南華大學科技學院資訊管理學系

碩士論文

Department of Information Management
College of Science and Technology
Nanhua University

Master Thesis

基於多維數據模型的動態大數據分析方法

A Multidimensional Data Model-based Analysis Approach for Dynamic Big Data

鄭嘉雄

Jia-Xiong Zheng

指導教授：邱宏彬 博士

Advisor: Hung-Pin Chiu, Ph.D.

中華民國 109 年 7 月

July 2020

南 華 大 學

科技學院資訊管理學系

碩 士 學 位 論 文

基於多維數據模型的動態大數據分析方法

A multidimensional data model-based analysis approach for dynamic big data

研究生：莊嘉雄

經考試合格特此證明

口試委員：林迪漢

陳淑宗榮

邱宏彬

指導教授：邱宏彬

系主任(所長)：陳信良

口試日期：中華民國 109 年 7 月 2 日

南華大學資訊管理學系碩士論文著作財產權同意書

立書人：鄭嘉雄 之碩士畢業論文

中文題目：

基於多維數據模型的動態大數據分析方法

英文題目：

A multidimensional data model-based analysis approach for dynamic big data

指導教授： 邱宏彬 博士

學生與指導老師就本篇論文內容及資料其著作財產權歸屬如下：

■ 共同享有著作權

□ 共同享有著作權，學生願「拋棄」著作財產權

□ 學生獨自享有著作財產權

學　　生：鄭嘉雄（請親自簽名）

指導老師：邱宏彬（請親自簽名）

中 華 民 國　109年 7 月 13 月

誌　　謝

首先，我要感謝我的指導教授—邱宏彬老師，在我對論文研究主題之定位還不是很清楚的時候，是邱老師一次次與我開會討論，慢慢將研究方向與研究重點定出來，且後續的研究過程中，邱老師也不斷針對研究上有問題的地方，適時建議我調整、修正之方向並闡述他的想法及觀點，從中讓我學習到用不同的角度來看待問題及處理問題思維。

接著要感謝陳張宗榮老師以及林迺衛老師兩位口試老師，在口試時建議我論文可以再調整、修改的更好的地方，讓這篇論文更加完善。也感謝南華大學健康資料收集、分析團隊，提供我軟、硬體設備以及模擬資料，讓我研究得以順利進行。

基於多維數據模型的動態大數據分析方法

學生：鄭嘉雄　　指導教授：邱宏彬

南華大學　資訊管理學系碩士班

摘　要

多維數據模型常應用於處理靜態資料分析工作，而面對資料快速增長的時代，該模型如何因應動態增長之資料是本研究欲探討之議題。本研究以全國高級中等以下學校學生體格生長資料分析工作作為動態資料增長之分析情境，並以分散式與集中式資料庫架構概念為基礎，分別設計、實作集中式與分散式之多維數據模型，再以9年的學生體格生長模擬資料做為測試資料集，將資料批次匯入兩種多維數據模型中並測試其查詢分析資料之效率。最後依據實驗結果探討，本研究認為集中式多維數據模型較適用於分析資料量少的情況，而分散式多維數據模型則適合處理分析資料量較大情況。

關鍵詞：多維數據模型、分散式資料庫架構、集中式資料庫架構

A multidimensional data model-based analysis approach for dynamic big data

Student: Jia-Xiong Zheng Advisor: Hung-Pin Chiu, Ph.D.

Department of Information Management
Nanhua University
Master Thesis

ABSTRACT

Multidimensional data model is often used to deal with the analysis of static data. In the face of the rapid growth of data, how this model responds to the dynamic growth of data is the subject of this study. This study takes the analysis of the physical growth data of students in senior secondary schools and below as the analysis scenario of dynamic data growth, and based on the concept of decentralized and centralized database architecture, designs and implements centralized and decentralized multidimensional data models respectively Then, the 9-year student physical growth simulation data is used as the test data set, and the batch of data is imported into two multi-dimensional data models and the efficiency of querying and analyzing the data is tested. Finally, based on the experimental results, this study believes that the centralized multidimensional data model is more suitable for the case of a small amount of analysis data, while the decentralized multidimensional data model is suitable for the case of a larger amount of analysis data.

