日的交易資料儲存,及日常交易資料查詢。	可提供長期的資料分析,以供決策者的參考。
儲存特性	線上交易的紀錄只維持在較短期間內的資料。	保存的資料可以回溯到數年以前。
系統特性	回應速度敏感、但要有不同查詢須另外設計程式。	回應速度較慢、使用者可以依查詢需要改變資料倉儲維度。
資料庫設計模式	實體關係圖、應用導向的線上交易處理。	星狀、雪花或星座綱要圖、主題分析導向的線上分析處理。
應用發展	及時資料的查詢與顯示,線上交易處理。	資料探勘分析、客戶關係管理、決策支援系統和線上分析處理。

本研究在資料倉儲的建構過程,採取的方式是直接由原始來源檔案轉入資料倉儲的多維度模型綱要的事實資料表與維度資料表中,並沒有設計中介資料庫來儲存三階正規化的資料表。選擇直接建構資料倉儲的主要原因在於:來源資料並沒有資料文字性屬性的紀錄。以水產加工製品資料倉儲為例,沒有設計資料表以及欄位以記錄各加工廠的基本資料,來源資料的格式與內容可見附錄一「水產加工製品資料倉儲的來源資料表」。原始檔案只有記錄每一個縣市內其水產加工廠的總生產值,故缺乏各加水產加工廠的基本資料情況下,直接匯入資料倉儲中儲存。

2.2.4 漁船資料倉儲的多維度模型

對於目前的漁船監控系統的運作,主要是透過漁船上的通訊設備,將漁船之資訊及時地回報給岸上的監控中心,以進行漁船管理。回報的資料包含兩大部分,一是船位資訊回報,二是漁獲量回報。對船位資訊回報而言,內容包括船位、航速、航向等數據。透過漁船上的通訊設備,將漁船之資訊及時地回報給岸上的監控中心。結合前文說明資料倉儲建構為「多維度模型」之優點,提供簡潔而有組織的資料倉儲架構,支援線上分析處理作業;因此為了有效整合漁船監控系統內三大資料庫的船位相關資訊,由王於 2003 年採用雪花綱要的多維度模型建構了「漁船資料倉儲」。[1]

漁船資料倉儲的多維度模型設計,是依據圖 2.2.5 之關聯式綱要中的架構進行分析,一共包含了七個維度:時間維度、漁區維度、漁獲種類維度、漁船資料維度、船公司維度、漁船種類維度與船主維度,以及兩項衡量值:漁獲量以及計數值,構成如圖 2.2.6 之雪花綱要圖。對於衡量值而言,漁獲量記載了漁業活動之漁獲重量,以公斤為單位,計數值是為了分析各漁業活動發生(或回報)的次數,有助於管理單位了解各漁區漁業活動之頻繁性。關於漁船資料倉儲在資料取得上與設計上的缺陷,將於第 3.1.3 小節中說明。



圖 2.2.5 漁船資料倉儲的三階正規化關聯網要

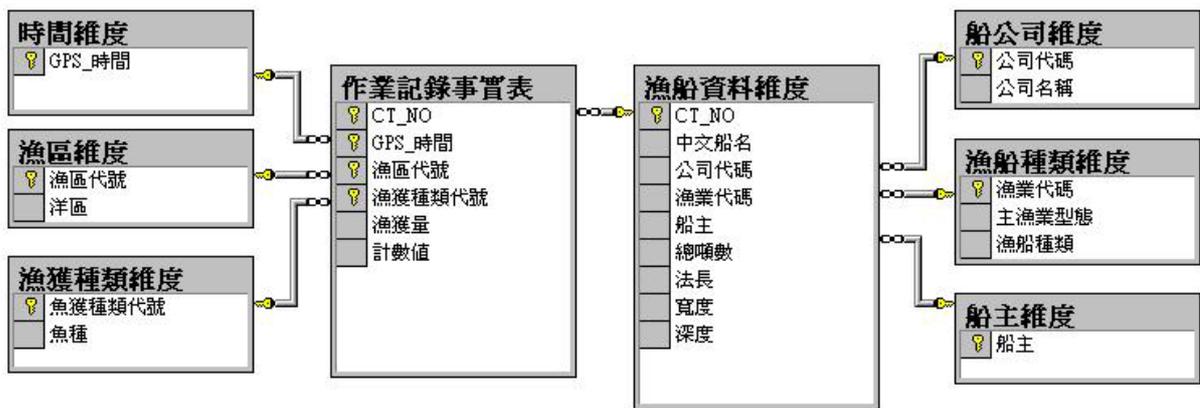


圖 2.2.6 漁船資料倉儲的雪花綱要



2.3 資料擷轉

對資料倉儲建置過程來說，在設計完成資料倉儲的儲存架構後，下一步是將實體資料匯入資料倉儲中，根據歷史經驗平均 80% 的工作時間都花在資料擷轉的工作上，即 ETL。這包含了資料擷取(Data Extraction)、資料轉換(Data Transformation)與資料載入(Data Loading)等工作。首先，資料萃取乃是從特定的來源資料庫中讀取並萃取出所需的資料；其次，資料轉換則是將前述萃取出來的資料，運用規則、索引欄位、或與其他資料相結合後，轉換成所需要的狀態；最後，資料載入則是將轉換後的資料，載入到一個指定的新的或原已存在的資料庫中。因此 ETL 是資料倉儲建置過程中最重要，也是困難度最高的工作。

對本論文使用的 ETL 工具而言，這是採用 MS SQL Server 2000 中的資料轉換服務(Data Transformation Service, DTS)負責處理所有資料轉換的部分。由於原始檔案系統中的資料未經過三階正規化，使得許多交易或供應商資料多有重複或空值，或者是輸入人員的操作疏失，或是未依照欄位要求輸入正確的資料，導致資料必須經過萃取、檢驗、重整、轉換與傳送等多道繁瑣的程序之後才能將原始檔案系統中的資料順利更新到三階正規化後的關聯式資料庫中[11]。能夠作為 DTS 資料來源，可以在任何 OLE DB Provider、ODBC、或純文字檔案格式之間設定連線，執行匯入、匯出、與轉換資料的功能，見圖 2.3.1。使用資料轉換服務時必須注意四個步驟，即設定資料來源，設定資料目標，設定轉換方式如是否用程式語言修改資料格式。

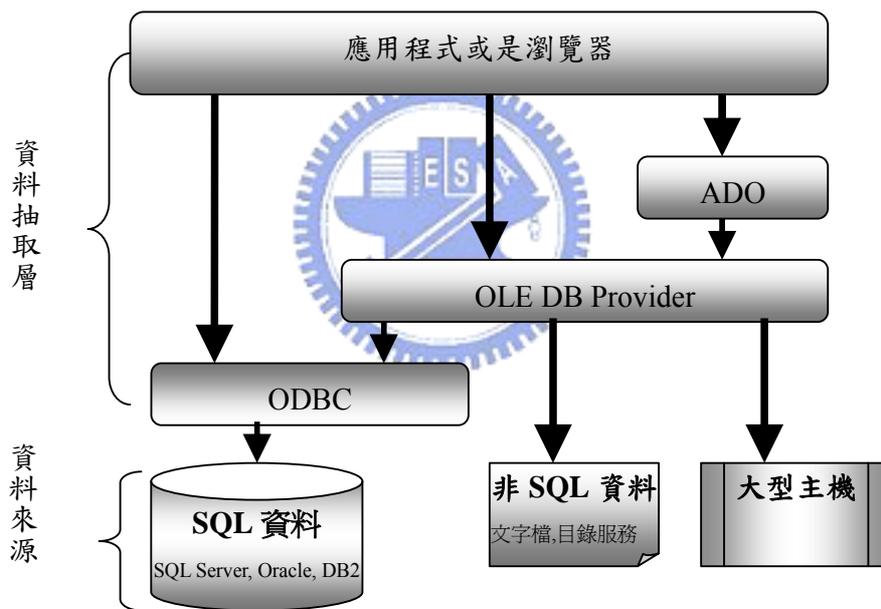


圖 2.3.1 資料轉換服務連接異質來源資料的方式

如圖 2.3.2 所示，由圖中可以看到欲將圖左的來源資料轉移到目的端的三階正規化資料庫，首先需要針對資料來源進行資料萃取的步驟，而資料萃取可以確保資訊的正確性、唯一性與完整性，在資料萃取之後，接著利用資料轉換服務的功能執行 SQL 命令敘述、或轉換資料庫物件及使用 VBScript、Java、COM 等程式語法將資料轉換為所需要的形式，此為資料淨化與資料轉換的階段，資料淨化的主要目的在確保資料的一致性，需要挑出所有不符合三階正規化的資料並加以處理，完成資料淨化的步驟之後，還不一定能夠將資料轉入目的端資料庫，因為某些欄位中的資料必須加以修改才能被轉換。例如需將原本民國的年度加上 1911 轉換為西元的年度。另外，各個欄位的資料也可能需要加以分割或合併，例如將地址前三個字取出來作為該工廠的設籍縣市，把所有有欄位皆經過資料轉換處理過，最後才能順利的將原始的文字檔案資料轉入三皆正規化資料庫，值得一提的是在 DTS 資料轉換服務中提供了呼叫外部程式的元件，此元件可執行.BAT 或.EXE 檔，使 DTS 能與外部程式相結合，

在所有的 DTS 資料轉換工作皆完成之後，DTS 可以設定郵件傳送工作，讓資料庫的管理者能夠知道資料轉換服務成功與否，最後再利用 DTS 的程式語法設定自動轉換的排程。

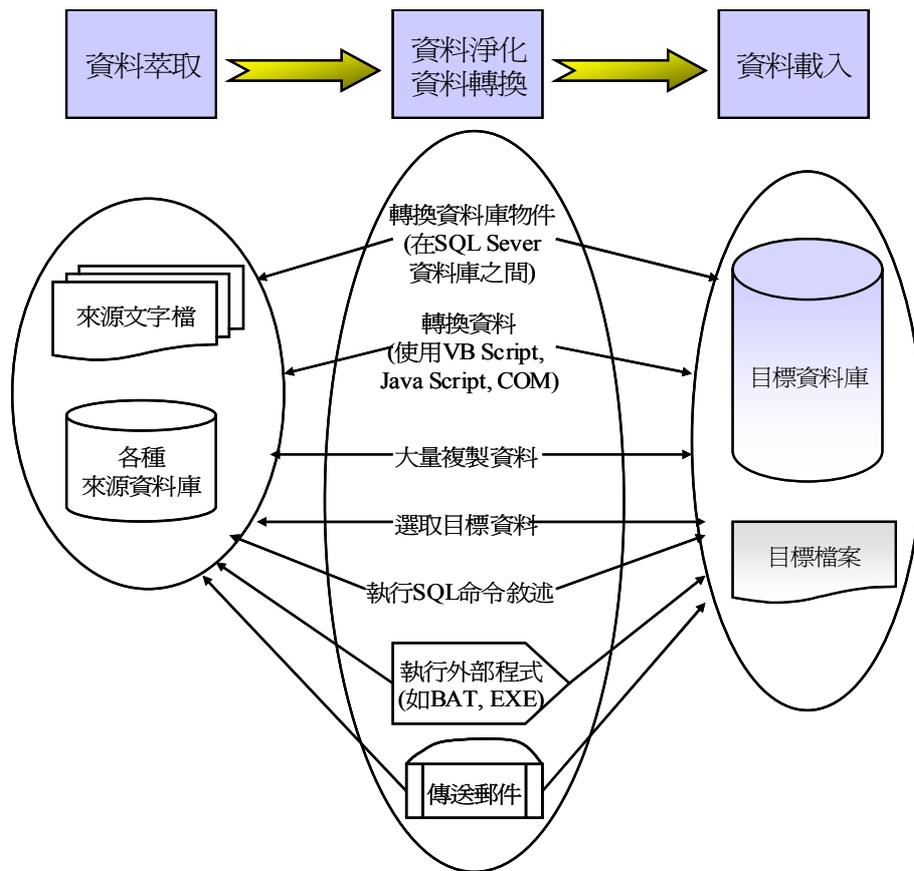


圖 2.3.2 資料轉換服務功能及資料擷轉流程圖

2.4 企業智慧系統架構

由加拿大 Cognos 公司所提出的商業情報系統(COGNOS Business Intelligence, COGNOS B.I.)[17]的主要架構如圖 2.4.1 所示，這包括 PowerPlay Web Server，Impromptu 及 Visualizer (VIP)三套應用軟體，各軟體之運用關係分別敘述如下：

1. 對本研究建構漁業供應鏈的漁業資訊分享熱線而言，則用資料庫管理系統 SQL Server 2000 裡的資料轉換服務來取代其決策流(Decision Stream)。
2. 以軟體間關聯性來看，對 Impromptu 軟體而言，這是以資料倉儲為資料來源並利用 Impromptu Administrator 建立資料目錄(Information Catalog)，並存成*.cat 檔以成為製作報表的基礎，而完成的報表可存成*.IMR(Impromptu Report)檔以供 Impromptu User 版報表瀏覽使用。利用 Impromptu 管理員所製作完成的報表可另存為*.IQD(Impromptu Query Definition)檔以作為製作超方體(Hyper Cube)工具 PowerPlay Transformer 的主要來源。建立好的超方體可存成*.MDC(Multi-Dimension Cube)檔，其可利用 PowerPlay User 版或 PowerPlay Web 版來進行多維度的分析。若不為 Cognos 所建立的超方體，則可利用 PowerPlay Connect 來作連結，如此一樣可被 PowerPlay User 版或 PowerPlay Web 版所接受。表 2.2.3 為 Impromptu、PowerPlay 及 Visualizer 三個軟體的版本及其用途比較。
3. Impromptu 軟體為 COGNOS 推出製作報表的最佳工具，資料來源是來自資料庫的連線與查詢。使用者可利用 Impromptu 查詢即時的資料或彙總異質資料，如關聯式資料庫的實際資料、Excel 格式的預算資料及其他格式的市場統計資料，製作更具實用性的比較性報表。
4. 對 PowerPlay 軟體而言，為線上分析處理工具，提供決策者綜觀角度如全年度的營業額，或可細微逐層分析到各筆交易明細資料。決策者並可彈性且快速地切換分析角度並且依時間序列、產品線或區域來分析資料。
5. 對 Visualizer 軟體而言，是提供高階經營決策人員所需的企業績效衡量指標(Key Performance Indicator,KPI)。Visualizer 以視覺化的方式，例如燈號、儀表版等圖形界面，讓高階主管一目了然地掌握目前公司的營運績效，從巨觀的角度立即發現企業問題，再更進一步去分析及解決問題。

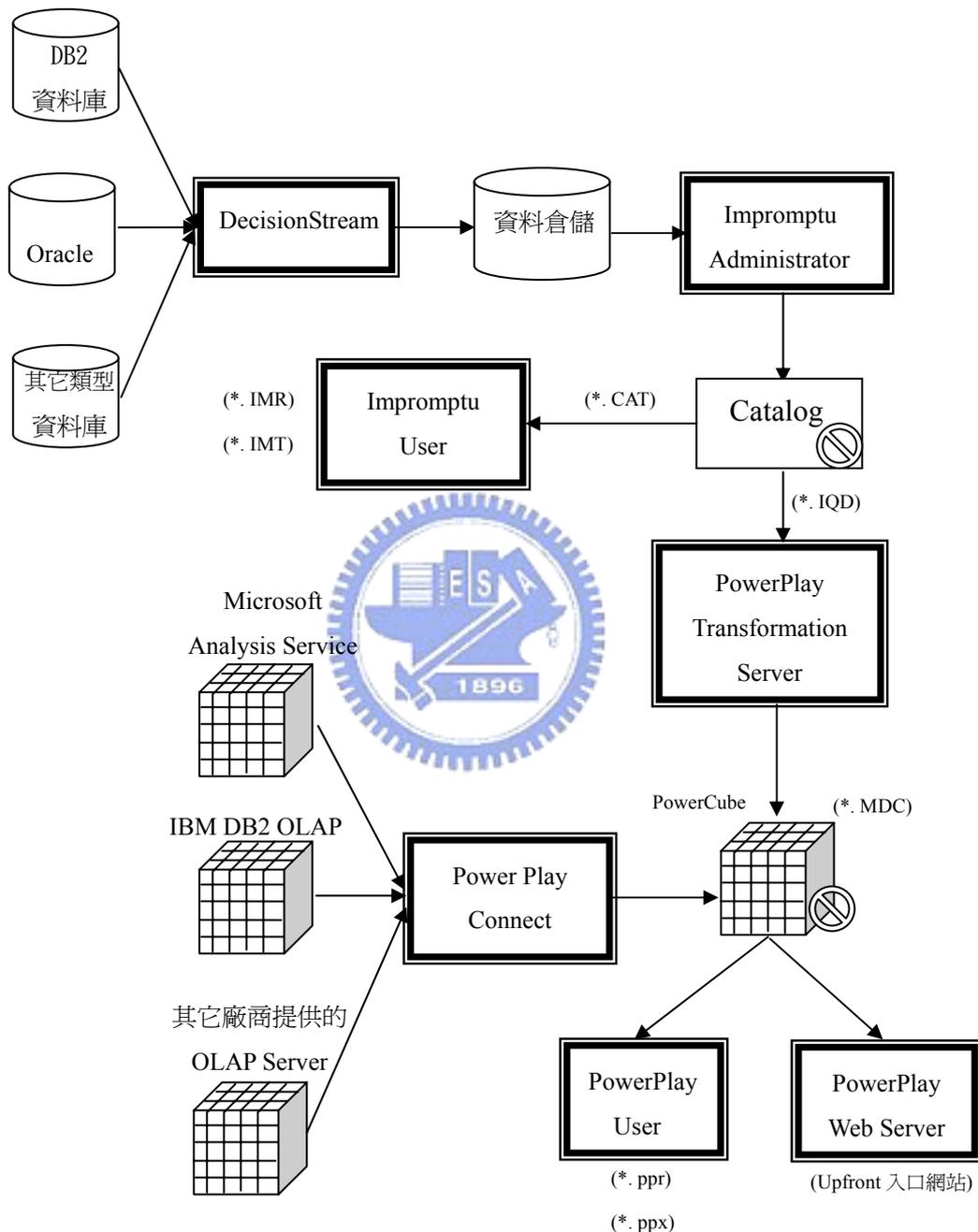
表 2.4.1 Impromptu、PowerPlay、Visualizer 的版本與功能比較[17]

比較項目		Impromptu	PowerPlay	Visualizer
產品定位		互動式報表查詢工具	線上分析處理工具	視覺化分析處理工具
產品主要用途	ADM 版	製作 Catalog 及權限管理	製作超方體及間接權限管理	無
	USER 版	報表的瀏覽及製作	超方體瀏覽(多維度分析與處理)	做視覺化的地圖，可供使用者來瀏覽資訊
	WEB 版	可將報表的瀏覽提升至 Internet 層次	可將超方體的瀏覽提升至 Internet 層次	可將地圖的瀏覽提升至 Internet 層次
使用人員	ADM 版	資訊中心之人員(開發版)	資訊中心之人員(開發版)	無
	USER 版	各階層之人員	各階層之人員	各階層之人員
	WEB 版	各階層之人員	各階層之人員	各階層之人員

綜合上述的分析，整體的使用順序，可由 PowerPlay 提供資料探勘的功能以作出決策，或由 Visualizer 看到企業各個構面的績效指標，再進一步穿透(Drill-through)至 PowerPlay 及 Impromptu 看詳細報表。從使用者需求的角度來比較，依據不同主管的需求，Visualizer

提供高階主管綜觀全局的 KPI 營運指標、PowerPlay 提供中階主管所需的多維度分析資訊、而 Impromptu 提供基層主管監督控制用的日常報表。

對權限控管而言，Impromptu 是利用 Impromptu 管理員進行權限控管，其目的在於限制不同階級人員所能接觸的資料目錄及報表內容。在 PowerPlay Web Server 軟體而言，則統一由 Access Manager 作全面的權限控管，其中包括不同超方體的權限設定，及在 Upfront 的使用者行為設定等權限。



⊗ 權限管控：在 Impromptu 的 Catalog 管理資料庫的資料擷取權限，或在 PowerCube 設定讀取 Cube 資料權限

圖 2.4.1 Cognos 業務情報系統架構[30]

2.5 線上分析挖掘

由於資料倉儲技術逐漸普及化，以及線上分析處理技術的網路化，線上分析挖掘技術也隨之興起。線上分析挖掘有助於使用者利用資料倉儲進行線上學習，進而將大量的交易資訊裡產生所需要的專業知識並能加以應用的學問，例如業務情報(Business Intelligence, BI) 或稱為商業智慧[6]。一般而言，線上分析挖掘技術理得學習理論，可分為教導型學習(Supervised Learning)及非教導型學習(Unsupervised Learning)兩種[12]：前者是指教導下學習並且培養出如何將新事物歸類的的能力，所以又稱為「歸類分析」(Classification Analysis)；後者是指在自我學習狀況下培育出如何將事物分群的能力，所以又稱為「分群分析」(Clustering Analysis)。本研究透過分群分析找尋優良的漁船單位，並且使用歸類分析找出高漁獲量法則。以下分別說明分群分析以及歸類分析等概念。此節共分成兩小節，第 2.5.1 小節介紹分群分析，第 2.6.2 小節介紹歸類分析。

2.5.1 分群分析

對分群分析而言，是依照事物的相似性(Similarity)來分群，好的分群結果是同群內的事物具有高度相似性，而各群間的相異度性高。目前分群分析可以處理不同測量尺度的變數包括：名目尺度(Normal Scale)、順序尺度(Ordinal Scale)、區間尺度(Interval Scale)和比例(Ratio Scale)的變數。

表 2.5.1 分群分析的步驟以及其方法

分群分析之步驟	1. 選擇	相似性指標	歐式距離、歐式距離平方、馬氏距離、馬可夫斯基距離				
		分群演算法	重心法、近距法、遠距法、均距法、華德法				
		分群技術	區分法 (Partitioning Method)	網格法 (Grid-based Method)	模式法 (Hierarchy Method)	密度法 (Density-Based Method)	階層式分群 (Hierarchy Method)
	例如：	K-means、 K-medoid	STING、 Wave Cluster	Conceptual clustering、 COBWEB	DBSCAN、 OPTICS	AGNES、 DIANA	
	2. 選擇	分群的個數					
3. 解釋	各群的特性						

表 2.5.1 說明分群分析大致分成三個步驟，第一步是做選擇包含了選擇「相似性指標」、「分群演算法」、「分群技術」。第二步是根據第一步的結果或是自己的判斷，選擇適當的分群個數。第三步解釋各群體的特性以及分群的結果之討論。由於第一步就是要選擇不同衡量相似性的指標，並且因不同測量尺度的變數需使用不同的演算法，首先必須瞭解測量尺度的意義。其中名目尺度的變數值之間沒有大小、優劣的關係，例如產品編號、性別、布魯林值等；因此「名目變數」計算相似度的方式，二元名目變數的相似度計算見圖 2.5.1(a)，名目變數之值高於 2 種情形的相似度計算圖 2.5.1 (b)。

1. 先將各名目變數以1和0表示其值。此處舉2個二元變數為例。
2. 計算各名稱變數組合後出現的次數，此處以a,b,c和d表達。
3. 依據兩變數出現機率的情形，計算相似度。

二元變數出現情形的紀錄表：

		Object j		
		1	0	sum
Object i	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	p

Simple matching coefficient (兩變數出現的機率均等):

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

Jaccard coefficient (兩變數出現的機率相異):

$$d(i, j) = \frac{b+c}{a+b+c}$$

(a) 二元名目變數相似度計算

超過兩種狀態的名目變數, 例如: 產品項目有花生罐頭、鮪魚罐頭、辣椒、鳳梨罐頭等

第一種方法: 簡單配對法 (Simple matching)

m : 符合該變數的個數, p : 所有變數的總個數

$$d(i, j) = \frac{p-m}{p}$$

第二種方法: 使用大量的二元變數來描述

如果名目變數有M種狀態, 一共會有 2^M 結果。

(b) 超過二元變數值的名目變數相似度計算

圖 2.5.1 名目變數相似度的計算

順序尺度的變數值間有大小或優劣關係, 但無法精確描述其差距, 例如團體中的名次, 第一名優於第二名、第二名又優於第三名, 但是第一、二、三名間的差距又各有差異。「順序變數」在計算相似度時影響十分大, 故使用圖 2.5.2 的方法解決上述問題後, 才能作為計算物體間的相似度。

順序變數可以是離散或連續的

轉換成 Z_{if} 變數後, 才可以使用區間變數的相似性衡量方法:

1. 以 X_{if} 代表原有排名 $X_{if} \in \{1, \dots, M_f\}$

2. 將所有的順序變數經過式子(1)的轉換, 使所有的值均介於0~1之間

$$z_{if} = \frac{r_{if} - 1}{M_f - 1} \dots\dots (1)$$

3. Z_{if} 可以使用衡量區間變數相似性的方式, 以求出名目變數間的相似度

圖 2.5.2 順序變數相似度的計算

區間尺度的變數值間有大小、優劣的關係, 彼此之間的差距具有意義, 例如速度差距。衡量相似性指標的方法眾多, 例如歐式距離(Euclidean Distance)、歐式距離平方(Squared Euclidean Distance)、馬氏距離(Manhattan Distance)、馬可夫斯基距離(Minkowski Distance)等, 以圖 2.5.3 說明各指標的定義。

- 計算物體、物體之間的相似度（距離越小，相似度越高）：

1. **Minkowski Distance:**

$$i = (X_{i1}, X_{i2}, \dots, X_{ip}), \quad j = (X_{j1}, X_{j2}, \dots, X_{jp}),$$

物體*i*、物體*j*具有

個維度，**q**為任意的正整數

$$d(i, j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \dots + |x_{ip} - x_{jp}|^q)}$$

(1.1) If $q = 1$, $d(i, j)$: **Manhattan distance**

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

(1.2) If $q = 2$, $d(i, j)$: **Euclidean distance**

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

2. 距離特性：

$$d(i, j) \geq 0$$

$$d(i, i) = 0 \quad d(i, j) = d(j, i)$$

$$d(i, j) \leq d(i, k) + d(k, j)$$

圖 2.5.3 區間變數相似度衡量指標的定義

如表 2.5.1 所示，分群分析的技術主要可以分成下列五種方式：區分法、網格法、模式法、密度法及階層式分群。前面四種分群技術屬於「非階層式分群」(Nonhierarchical Clustering)，第五種屬於「階層式分群」(Hierarchical Clustering)，另外一種技術稱為「二階段式分群」(Two-stage Clustering)是結合階層式與非階層式分群技術的優點，因此分群分析可以概括分成三大類[32]。由於分群分析的理論眾多，本文僅介紹此三大種類技術的原理。

在選定相似性指標以及分群演算法後，對階層式分析而言，先以各點為單群點，然後取最接近的兩點形成一群，再重複此步驟直到所有的點均納入同一群為止。階層式分群的過程如圖 2.5.4 所示，以歐式距離平方為衡量相似性指標，並且以華得法作為分群演算法，圖形 Y 軸代表 487 種個例作階層式分群分析，圖形 X 軸代表各群內所有樣本點和該群重心的距離平方和，同一群的樣本點是來自平方和誤差(Error Sum of Squares, ESS)最小的組合結果。分群分析的下一步驟是挑選分群的個數，階層式分群分析會列出所有分群的結果，由分析者自行挑選分群數。統計學家提出均方根標準差(Root-Mean-Square Standard Deviation, RMSSTD)、平方和比(R-Square, RSQ)、半平方和比(Semipartial R-Square, SPRSQ)和群距(Distance between Clusters, CD)四個評量指標來協助分群數決定的過程。各指標的意義以及建議值見表 2.5.2 詳細說明。

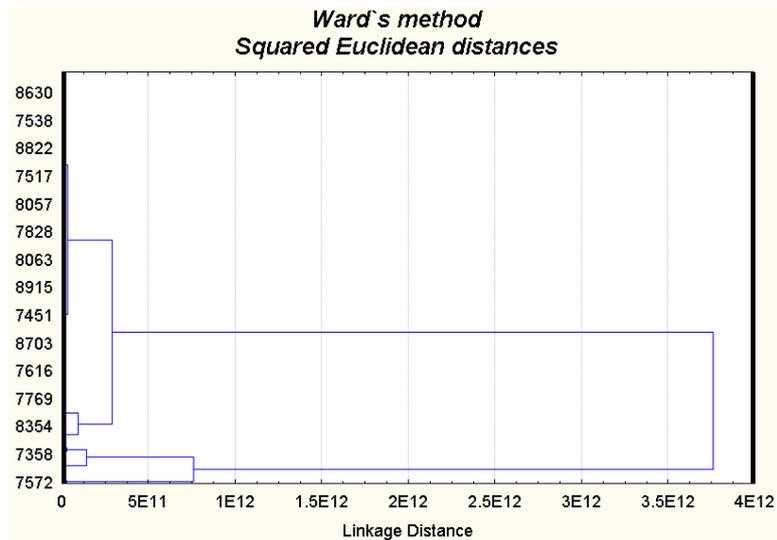


圖 2.5.4 階層式分群的過程

表 2.5.2 階層式分群分析的決定分群數之評量指標

評量指標	物理意義	建議值
均方根標準差 (RMSSTD)	新群內的樣本相似度	值越小，群內相似度越高，故越好。
平方和比(RSQ)	$\frac{\text{群間的平方和}}{\text{群內平方和} + \text{群間平方和}}$ ，代表群間相異度。	值越大，各群間的相異度越高，故越好。
半平方和比 (SPRSQ)	兩群組結合成一個新群組後，損失的群內相似度。	值越小，新群內相似度越高，故越好。
群距(CD)	兩群合併時的，兩個群重心的歐式距離。	值越小，欲結合的兩群相似度越高，故越好。

在選定相似性指標以及分群演算法後，對非階層式分析而言，首先要決定分群數 k 。接著，任選 k 個樣本作為起始群心後，將資料分配至距離最近的代表點群；下一步重新計算新群重心以形成新的代表點，重複第二步驟直到代表點收斂。分群的過程如圖 2.5.5。

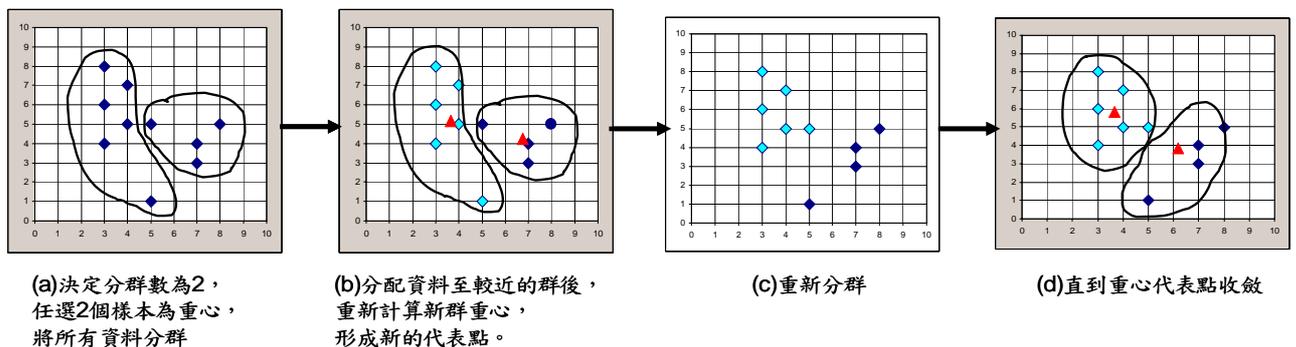


圖 2.5.5 非階層式分群的過程

由於階層式分群分析的缺點是已分群的樣本點無法再重新分群，也就是說當樣本一旦被放在同一群內，則以後的步驟必定仍在同一群內。非階層式分析的缺點則是事先必須知道分群個數 k 值，然而實務上往往一開始分析者並不能掌握開資料合適的分群個數。綜合上述分析，將階層式和非階層式分群分析互補，利用二階段式分群分析來進行：先以階層式分群分析得到 k 值和各群重心，接著再執行非階層分群直到各群重心收斂。優點是分析者可依據第一階段的分析決定 k 值，並利用第二階段的分析將樣本點做任意組合的分群。

2.5.2 歸類分析

歸類分析在線上分析挖掘的技術裡，是屬於教導型的學習。對歸類分析而言，這包含五個主要元素：未知樣本（即資料）、訓練樣本(Training Data)、測試樣本(Testing Data)、專家(Model)及評審(Adjuster)。如圖 2.5.6 的歸類分析架構圖所示，歸類分析中專家的建立可分為兩個時期：第一個是模式建構期—用訓練樣本(a)來產生專家(b)，第二個時期是模式測試期—用測試樣本(c)來估計專家(b)的正確率。由此可知訓練樣本和測試樣本互相獨立，否則在專家內的模式可能會發生過份強調(Overfitting)某一個在母群體中不顯著的現象[31]。當一個未知分類結果的樣本(d)出現後，可先用專家來判定此新資料的類別為何，等新資料通過專家的測試後再送到評審接受真正的考驗（以本研究中的評審，即漁船公司的魚貨量），這時新資料通過檢驗的機率就很大。目前建立專家的方法主要有決策樹、貝氏推理、神經網路及統計多變量的判別分析等多種，如表 2.5.3 所示。

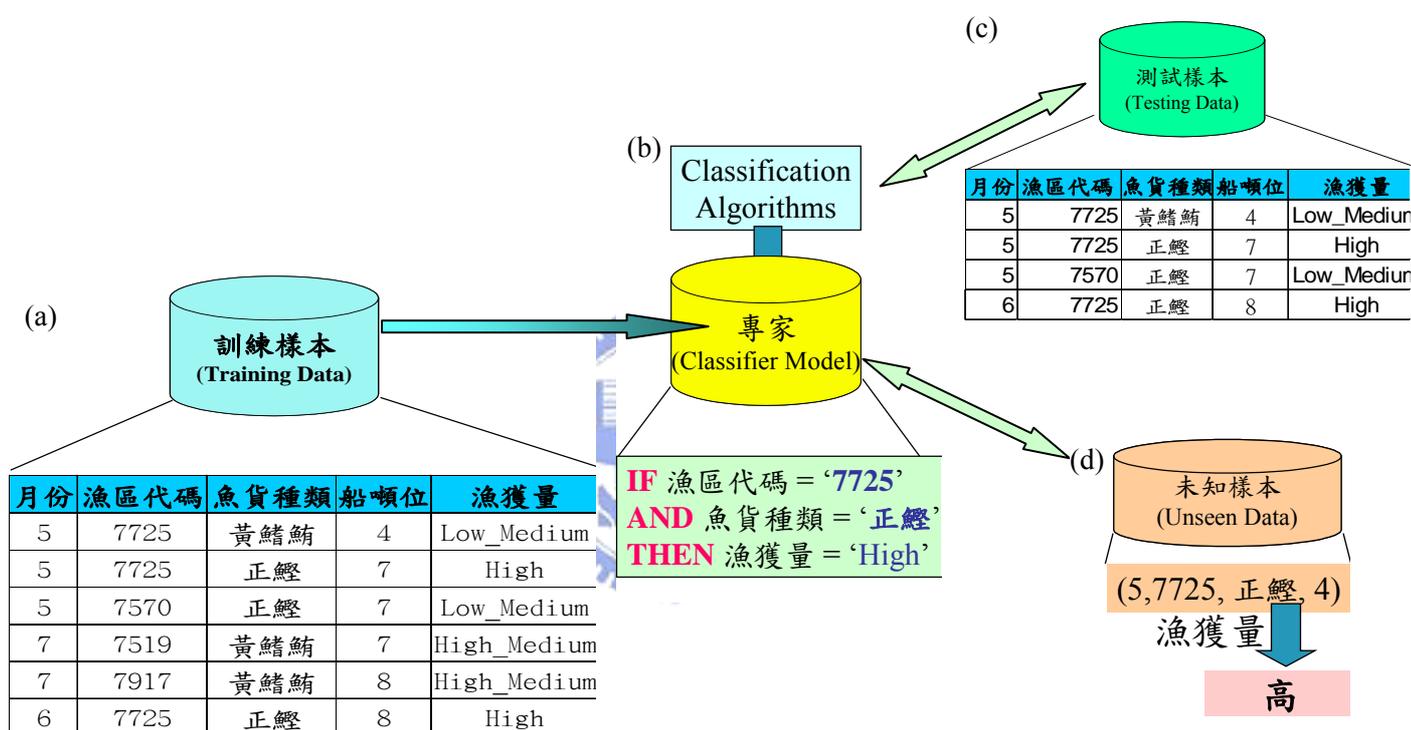


圖 2.5.6 歸類分析的架構

對決策樹而言，這方法又可以細分為(Interactive Dichotomiser 3)、CHAID(Chi-squared Automatic Interaction Detection)、CART(Classification and Regression Tree)等技術[25]。由表 2.5.3 可看出這些技術最大的差異在於選擇屬性的準則。ID3 和 C4.5 技術由 J.R. Quinlan 於 1986 和 1993 提出 [22][23]，主要是依據亂度的大小為準則，ID3 是 C4.5 的早期版本。本文將說明 ID3、C4.5 決策樹的歸類分析技術之原理，並且以圖 2.5.6(a)的訓練樣本為例，結果如圖 2.5.7 所示。

表 2.5.3 線上學習挖掘的學習方法

學習方式 (線上分析挖掘)	主要方法	備註	
教導型學習 (歸類分析)	決策樹(Decision Tree)	選擇屬性的方法 (Attribute selection)	選擇屬性的準則
		C4.5	增益比(Gain Ratio)
		CART	吉尼指數(Gini Index)
		CHAID	卡方檢定(Chi-square Test)
		ID3	資訊增益(Information Gain)
	貝氏推理(Bayesian Reasoning)		
神經網路(Neural Network)			
判別分析(Discriminant Analysis)			

決策樹的架構如同一個樹狀的流程圖，可以分為三個部分：分支(Branch)、中間節點(Internal node)及葉節點(Leaf node)。以圖 2.5.7 為例，中間節點代表的是屬性，故又稱屬性節點，有”魚貨種類”、”月份”及”漁區代碼”。分支代表的是該中間節點分類後的值，如通過魚貨種類節點後得到”黃鰭魚”、”正鯉”，通過月份節點後得到”5”、”7”。葉節點代表樣本最後分類所歸屬的類別(Class)，故又稱末端屬性節點，本例有”Low_Medium”、”High_Medium”及”High”。當樣本通過決策樹時，屬性節點會將樣本集合拆成樣本子集合，而每一個分支就是一個樣本子集合。最後當樣本的所有點到達葉節點時，則各點會歸屬到某個固定類別。從亂度的角度分析，決策樹越上方其亂度越大，此為決策樹的特性。

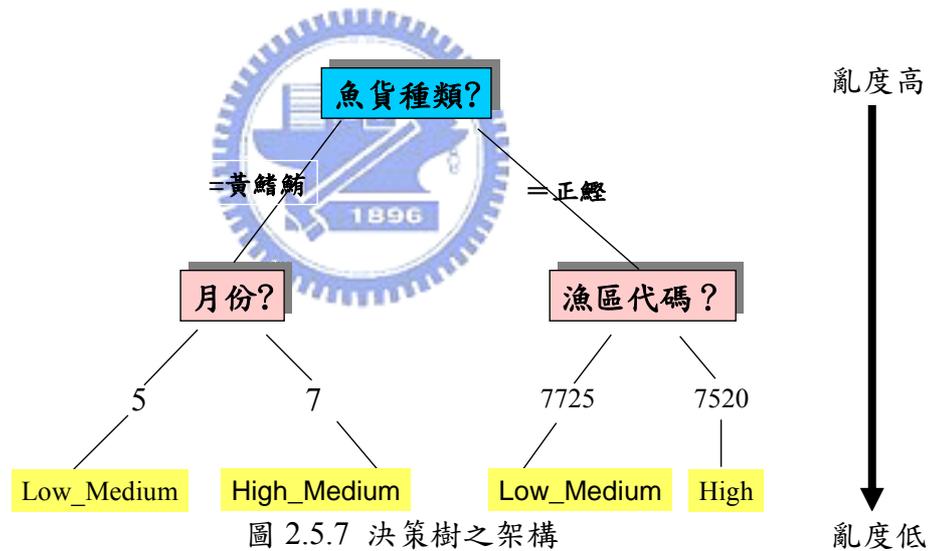


圖 2.5.7 決策樹之架構

因此決策樹的形成，利用 ID3/C4.5 技術是把決策樹由高到低降低其最大亂度為理論基礎，使用圖 2.5.8 的步驟以遞迴的方式不斷地切割訓練樣本，每次完成圖 2.5.8 的步驟後決策樹則多一個層級。ID3/C4.5 選擇資訊增益最大的屬性作節點，其目的是希望建構的決策樹越小越好，也就是經過較少的屬性測試就能將所有樣本歸類。

1. 假設一個樣本S，共有k個屬性(即Attribute Number = k)。
2. 對末端屬性而言，將S分成P和N類別(Class)，
分別包含的樣本數為Class (P,N)=(p,n)
總資訊量： $I(p,n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$
3. S其中一個屬性為A，可將樣本分成 $\{S_1, S_2, \dots, S_v\}$, v個集合(分支)
S_i的資訊量為I_i:
 1. 末端屬性將S_i分成p_i個樣本屬於P類別
n_i個樣本屬於N類別
 2. $I_i = I(p_i, n_i) = -\frac{p_i}{p_i+n_i} \log_2 \frac{p_i}{p_i+n_i} - \frac{n_i}{p_i+n_i} \log_2 \frac{n_i}{p_i+n_i}$
 3. 屬性A的亂度為 = $E(A) = \sum_{i=1}^v \frac{p_i+n_i}{p+n} I(p_i, n_i)$
 4. 屬性A的資訊增量 = $\text{Gain}(A) = I(p, n) - E(A)$
4. 將k個屬性的資訊增量排序，挑選最大的資訊增量做節點。

圖 2.5.8 決策樹形成之步驟



2.6 序列模式挖掘

每一次進行遠洋漁獲的作業，無論在燃料、人事、時間等各項成本的投入均不低。因此漁船公司會依據以往的遠洋出海經驗，分派公司所屬的船隻在某段時間到某個海域進行捕撈，希望藉由過往經驗的法則得到較高的漁獲量。漁船接收到公司發佈的消息，在該時間點到達該海域，但是由何處開始航行以及移動方向為何來捕撈，對於船長而言是一項大問題。無論是漁船公司或是船長，在整體魚貨供應鏈中(Seafood Supply Chain)都是屬於資訊相對弱勢的族群，面對如何選擇有效航行路徑以協助捕撈作業進行，都是需要從有系統的資料中運用資料探勘技術挖掘出知識。本研究使用序列模式挖掘來找出高漁獲量漁船航行的軌跡模式，分析的數據來自漁船的船位資料。以下分別說明序列模式挖掘(Mining Sequential Patterns)以及應用分析工具的架構等觀念。此節共分成兩小節。其中第 2.6.1 小節說明序列模式，第 2.6.1 小節說明序列模式挖掘的應用工具—WUM(Web Utilization Miner)。

2.6.1 序列模式

序列模式分析用來發現時間性依序出現的事件之間彼此的關聯性[30]。例如藥物誤用的後續反應模式：病人依照醫生處方服用兩種以上的藥物，經過一段時間後由於藥物之間產生的交互作用，引起病人一連串不正常的生理反應，因此醫生可以依據病人各症狀發病的順序來判斷如何診治以及其原因。此依序出現的生理反應就可以使用序列模式分析發現發病的模式(Sequential Patterns)。由上述說明可知，序列模式可視為關聯性(Association)的特例，只是多了時間維度的考量，並且在出現位置以及時間先後性上有因果間的關聯(Cause-effect Association)，因此這種挖掘知識的方法屬於時間模式挖掘(Temporal Data Mining)的一種。

表 2.6.1 挖掘序列模式的方法與步驟

Step	Phase name	Description
1	Sort	對原始的資料庫進行排序。
2	Litemset	找出符合條件（即 Minimum Support）的項目集合(Itemset)，稱為「大項目集合」（又稱 Large Itemset）。
3	Transformation	將項目集合轉換成整數值。
4	Sequence	找出符合條件（即 Minimum Support）的序列，即「大序列」。
5	Maximal	從「大序列」找出「最大序列」即「序列模式」。
名詞解釋		意義
項目(Item)		各產品品名。例如：麵包、果醬都是交易的項目。
支持度(Support)		某一種項目集合，出現在資料庫的比例或是次數。可以作為序列模式分析時規範的依據。
項目集合(Itemset)		所有項目所構成的集合，不可以是空集合。例如：(麵包，果醬)。
大項目集合(Large Itemset)		又稱為項目集合，滿足所定義的 Minimum Support 限制的項目集合。
序列(Sequence)		所有項目集合的序列紀錄。例如：{(柳橙汁)，(麵包，果醬)} 代表排序後的顧客交易資料，因此又稱為顧客序列(Customer-sequence)。
大序列(Large Sequence)		滿足所定義的 minimal support 限制的序列，又稱 frequent sequence。
最大序列(Maximal Sequence)		不被其他序列所包含的序列，即不包含在其他序列中。
序列模式(Sequential Patterns)		符合 minimal support 的最大序列。
序列長度(Length of Sequence)		序列中包含的項目集合數目。例如：{(柳橙汁)，(麵包，果醬)}，此序列包含該名顧客兩次交易的資料，其長度為 2，又稱為 2-sequence。

從交易紀錄型態來比較關聯性模式與序列模式，可以得知由前者模式的交易資料庫分析的是哪些商品會被一起購買，這是屬於挖掘交易內的型態(Intra-transaction Patterns)；而後者模式的交易資料庫分析的是顧客購買的商品順序，這是屬於挖掘交易間的型態(Inter-transaction patterns)，後者的模式分析主要是找出顧客購買這些商品後，接下來會購買的商品為何。以 Agrawal 和 Srikant(1995)[28]的例子說明挖掘序列模式的關聯規則之方法以及步驟，序列模式的步驟說明整理以及專有名詞的解釋裡於表 2.6.1。

首先在表 2.6.2 顧客交易的資料庫中，選擇交易時間、顧客編號及購買的產品項目三項欄位。第一步以顧客編號為主要排序依據，交易時間為次要排序依據而得到如表 2.6.3 的結果。第二步。依據最小支持度的限制下，找出所有的大項目集合，結果如表 2.6.4 所示。首先定義出計算支持度的單位，不同的單位計算出來的支持度值會有所不同。以本案例來說，可以用顧客編號為計算單位，也可以用交易記錄為記錄單位。假設一名顧客購買產品的 2 次記錄為{(40), (40, 70)}，就前者為計算單位的算法是認為產品 40 的序列出現 1(次)，也就是支持度值為 1；就後者為計算單位的算法是為計算單位的算法是認為產品 40 分屬兩次交易，其序列出現 2(次) 也就是支持度值為 2。接著產生大項目集合的候選者(Candidate)後，依據 Apriori-like 演算法，找出所有的大項目集合。最後以整數為代碼來對應所有的大項目集合，這會提高日後處理的便利性。以本案例而言，規範支持度的閾值(Support Threshold)為 25%，用顧客編號為支持度的計算單位，資料庫中共有 5 筆顧客交易紀錄，由 $\text{Support} = 5 * 25\% = 1.25$ 得知大項目集合至少要有 2 名顧客的序列，就會得到表 2.6.4 的結果。

表 2.6.2 顧客交易資料庫原始資料

交易時間	顧客編號	購買產品項目的編號
June 10 '93	2	10,20
June 12 '93	5	90
June 10 '93	2	30
June 10 '93	2	40,60,70
June 10 '93	4	30
June 10 '93	3	30,50,70
June 10 '93	1	30
June 10 '93	1	90
June 10 '93	4	40,70
June 10 '93	4	90

表 2.6.3 顧客交易資料庫排序後的資料

顧客編號	交易時間	購買產品項目的編號
1	June 25 '93	30
1	June 30 '93	90
2	June 10 '93	10, 20
2	June 15 '93	30
2	June 20 '93	40, 60, 70
3	June 25 '93	30, 50, 70
4	June 25 '93	30
4	June 25 '93	40, 70
4	June 25 '93	90
5	June 25 '93	90

表 2.6.4 由顧客交易資料庫中建立「大項目集合」以及其對應代碼

大項目集合	對應代碼
(30)	1
(40)	2
(70)	3
(40 70)	4
(90)	5

經過上述兩步驟後，第三步依序進行兩階段的轉換工作，首先由排序後的交易資料中刪除不是大項目序列的顧客序列；接著將剩下的顧客序列（包含各大項目集合）轉換成對應的整數代碼，步驟和過程如表 2.6.5 所示。例如顧客 2 的項目集合(10 20) 及(60)不屬於大項目集合，在第一階段的轉換中直接刪除，並且在第二階段終將(40 70)轉換成{(40), (70), (40,70)}再對應到整數代碼{2, 3, 4}。

表 2.6.5 顧客交易資料庫的兩階段轉換

顧客編號	原始顧客序列	第一階段轉換後的顧客序列	第二階段轉換後的顧客序列
1	<(30)(90)>	<{(30)} {(90)}>	<{1} {5}>
2	<(10 20)(30)(40 60 70)>	<{(30)} {(40),(70), (40,70)}>	<{1} {2,3,4}>
3	<(30 50 70)>	<{(30),(70)}>	<{1,3}>
4	<(30)(40 70)(90)>	<{(30)} {(40),(70),(40,70)} {(90)}>	<{1} {2,3,4} {5}>
5	<(90)>	<{(90)}>	<{5}>

*陰影代表每一個大項目集合

第四步驟利用大項目集合找出序列。從 Agrawal 和 Srikant(1995)提出 AprioriAll 演算法開始，「序列模式的挖掘方法就和關聯性法則(Association Rule)有著很深的關係。因為在尋找序列模式（包含尋找序列→大序列→最大序列→序列模式）的過程，與尋找關聯性法則的大項目集合是相類似的。兩者不同點只是在關聯性法則中的大項目集合指示將交易中的項目(Item)結合在一起；而序列模式必須考量項目出現的先後性或是路徑的順序性。

分別比較圖 2.6.1(a)與圖 2.6.1 (b)，前者為建立關聯性法則的演算法；前者為建立序列模式的演算法。使用的方法依據 Apriori-like 演算法，一共分為兩個步驟，再以遞回的方式重複此兩個步驟。第一步先產生大序列的「候選者」(Candidate)，也就是先從 Apriori-generate Function 所取得的 C_k (序列長度為 k 的大序列候選者)。第二步逐一計算每一個資料庫中的序列其支持度值，如果在支持度閾值之上，就產生 L_k (序列長度為 k 的大序列)，此步驟會隨著資料量的增加而產生越多的「候選者」，花費的時間越長。然後再利用 Apriori-generate Function 產生下一個 C_{k+1} ，再找出 L_{k+1} ，.....，如此循環操作，直到找到所有長度的大序列為止。產生「候選者」方法的方式，在關聯性法則或是序列模式的探勘上，兩者十分類似，分別見圖 2.6.1(c)與圖 2.6.1 (d)說明。

Join Step: C_k is generated by joining L_{k-1} with itself

Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset

Pseudo-code:

```
 $C_k$ : Candidate itemset of size k  
 $L_k$ : frequent itemset of size k  
 $L_1 = \{\text{frequent items}\};$   
for ( $k = 1; L_k \neq \emptyset; k++$ ) do begin  
   $C_{k+1}$  = candidates generated from  $L_k$ ;  
  for each transaction  $t$  in database do  
    increment the count of all candidates in  $C_{k+1}$   
    that are contained in  $t$   
   $L_{k+1}$  = candidates in  $C_{k+1}$  with min_support  
  end  
return  $\cup_k L_k$ ;
```

(a) The Apriori Algorithm[24]建立關聯性法則的項目集合

Join Step: C_k is generated by joining L_{k-1} with itself

Prune Step: Any (k-1)-sequences that is not frequent cannot be a subset of a frequent k-sequences

Pseudo-code:

```
 $C_k$ : Candidate itemset of size k  
 $L_k$ : frequent itemset of size k  
 $L_1 = \{\text{large 1-sequences}\};$  //Result of itemset phase(Step2)  
for ( $k = 1; L_{k-1} \neq \emptyset; k++$ ) do begin  
   $C_{k+1}$  = candidates generated from  $L_k$ ;  
  for each customer-sequence  $c$  in the database do  
    increment the count of all candidates in  $C_{k+1}$   
    that are contained in  $c$   
   $L_{k+1}$  = Candidates in  $C_{k+1}$  with min_support  
  end  
return Maximal Sequences in  $\cup_k L_k$ ;
```

圖 2.6.1(b) The AprioriAll Algorithm[28]

Suppose the items in L_{k-1} are listed in an order

Step 1: self-joining L_{k-1}

```
insert into  $C_k$   
select  $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$   
from  $L_{k-1} p, L_{k-1} q$   
where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} <$   
   $q.item_{k-1}$ 
```

Step 2: pruning

```
forall itemsets  $c$  in  $C_k$  do  
  forall (k-1)-subsets  $s$  of  $c$  do  
    if ( $s$  is not in  $L_{k-1}$ ) then delete  $c$  from  $C_k$ 
```

(c) 關聯性法則中產生候選者的方法[24]

Suppose the litemset in L_{k-1} are listed in an order

Step 1: self-joining L_{k-1}

insert into C_k

select $p.litemset_1, p.litemset_2, \dots, p.litemset_{k-1}, q.litemset_{k-1}$

from $L_{k-1} p, L_{k-1} q$

where $p.litemset_1=q.litemset_1, \dots,$

$p.litemset_{k-2}=q.litemset_{k-2},$

$p.litemset_{k-1} < q.litemset_{k-1}$

Step 2: pruning

forall sequences c in C_k do

forall $(k-1)$ -subsequences s of c do

if (s is not in L_{k-1}) then delete c from C_k

(d) 序列模式挖掘中產生候選者的方法[28]

圖 2.6.1 序列模式與關聯性法則的比較

最後進行到第五步驟，從所有的大序列中，找出「最大序列，方法見圖 2.6.6 的說明。圖中的「 n 」指的是最長的序列長度。尋找最大序列是首先由最長的序列開始(即 n -sequence)，然後將所有的次序列(即包含在最大序列中的次序列)刪除，依次遞回進行刪除步驟直到序列長度為 2 則停止。就本案例而言，最大序列為 ($\{1\} \{2\} \{3\} \{4\}$)，也就是所謂的「序列模式」。

S : All larges sequences in the sequence phase

n : the length of the longest sequence

for ($k = n; k > 1; k --$) **do**

for each k -sequence s_k **do**

delete from S all subsequences of s_k

圖 2.6.2 尋找最大序列的方法[28]

上述挖掘出序列模式的五個步驟，使用的是 Apriori-like 演算法。發生的問題有幾項：

1. 重複產生候選者，大量消耗記憶體的儲存空間。

每產生一組相同序列長度的候選者時，都需要重新掃描過整個資料庫的資料。不但增加許多掃描的時間，另外在反覆處理候選者以及大序列的過程中，由於十分佔用記憶體的儲存空間，經常會發生記憶體不足而無法進行序列模式分析。

2. 缺少時間的組合[27][30]。

所謂的時間組合是，將某一段時間內將所有觀察的項目做序列模式分析，即使項目來自於不同的交易記錄，這就是所謂的滑動時間組合(Incorporate Sliding Time Windows)。

3. 無法挖掘出稀少性的序列事件。

由於在序列→大序列的挑選過程，是利用出現次數作為門檻，對於某些分析的情境是屬於稀少發生的狀況，而這序列又是顯著性的影響一個系統下，使用上述方法做序列模式挖掘就無法挖掘出有用的知識。例如故障序列模式的發現。假設一個原件 C 的安裝後總是接著發生機械 A 的故障故障，故障事件的發生在一個系統中並不經常發生，那麼即使 C,A 兩者之間存在強的因果關係，也就是有高信度 (Confidence) 的情況下，也會因為其支持度太低而在產生大序列的過程，將 CA 序列予以刪除。另一種改善的方法是下降支持度的閾值來做序列模式挖掘，缺點是很難從大量的序列中找出 CA 事件。

依據以上三個缺點，在實務上的應用已有許多新的演算法來解決此問題，例如 PlanMine

Algorithm、SPADE Algorithm、MINT—mining language 和 MiDAS 等方法，下一個小節將說明 WUM 以及其包含的挖掘語言 MINT。

2.6.2 序列模式挖掘的應用工具

WUM 全名是 Web Utilization Miner，是一項挖掘序列模式的工具。而設計的目的在於研究網路瀏覽者瀏覽網頁的行為模式，依據其瀏覽路徑的順序，找出一般使用者以及顧客的瀏覽模式 (Navigation Patterns)，即序列模式。該工具由德國柏林洪保特大學 (Humboldt-University Berlin) 經濟科學院所發展出來的研發系統，圖 2.6.3 為 WUM 的開發環境和架構。WUM 依據使用功能可以分成 4 大塊模組(modules)，分別為「WUM_prep」、「WUM_agService」、「WUM_gseqm」及「WUM_visualizer」。鬚線箭頭是 WUM 使用者與 WUM 系統間的溝通，可以是傳達的觀念、查詢語法、網站的架構等溝通。實線箭頭代表的是資訊流，也就是模組之間、使用者與系統之間傳遞的資料。

第一塊模組 WUM_prep 負責兩項任務：一個是做資料淨化，移除不相關或是多餘的資料欄位。例如研究目的在於尋找瀏覽者的瀏覽路徑，只需要保留路徑紀錄，因此擷取網頁圖片、註解記錄或是失敗的登入等訊息應該予以去除。經過此模組的處理後，可以移除掉非人類瀏覽行為的紀錄檔，避免處理大量沒意義的資料而可以提高資料處理的效率。WUM_prep 另一個負責的任務是可以輸入不同時間長度的限制條件，依據每一個議期 (Session) 記錄，組織起瀏覽者登入網站的行為，並且在長時間的瀏覽記錄下，辨別個別使用者不同時間點進入網站後的記錄。例如瀏覽者如果超過預設的連線時間，則判定為一個新的議期。

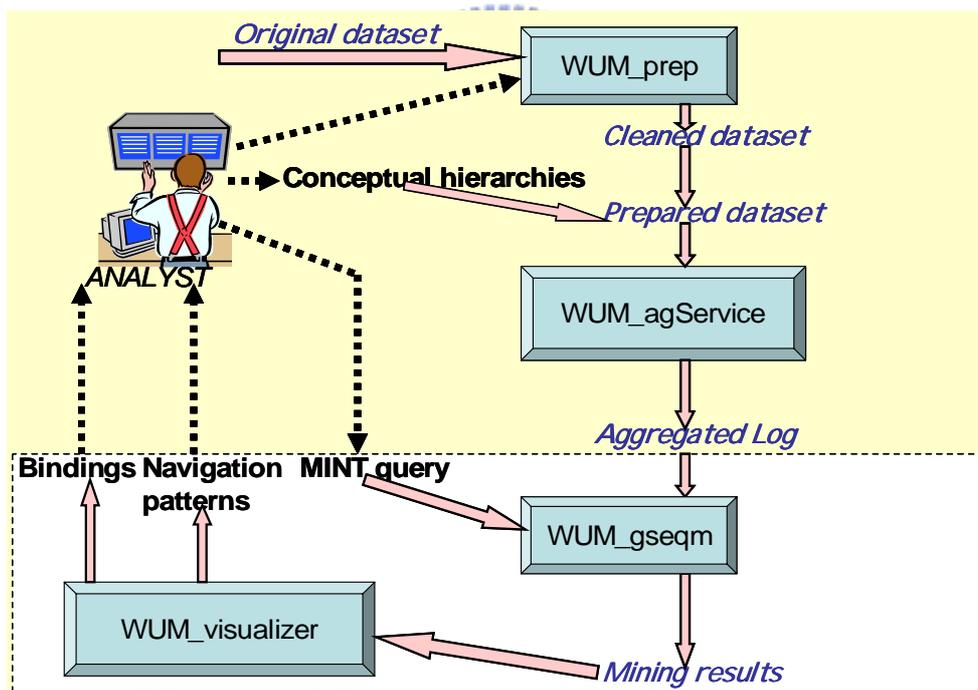


圖 2.6.3 WUM 開發環境以及架構[33]

第二塊模組 WUM_agService 在收到淨化處理與切割後的資料，將每一次的議期視為每一個新的交易，最後處理成「瀏覽途徑」(Visitor Trail)，也就是指所有按照瀏覽先後順序排列的網頁，即序列，說明了使用者如何該網站以及如何該在網頁間移動。因此 WUM_agService 將瀏覽者每一個議期內的瀏覽記錄收集並且集成一個樹狀結構，此時可得到彙集後的資料也就是「Aggregated Log」，而其內形成各個接層式架構的資料稱為「瀏覽樹」(Aggregated Tree)。以下用圖 2.6.4 說明 Aggregated Log 的結構[26]：主要有兩大原件構成，網頁 (Page) 與節點 (Node)。「網頁」指的是一個 HTML-file，包含的資訊含有網址 (URL)、網頁標題、網頁內容、網頁擁有者以及造訪網頁總次數 (The Total Number of Assesses)，這裡包含所有

資料庫紀錄的網頁，不見得是出現在序列的網頁，所以最少有一位瀏覽者的紀錄。「節點」是位於瀏覽樹上一個瀏覽分支的點，一個節點對應到一張網頁，由於瀏覽者有可能在網站上來回某些網頁，瀏覽途徑有可能是一個循環。因此節點包含的資訊內容包含了在該分支出現的該網頁次數(Occurrences)及在該分支上的瀏覽次(Supports)。

舉例來說，圖 2.6.4 有六種瀏覽途徑，以 a 網頁為首頁的瀏覽途徑分別為 1、4 和 5，因此 a 網頁的總瀏覽次數=8+ 3+ 10 = 21，使用節點「(a,1),21」；同樣的情形 b 網頁為首頁的瀏覽途徑分別為 2、3 和 6，總瀏覽次數=2+ 7+ 4 = 13，使用節點「(b,1),13」。在瀏覽途徑 2、6 中，有回到 b 網頁瀏覽的紀錄，發生的次數為 6 次，因此由節點「(b,2),6」，可以得知這是第二次造訪 b 網頁而且一共有 6 次紀錄。節點網頁的符號「^」代表的是根節點，以本例而言一共有 340 次進入該網站的紀錄。

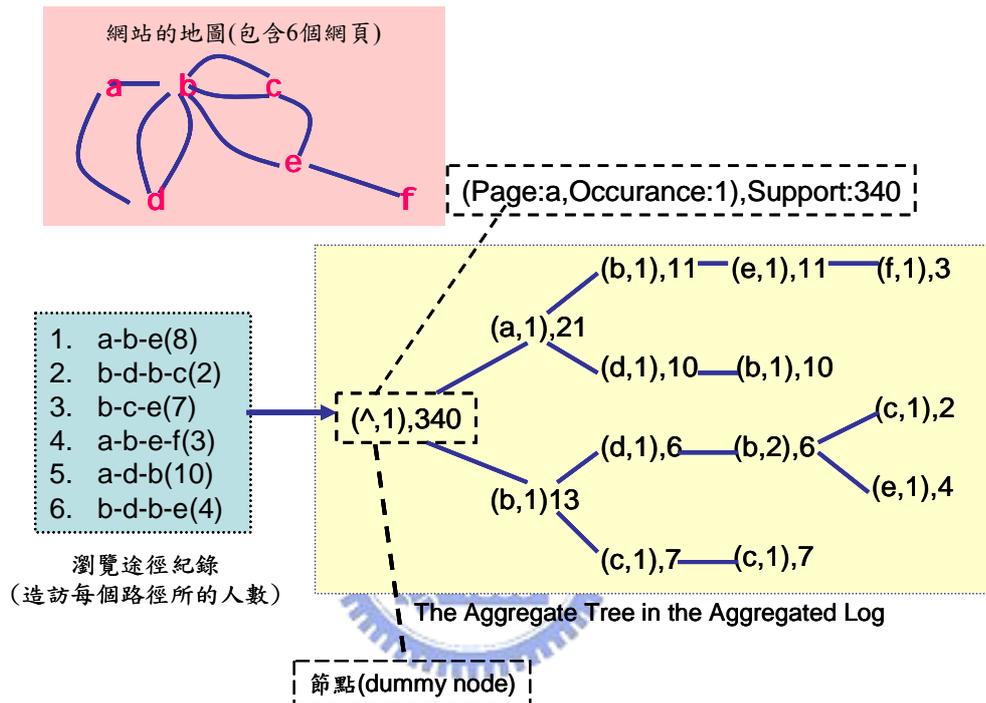


圖 2.6.4 Aggregated Log 所包含的架構

第三塊模組 WUM_gseqm，負責挖掘瀏覽者的序列模式，也就是瀏覽模式，使用的方式是運用「MINT-query」以及包含的「g-sequence」。在挖掘大序列時，大序列中的各項目在實際情況中是未必緊鄰的狀況[27]。例如六張網頁 A, B, C, D, E, F 依據使用的紀錄得到下列序列{(A,B,C,D), (B,A,E,D), (A,B,C,E,F,D)}。可以得知 AD 出現 3 次，為大序列，即使 A、D 之間在瀏覽途徑中並沒有彼此相鄰的情形，也就是 A 網頁不會直接跳到 D 網頁。因此為了在日後解讀瀏覽模式時不會發生錯誤，並且能得到更詳記得結果，WUM 使用 g-sequence，容許探討大序列間的項目之路徑。例如：**g-sequence B*[1; 5] D**，代表 B 網頁和 D 網頁之間還經過 1~5 張瀏覽的網頁，倘若使用 MINT 查詢後，得到的瀏覽模式如圖 2.6.5。

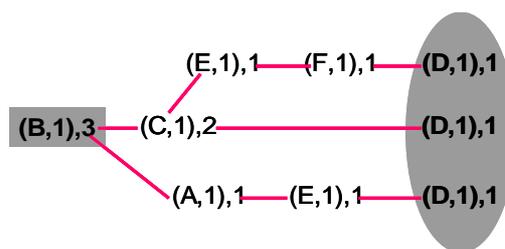


圖 2.6.5 g-sequence B*[1; 5] D 的瀏覽模式

MINT 是挖掘序列模式的語法(A Mining Query Language)，由 WUM 使用者自行定義有興趣的標的（例如配合 g-sequence 的使用）以及設定支持度閾值。因此含有下列四大挖掘功能，依照順序每個功能都是上一個功能的延伸，以下舉例分別說明之：

1. 查詢簡單的統計資訊。

目標是找出瀏覽某張網頁後，接下來會再瀏覽的下一張網頁為何。因此使用圖 2.6.6 的語法，查詢的目標是「找出瀏覽過 CORBA 的網址 a，網址 a 至少有 100 次的紀錄，下一步至少有 10% 的機會會瀏覽的網址為何？」，結果回得到所有符合條件的 a 網址與 b 網址。

```
SELECT a.url , b.url
FROM NODE AS a b, TEMPLATE a*b
WHERE a.support > 100
AND a.title LIKE "%Corba"
AND b.support / a.support > 0.1
```

圖 2.6.6 a MINT query(1)

2. 查詢出新的瀏覽樹。

由圖 2.6.6 的查詢結果，並不能得知網址 a 和網址 b 之間的瀏覽途徑究竟是如何移動的資訊。因此圖 2.6.7 用相同的查詢，加入模版(template)名稱為 t，找出所有起點為 a 網址，終點為 b 網址以及其所包含的路徑。

```
SELECT t
FROM NODE AS a b, TEMPLATE a* [1;5] b AS t
WHERE a.support > 100
AND a.title LIKE "%Corba"
AND b.support / a.support > 0.1
```

圖 2.6.7 a MINT query(2)

3. 結合目標瀏覽樹而成為一個瀏覽模式。

由圖 2.6.7 查詢的結果可視為是一個個的瀏覽模版，圖 2.6.8 進一步使用 GLUE 功能，目的是結合所有的模版，形成一個瀏覽模式。再加入 GLUE 後，會將瀏覽樹上所有具有相同值的節點結合在一起，也就是將大序列轉變成最大序列，新的節點其支持度為原來所有節點的總和。

```
SELECT GLUE(t)
FROM NODE AS a b, TEMPLATE a* [1;5] b AS t
WHERE a.support > 100
AND a.title LIKE "%Corba"
AND b.support / a.support > 0.1
```

圖 2.6.8 a MINT query(3)

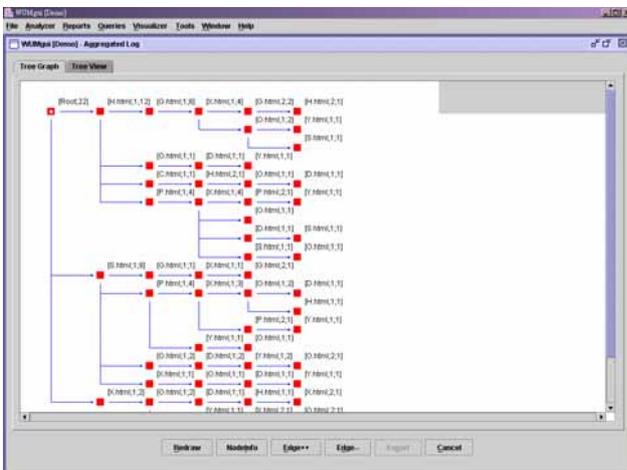
4. 尋找來回瀏覽的網頁。

Tauscher 和 Greenberg(1997) 對於瀏覽者在網站的研究中發現，在一個議期內約有 58% 的網頁會被再次瀏覽。因此 MINT 查詢節點出現次數的資訊，來找出重複被瀏覽的網頁。以圖 2.6.9 的語法為例，則是詢問網址 a，而且再次瀏覽的比例要高於 10%，其餘條件則和上面例題相同。

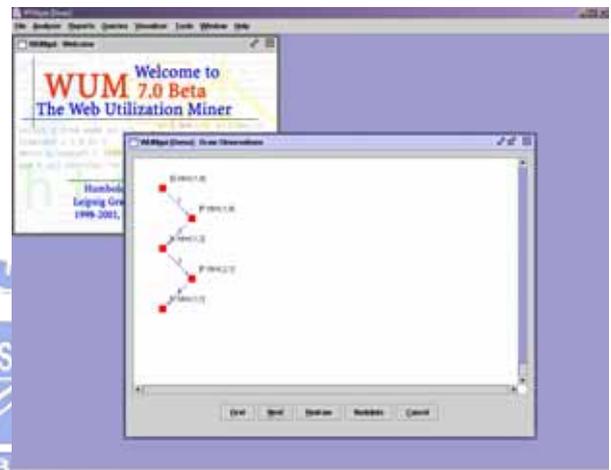
```
SELECT b
FROM NODE AS a b, TEMPLATE *a* b
WHERE a.page = b.page
AND a.title LIKE "%Corba"
AND a.support > 100
AND a.occurance = 1
AND b.occurance = 2
AND b.support / a.support > 0.1
```

圖 2.6.9 a MINT query(4)

第四塊模組 WUM_visualizer 主要是將結果模組二以及模組三的結果以視覺化的方式呈現。以圖 2.6.10(a)即模組二匯集的 Aggregated Log，如果經過模組三以 MINT 查詢後，也可呈現瀏覽模式的架構。另外也可以針對資料庫做每一個序列的觀察，或是針對每一位網站瀏覽者的瀏覽路徑做觀察，分別見圖 2.6.10(b)所示。



(a) Aggregated Log 的一部份



(b) 序列的視覺化呈現

圖 2.6.10 序列的觀察

第三章 漁業資訊分享熱線的資料倉儲設計與實作

在建構漁業資訊分享熱線之前，首先必須建構不同主題的資料倉儲，作為線上分析處理的資料來源。本研究建構了四個主題的資料倉儲，內容涵蓋了水產加工製品的生產量與產值，養殖漁業的生產量、產值與面積，遠洋漁獲生產量，以及遠洋漁獲在國外銷售的產量、產值並結合了該漁船的船位資訊。本章節介紹各資料倉儲的設計與實作，在第 3.1 節針對來源資料的性質作描述，並且分析原始系統的各项問題點，這包含了三大系統的資料之描述。在第 3.2 節介紹 遠洋魚貨銷售與漁船軌跡資料倉儲的設計與實作，在第 3.3 節介紹魚貝苗養殖資料倉儲的設計與實作，在第 3.4 節介紹水產加工製品資料倉儲的設計與實作，在第 3.5 節介紹遠洋漁獲資料倉儲的設計與實作，在第 3.6 節說明各資料倉儲的研究限制。

3.1 來源資料的性質與問題描述

3.1.1 漁業統計資料庫系統

本研究建立資料倉儲的主要來源是漁業署的漁業統計資料庫，而該系統使用 IBM DB2 Universal Database，資料庫系統的網路架構屬於胖從端(Fat Server)的「終端機架構」。資料庫中的資料包含了「台灣區漁會及會員人數的統計」、「台閩地區漁戶數以及漁業從業人數的統計」、「漁船數量統計」(包括動力漁船數、新造動力漁船數、無動力舢舨數、動力漁筏數、無動力漁筏數)、「漁業生產量統計」(包括遠洋漁業、近海漁業、沿岸漁業、養殖漁業)、「重要魚市場魚貨供銷量及價格」、「水產品進口與出口貿易量值」、「水產養殖面積」、「魚貝苗的產量及價值」、「台閩地區水產製品量與價值」(包括食用、非食用、罐頭製品)、「台閩地區水產加工廠數統計」、「製冰冷藏及凍結概況統計」、「遭難漁船數與漁民數統計」以及「漁業固定投資情形資料統計」。合計有 13 大主題，描述資料屬性的文字資料表共有 47 張，例如行政地區、漁會單位、魚種類等；而彙總型的數值資料表則含有 43 張。因此該統計資料庫至少可以建置 13 個以上 43 個以下的資料倉儲。而本研究選定其中漁業生產量統計、魚貝苗的產量及價值以及水產製品量與價值三個主題，並且結合了其他漁業的資料庫系統。

漁業統計資料庫的資料是來自各鄉鎮區公所以及各漁會的調查資料，送至各縣市政府資料庫進行單機作業的整理後，由各縣市政府每月送傳文字壓縮檔(檔案的格式為 *.001)至漁業署的漁業統計資料庫作新增、刪除、修改的資料管理，流程如圖 3.1.1 所示，而各縣市政府傳送的文字檔如圖 3.1.2 所示。

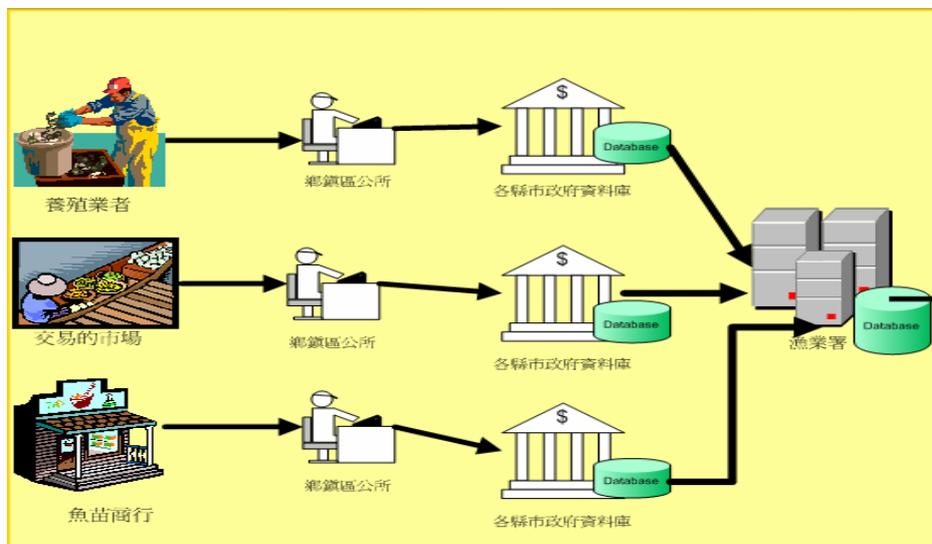


圖 3.1.1 漁業統計資料庫之資料來源

檔案名	編碼	格式	說明
001	0004	000	0091001 0091102 0 1 005 000 4 00200 40,000 0.0,000 40,000 0
002	0004	000	0091001 0091102 0 1 005 000 5 00200 52,000 0.0,000 52,000 0
003	0004	000	0091001 0091102 0 1 005 000 6 00200 42,000 5.0,000 57,000 0
004	0004	000	0091001 0091102 0 1 005 000 7 00200 30,000 5.0,000 42,000 0
005	0004	000	0091001 0091102 0 1 002 000 1 00200 95,000 15.0,000 120,000 0
006	0004	000	0091001 0091102 0 1 004 000 1 00200 30,000 5.0,000 42,000 0
007	0004	000	0091001 0091102 0 1 004 000 2 00200 35,000 0.0,000 42,000 0
008	0004	000	0091001 0091102 0 1 004 000 3 00200 42,000 5.0,000 45,000 0
009	0004	000	0091001 0091102 0 1 005 000 1 00200 30,000 5.0,000 42,000 0
010	0004	000	0091001 0091102 0 1 005 000 2 00200 20,000 0.0,000 35,000 0
011	0004	000	0091001 0091102 0 1 005 000 0 00200 95,000 15.0,000 120,000 0
012	0004	000	0091001 0091102 0 1 005 000 1 00200 0.0,000 0.0,000 0.0,000 0
013	0004	000	0091001 0091102 0 1 006 000 1 00200 30,000 5.0,000 45,000 0
014	0004	000	0091001 0091102 0 1 006 000 2 00200 35,000 0.0,000 42,000 0
015	0004	000	0091001 0091102 0 1 006 000 3 00200 45,000 0.0,000 70,000 0
016	0004	000	0091001 0091102 0 1 006 000 4 00200 40,000 0.0,000 55,000 0
017	0004	000	0091001 0091102 0 1 006 000 5 00200 42,000 5.0,000 52,000 0
018	0004	000	0091001 0091102 0 1 006 000 6 00200 30,000 5.0,000 45,000 0
019	0004	000	0091001 0091102 0 1 006 000 7 00200 95,000 15.0,000 120,000 0
020	0004	000	0091001 0091102 0 1 007 000 1 00200 32,000 5.0,000 40,000 0
021	0004	000	0091001 0091102 0 1 007 000 2 00200 25,000 0.0,000 30,000 0
022	0004	000	0091001 0091102 0 1 007 000 3 00200 40,000 0.0,000 45,000 0
023	0004	000	0091001 0091102 0 1 007 000 4 00200 50,000 0.0,000 55,000 0
024	0004	000	0091001 0091102 0 1 007 000 5 00200 42,000 5.0,000 50,000 0
025	0004	000	0091001 0091102 0 1 008 000 1 00200 0.0,000 0.0,000 0.0,000 0
026	0004	000	0091001 0091102 0 1 100 000 1 00200 30,000 5.0,000 45,000 0
027	0004	000	0091001 0091102 0 1 100 000 2 00200 20,000 0.0,000 30,000 0
028	0004	000	0091001 0091102 0 1 100 000 3 00200 45,000 0.0,000 70,000 0
029	0004	000	0091001 0091102 0 1 100 000 4 00200 40,000 0.0,000 55,000 0
030	0004	000	0091001 0091102 0 1 100 000 5 00200 42,000 5.0,000 48,000 0
031	0004	000	0091001 0091102 0 1 100 000 6 00200 30,000 5.0,000 45,000 0
032	0004	000	0091001 0091102 0 1 100 000 7 00200 30,000 0.0,000 45,000 0
033	0004	000	0091001 0091102 0 1 100 000 8 00200 20,000 5.0,000 40,000 0
034	0004	000	0091001 0091102 0 1 100 000 9 00200 45,000 0.0,000 70,000 0
035	0004	000	0091001 0091102 0 1 100 000 0 00200 40,000 5.0,000 48,000 0

圖 3.1.2 各縣市政府傳送的文字檔

然而漁業統計資料庫的資料記錄方式以及資料表的設計，並無法有效、正確、快速地從該資料庫中萃取出資料以進行後續加值應用。原因有下列幾點：

1. 設計用作整合性資料儲存體系為「資料倉儲」會較「線上交易處理」系統佳。

如前言所述，漁業統計資料庫的資料含蓋漁業各項經營的歷史資料，更適合使用資料倉儲系統。相對於線上分析處理的技術，線上交易處理（以下簡稱 OLTP），其主要是架構於傳統資料庫系統之上，其所儲存的資料為即時、最新的資料。OLTP 目的在於精確地回應使用者所下達的每一個查詢指令，且 OLTP 系統中的資料主要是提供組織支援日常交易紀錄之用，用來處理同時進入資料庫的數百或數千筆交易。例如：

- 記錄來自銷售點終端機或 Web 站台所下的訂單。
- 追蹤每日某家漁市場交易中的金額與狀態。
- 記錄漁會會員的基本資料。
- 記錄漁船公司營業執照的持有人。

由於 OLTP 系統中，存取相對較 OLAP 小量的資料，因此個別交易可以迅速完成。雖然 OLTP 系統在記錄維持每日營運所需資料的功能十分優異，OLTP 在資料組織的方式上卻不能簡易、快速地提供管理者在規劃組織工作時所需的資訊。管理者需要的是摘要資訊，從歷史紀錄中分析影響組織或團隊的因素與未來趨勢。他們需要找出影響組織成功的關鍵因素，以及如何調整這些因素來增進企業的成就。例如瞭解每一季、每一年的趨勢如何影響漁獲量、魚貨銷售價，以決策、執行某項工作需要投入多少的人力與資源。

故本研究的主要方向是將漁業統計資料庫的各大主題，實作為資料倉儲系統。

2. 資料庫的資料參考完整性 (Referential Integrity) 不足

資料參考完整性的缺失來自於下來兩點：第一點，本研究的來源資料庫總共有 135 張資料表，包括了 1283 個欄位，538 個主索引建 (PRIMARY KEY)。資料庫的設計雖屬於關聯性資料庫模型 (Relational Database Model)，然而由於資料表欄位的設計 (包括資料型別、欄位特殊限制等) 不當，導致在不同資料表間即使某些欄位具有關聯性，欄位值上無法確保索引鍵數值在資料表內是擁有一致性的。

第二點，該資料庫系統內屬於彙總型的數值資料表含有 43 張，部分資料表的主題是有相關性的。例如同為「水產加工製品產量與產值」主題的數值資料表就包含 2 張 (資料表名稱分別為 fipmard 以及 fimmar)，各張資料表彼此獨立，欄位記錄的方式也不同。此系統下每一張資料表都是一個檔案，為了維護該資料庫系統的運作，必須撰寫各種程式碼來做資料的新增、更新和刪除紀錄。倘若資料庫系統逐漸龐大時，找尋相關的紀錄時就必須蒐集各張獨立的彙總型資料表。另一方面，彙總型資料表與描述資料屬性的文字資料表之間，其關聯處理也必須要設計程式碼。因此在漁業統計資料庫中就包含了 418 條的程式碼用來處理各資料

表間的關聯性以及資料表的維護，卻仍有數值資料的不一致的情形出現。這種資料庫模型的设计會造成系統維護的複雜度提高；在輸入或刪除記錄時，資料表之間的定義關聯參考到不存在的數值；倘若索引鍵值變更了，資料庫中所有指向它的參考並無法做一致地變更。

因此為了強制資料完整性(Data Integrity)，確保資料庫內的資料正確性與品質，本研究使用「三階正規化」技術以達到表 3.1.1 目的：

表 3.1.1 資料完整性

實體完整性 (Entity Integrity)	透過 UNIQUE 條件約束、PRIMARY KEY 條件約束或 IDENTITY 屬性的強制識別碼，定義一資料列作為特定資料表的唯一實體。
參考完整性 (Referential Integrity)	透過外部索引鍵與主索引鍵的關聯，或外部索引鍵與唯一索引鍵的關聯為基礎（也就是透過 FOREIGN KEY 和 CHECK 條件約束），在輸入或刪除記錄時，保存資料表之間的定義關聯。
值域完整性 (Domain Integrity)	包括限制類型（透過資料型別）、格式（透過 CHECK 條件約束和規則）、或可能數值的範圍（透過 FOREIGN KEY 條件約束、CHECK 條件約束、DEFAULT 定義、NOT NULL 定義和規則），維持特定資料行的項目有效性。

3. 使用多個欄位來代表一種屬性，需要極大的儲存空間並且易發生錯誤。

以「魚貝苗產量與產值」資料表為例，在此資料表中需要有五個欄位才能代表一個鄉鎮行政地區。如此的設計在系統資料量越大時，會造成幾個缺點：首先，多個欄位代表一個地區，造成空間上的浪費，儲存資料會佔去相當大的空間。其次，多個欄位之間並沒有解釋該屬性的階層關係（Concept Hierarchies）；所以無法表達出該屬性從低階層次到高階層次的序列。以行政地區為例，我們無法得知台北縣與三重市之間，鄉鎮與縣市的附屬關係；無法得知新竹縣與桃園縣之間，代表同一縣市等級；無法分辨台灣與金門馬祖之間，本島與外島的差異狀況。第三，一個屬性由多個欄位表示，容易造成輸入時的人為誤差。第四，從詳細的事實資料表計算出摘要資料的過程（稱為彙總，Aggregation），因為維度缺乏的階層關係，會發生重複計算的現象，因而導致錯誤的彙總值。

本研究將使用編碼技術與「資料轉換服務」技術，重新為該屬性的階層關係作描述。

3.1.2 漁船管理系統(VAS)

漁船管理系統(Vessel Administrative System, VAS)係由漁業署為了統一管理漁船、船員而建構的資料庫。資料庫總共包含三大部分，分別是 1.漁船管理資料 2.船員管理資料 3.漁船違規資料。本研究使用「漁船管理資料」與其他資料庫系統進行整合。

漁管資料庫包含主要的漁船資訊，共可分為兩大類，1.漁船船籍資料及 2.漁船規格資料。對漁船船籍資料而言，包含了漁船統一編號 CT_NO、漁船設籍縣市、漁船中英文船名、漁船種類、漁船船主、漁船所屬船公司以及船公司資訊。對漁船規格而言，包括了漁船的總噸數、主馬力與副馬力等。

漁船管理系統資料庫主要問題來自於資料表欄位未經過正規化，資料庫屬於「檔案結構資料庫(File Structure)」而非關聯式資料庫，因而造成儲存空間的浪費。其次資料編碼方式相異於其他資料庫系統，不易和其他資料庫做連結。因此本研究將使用「三階正規化」與「資料轉換服務」技術，前項技術用於資料倉儲中維度資料表的設計，後者技術則用於重新為部分欄位屬性制訂新的編碼方式方能連結異質資料庫系統。

3.1.3 漁船監控系統(VMS)與漁船資料倉儲(VDW)

台灣所使用的「漁船監控系統」，根據遠近洋作業形態之不同，分別使用三套不同的系統，回報的各資料庫管理系統分屬於不同單位做管理。方式計有 1.衛星即時船位通報系統(如 ARGOS 及 Inmarsat)、2.GPS & SSB 船位自動回報系統、以及 3.人工 SSB 船位回報系統等三種。其中 Inmarsat 及 ARGOS 系統是採用人造衛星的方式將漁船回報資料傳回地面接收站，而 SSB 系統則是使用機器回報方式將漁船資料回報到岸上的漁業電台；各接收站在接收資料後透過 Internet 將漁船回報資料存到各資料庫管理系統。

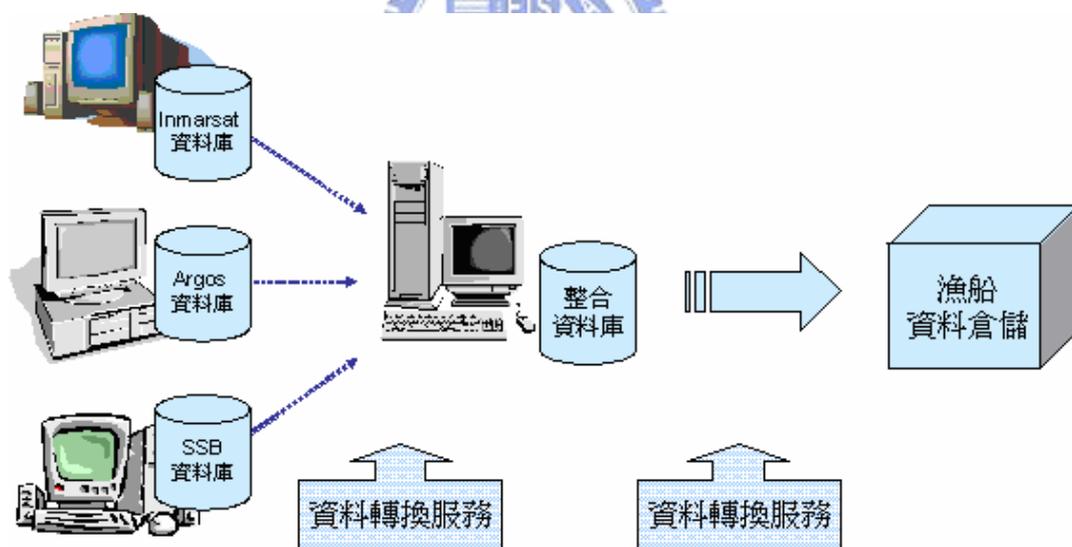


圖 3.1.3 VDW 建構流程[1]

由於各單位所使用的資料庫管理系統也不完全相同，使得三個系統的欄位資料庫屬性存在著分歧。在這樣的情況之下要透過分散式資料庫系統的網路架構來執行，出現管理及技術上的問題存在。有鑑於此，由前人建立一資料庫傳輸界面，蒐集分散漁各資料庫之資料，配合資料庫三階正規化技術、資料轉換服務技術以及倉儲技術資料，建構出一套共包含三階層的整合漁船監控系統資料庫，也就是「漁船資料倉儲」(Vessel Data Warehouse)，如圖 3.1.3 所示[1]。

漁船資料倉儲主要問題來自於下列幾點：

1. 缺乏真實的漁獲量紀錄

漁船資料倉儲系統的漁獲量資訊仰賴於各單位漁船的回報，目前仍無良好的制度設計使得漁民在該所在位置回報正確的漁獲量。本研究在制度未齊全的過渡時期所採取的措施，擬以結合漁業統計資料庫系統，使用該漁船近期在國外基地的銷售量作為其漁獲量。

2. 漁區範圍廣大，分類卻不夠仔細

漁業管理單位將全球各大海域以每 5 度經度和每 5 度緯度劃分為 2304 個漁區代碼。然而漁船資料倉儲系統僅將 2304 個漁區代碼劃分為 6 種海域，對於資料屬性的描述有不足之處。本研究將利用漁區劃分了 17 種海域；以及根據「中華民國對外漁業合作發展協會」繪製的「中華民國遠洋漁業漁場分佈圖」利用漁區劃分了 24 種漁獲區。本研究在 3.2 節會有詳細的描述。



3.2 遠洋魚貨銷售與漁船軌跡資料倉儲的設計與實作

為了有效的建立漁船動態和靜態的資料追蹤以控管漁船的行蹤，政府結合了人造衛星技術、無線電技術、資料庫技術而建立三大系統，即靜態的「漁船管理系統」以記錄漁船及其所屬單位資料，動態的「漁船監控系統」以即時獲得船位資訊，以及整合前述兩大系統的「漁船資料倉儲」。漁船在海上的紀錄除了可以進行船位管理，倘若能夠結合魚貨銷售的資訊，可進一步提供產業界獲取漁業捕獲的專業知識以產生競爭優勢。例如漁船行經軌跡為何可以有較高的漁獲量。因此本研究結合漁業銷售資訊、「漁船資料倉儲」建構「漁業資訊分享熱線」，使用的資訊技術包括三階正規化資料庫、雪花式綱要資料倉儲、資料轉換服務、線上分析處理等技術。可提供漁業管理單位在漁業資訊以及漁船管理能做有效率的查詢，並可利用線上分析處理進行資料探勘，建構漁業決策支援系統。本小節說明「遠洋魚貨銷售與漁船軌跡資料倉儲」的設計與實作。

3.2.1 遠洋魚貨銷售與漁船軌跡資料倉儲的多維度資料模式

由圖 3.2.1 可知，建構本資料倉儲的資料來源來自於三大系統，分別是「漁船資料倉儲」提供漁船船位資訊，「遠洋魚貨銷售記錄表」提供各漁船進行魚貨交易的時間、交易金額等交易記錄，「漁船管理系統」提供漁船本身資訊，如所屬船公司、漁船噸數等紀錄。因為資料來源所屬不同系統，經過一連串複雜的資料轉換服務後，存入魚貨銷售與漁船軌跡資料倉儲中[10]。底下先說明輸入資料倉儲的資料來源格式，接著說明輸出資料倉儲其雪花綱要的各組成資料表，在下一小節中在詳細說明如何使用資料轉換服務技術將各資料庫的資料匯入至資料倉儲中。

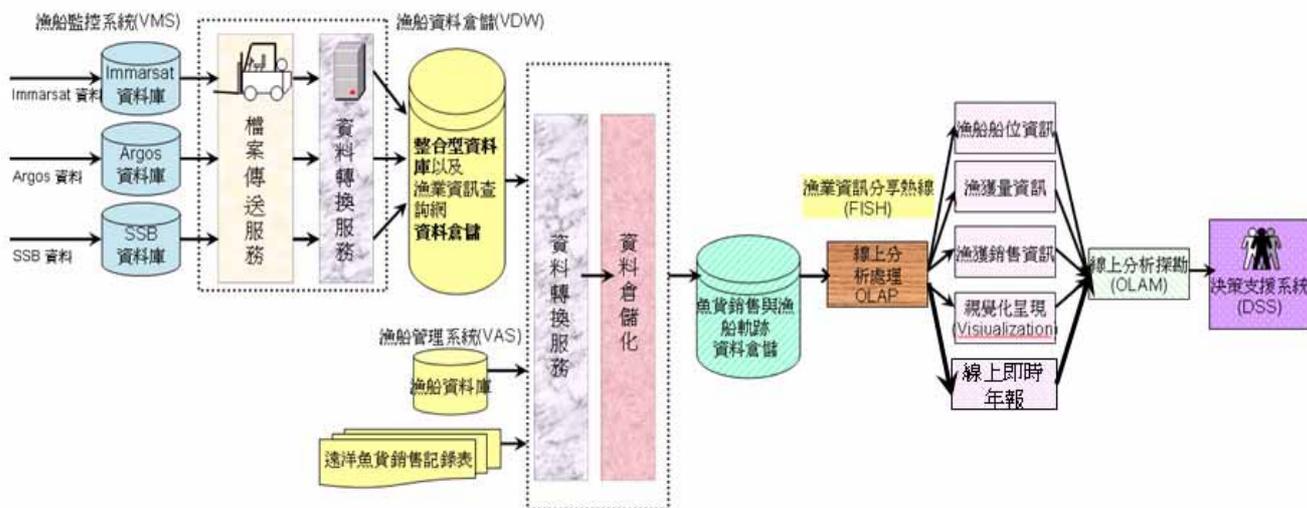


圖 3.2.1 遠洋魚貨銷售與漁船軌跡資料倉儲的架構與資料來源

對於本系統的輸入資料而言，一共有 11 張未經過三階正規化的資料表資料表。分別是來自漁船資料倉儲的「漁船軌跡事實表」共 1 張；來自魚貨銷售統計系統的「魚貨銷售紀錄表 1」、「魚貨銷售紀錄表 2」、「船公司或代理商資料表」、「國外基地資料表」、「漁船資料資料表」、「銷售國別資料表」、「作業海域資料表」、「魚種資料表」、「漁業作業種類資料表」共 9 張；以及來自漁船管理系統的「漁船明細資料表」1 張。各來源資料表的屬性規格分別列在附錄一的附錄表 A.1.1.1~表 A.1.1.11 中。對於本系統的輸出而言，是建構資料倉儲的綱要 (Schema)。綱要規劃的方式是採用多維度資料模型 (Multidimensional Data Model) 的雪花綱要 (Snowflake Schema)。本研究共設計 11 個維度，分別為船主維度、漁船公司維度、國外基地維度、作業海域維度、漁船噸位維度、漁船所屬地區維度、漁業別維度、漁獲區維度、

魚種類維度、銷售國別維度以及時間維度；衡量值有 5 個，分別是交易的數量、交易金額、單位重量的平均價格、經度以及緯度，如果加上延伸出去的維度表則包含 18 張維度資料表，如圖 3.2.2 所示。在附錄二表 A.2.1~表 A.2.11 表中說明事實資料表與各維度資料表的欄位。由於「時間」維度常被資料倉儲視為理所當然，故其圖表的內容在此處沒有表達。

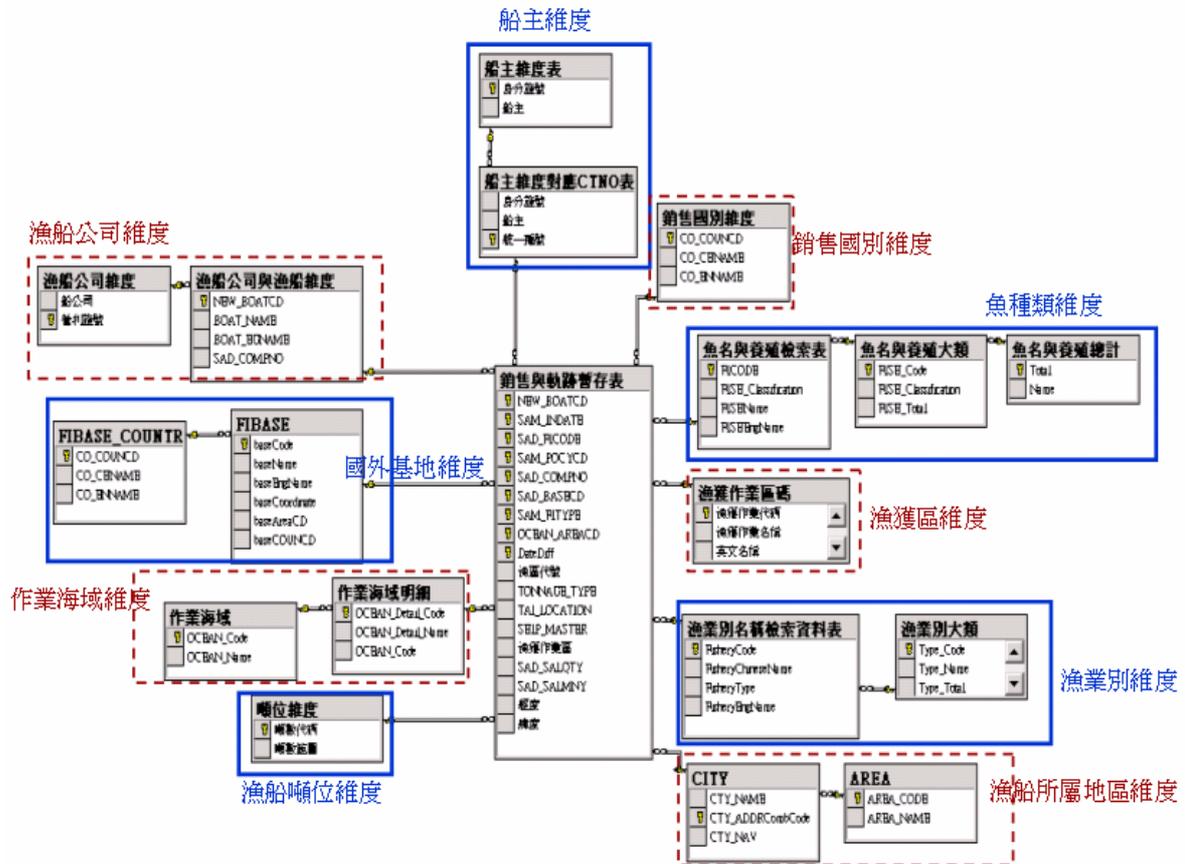


圖 3.2.2 遠洋魚貨銷售與漁船軌跡資料倉儲的雪花網要

對事實資料表而言，這裡儲存了漁船出海的軌跡記錄、航海日期、魚貨交易日期、漁獲種類、銷售的國家、進行交易的國外基地、漁船作業海域、銷售量、銷售金額、單位重量的價格等重要資料。如本文前言所述，由於漁船回報漁獲量發生了隱藏性資訊的情況，無法由漁船監控系統獲得正確有效的漁獲量數據。因此本研究利用進行魚貨交易的銷售量來估計漁船航行時的漁獲量。方法是依據各漁船的魚貨交易日期以及魚貨交易數量，反推找尋最近天數的航海日期，以魚貨交易數量作為航海時漁獲量的估計值。使用語法將於下一節資料轉換服務中說明。

本研究對於建構各維度資料表的來源共分成四大部分，考量資料來源使用哪一個系統主要是基於資料的完整性與充足性。第一部分來自魚貨銷售統計系統的來源資料表，有國外基地維度、魚種類維度、銷售國別維度、時間維度。第二部分來自於漁船管理系統，有船主維度、漁船公司維度、漁船所屬地區維度、漁業別維度。第三部分來自於漁船資料倉儲，有時間維度。第四部分來自於依據漁船航行軌跡之經度與緯度，或是依據漁船登記的屬性而自行建立的維度，有作業海域維度、漁獲區維度、漁船噸位維度。其中時間維度的建構使用 2 個來源，一個是漁船航行海上的時間，另一個是魚貨交易的時間。表 3.2.1 則說明了事實資料表及各維度表屬性內容的資料來源以及是否需要轉換資料格式。本系統建構事實資料表以及各維度表的方式共可分三大方向，一個是維度表的三階正規化，一個是使事實資料表具有統一格式的資料，另一個方向是建構各維度完整的資料以利線上分析處理；前兩大方向皆須藉由資料轉換服務來完成，最後一個方向需要藉由人工查詢與手動輸入的方式來完成。

表 3.2.1 資料倉儲參數

資料表名稱	屬性名稱	資料來源	資料格式轉換
魚貨銷售與軌跡事實資料表	NEW_BOATCD	漁船明細表的統一編號	需要
	SAM_INDATE	魚貨銷售紀錄表 1 的 SAM_INDATE，漁船軌跡事實表的 GPS_時間	需要
	SAD_FICODE	魚貨銷售紀錄表 2 的 SAD_FICODE	不需要
	SAM_POCYCD	魚貨銷售紀錄表 1 的 SAM_POCYCD	不需要
	SAD_COMPNO	漁船明細表的營利證號	需要
	SAD_BASECD	魚貨銷售紀錄表 1 的 SAM_BASECD	需要
	SAM_FITYPE	魚貨銷售紀錄表 1 的 SAM_FITYPE	需要
	OCEAN_AREACD	漁船軌跡事實表的經度、緯度	需要
	DateDiff	魚貨銷售紀錄表 1 的 SAM_INDATE 和漁船軌跡事實表的 GPS_時間，兩者差異的時間	需要
	漁區代號	漁船軌跡事實表的漁區代號	不需要
	TONNAGE_TYPE	漁船明細表的總噸數	需要
	TAI_LOCATION	漁船明細表的設籍縣市	需要
	SHIP_MASTER	漁船明細表的身分證號	需要
	漁獲作業區	漁船軌跡事實表的經度、緯度	需要
	SAD_SALQTY	魚貨銷售紀錄表 2 的 SAD_SALQTY	不需要
	SAD_SALMNY	魚貨銷售紀錄表 2 的 SAD_SALMNY	不需要
	經度	漁船軌跡事實表的經度	不需要
緯度	漁船軌跡事實表的緯度	不需要	
船主維度			
1. 船主維度對應 CTNO 表	統一編號	漁船明細表的漁船船號	需要
	身分證號	漁船明細表的身分證號	不需要
2. 船主維度表	身分證號	漁船明細表的身分證號	不需要
	船主	漁船明細表的船主	需要
漁船公司維度			
1. 漁船公司與漁船維度	NEW_BOATCD	漁船明細表的漁船船號	需要
	BOAT_NAME	漁船明細表的中文船名	不需要
	BOAT_EGNAME	漁船明細表的英文船名	不需要
	SAD_COMPNO	漁船明細表的營利證號	需要
2. 漁船公司維度	營利證號	漁船明細表的營利證號	需要
	船公司	漁船明細表的船主	需要
國外基地維度			
1. FIBASE	baseCode	國外基地的 ba_basecd	不需要
	baseName	國外基地的 ba_basnm	不需要
	baseEngName	自行建立	不需要
	baseCoordinate	自行建立	不需要
	baseAreaCD	自行建立	不需要
	baseCOUNCD	自行建立	不需要
2. FIBASE_COUNTRY	CO_COUNCD	自行建立	不需要
	CO_CHNAME	自行建立	不需要
	CO_ENNAME	自行建立	不需要
作業海域維度			
1. 作業海域明細	OCEAN_Detail_Code	自行建立	不需要
	OCEAN_Detail_Name	自行建立	不需要
	OCEAN_Code	自行建立	不需要
2. 作業海域	OCEAN_Code	自行建立	不需要
	OCEAN_Name	自行建立	不需要

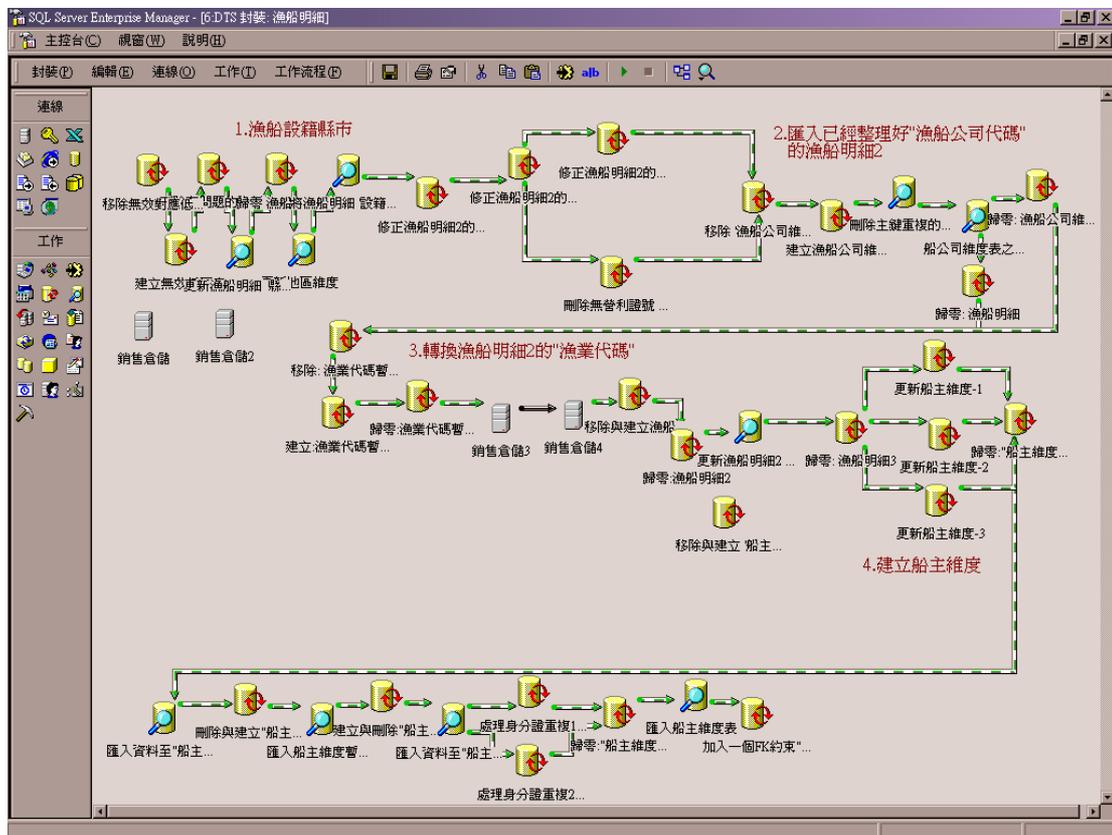
(續)表 3.2.1

漁船噸位維度			
作業海域明細	OCEAN_Detail_Code	自行建立	不需要
	OCEAN_Detail_Name	自行建立	不需要
漁船所屬地區維度			
1.CITY	CTY_ADDRCombCode	DB2ADMIN_ADDR TYPE,NAV,CNT,CTY	需要
	CTY_NAME	DB2ADMIN_ADDR 的 NAME	需要
	CTY_NAV	DB2ADMIN_ADDR 的 NAV	需要
2.AREA	AREA_CODE	DB3ADMIN_ADDR 的 NAV	需要
	AREA_NAME	DB2ADMIN_ADDR 的 NAME	不需要
漁業別維度			
1.漁業別名稱檢索資料表	FisheryCode	FITYPEM 的 TYM_FITYCD	不需要
	FisheryChineseName	FITYPEM 的 TYM_FITYPE	不需要
	FisheryEngName	FITYPEM 的 TYM_ENNAME	不需要
	FisheryType	自行建立	不需要
2.漁業別大類	Type_Code	自行建立	不需要
	Type_Name	FITYPEM 的 TYM_FISECT	不需要
漁獲區維度			
漁獲作業區碼	漁獲作業代碼	自行建立	不需要
	漁獲作業名稱	自行建立	不需要
	英文名稱	自行建立	不需要
魚種類維度			
1.魚名與養殖檢索表	FICODE	FIFISHM 的 FIM_FICODE	不需要
	FISH_Classification	自行建立	不需要
	FISHName	FIFISHM 的 FIM_FISHCH	不需要
	FISHEngName	FIFISHM 的 FIM_FISHEN	不需要
2.魚名與養殖大類	FISH_Code	自行建立	不需要
	FISH_Classification	自行建立	不需要
	FISH_Total	自行建立	不需要
3.魚名與養殖總計	Total	自行建立	不需要
	Name	自行建立	不需要
銷售國別維度			
銷售國別維度	CO_COUNCD	FISOUN 的 CO_COUNCD	不需要
	CO_CHNAME	FISOUN 的 CO_CHNAME	不需要
	CO_ENNAME	FISOUN 的 CO_ENNAME	不需要
時間維度			
	年	魚貨銷售紀錄表 1 的 SAM_INDATE, 漁船軌跡事實表的 GPS_時間	需要
	季	魚貨銷售紀錄表 1 的 SAM_INDATE, 漁船軌跡事實表的 GPS_時間	需要
	月	魚貨銷售紀錄表 1 的 SAM_INDATE, 漁船軌跡事實表的 GPS_時間	需要
	日	魚貨銷售紀錄表 1 的 SAM_INDATE, 漁船軌跡事實表的 GPS_時間	需要

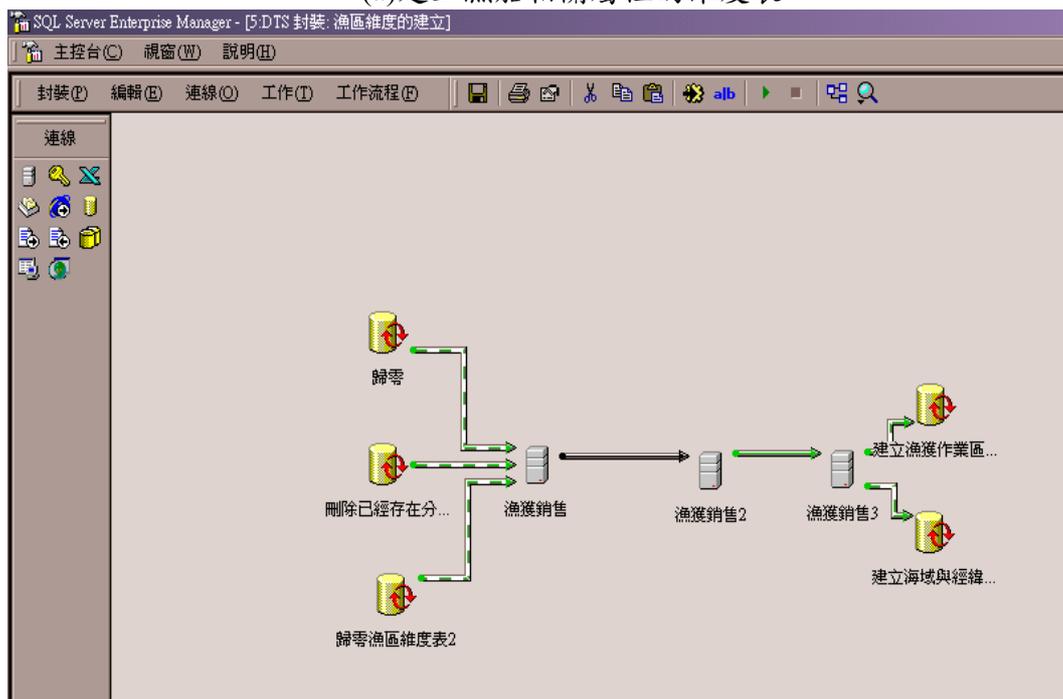
3.2.2 遠洋魚貨銷售與漁船軌跡資料倉儲的資料轉換服務技術

由前文可知建構漁業資訊查詢網需要三層資料庫，分別為漁船監控系統的資料庫、漁船資料倉儲的資料庫以及漁業統計資料庫，目標是建構魚貨銷售與漁船軌跡資料倉儲。因此核心技術在於利用資料倉儲技術，來建構規格一致的三階正規化的整合性資料。為了使這三層資料庫中的資訊能流通，蒐集分散於各資料庫的資料，因此在各層使用不同的資訊傳遞技術，分別為第一、二、三層資料庫間的檔案傳送服務及資料轉換服務技術，三層資料庫間的資料轉換服務技術及資料倉儲化技術。建立圖 4 資料倉儲雪花綱要，各張資料表的屬性內容

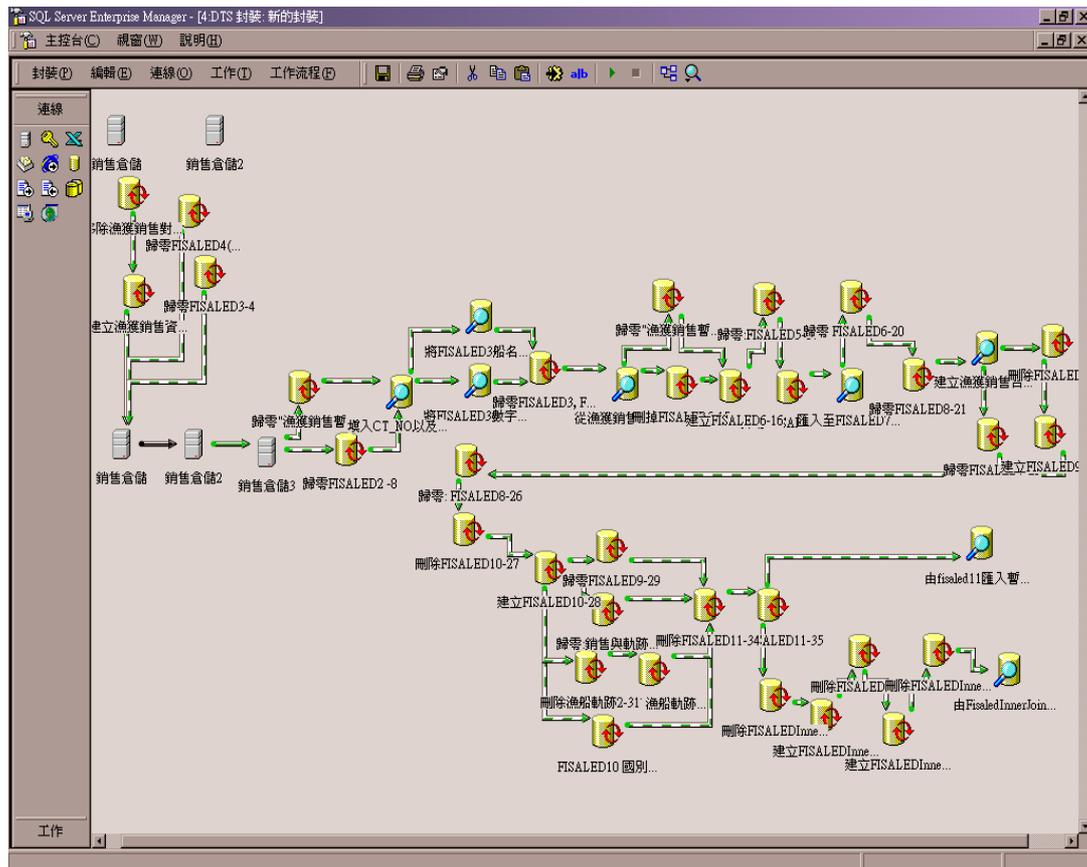
均可以透過資料轉換服務來產生，其中包含了資料的篩選、建立與格式統一。由於本研究採用微軟公司 SQL Sever 2000 軟體來實作資料倉儲，並且以該軟體提供的資料轉換服務做為實作工具。本研究共包含了三大資料轉換服務模組，各模組的目的分別是建立事實資料表、建立漁船相關屬性的維度資料表以及建立漁船航行軌跡相關屬性的維度資料表。包含四個步驟：設定資料來源、設定資料目的地、設定轉換方式(例如可用程式來修改資料格式)、設定執行方式[11]。其資料轉換服務的畫面如圖 3.2.3(a)、5(b)、5(c)所示。



(a) 建立漁船相關屬性的維度表



(b) 建立漁船航行軌跡相關屬性的維度表



(c) 建立事實資料表

圖 3.2.3 資料轉換服務

對於建立漁船相關屬性的維度資料表，如圖 3.2.3(a)，依照處理的先後順序包含下列 4 個維度：漁船所屬地區維度、漁船公司維度、漁業別維度、船主維度。圖 3.2.4 則說明了程式設計的流程以及邏輯架構。以漁船公司維度建立為例，目標是由來源資料建立三階正規化的漁船公司維度表，以公司營利證號作為主鍵，並且使該維度可以進行更新的能力。

資料的來源是漁船明細表中的「船公司」與「營利證號」2 個欄位。首先假設此兩個欄位的存在狀態，有 4 種可能，分別為(船公司，營利證號)是(空值，空值)、(空值，非空值)、(非空值，空值)以及(非空值，非空值)。前面兩種情況，可以利用圖 3.2.5(a)(b)提供的 SQL 語法將欄位值作更新。第三種情況由於並不知道該公司的營利證號，最好的方式是到經濟部商業司查詢各公司的登記資料，如圖 3.2.6(a)(b)所示的入口網站以及查詢結果，並將所得的資料以手動的方式輸入至資料表中。

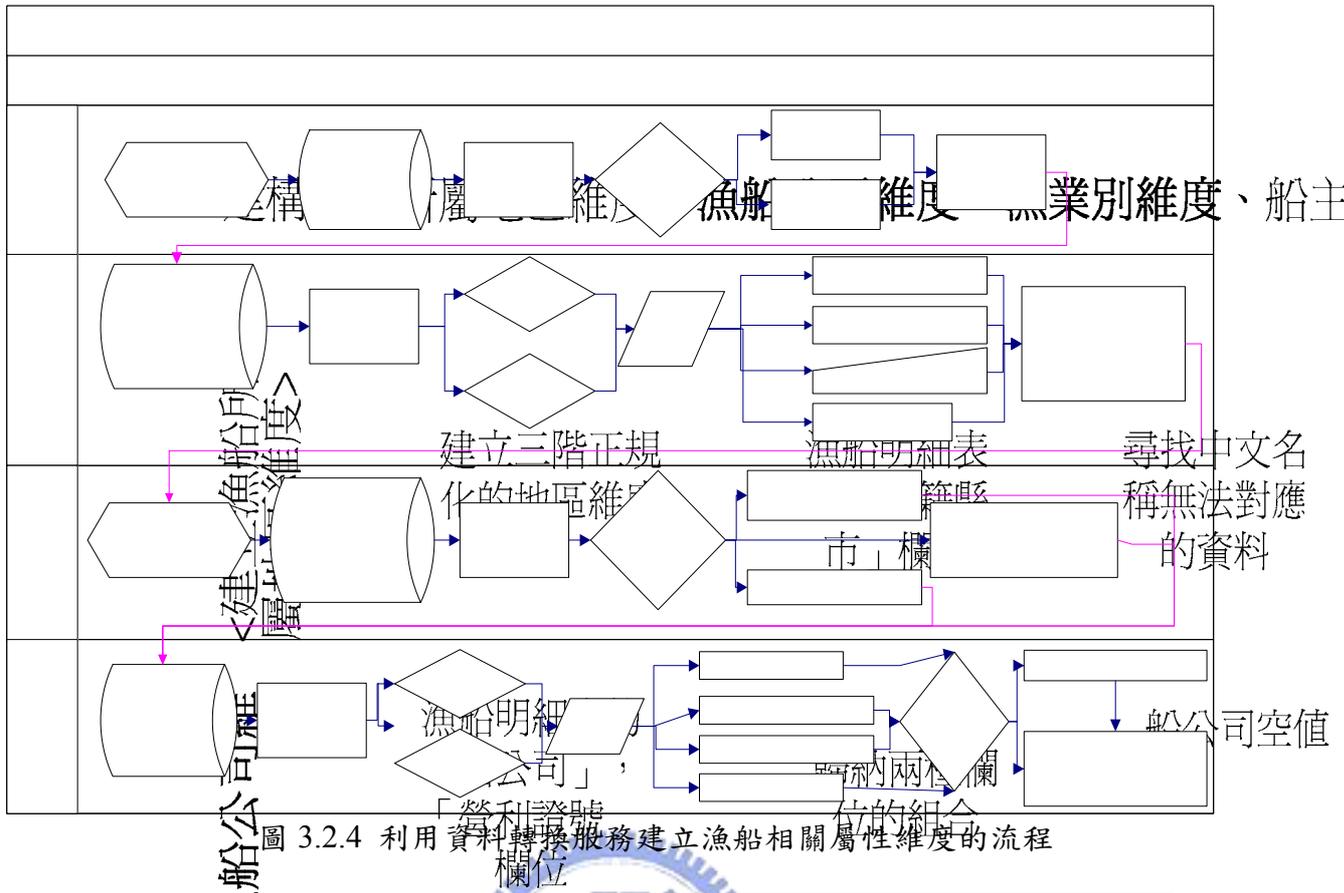


圖 3.2.4 利用資料轉換服務建立漁船相關屬性維度的流程



(a) 欄位均為空值的更新



(b) 船公司欄位空值的更新

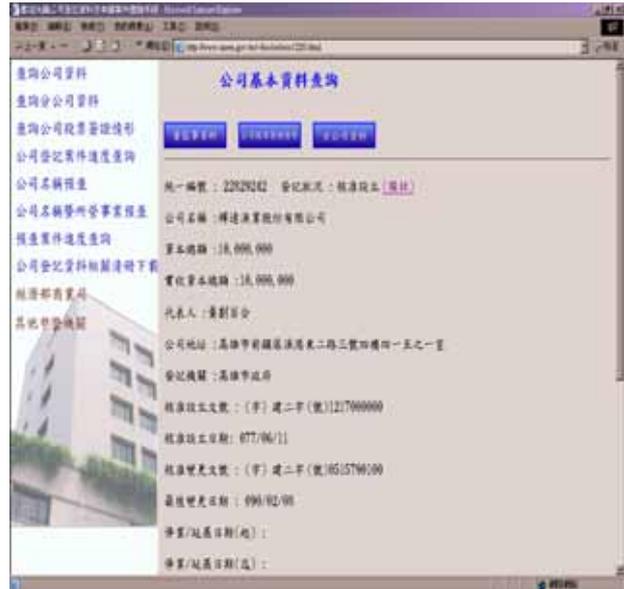
圖 3.2.5 SQL 更新

手動輸入的情況下，可以不予以攝身份證號，程式仍會進行後續步驟，至到最後步驟將主鍵為空值的資料作預設更新的動作，順利完成系統目標。利用圖 3.2.7(a)的「SQL 查詢」將整理好的欄位資料擷取至維度暫存表中，再利用圖 3.2.7(b)的「SQL 查詢」找出發生主鍵重複的資料後，將重複性的資料從來源資料表中刪除，如圖 3.2.7(c)的欄位資料轉換圖，以及圖 3.2.7(d)的程式碼。經過上一步驟，使得維度暫存表內的資料內容是新增的而且主鍵不重複的資料。接下來的步驟，目標是使漁船公司維度表可以加入更新後的資料，資料的來源是維度暫存表，目的則是漁船維度表，見圖 3.2.8(a)及圖 3.2.8(b)，以及圖 3.2.8(c)的程式碼。為了每一次程式的運轉均能夠執行以及不用處理已經更新過的資料，故將維度暫存表的資料予

以清除，如圖 3.2.8(d)所示，這時漁船公司維度表則順利建立。



(a)查詢公司登記資料



(b)查詢結果

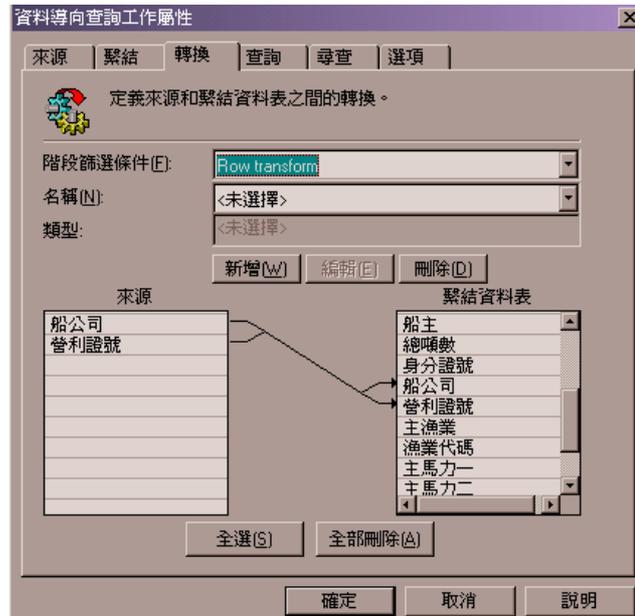
圖 3.2.6 經濟部商業司查詢系統



(a) 資料擷取



(b) 找出發生主鍵重複的資料



(c) 欄位對應圖

圖 3.2.7 處理主鍵重複的步驟

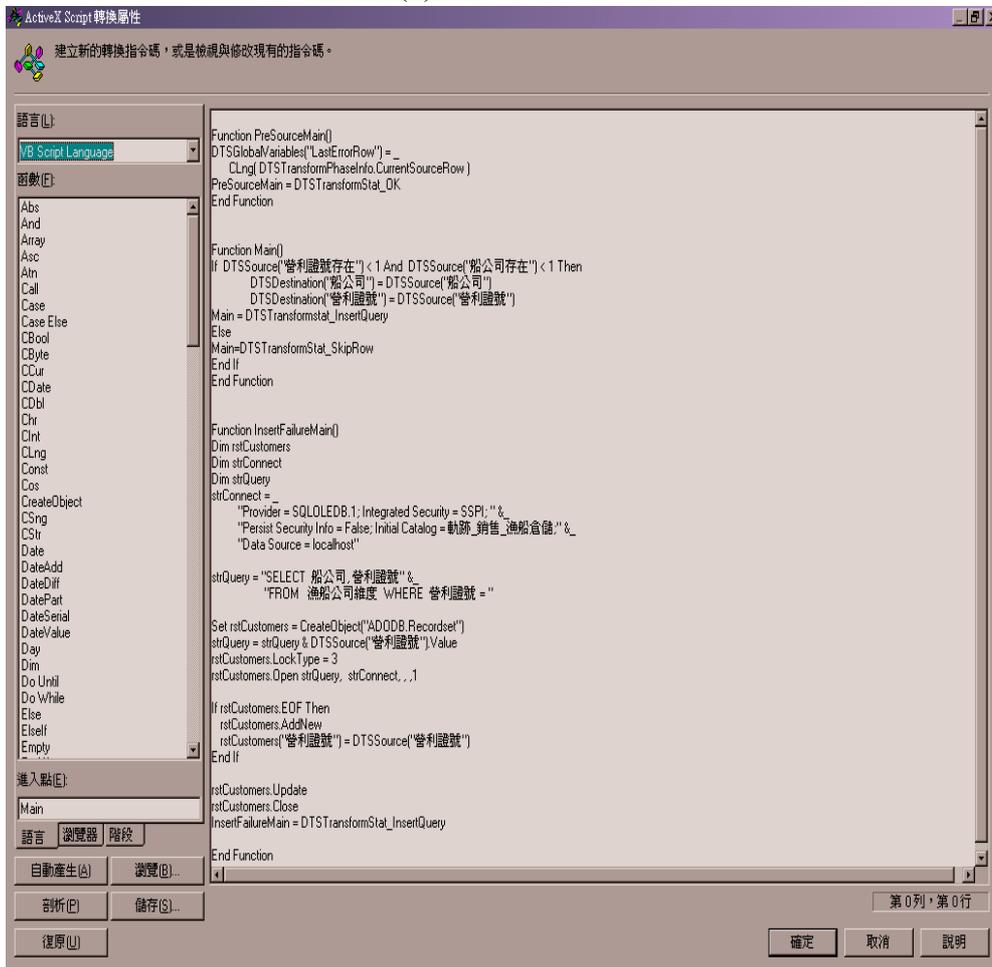
對於建立漁船航行軌跡相關屬性的維度資料表，如圖 3.2.3(b)，主要目標是利用經度及緯度的劃分以建立階層架構(Concept Hierarchy)的維度，其中包括下列 2 個維度：作業海域維度以及漁獲區維度，如圖 3.2.9 所示 2 種維度代表的地理意義。本研究依據漁業管理單位制訂的地圖[3]建立各維度內的項目(Category)。



(a) 資料擷取



(b) 欄位對應圖



(c) 程式碼

圖 3.2.8 更新維度表的步驟

程式實作的設計流程以及邏輯架構見圖 3.2.10。首先利用圖 3.2.11(a)(b)提供的預存程序 (stored procedure)依照經度以及緯度範圍各分割了 72 格水平區域代碼及 32 格垂直區域代碼並存放在 2 張表格中，如圖 3.2.11(c)(d)所示。接著使用 SQL 語法將兩張表格作交叉聯結(cross join)，而得到 2304 塊的漁區代碼。下一步驟就作業海域維度而言，由地理上對各大海洋範圍的定義而劃分出 17 大海洋，以建立東北太平洋海域為例使用的語法見圖 3.2.12。



圖 3.2.9 維度代表的地理意義¹

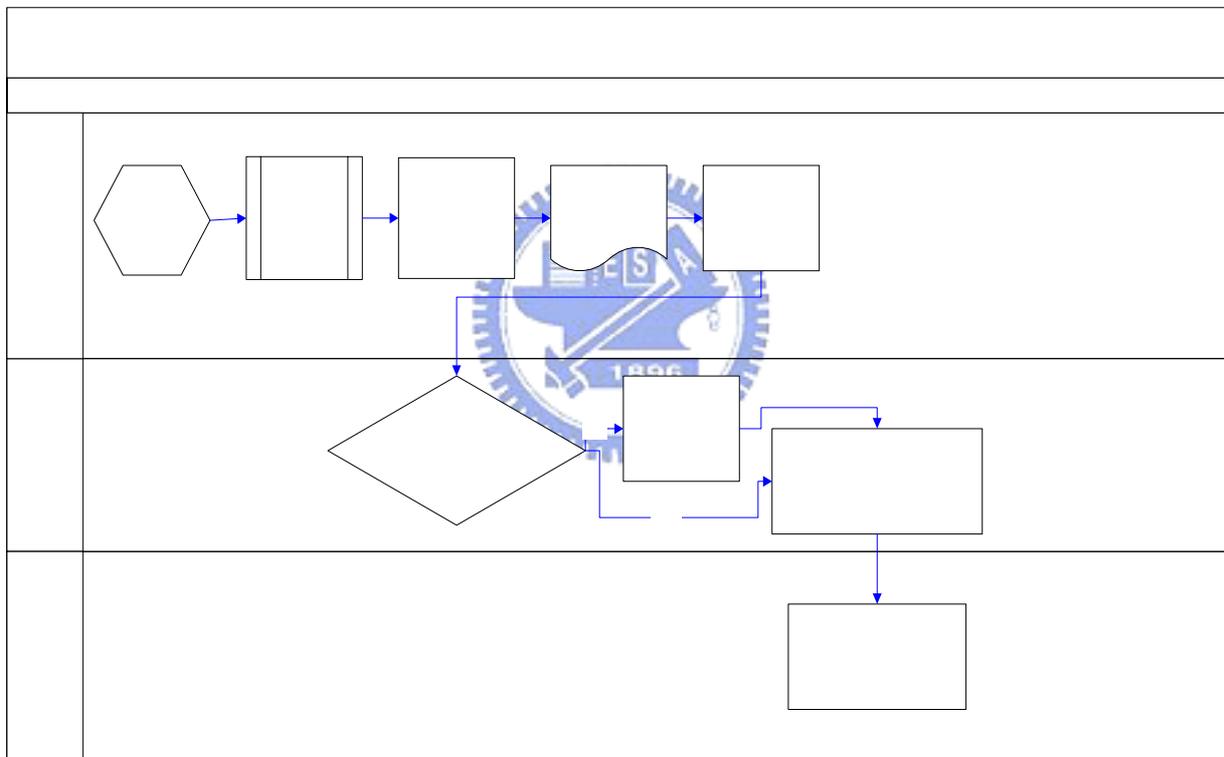


圖 3.2.10 利用資料轉換服務及預存程序建立漁船航行軌跡相關屬性維度的流程

<漁船航行軌跡相關屬性的維度>

建立漁區代碼

所有
資料
表
歸零

分別將經
度、緯度
畫成72區
和32區

結合經度與
緯度，得到
2304格漁
區

參考地圖
劃分漁區
制訂範

¹ 本地圖由航空測量及遙測探測學會於1996年6月編制



(a) 建立經度的預存程序



(b) 建立緯度的預存程序

Horizontal_Area	begin_code	end_code
02	0	4.9999
04	5	9.9999
06	10	14.9999
08	15	19.9999
10	20	24.9999
12	25	29.9999
14	30	34.9999
16	35	39.9999
18	40	44.9999
20	45	49.9999
22	50	54.9999
24	55	59.9999
26	60	64.9999
28	65	69.9999
30	70	74.9999
32	75	79.9999
34	80	84.9999
36	85	89.9999
38	90	94.9999
40	95	99.9999
42	100	104.9999
44	105	109.9999
46	110	114.9999
48	115	119.9999
50	120	124.9999
52	125	129.9999
54	130	134.9999
56	135	139.9999
58	140	144.9999
60	145	149.9999
62	150	154.9999
64	155	159.9999
66	160	164.9999
68	165	169.9999
70	170	174.9999
72	175	179.9999
01	0	-4.9999
03	-5	-9.9999
05	-10	-14.9999
07	-15	-19.9999
09	-20	-24.9999
11	-25	-29.9999
13	-30	-34.9999
15	-35	-39.9999

(c) 經度劃分與建立代碼

Vertical_Area	begin_code	end_code
73	0	4.9999
75	5	9.9999
77	10	14.9999
79	15	19.9999
81	20	24.9999
83	25	29.9999
85	30	34.9999
87	35	39.9999
89	40	44.9999
91	45	49.9999
93	50	54.9999
95	55	59.9999
97	60	64.9999
99	65	69.9999
01	70	74.9999
03	75	79.9999
05	80	84.9999
07	85	89.9999
09	90	94.9999
11	95	99.9999
13	100	104.9999
15	105	109.9999
17	110	114.9999
19	115	119.9999
21	120	124.9999
23	125	129.9999
25	130	134.9999
27	135	139.9999
29	140	144.9999
31	145	149.9999
33	150	154.9999
35	155	159.9999
37	160	164.9999
39	165	169.9999
41	170	174.9999
43	175	179.9999
45	180	184.9999
47	185	189.9999
49	190	194.9999
51	195	199.9999
53	200	204.9999
55	205	209.9999
57	210	214.9999
59	215	219.9999
61	220	224.9999
63	225	229.9999
65	230	234.9999
67	235	239.9999
69	240	244.9999
71	245	249.9999
73	250	254.9999
75	255	259.9999
77	260	264.9999
79	265	269.9999
81	270	274.9999
83	275	279.9999
85	280	284.9999
87	285	289.9999
89	290	294.9999
91	295	299.9999
93	300	304.9999
95	305	309.9999
97	310	314.9999
99	315	319.9999
01	320	324.9999
03	325	329.9999
05	330	334.9999
07	335	339.9999
09	340	344.9999
11	345	349.9999
13	350	354.9999
15	355	359.9999
17	360	364.9999
19	365	369.9999
21	370	374.9999
23	375	379.9999
25	380	384.9999
27	385	389.9999
29	390	394.9999
31	395	399.9999
33	400	404.9999
35	405	409.9999
37	410	414.9999
39	415	419.9999
41	420	424.9999
43	425	429.9999
45	430	434.9999
47	435	439.9999
49	440	444.9999
51	445	449.9999
53	450	454.9999
55	455	459.9999
57	460	464.9999
59	465	469.9999
61	470	474.9999
63	475	479.9999
65	480	484.9999
67	485	489.9999
69	490	494.9999
71	495	499.9999
73	500	504.9999
75	505	509.9999
77	510	514.9999
79	515	519.9999
81	520	524.9999
83	525	529.9999
85	530	534.9999
87	535	539.9999
89	540	544.9999
91	545	549.9999
93	550	554.9999
95	555	559.9999
97	560	564.9999
99	565	569.9999
01	570	574.9999
03	575	579.9999
05	580	584.9999
07	585	589.9999
09	590	594.9999
11	595	599.9999
13	600	604.9999
15	605	609.9999
17	610	614.9999
19	615	619.9999
21	620	624.9999
23	625	629.9999
25	630	634.9999
27	635	639.9999
29	640	644.9999
31	645	649.9999
33	650	654.9999
35	655	659.9999
37	660	664.9999
39	665	669.9999
41	670	674.9999
43	675	679.9999
45	680	684.9999
47	685	689.9999
49	690	694.9999
51	695	699.9999
53	700	704.9999
55	705	709.9999
57	710	714.9999
59	715	719.9999
61	720	724.9999
63	725	729.9999
65	730	734.9999
67	735	739.9999
69	740	744.9999
71	745	749.9999
73	750	754.9999
75	755	759.9999
77	760	764.9999
79	765	769.9999
81	770	774.9999
83	775	779.9999
85	780	784.9999
87	785	789.9999
89	790	794.9999
91	795	799.9999
93	800	804.9999
95	805	809.9999
97	810	814.9999
99	815	819.9999
01	820	824.9999
03	825	829.9999
05	830	834.9999
07	835	839.9999
09	840	844.9999
11	845	849.9999
13	850	854.9999
15	855	859.9999
17	860	864.9999
19	865	869.9999
21	870	874.9999
23	875	879.9999
25	880	884.9999
27	885	889.9999
29	890	894.9999
31	895	899.9999
33	900	904.9999
35	905	909.9999
37	910	914.9999
39	915	919.9999
41	920	924.9999
43	925	929.9999
45	930	934.9999
47	935	939.9999
49	940	944.9999
51	945	949.9999
53	950	954.9999
55	955	959.9999
57	960	964.9999
59	965	969.9999
61	970	974.9999
63	975	979.9999
65	980	984.9999
67	985	989.9999
69	990	994.9999
71	995	999.9999
73	1000	1004.9999
75	1005	1009.9999
77	1010	1014.9999
79	1015	1019.9999
81	1020	1024.9999
83	1025	1029.9999
85	1030	1034.9999
87	1035	1039.9999
89	1040	1044.9999
91	1045	1049.9999
93	1050	1054.9999
95	1055	1059.9999
97	1060	1064.9999
99	1065	1069.9999
01	1070	1074.9999
03	1075	1079.9999
05	1080	1084.9999
07	1085	1089.9999
09	1090	1094.9999
11	1095	1099.9999
13	1100	1104.9999
15	1105	1109.9999
17	1110	1114.9999
19	1115	1119.9999
21	1120	1124.9999
23	1125	1129.9999
25	1130	1134.9999
27	1135	1139.9999
29	1140	1144.9999
31	1145	1149.9999
33	1150	1154.9999
35	1155	1159.9999
37	1160	1164.9999
39	1165	1169.9999
41	1170	1174.9999
43	1175	1179.9999
45	1180	1184.9999
47	1185	1189.9999
49	1190	1194.9999
51	1195	1199.9999
53	1200	1204.9999
55	1205	1209.9999
57	1210	1214.9999
59	1215	1219.9999
61	1220	1224.9999
63	1225	1229.9999
65	1230	1234.9999
67	1235	1239.9999
69	1240	1244.9999
71	1245	1249.9999
73	1250	1254.9999
75	1255	1259.9999
77	1260	1264.9999
79	1265	1269.9999
81	1270	1274.9999
83	1275	1279.9999
85	1280	1284.9999
87	1285	1289.9999
89	1290	1294.9999
91	1295	1299.9999
93	1300	1304.9999
95	1305	1309.9999
97	1310	1314.9999
99	1315	1319.9999
01	1320	1324.9999
03	1325	1329.9999
05	1330	1334.9999
07	1335	1339.9999
09	1340	1344.9999
11	1345	1349.9999
13	1350	1354.9999
15	1355	1359.9999
17	1360	1364.9999
19	1365	1369.9999
21	1370	1374.9999
23	1375	1379.9999
25	1380	1384.9999
27	1385	1389.9999
29	1390	1394.9999
31	1395	1399.9999
33	1400	1404.9999
35	1405	1409.9999
37	1410	1414.9999
39	1415	1419.9999
41	1420	1424.9999
43	1425	1429.9999
45	1430	1434.9999
47	1435	1439.9999
49	1440	1444.9999
51	1445	1449.9999
53	1450	1454.9999
55	1455	1459.9999
57	1460	1464.9999
59	1465	1469.9999
61	1470	1474.9999
63	1475	1479.9999
65	1480	1484.9999
67	1485	1489.9999
69	1490	1494.9999
71	1495	1499.9999
73	1500	1504.9999
75	1505	1509.9999
77	1510	1514.9999
79	1515	1519.9999
81	1520	1524.9999
83	1525	1529.9999
85	1530	1534.9999
87	1535	1539.9999

而就漁獲區維度而言，最後建立了 24 大漁獲區，必須事先從地圖上計算該漁獲區所涵蓋的經緯度範圍，並且需處理下列兩種問題：同一塊範圍內會出現漁獲區重疊的情形以及同一個漁獲區分佈於分隔的海域。針對前者情況，如由圖 3.2.9 中可以得知，大目鮪魚區和長鰭鮪魚區在東經 0~10、南緯 20~25 之間有重疊的現象，而不是一塊區域僅有一個漁獲區代碼。故本研究將發生重疊的區域採行編碼的方式是組合所有涵蓋的漁獲區代碼，最高重疊情形是一個漁區來自 3 種漁獲區的組合，見表 3.2.2(a)所整理的情形。此時在撰寫程式必須分成 3 個階段，目的就是為了將重疊的 3 個漁獲區代碼安排在 3 個資料欄中，才能在下一步進行漁獲區代碼的組合，處理的經過見圖 3.2.13(a)(b)解決第一種情況的問題。至於針對後者情況，則需找出各漁獲區所有分佈的範圍，以南方黑鮪漁獲區為例，位於東南大西洋以及南印度洋均有其漁場，整合兩個漁場使用的語法見圖 3.2.13(c)。其中分佈區域最多的漁獲區為分佈 6 大區域的長鰭鮪魚區，見表 3.2.2(b)所整理的情形，以及圖 3.2.13(d)最後建立的資料表。

緯度起點	緯度終點	經度起點	經度終點	漁獲區代碼	漁獲區代碼	漁獲區代碼
-10	-14.9999	-130	-134.9999	ALB	YFT	BET
-10	-14.9999	-135	-139.9999	ALB	YFT	BET
-10	-14.9999	-140	-144.9999	ALB	YFT	BET
-10	-14.9999	-145	-149.9999	ALB	YFT	BET
-10	-14.9999	-150	-154.9999	ALB	YFT	BET
-10	-14.9999	-155	-159.9999	ALB	YFT	BET
-10	-14.9999	-160	-164.9999	ALB	YFT	BET
-10	-14.9999	-165	-169.9999	ALB	YFT	BET
-10	-14.9999	-170	-174.9999	ALB	YFT	BET
-10	-14.9999	-175	-179.9999	ALB	YFT	BET
-10	-14.9999	-180	-184.9999	ALB	YFT	BET
-15	-19.9999	0	4.9999	NULL	NULL	BET
-15	-19.9999	5	9.9999	NULL	NULL	BET
-15	-19.9999	10	14.9999	NULL	NULL	BET
-15	-19.9999	15	19.9999	NULL	NULL	BET
-15	-19.9999	20	24.9999	NULL	NULL	BET
-15	-19.9999	25	29.9999	NULL	NULL	BET
-15	-19.9999	30	34.9999	NULL	NULL	BET
-15	-19.9999	35	39.9999	NULL	NULL	BET
-15	-19.9999	40	44.9999	NULL	NULL	BET
-15	-19.9999	45	49.9999	NULL	NULL	BET
-15	-19.9999	50	54.9999	NULL	YFT	BET
-15	-19.9999	55	59.9999	NULL	YFT	BET
-15	-19.9999	60	64.9999	ALB	YFT	BET
-15	-19.9999	65	69.9999	ALB	YFT	BET
-15	-19.9999	70	74.9999	ALB	YFT	BET
-15	-19.9999	75	79.9999	ALB	YFT	BET
-15	-19.9999	80	84.9999	ALB	YFT	BET
-15	-19.9999	85	89.9999	ALB	YFT	BET
-15	-19.9999	90	94.9999	ALB	YFT	BET
-15	-19.9999	95	99.9999	NULL	YFT	NULL
-15	-19.9999	100	104.9999	NULL	YFT	NULL
-15	-19.9999	105	109.9999	NULL	YFT	NULL
-15	-19.9999	110	114.9999	NULL	YFT	NULL
-15	-19.9999	115	119.9999	NULL	YFT	NULL
-15	-19.9999	120	124.9999	NULL	YFT	NULL
-15	-19.9999	125	129.9999	NULL	YFT	NULL
-15	-19.9999	130	134.9999	NULL	YFT	NULL
-15	-19.9999	135	139.9999	NULL	YFT	NULL
-15	-19.9999	140	144.9999	NULL	YFT	NULL
-15	-19.9999	145	149.9999	NULL	YFT	NULL
-15	-19.9999	150	154.9999	NULL	YFT	NULL
-15	-19.9999	155	159.9999	ALB	YFT	NULL
-15	-19.9999	160	164.9999	ALB	YFT	NULL
-15	-19.9999	165	169.9999	ALB	YFT	NULL

(a) 範圍重疊的漁獲區

漁獲區代碼	漁獲區名稱	英文名稱
ALB	長鰭鮪魚區	Albacore
ALBEBET	長鰭鮪魚區和大目鮪魚區	Albacore, Bigeye tuna
ALBNFS	長鰭鮪魚區和赤魷魚區	Albacore, Neon flying squid
ALBYFT	長鰭鮪魚區和黃鰭鮪魚區	Albacore, Yellowfin tuna
ALBYFTBET	長鰭鮪魚區和黃鰭鮪魚區及大目鮪魚區	Albacore, Yellowfin tuna, Bigeye tuna
ASS	阿根廷魷魚區	Argentine shortfin squid
BET	大目鮪魚區	Bigeye tuna
BFT	黑鮪魚區	Bluefin tuna
JFS	美洲大赤魷魚區	Junco flying squid
NFS	赤魷魚區	Neon flying squid
OTH	未標示漁獲區	Other
SAU	秋刀魚區	Pacific Saury
SAUNFS	秋刀魚區和赤魷魚區	Pacific Saury, Neon flying squid
SBT	南方黑鮪魚區	Southern bluefin tuna
SBYFT	南方黑鮪魚區和黃鰭鮪魚區	Southern bluefin tuna, Yellowfin tuna
SKJ	正鰭鮪魚區	Skippack
SKJBET	正鰭鮪魚區和大目鮪魚區	Skippack, Bigeye tuna
SKJVFTBET	正鰭鮪魚區和大目鮪魚區及黃鰭鮪魚區	Skippack, Yellowfin tuna, Bigeye tuna
TRA	拖網漁區	Trawling
TRAYFT	拖網漁區和黃鰭鮪魚區	Trawling, Yellowfin tuna
TRAYFTBET	拖網漁區和黃鰭鮪魚區及大目鮪魚區	Trawling, Yellowfin tuna, Bigeye tuna
WFS	紐西蘭魷魚區	Wellington flying squid
YFT	黃鰭鮪魚區	Yellowfin tuna
YFTBET	黃鰭鮪魚區和大目鮪魚區	Yellowfin tuna, Bigeye tuna

(b) 完成漁獲區代碼的組合

```

1 **南方黑鮪第一區
2 ElseIf (CInt(DTSSource("V_begin_code")) > -35 And CInt(DTSSource("V_end_code")) < -30_
3 And CInt(DTSSource("H_begin_code")) > 25 And CInt(DTSSource("H_end_code")) < 50)_
4 Then
5     DTSDestination("漁獲區代碼") = "SBT"
6 ElseIf (CInt(DTSSource("V_begin_code")) > -40 And CInt(DTSSource("V_end_code")) < -35_
7 And CInt(DTSSource("H_begin_code")) > 20 And CInt(DTSSource("H_end_code")) < 50)_
8 Then
9     DTSDestination("漁獲區代碼") = "ASS"
10 ElseIf (CInt(DTSSource("V_begin_code")) > -45 And CInt(DTSSource("V_end_code")) < -40_
11 And CInt(DTSSource("H_begin_code")) > 15 And CInt(DTSSource("H_end_code")) < 45)_
12 Then
13     DTSDestination("漁獲區代碼") = "SBT"
14 ElseIf (CInt(DTSSource("V_begin_code")) > -50 And CInt(DTSSource("V_end_code")) < -45_
15 And CInt(DTSSource("H_begin_code")) > 25 And CInt(DTSSource("H_end_code")) < 40)_
16 Then
17     DTSDestination("漁獲區代碼") = "SBT"
18
19 **南方黑鮪第二區
20 ElseIf (CInt(DTSSource("V_begin_code")) > -40 And CInt(DTSSource("V_end_code")) < -35_
21 And CInt(DTSSource("H_begin_code")) > 95 And CInt(DTSSource("H_end_code")) < 110)_
22 Then
23     DTSDestination("漁獲區代碼") = "SBT"
24 ElseIf (CInt(DTSSource("V_begin_code")) > -45 And CInt(DTSSource("V_end_code")) < -40_
25 And CInt(DTSSource("H_begin_code")) > 75 And CInt(DTSSource("H_end_code")) < 110)_
26 Then
27     DTSDestination("漁獲區代碼") = "SBT"
28 ElseIf (CInt(DTSSource("V_begin_code")) > -50 And CInt(DTSSource("V_end_code")) < -45_
29 And CInt(DTSSource("H_begin_code")) > 75 And CInt(DTSSource("H_end_code")) < 105)_
30 Then
31     DTSDestination("漁獲區代碼") = "SBT"

```

(c) 以語法整合分散各地的漁獲區

漁獲區代碼	漁獲區代碼	緯度起點	緯度終點	經度起點	經度終點
7446	OTH	0	-4.9999	110	114.9999
7447	OTH	0	-4.9999	-115	-119.9999
7448	OTH	0	-4.9999	115	119.9999
7449	OTH	0	-4.9999	-120	-124.9999
7450	OTH	0	-4.9999	120	124.9999
7451	OTH	0	-4.9999	-125	-129.9999
7452	OTH	0	-4.9999	125	129.9999
7453	YFTBET	0	-4.9999	-130	-134.9999
7454	TRA	0	-4.9999	130	134.9999
7455	ALBYFTBET	0	-4.9999	-135	-139.9999
7456	OTH	0	-4.9999	135	139.9999
7457	ALBYFTBET	0	-4.9999	-140	-144.9999
7458	SKJBET	0	-4.9999	140	144.9999
7459	ALBYFTBET	0	-4.9999	-145	-149.9999
7460	SKJBET	0	-4.9999	145	149.9999
7461	ALBYFTBET	0	-4.9999	-150	-154.9999
7462	SKJVFTBET	0	-4.9999	150	154.9999
7463	ALBYFTBET	0	-4.9999	-155	-159.9999
7464	SKJVFTBET	0	-4.9999	155	159.9999
7465	ALBYFTBET	0	-4.9999	-160	-164.9999
7466	ALBYFTBET	0	-4.9999	160	164.9999
7467	ALBYFTBET	0	-4.9999	-165	-169.9999
7468	ALBYFTBET	0	-4.9999	165	169.9999
7469	ALBYFTBET	0	-4.9999	-170	-174.9999
7470	ALBYFTBET	0	-4.9999	170	174.9999
7471	ALBYFTBET	0	-4.9999	-175	-179.9999
7472	ALBYFTBET	0	-4.9999	175	179.9999
7501	OTH	5	9.9999	0	-4.9999
7502	OTH	5	9.9999	0	4.9999
7503	OTH	5	9.9999	-5	-9.9999
7504	OTH	5	9.9999	5	9.9999
7505	OTH	5	9.9999	-10	-14.9999
7506	OTH	5	9.9999	10	14.9999
7507	OTH	5	9.9999	-15	-19.9999
7508	OTH	5	9.9999	15	19.9999
7509	YFT	5	9.9999	-20	-24.9999
7510	OTH	5	9.9999	20	24.9999
7511	YFT	5	9.9999	-25	-29.9999
7512	OTH	5	9.9999	25	29.9999
7513	YFT	5	9.9999	-30	-34.9999
7514	OTH	5	9.9999	30	34.9999
7515	YFT	5	9.9999	-35	-39.9999
7516	OTH	5	9.9999	35	39.9999
7517	YFT	5	9.9999	-40	-44.9999

(d) 完成以經緯度劃分的漁獲區

圖 3.2.13 建立漁獲區維度的語法

建立事實資料表一共有 42 個步驟，如圖 3.2.3(c)所示。圖 3.2.14 則說明程式設計在各階

段中的主要目標，在於以漁業統計資料庫中的遠洋魚貨銷售記錄為中心作為事實資料表 (Fact table)，加入漁船相關屬性以及漁船航行軌跡相關屬性的 11 個維度，並且將資料淨化、格式統一化。

表 3.2.2(a) 漁獲區重疊的狀況統計

同一區域出現的漁獲區數目(個漁獲區)	重疊情形數目(種情形)
1	12
2	8
3	3

表 3.2.2(b) 漁獲區分散的狀況統計

同一漁獲區分散於各大海域的數目(次)	出現的情形
1	6
2	2
3	3
6	1

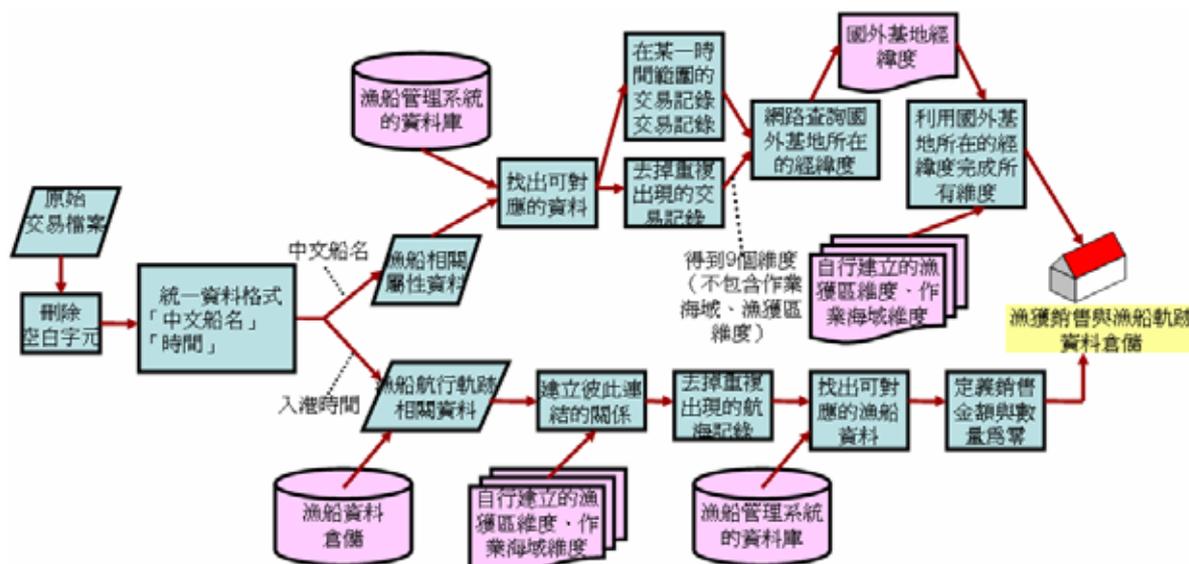


圖 3.2.14 利用資料轉換服務建立事實資料表的流程

由於本系統的維度來自不同的資料庫並且受限於來源資料的格式(可參考附錄一)無法直接建立各維度，首先分呈兩大方向與各維度作連結，一個是以「中文船名」對漁船相關屬性的維度作連結，另一個是以「時間」對漁船航行軌跡相關屬性的維度作連結。此兩種情況均需將格式不一的資料予以統一化、刪除內容不齊全或記錄錯誤的資料後，才能在下一步驟事實資料表中建構其他的維度並且以代碼取代文字敘述的維度。就第一個方向來說，中文船名來自於原始檔案所記錄的「SAM_BOATCD」(船號)，雖然為代碼卻是格式不一，歸納起來共有 6 種格式，表 3.2.3 為整理原始檔案的結果以及採取的措施。以中文船名多一個「號」字為例，會使得該筆交易紀錄無法找到對應的漁船以及其相關屬性維度，因此必須去掉「號」字，使用的語法見圖 3.2.15(a)、(b)。就第二個方向來說，「時間」來自於原始檔案的欄位「SAM_INDATE」(入港日期)，歸納起來一共高達 7 種格式的資料記錄，表 3.2.4 為整理原始檔案的結果以及採取的措施。採取下列措施的目的是為了日後將該欄位的資料型別轉為時間格式「datetime」，因此將日期均轉為如同「1999/07/11」的紀錄方式。以記錄時間的格式

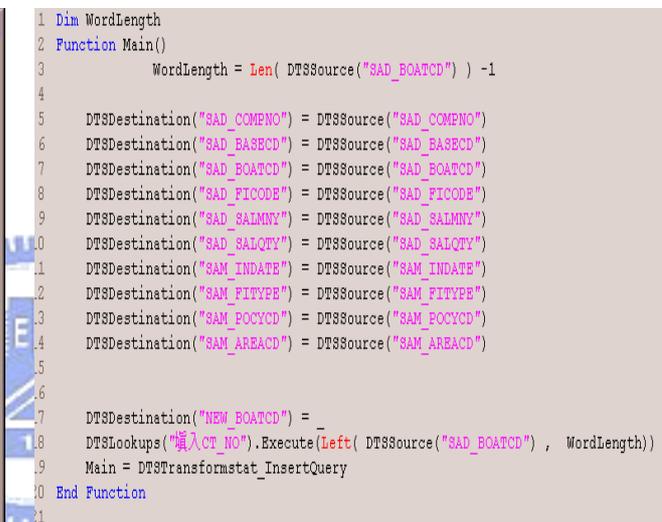
「民國/月份/日期」為例，首先辨識來源檔案是屬於此種格式後，需要將民國轉為西元，並且使用分隔符號「/」作為年、月、日的區隔，使用的語法見圖 3.2.16(a)(b)。此兩大方向的完成，由表 3.2.5 得知已進行了 12 道步驟。

表 3.2.3 不同格式的中文船名

中文船名格式描述	舉例	採取措施(目標格式：中文船名)
中文船名	台盈 6	不更動
中文船名+號	台盈 6 號	去掉"號"
中文船名+?	台盈 6?	去掉"?"
CT+數字	CT33297	去掉"CT"，並尋查其中文船名
CT+中文船名	CT 台盈 6	去掉"CT"
CT+中文船名+號	CT 台盈 6 號	去掉"CT"，去掉"號"



(a) 找尋目標格式的資料



(b) 去掉「號」字

圖 3.2.15 中文船名格式的統一

表 3.2.4 不同格式的時間欄位

時間資料位元數	時間格式描述	舉例	採取措施 (目標格式:1999/07/11)
6	民國/月份/日期	800113	加入"1911" & "/" & "/"
6	西元末兩位/月份/日期	910113	加入"1900" & "/" & "/"
7	空白+民國/月份/日期	_910113	取末六位，加入"1911" & "/" & "/"
7	0/民國/月份/日期	0800113	取末六位，加入"1911" & "/" & "/"
7	0/西元末兩位/月份/日期	0910113	取末六位，加入"1900" & "/" & "/"
7	民國/0/月份/日期	8000113	開頭不為 0，取前兩位，加入"1911"，末 4 位加入"/" & "/"

```

1 SELECT      SAD_COMPNO, SAD_BASECD, SAD_BOATCD, SAD_FICODE, SAD_SALQTY,
2             SAD_SALMNY, SAM_FITYPE, SAM_POCYCD, SAM_AREACD, SAM_INDATE,
3             SAD_PROCYM
4 FROM        FISALED4
5 WHERE       (RIGHT(SAM_INDATE, 2) <= '31') AND
6             (RIGHT(SAD_PROCYM, 2) <= '12') AND
7             (RIGHT(SAM_INDATE, 2) <> '00') AND
8             (RIGHT(SAD_PROCYM, 2) <> '00')

```

(a)選擇符合格式的來源檔案

```

1 Dim WordLength
2 Dim MyPosition
3 Function Main()
4
5     If Len( DTSSource("SAD_PROCYM") ) = 5 Then
6         DTSDestination("SAD_COMPNO") = DTSSource("SAD_COMPNO")
7         DTSDestination("SAD_BASECD") = DTSSource("SAD_BASECD")
8         MyPosition = Len( DTSSource("SAD_BOATCD") ) -2
9         WordLength = Len( DTSSource("SAD_BOATCD") ) -1
10        If Left( DTSSource("SAD_BOATCD") , 2 ) = "CT" Then
11            DTSDestination("SAD_BOATCD") = Mid( DTSSource("SAD_BOATCD") , 3 , MyPosition )
12            ElseIf Right( DTSSource("SAD_BOATCD") , 1 ) = "?" Then
13                DTSDestination("SAD_BOATCD") = Left( DTSSource("SAD_BOATCD") , WordLength )
14            Else
15                DTSDestination("SAD_BOATCD") = DTSSource("SAD_BOATCD")
16            End If
17        DTSDestination("SAD_FICODE") = DTSSource("SAD_FICODE")
18        DTSDestination("SAD_SALMNY") = DTSSource("SAD_SALMNY")
19        DTSDestination("SAD_SALQTY") = DTSSource("SAD_SALQTY")
20
21        DTSDestination("SAM_INDATE") = 1911 + _
22            Mid(DTSSource("SAD_PROCYM") , 2 , 2 ) & _
23            "/" & _
24            Mid(DTSSource("SAD_PROCYM") , 4 , 2 ) & _
25            "/" & _
26            Right( DTSSource("SAM_INDATE") , 2 )
27
28        DTSDestination("SAM_FITYPE") = DTSSource("SAM_FITYPE")
29        DTSDestination("SAM_POCYCD") = DTSSource("SAM_POCYCD")
30        DTSDestination("SAM_AREACD") = DTSSource("SAM_AREACD")
31        Main = DTSTransformStat_OK
32    Else
33        Main = DTSTransformStat_SkipRow
34    End If
35 End Function

```

(b)轉換成為目標的時間格式

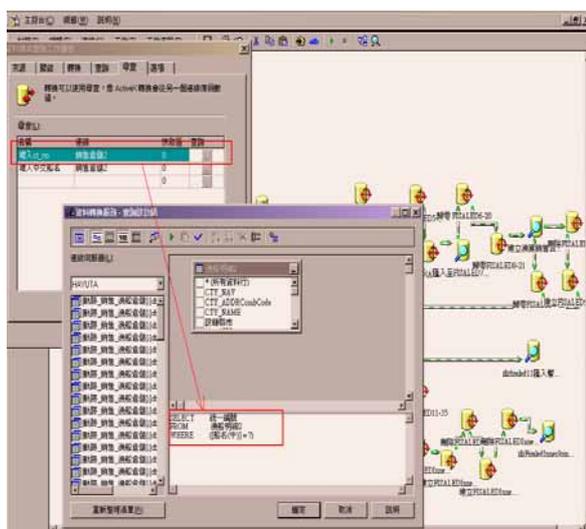
圖 3.2.16 時間格式的統一

程式設計的流程以及階段性完成的任務說明於表 3.2.5，最後建立的事實資料表資料涵蓋範圍由 1999/1/1~2001/12/31 經過 42 個步驟建構出 11 個維度，一共 16 多萬筆的資料。在檢查資料是否重複或是整合分散的交易紀錄上，以建立目標資料表 FISALED6、FISALED7 為例。

表 3.2.5 建構事實資料表的步驟

目標資料表	執行步驟	筆數	來源資料表	執行步驟數	維度數目
FISALED2	inner join	166384	FISALED & FISALEM	3	2
FISALED4	刪除空白字元	166384	FISALED2	3	2
FISALED3	統一 6 種時間格式，6 種中文船名	89192 ->58898	FISALED4, 漁船明細表	6	2
魚貨銷售暫存表	將中文名稱轉換為漁船的統一編號	58898	FISALED3, 漁船明細表 2	2	3

目標資料表	執行步驟	筆數	來源資料表	執行步驟數	維度數目
FISALED5	從魚貨銷售暫存表,找出可轉換成 datetime 欄位的資料	58657	魚貨銷售暫存表	1	4
FISALED6	整理每一筆的交易資料,不要有重複紀錄	58567	FISALED5	3	4
FISALED7	加總同一筆交易記錄的金額與數量	55928	FISALED6	3	4
FISALED8	插入正確的漁船公司代碼,漁業別代碼,漁船所屬地區代碼,船主代碼	55928	FISALED7, 漁船明細表 2	3	8
FISALED9	整理每一筆的交易資料,不要有重複紀錄	55906	FISALED8	3	8
FISALED10	建立每一筆交易記錄的主鍵,以及 9 個維度,分別為船主維度、漁船公司維度、國外基地維度、漁船噸位維度、漁船所屬地區維度、漁業別維度、魚種類維度、銷售國別維度以及時間維度	55906	FISALED9	3	9
FISALED11	檢查各維度必須能參照各維度資料表	14082	FISALED10	3	9
銷售與軌跡暫存表	一種是航海紀錄,另一種是魚貨交易記錄	152219+ 14082	FISALED11, 漁船軌跡 2	9	11
完成	事實資料表	166301 筆	共 42 步驟		11 個維度



(a) 建立查詢項目

```

1 Dim WordLength
2 Function Main()
3     If DTSSource("存在中文船名") > 0 Then
4         DTSDestination("NEW_BOATCD") =
5             DTSLookups("填入ct_no").Execute(DTSSource("SAD_BOATCD"))
6         DTSDestination("SAD_BOATCD") = DTSSource("SAD_BOATCD")
7         DTSDestination("SAM_AREACD") = DTSSource("SAM_AREACD")
8         DTSDestination("SAM_POCYCD") = DTSSource("SAM_POCYCD")
9         DTSDestination("SAM_FITYPE") = DTSSource("SAM_FITYPE")
10        DTSDestination("SAM_INDATB") = DTSSource("SAM_INDATB")
11        DTSDestination("SAD_SALQTY") = DTSSource("SAD_SALQTY")
12        DTSDestination("SAD_SALMNY") = DTSSource("SAD_SALMNY")
13        DTSDestination("SAD_FICODE") = DTSSource("SAD_FICODE")
14        DTSDestination("SAD_BASECD") = DTSSource("SAD_BASECD")
15        DTSDestination("SAD_COMPNO") = DTSSource("SAD_COMPNO")
16
17        Main = DTSTransformStat_InsertQuery
18        Else
19            Main=DTSTransformStat_SkipRow
20        End If
21    End Function

```

(b) 將來源資料作轉換

圖 3.2.17 利用中文名稱尋查漁船代碼

在完成時間格式以及中文船名格式的統一後,下一步利用圖 3.2.17 (a)、(b) 的查詢語法,將漁船的中文名稱轉為代碼,以較佳的資料格式儲存。接下來整理資料表去除重複出現的筆數,並且將同一筆交易記錄的金額與數量作加總,使用的 SQL 語法見圖 3.2.18(a)、(b)。

```

1 SELECT          COUNT(*) AS 出現次數, NEW_BOATCD, SAM_INDATE, SAD_FICODE,
2
3                SAM_POCYCD, SAD_COMPNO, SAD_BASECD,
4 SAM_FITYPE, SAM_AREACD,
5                SAD_SALMNY
6 AS SAD_SALMNY, SAD_SALQTY AS SAD_SALQTY
7 INTO FISALED6
8 FROM            FISALED5
9 GROUP BY NEW_BOATCD, SAM_INDATE, SAD_FICODE, SAM_POCYCD, SAD_COMPNO,
10
11              SAD_BASECD, SAM_FITYPE, SAM_AREACD,
12 SAD_SALMNY,
13              SAD_SALQTY
14 ORDER BY
15 COUNT(*) DESC, NEW_BOATCD, SAM_INDATE, SAD_FICODE
16

```

(a)使資料不重複出現

```

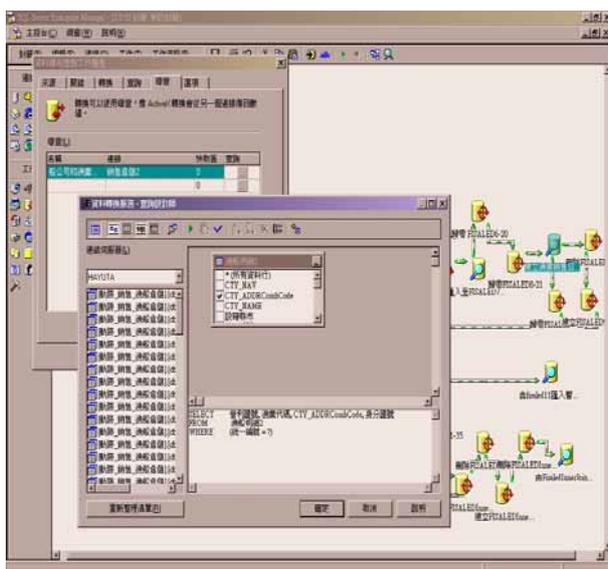
1 SELECT          COUNT(*) AS 交易數量, NEW_BOATCD, SAM_INDATE, SAD_FICODE,
2                SAM_POCYCD, SAD_COMPNO, SAD_BASECD, SAM_FITYPE, SAM_AREACD,
3                SUM(SAD_SALMNY) AS SAD_SALMNY, SUM(SAD_SALQTY)
4                AS SAD_SALQTY
5 FROM            FISALED6
6 GROUP BY NEW_BOATCD, SAM_INDATE, SAD_FICODE, SAM_POCYCD, SAD_COMPNO,
7                SAD_BASECD, SAM_FITYPE, SAM_AREACD
8 ORDER BY COUNT(*) DESC, NEW_BOATCD, SAM_INDATE, SAD_FICODE
9
10

```

(b)加總同一筆交易記錄的金額、數量

圖 3.2.18 整合每一筆交易資料

至於在增加維度的方式上，以建立目標資料表 FISALED8 為例，使用圖 3.2.19 (a)、(b) 的查詢語法，利用漁船代碼尋查到漁船公司代碼、漁業別代碼、漁船所屬地區代碼以及船主代碼，該步驟完成後可擴建 4 個維度。最後匯入至事實資料表時，必須解決兩個問題。第一個問題來自於原始交易記錄資料表中，並沒有經度與緯度的航海紀錄，匯入事實資料表會造成主鍵有空值的錯誤。如圖 3.2.14 的程式設計說明流程，以圖 3.2.20 的查詢建構某一國家銷售基地的經緯度位置於維度資料表中，再利用尋查的功能將該漁船當時的經緯度位置記錄在事實資料表中。第二個問題來自於漁船航行軌跡的紀錄中只有包括 3 維度：時間、作業海域以及漁獲區維度，直接匯入事實資料表中會造成多項主鍵欄位為空值的錯誤。



(a)建立查詢項目

```

1 Function Main()
2 Dim queryArray
3 queryArray = DTSLookups("船公司和漁業代碼和縣市和船主").Execute(DTSSource("NEW_BOATCD"))
4 DTSDestination("NEW_BOATCD") = DTSSource("NEW_BOATCD")
5 DTSDestination("SAM_INDATE") = DTSSource("SAM_INDATE")
6 DTSDestination("SAD_FICODE") = DTSSource("SAD_FICODE")
7 DTSDestination("SAM_POCYCD") = DTSSource("SAM_POCYCD")
8 DTSDestination("SAD_COMPNO") = queryArray (0)
9 DTSDestination("SAD_BASECD") = DTSSource("SAD_BASECD")
10 DTSDestination("SAM_FITYPE") = queryArray (1)
11 DTSDestination("SAM_AREACD") = DTSSource("SAM_AREACD")
12 DTSDestination("SAD_SALMNY") = DTSSource("SAD_SALMNY")
13 DTSDestination("SAD_SALQTY") = DTSSource("SAD_SALQTY")
14 DTSDestination("TAI_LOCATION") = queryArray (2)
15 DTSDestination("SHIP_MASTER") = queryArray (3)
16
17 DTSDestination("TONNAGE_TYPE") = Left( DTSSource("NEW_BOATCD"), 1 )
18
19 Main = DTSTransformstat_InsertQuery
20 End Function
21

```

(b) 將來源資料作轉換

圖 3.2.19 擴增維度所使用的方法

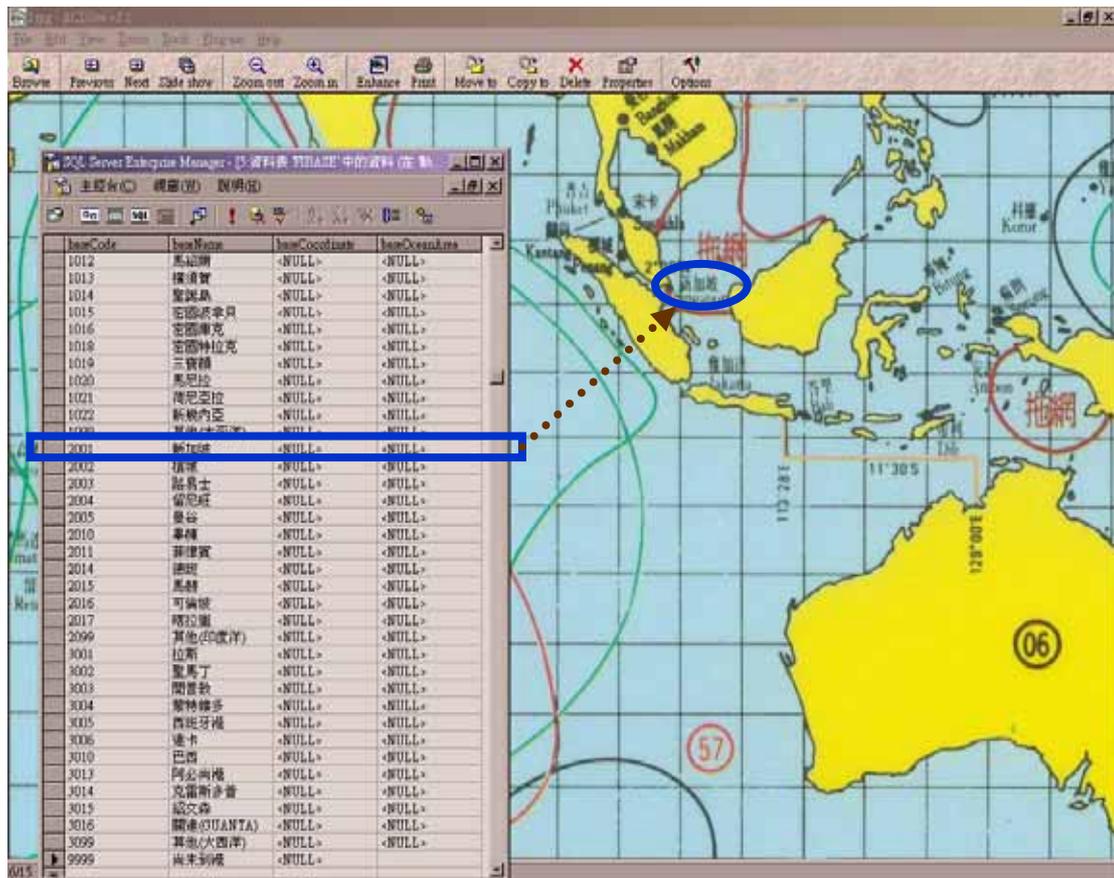


圖 3.2.20 以地圖建立銷售國別的維度資料表²

因此採取的方式：首先利用漁船代碼以及航海時間對銷售記錄表作聯結；接著定義每一趟漁船出海紀錄其航海時間至少不晚於至市場交易的時間；最後在考量漁船航海時並沒有在市場交易的紀錄，因此將銷售金額和數量記錄為零值而非空值(Zero, Not Null Value)。使用的語法見圖 3.2.21(a)、(b)。經過上述的各步驟後，資料轉換服務的過程不會因為一個小錯誤而無法執行所有程序，完整正確的事實資料表得以建立。

```

1 SELECT      FISALED11.NEW_BOATCD, FISALED11.SAM_INDATE, FISALED11.SAD_FICODE,
2             FISALED11.SAM_POCYCD, CONVERT(varchar(3), FISALED11.SAM_FITYPE)
3 AS SAM_FITYPE, FISALED11.SAD_BASECD, FISALED11.SAD_COMPNO,
4             FISALED11.SAD_SALMNY, FISALED11.SAD_SALQTY,
5             FISALED11.SHIP_MASTER, FISALED11.TAI_LOCATION,
6             FISALED11.TONNAGE_TYPE, 漁船軌跡2.GPS_時間, 漁船軌跡2.漁區代號,
7             漁船軌跡2.經度, 漁船軌跡2.緯度, 漁船軌跡2.海域洋區代碼,
8             漁船軌跡2.漁獲作口區代碼, DATEDIFF(week, 漁船軌跡2.GPS_時間,
9             FISALED11.SAM_INDATE) AS 相差週數, DATEDIFF([day], 漁船軌跡2.GPS_時間,
10            FISALED11.SAM_INDATE) AS 相差天數
11 INTO
12           FISALEDInnerJoin
13 FROM      FISALED11 INNER JOIN
14           漁船軌跡2 ON FISALED11.NEW_BOATCD = 漁船軌跡2.CT_NO AND
15           FISALED11.SAM_INDATE >= 漁船軌跡2.GPS_時間
16 WHERE     (DATEDIFF(week, 漁船軌跡2.GPS_時間, FISALED11.SAM_INDATE) <= 2) AND
17           (漁船軌跡2.GPS_時間 IS NOT NULL) AND (漁船軌跡2.漁區代號 IS NOT NULL) AND
18           (漁船軌跡2.經度 IS NOT NULL) AND (漁船軌跡2.緯度 IS NOT NULL) AND
19           (漁船軌跡2.海域洋區代碼 IS NOT NULL) AND
20           (漁船軌跡2.漁獲作口區代碼 IS NOT NULL)

```

(a)利用內部聯結的方式

²本地圖由航空測量及遙程探測學會於1996年6月編制

```

1 Function Main()
2     DTSDestination("NEW_BOATCD") = DTSSource("NEW_BOATCD")
3     DTSDestination("SAM_INDATE") = DTSSource("GPS_時間")
4     DTSDestination("SAD_FICODE") = DTSSource("SAD_FICODE")
5     DTSDestination("SAM_POCYCD") = DTSSource("SAM_POCYCD")
6     DTSDestination("SAD_COMPNO") = DTSSource("SAD_COMPNO")
7     DTSDestination("SAD_BASECD") = DTSSource("SAD_BASECD")
8     DTSDestination("SAM_FITYPE") = DTSSource("SAM_FITYPE")
9     DTSDestination("DateDiff") = DTSSource("相差天數")
10    DTSDestination("漁區代號") = DTSSource("漁區代號")
11    DTSDestination("OCEAN_AREACD") = DTSSource("海域洋區代碼")
12    DTSDestination("漁獲作業區") = DTSSource("漁獲作業區代碼")
13        DTSDestination("SHIP_MASTER") = DTSSource("SHIP_MASTER")
14        DTSDestination("TAI_LOCATION") = DTSSource("TAI_LOCATION")
15        DTSDestination("TONNAGE_TYPE") = DTSSource("TONNAGE_TYPE")
16
17    DTSDestination("SAD_SALQTY") = 0
18    DTSDestination("SAD_SALMNY") = 0
19    DTSDestination("經度") = DTSSource("經度")
20        DTSDestination("緯度") = DTSSource("緯度")
21    Main = DTSTransformstat_InsertQuery
22 End Function

```

(b)將交易記錄設定為零值

圖 3.2.21.將漁船航海軌跡記錄存入事實資料表



3.3 魚貝苗養殖資料倉儲的設計與實作

台灣的水產養殖，由早期看天田式的淺坪虱目魚養殖，已發展至利用先進科技的室內自動化超高密度循環水養殖，也由高山的冷水性魚類養殖分布到外海的箱網養殖，為沿海農村經濟結構重要之一環，對於繁榮漁村經濟與安定三萬養殖漁家的生活，有其不可磨滅的貢獻。[5]

但是由於快速而盲目的成長，而埋下產銷秩序失衡與生產環境惡化的問題。因此為了養殖產業的轉型調整與配合產業需求，政府積極於產業組織的再造、生產環境的改善，以及產業結構的調整，使養殖漁業達到永續發展的目標。此時政府管理單位需要正確、及時的產銷資訊，以制訂各項在水資源方面等的環境改善政策。利用本研究所建立的「魚貝苗資料倉儲」，建構「漁業資訊分享熱線」，使用的資訊技術包括三階正規化資料庫、雪花式綱要資料倉儲、資料轉換服務、線上分析處理。本小節將說明「魚貝苗資料倉儲」的設計與實作。

對於本系統的輸入資料而言，一共有 3 張未經過三階正規化的資料表資料表。分別是來自漁業統計資料庫系統的「魚貝苗產量及價值資料表」、「行政地區資料表」、「魚貝苗種類資料表」共 3 張。各來源資料表的屬性規格分別列在附錄一的附錄表 A.1.2.1~1.2.3 中。

對於本系統的輸出而言，是建構資料倉儲的綱要。綱要規劃的方式是採用多維度資料模型的雪花綱要。雪花綱要的軸心為事實資料表，主要是由數個維度及衡量值所構成。本研究「魚貝苗資料倉儲」共設計 4 個維度，分別為作業方式維度、魚貝苗種類維度、產地維度（即行政地區）以及時間維度；衡量值有 7 個，分別是產出數量、產出重量、單位數量的平均售格、單位重量的平均售格、產值、養殖場家數以及養殖場面積，一共包含 7 張維度資料表以及 1 張事實資料表，雪花綱要如圖 3.3.1 所示。在表 3.3.1 中說明各維度的階層設計與關係，並且在表 3.3.2~3.3.6 表中說明事實資料表與各維度資料表的欄位。

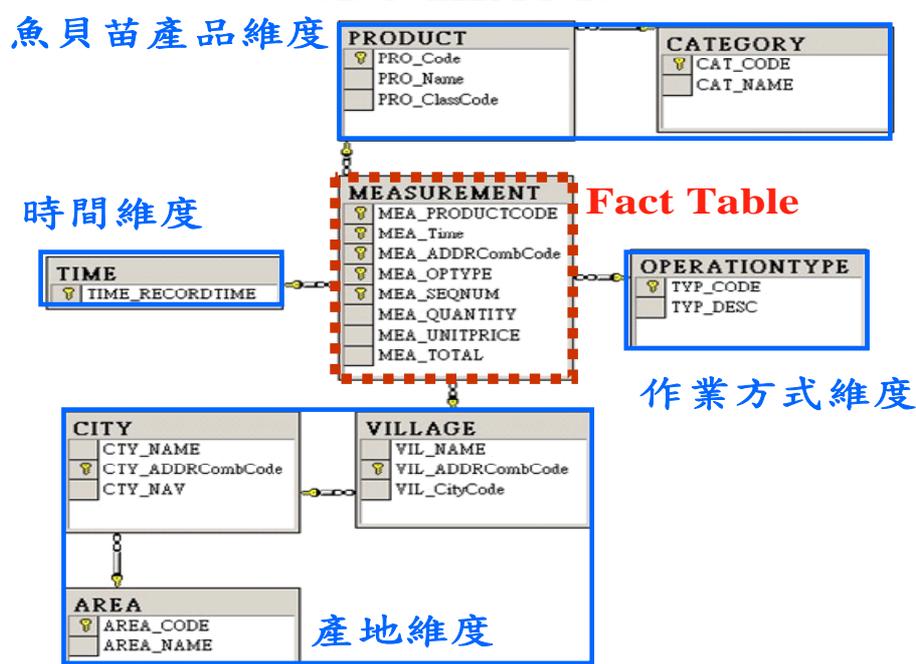


圖 3.3.1 魚貝苗資料倉儲的雪花綱要

表 3.3.1 魚貝苗資料倉儲各維度的資料階層關係

維度	Level 1	Level 2	Level 3
作業方式維度	作業方式-		
魚貝苗種類維度	種苗分類名類+	各種苗名類-	
產地維度	行政單位+	各縣市政府、直轄市+	鄉、鎮、市或漁會-
時間維度	年+	季+	月-

+：表示階層可再往下延伸

-：表示階層已到最底層

各維度均以最低階層（數字越多代表階層越低，Level1 最高）與事實資料表做聯結

表 3.3.2 事實資料表

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值	備註
◎	◎	MEA_PRODUCTCODE	魚貝苗產品種類	varchar	4		
◎	◎	MEA_Time	時間	datetime	8		
◎	◎	MEA_ADDRCombCode	產地	varchar	8		
◎	◎	MEA_OPTYPE	作業方式	varchar	1		
◎		MEA_SEQNUM	序號	varchar	4		
		MEA_QUANTITY1	生產數量	int	4	◎	產量(隻,尾,個)
		MEA_QUANTITY2	生產重量	int	4	◎	產量(公噸)
		MEA_UPRICE1	單位數量的平均售價	float	8	◎	單位：千元/每千個
		MEA_UPRICE2	單位重量的平均售價	float	8	◎	單位：千元/每公噸
		MEA_TOTAL	產值	int	4	◎	單位：千元
		MEA_RQY	養殖場家數	int	8	◎	單位：家
		MEA_SQA	養殖場面積	float	8	◎	單位：公頃

表 3.3.3 作業方式維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		TYP_CODE	作業方式編碼	varchar	1	
		TYP_DESC	作業方式名稱	varchar	6	

表 3.3.4 魚貝苗種類維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		PRO_CODE	魚貝產品種苗編碼	varchar	4	
		PRO_NAME	魚貝產品種苗名稱	varchar	12	
	◎	PRO_ClassCode	魚貝產品種苗分類編碼	varchar	1	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		CAT_CODE	魚貝產品種苗分類編碼	varchar	1	
		CAT_NAME	魚貝產品種苗分類名稱	varchar	10	

表 3.3.5 產地維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		VIL_ADDRCombCode	鄉、鎮、市或漁會編碼	varchar	8	
	◎	VIL_CityCode	所屬的各縣市政府編碼	varchar	8	
		VIL_NAME	鄉、鎮、市或漁會名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		CTY_ADDRCombCode	所屬的各縣市政府編碼	varchar	8	
	◎	CTY_NAV	所屬省單位編碼	varchar	1	
		CTY_NAME	所屬的各縣市政府名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		AREA_CODE	所屬省單位編碼	varchar	1	
		AREA_NAME	所屬省單位名稱	varchar	10	

表 3.3.6 時間維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		TIME_RECORDTIME	時間	datetime	8	



3.4 水產加工製品資料倉儲的設計與實作

水產品經過適當的加工處理，可以有效利用漁獲物，並且提高其經濟價值，進一步具有調節供需及穩定魚價之功能。根據 2002 年的漁業統計年報，台灣目前水產加工廠有 695 家，其中 20% 的加工廠如冷凍及罐頭加工廠較具規模外，其餘 80% 屬於小型或家庭式的水產加工廠。民國 76 年水產加工品外銷產值達 328 億元，為歷年最高值，佔當年水產外銷總值之 72%，農產外銷總值之 28%，水產加工在我國之農水畜產業發展過程中扮演重要之角色。近年來水產加工品外銷產量有衰退現象，影響成因有：與開發中國家的競爭、勞力與製造成本的提高、臺幣升值等因素。

因此為了穩定水產加工業基礎，繼續維持水產加工品在國內與國外的市場佔有率，政府必須結合民間業者透過「品牌認證」、「開發新的水產加工產品」，以達到增加水產品附加價值、促進漁業產銷平衡、穩定魚價等目標。利用本研究所建立的「水產加工製品資料倉儲」，建構「漁業資訊分享熱線」，透過使用下列資訊技術：包括三階正規化資料庫、星座綱要資料倉儲、資料轉換服務、線上分析處理等；使政府與民間業者在品牌認證的制度規範、新產品的開發或是消費者喜好的趨勢研究上，擁有更多有利、正確以及方便的資訊，以協助產品銷售增加。本小節將說明「水產加工製品資料倉儲」的設計與實作。

對於本系統的輸入資料而言，一共有 5 張未經過三階正規化的資料表資料表。分別是來自漁業統計資料庫系統的「水產加工製品產量及價值資料表 1」（包含魚種類別）、「水產加工製品產量及價值資料表 2」（月別）、「水產生物種類資料表」及「行政地區資料表」共 5 張。各來源資料表的屬性規格分別列在附錄一的附錄表 A.1.3.1~A.1.3.5 中。其中水產加工製品產量及價值資料表 1（以下簡稱資料表 1）與水產加工製品產量及價值資料表 2（以下簡稱資料表 2）的差別，在於資料表 1 多了「水產生物種類」來對產品做描述；資料表 1 的時間記錄以「年度」為單位，資料表 2 的時間記錄以「月份」為單位，表 3.4.1 列出相異處。兩張資料表的差異造成資料表的數值無法做合併成 1 張事實資料表，但是可以用本文下述方式取代。

表 3.4.1 來源資料表的差異

		水產加工製品產量及價值資料表 1	水產加工製品產量及價值資料表 2
相 異 處	1	使用「水產生物種類」對加工製品做描述。但是該項欄位屬性可以為空值，故非每一筆資料記錄均有描述。	無使用「水產生物種類」對加工製品做描述。
	2	資料記錄時間以「年度」為單位。	資料記錄時間以「月份」為單位。
	3	編制「時間」、「行政地區」、「產品」與「序號」作為主鍵。倘若缺少序號欄位，會導致資料發生重複而無法辨識紀錄。	編制「時間」、「行政地區」與「產品」作為主鍵。因此無法與資料表 1 利用外部聯結(Outer Join)進行合併。
	4	總彙的數值欄位有 3 種。	總彙的數值欄位僅有 2 種。

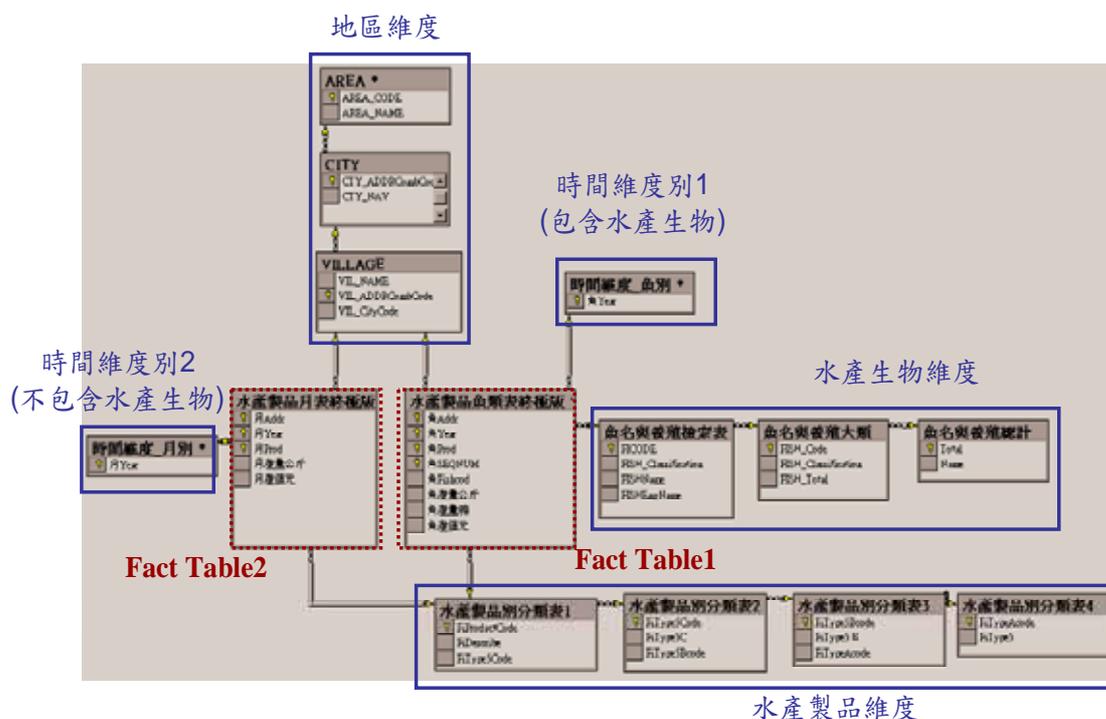


圖 3.4.1 水產加工製品資料倉儲的星座綱要

對於本系統的輸出而言，是建構資料倉儲的綱要。綱要規劃的方式是採用多維度資料模型的星座綱要。星座綱要綱要的軸心為兩張以上的事實資料表，主要是由數個維度及衡量值所構成。本研究「水產加工製品資料倉儲」共設計 10 張維度資料表以及 2 張事實資料表。為了使衡量值不會因為交叉查詢而導致數值錯誤，故將時間維度分割成 2 種時間維度，分別代表兩張事實資料表。因此一共包含 5 個維度，分別為時間維度 1（包含水產生物）、時間維度 2（不包含水產生物）、水產生物種類維度、水產製品種類維度以及地區維度；衡量值有 5 個，分別是年底魚製加工品產出重量、年底魚製加工品產出數量、年底魚製加工品產值、每月加工品產出重量及每月加工品產值，雪花綱要如圖 3.4.1 所示。在表 3.4.2 中說明各維度的階層設計與關係，並且在表 3.4.3~3.4.9 表中說明事實資料表與各維度資料表的欄位。

表 3.4.2 水產加工製品資料倉儲之各維度的資料階層關係

維度	Level 1	Level 2	Level 3	Level 4
時間維度 1 (包含水產生物)	年+	季+	月-	
時間維度 2 (不包含水產生物)	年-			
水產生物種類維度	水產生物總類+	水產生物分類+	水產生物名稱-	
水產製品種類維度	水產製品總類+	水產製品分類+	水產製品細項分類+	水產製品名稱-
地區維度	行政單位+	各縣市政府、直轄市+	鄉、鎮、市或漁會-	

+：表示階層可再往下延伸

-：表示階層已到底層

各維度均以最低階層（數字越多代表階層越低，Level1 最高）與事實資料表做聯結

表 3.4.3 事實資料表 1

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值	備註
◎	◎	魚 Addr	地區維度	varchar	8		
◎	◎	魚 Year	時間維度 1	datetime	8		
◎	◎	魚 Prod	水產製品維度	varchar	4		
◎		魚 SEQNUM	序號	varchar	3		
	◎	魚 Fishcod	水產生物維度	varchar	4	◎	
		魚產量公斤	生產重量	float	8	◎	產量(公斤)
		魚產量箱	生產數量	float	8	◎	產量(箱數)
		魚產值元	產值	float	8	◎	產值(元)

表 3.4.4 事實資料表 2

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值	備註
◎	◎	月 Addr	地區維度	varchar	8		
◎	◎	月 Year	時間維度 1	datetime	8		
◎	◎	月 Prod	水產製品維度	varchar	4		
		月產量公斤	生產重量	float	8	◎	產量(公斤)
		月產值元	產值	float	8	◎	產值(元)

表 3.4.5 時間維度 1

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		魚 Year	時間維度 1	datetime	8	

表 3.4.6 時間維度 2

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		月 Year	時間維度 2	datetime	8	

表 3.4.7 水產生物種類維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FICODE	水產生物編碼	varchar	4	
	◎	FISH_Classification	水產生物分類編碼	varchar	4	◎
		FISHName	水產生物名稱(中文)	varchar	12	◎
		FISHEngName	水產生物名稱(英文)	varchar	100	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FISH_Code	水產生物分類編碼	varchar	4	
		FISH_Classification	水產生物分類名稱	varchar	50	
	◎	FISH_Total	水產生物總類編碼	varchar	4	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		Total	水產生物總類編碼	varchar	4	
		Name	水產生物總類名稱	varchar	10	◎

表 3.4.8 水產製品種類維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FiProductCode	水產製品編碼	varchar	4	
		FiDescribe	水產製品名稱	varchar	20	◎
	◎	FiType5Code	水產製品細項分類編碼	varchar	2	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FiType5Code	水產製品細項分類編碼	varchar	2	
		FiType5C	水產製品細項分類名稱	varchar	12	◎
	◎	FiType5Bcode	水產製品分類編碼	varchar	3	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FiType5Bcode	水產製品分類編碼	varchar	3	
		FiType5 B	水產製品分類名稱	varchar	16	◎
	◎	FiTypeAcode	水產製品總類編碼	varchar	3	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FiTypeAcode	水產製品總類編碼	varchar	3	
		FiType5	水產製品總類名稱	varchar	10	

表 3.4.9 地區維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		VIL_ADDRCombCode	鄉、鎮、市或漁會編碼	varchar	8	
	◎	VIL_CityCode	所屬的各縣市政府編碼	varchar	8	
		VIL_NAME	鄉、鎮、市或漁會名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		CTY_ADDRCombCode	所屬的各縣市政府編碼	varchar	8	
	◎	CTY_NAV	所屬省單位編碼	varchar	1	
		CTY_NAME	所屬的各縣市政府名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		AREA_CODE	所屬省單位編碼	varchar	1	
		AREA_NAME	所屬省單位名稱	varchar	10	

3.5 遠洋漁獲資料倉儲的設計與實作

本研究主要以遠洋漁業的業者在國內各魚市場銷售情形為研究對象。我國遠洋漁業近年來漁獲量已超過 82 萬公噸，佔我國漁業總產量之 49% 以上，作業漁場遍布三大洋之公海漁場，主要漁法包括鮪延繩釣、鯷鮪圍網、拖網、魷釣及秋刀魚棒受網等，為我國十分重要的產業。

由於世界各沿海國實施 200 浬經濟海域及公海資源共管之趨勢；再加上漁業捕獲技術的進步卻缺乏有效的管理規範，出現海洋資源逐漸衰竭的危機，使得我國遠洋漁業作業漁場以及漁獲作業進行遭受重大衝擊。民間遠洋產業的發展需要政府政策的協助，政府第一步必須因應國際「責任制漁業」之相關漁業管理規範，為的是積極參與國際漁業組織，擴大與各沿岸國家的漁業合作，確保我遠洋漁船之傳統作業漁場與權益。為了達到上述目標，政府單位就必須能掌握各項漁業資訊，例如國內業者的漁獲能力、漁業投資或是漁業資源調查，才能進行有效的管理。

本研究為此建立了「遠洋漁獲資料倉儲」，建構在「漁業資訊分享熱線」上；並且透過使用下列資訊技術：包括三階正規化資料庫、雪花網要資料倉儲、資料轉換服務、線上分析處理等；使政府與民間業者能善盡國際漁業責任、調整遠洋漁業經營規模及漁獲能力、魚貨銷售市場的開發或是消費者喜好的趨勢研究，以協助我國遠洋的永續經營。本小節將說明「遠洋漁獲資料倉儲」的設計與實作。

對於本系統的輸入資料而言，一共有 6 張未經過三階正規化的資料表資料表。分別是來自漁業統計資料庫系統的「遠洋漁業生產量調查資料表」、「區漁會資料表」、「魚市場資料表」、「漁獲生物種類資料表」、「漁業作業種類資料表」及「作業海域資料表」共 6 張。各來源資料表的屬性規格分別列在附錄表 A.1.4.1~1.4.6 中。其中在「區漁會」與「魚市場」兩者的關係，調查「魚市場」是由各區漁會所經營的單位。故在設計資料倉儲網要時，會將兩張獨立的資料表，透過編碼技術與資料轉換服務技術將「區漁會」與「魚市場」關係作聯繫。



圖 3.5.1 遠洋漁獲資料倉儲的星座網要

對於本系統的輸出而言，是建構資料倉儲的綱要。綱要規劃的方式是採用多維度資料模型的雪花綱要。本研究「遠洋漁獲資料倉儲」共設計 5 個維度，分別為銷售地維度（即行政地區）、漁獲生物種類維度、漁業作業種類維度、作業海域維度以及時間維度；衡量值有 3 個，分別是魚市場交易數量、自用及其他用途估計量、遠洋漁業生產量。因此本資料倉儲一共包含 12 張維度資料表以及 1 張事實資料表，雪花綱要如圖 3.5.1 所示。在表 3.5.1 中說明各維度的階層設計與關係，並且在表 3.5.2~3.3.7 表中說明事實資料表與各維度資料表的欄位。

表 3.5.1 遠洋漁獲資料倉儲之各維度的資料階層關係

維度	Level 1	Level 2	Level 3	Level 4
銷售地維度	行政單位+	各縣市政府、直轄市+	所屬的漁會+	魚市場-
漁獲生物種類維度	水產生物總類+	水產生物分類+	水產生物名稱-	
漁業作業種類維度	漁業作業總類+	漁業作業分類+	漁業作業名稱-	
作業海域維度	作業海域名稱-			
時間維度	年+	季+	月-	

+：表示階層可再往下延伸

-：表示階層已到最底層

各維度均以最低階層（數字越多代表階層越低，Level1 最高）與事實資料表做聯結

表 3.5.2 事實資料表

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值	備註
◎	◎	ADDR	銷售漁會	varchar	8		
◎	◎	ADDRmarket	銷售魚市場(銷售地維度)	datetime	8		
◎	◎	WorkType	漁業作業種類	varchar	4		
◎		YYMM	時間	varchar	3		
	◎	SEQNUM	序號	varchar	4	◎	
		FishTYpe	漁獲生物種類	float	8	◎	
		OceanType	作業海域	float	8	◎	
		TransQ	市場交易量	float	8	◎	產量(公斤)
		SelfQ	自用及其他估計量	float	9	◎	產量(公斤)

表 3.5.3 銷售地維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		MarketCode	魚市場名稱	varchar	10	
		MarketName	魚市場編碼	varchar	20	◎
	◎	MarketCty	所屬漁會編碼	varchar	8	◎
◎		VIL_ADDRCombCode	鄉、鎮、市或漁會編碼	varchar	8	
	◎	VIL_CityCode	所屬的各縣市政府編碼	varchar	8	
		VIL_NAME	鄉、鎮、市或漁會名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		CTY_ADDRCombCode	所屬的各縣市政府編碼	varchar	8	
	◎	CTY_NAV	所屬省單位編碼	varchar	1	
		CTY_NAME	所屬的各縣市政府名稱	varchar	14	
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		AREA_CODE	所屬省單位編碼	varchar	1	
		AREA_NAME	所屬省單位名稱	varchar	10	

表 3.5.4 漁獲生物種類維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FICODE	漁獲生物編碼	varchar	4	
	◎	FISH_Classification	漁獲生物分類編碼	varchar	4	◎
		FISHName	漁獲生物名稱 (中文)	varchar	12	◎
		FISHEngName	漁獲生物名稱 (英文)	varchar	100	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FISH_Code	漁獲生物分類編碼	varchar	4	
		FISH_Classification	漁獲生物分類名稱	varchar	50	
	◎	FISH_Total	漁獲生物總類編碼	varchar	4	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		Total	漁獲生物總類編碼	varchar	4	
		Name	漁獲生物總類名稱	varchar	10	◎

表 3.5.5 漁業作業種類維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		FisheryCode	漁業別編碼	varchar	3	
		FisheryChineseName	漁業別中文名稱	varchar	20	◎
		FisheryEngName	漁業別英文名稱	varchar	50	◎
	◎	FisheryType	漁業別大類編碼	varchar	3	◎
主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		Type_Code	漁業別大類編碼	varchar	3	
		Type_Name	漁業別大類名稱	varchar	16	
	◎	Type_Total	漁業別總類編碼	varchar	3	◎

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		Total	漁業作業總類代碼	varchar	3	
		Name	漁業作業總類名稱	varchar	4	◎

表 3.5.6 作業海域維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		OCEAN_Code	作業海域名稱	varchar	1	
		OCEAN_Name	作業海域編碼	varchar	14	

表 3.5.7 時間維度

主鍵	外鍵	欄位名稱	欄位意義	型態	長度	空值
◎		YYMM	時間	datetime	8	

3.6 各資料倉儲的研究限制

本研究由於源資料的記錄方式，使得資料倉儲設計時，出現該維度不夠詳細或是衡量值缺乏的情形。以表 3.6.1 列出本研究限制。

表 3.6.1 各資料倉儲的研究限制

	各資料倉儲的研究限制
遠洋魚貨銷售與漁船軌跡資料倉儲	<ol style="list-style-type: none"> 1. 因缺少各單位漁船的漁獲量紀錄（衡量值），故以該漁船近期內魚貨的銷售紀錄代替。 2. 各來源資料庫系統的編碼方式差異過大，以致於進行資料整合時流失許多無法對應的紀錄。建議將代表各單位漁船的欄位記錄（即「漁船編碼」）予以統一，在整合異質來源資料庫上可有顯著改善。
魚貝苗養殖資料倉儲	<ol style="list-style-type: none"> 1. 每筆資料紀錄是由「各鄉鎮市或漁會」匯總後的數據，而非來自「各養殖或捕撈單位」的數據。因此在「地區」維度上只能下挖至各鄉鎮市或漁會，完整性不足。倘若資料直接來自各養殖或捕撈單位，除了在維度的呈現上更詳盡，另外也能提供更多或更正確的衡量值與維度，例如：各養殖方式的面積（衡量值）、作業方式（維度）。 2. 「時間」維度只能下挖至「月」。若能下挖至「日」，則更能分析市場需求的波動性。
水產加工製品資料倉儲	<ol style="list-style-type: none"> 1. 每筆資料紀錄是由「各鄉鎮市」匯總後的數據，而非來自「水產加工廠」的數據。因此在「地區」維度上只能下挖至各鄉鎮市，完整性不足。倘若資料直接來自各水產加工廠，除了在維度的呈現上更詳盡，另外也能提供更多或更正確的衡量值與維度，例如：水產加工廠數量（衡量值）、水產加工廠投資金額（衡量值）、水產加工廠規模（衡量值）。 2. 未以魚種類描述的「時間」維度只能下挖至「年」，而以魚種類描述的「時間」維度只能下挖至「月」。若能下挖至「日」，則更能分析各項產品在市場需求的波動性。
遠洋魚貨資料倉儲	<ol style="list-style-type: none"> 1. 每筆資料紀錄是由「各魚市場」匯總後的數據，只有對「交易地」描述。倘若資料記錄能增加「供應單位」的描述，則有更多的市場情報。 2. 「時間」維度只能下挖至「月」。若能下挖至「日」，則更能分析市場需求以及產品季節性產出的波動性。 3. 缺少交易金額的紀錄（衡量值）。