Keywords: Multidimensional data model, Distributed database architecture, Centralized database architecture

目錄

圖目錄

表目錄

第一章 緒論

第一節　研究動機

隨著科技進步及資訊化的普及，資料庫系統已廣泛應用在產、官、學等界，利用資料庫系統協助資料的存放及管理工作已是相當普遍，目前絕大多數使用資料庫系統主要是應用在基本日常事務的處理，如訂單交易紀錄、管理等反覆性的日常性交易活動，此等應用稱之為「線上交易處理」(OLTP, Online transaction processing)。而 OLTP 主要會採用關聯式資料庫，優點在於高效率處理頻繁異動的資料及節省資料儲存空間。

現今各界都希望從過去累積而來的大量資料中挖掘有價值的規則或是現象，於是對資料分析工作之需求逐漸提升，面對大量歷史資料分析之工作，對於存放原始資料的 OLTP 資料庫因採用關聯式設計架構之高度正規化關係，資料多分散存於各資料表，若要完成資料分析工作，則需要進行大量資料表連接(Table join)才可取得分析結果，此舉會耗費大量的運算成本，造成分析效率不佳，因此對於大量資料分析之工作，一般會依照分析需求，有主題的建立資料倉儲並利用多維數據模型存放資料，該模型優勢在於能快速取得存放在各個維度的靜態分析結果資料，此應用稱之為「線上分析處理」(OLAP, Online analytical processing)。

然而資料倉儲有一「非易變性」(Non-volatile)之特性，意即當資料進入資料倉儲後就不應再加以變更，僅能固定且週期性的方式加入新的歷史資料。而面對於大量且隨時間動態增長資料，若持續性的將資料加入同一多維數據模型中，隨著資料增長，對於分析資料的查詢效率恐會受到影響。

隨著近年來資料量的快速成長，分散式資料庫之應用也快速的發展，開始從傳統集中式架構向分散式架構發展，從集中式存儲及計算走向分散式存儲及計算。因此本研究採用集中式及分散式之資料庫系統架構設計多維度資料模型，並且模擬持續性的將大量資料加入集中式模型與分散式模型，測試兩者架構模型對於資料查詢的效能表現，以及探討兩者對於動態大數據分析優勢及劣勢。

第二節　研究目的

本研究採用全國高級中等以下學生體格生長資料分析工作做為研究情境，依據該情境之分析需求設計集中式與分散式之多維數據模型，並將其實作以及測試分析表現。目

前全國高級中等以下學校之學生約莫有兩百多萬人，依教育部規定學校每學期必須對每位學生量測一次身高及體重並紀錄量測結果，且教育部每年會將近三年各校學生體格生長量測結果進行統計分析。此情境之學生體格生長資料每半年會增長兩百多萬筆，具有大量動態增長之特性，且每年會固定將資料進行分析，因此該情境符合本研究所要探討的情況。而因為全國高級中等以下學校學生之體格生長資料有個資安全問題，因此本研究採用以模擬資料方式做為測試資料集進行實驗，此測試資料集是以真實資料量體作為參照模擬而來。

本研究根據體格生長模擬資料集，設計資料分析之項目，進而依照須分析之項目設計並且實作集中式及分散式架構之多維數據模型，再批次匯入資料至兩模型中，並依序測試兩者資料查詢效率表現，希望達到以下目的：

1. 探討集中式架構之多維數據模型對於動態大數據分析效率表現。
2. 探討分散式架構之多維數據模型對於動態大數據分析效率表現。
3. 探討集中式架構及分散式架構之多維度資料模型在動態大數據分析應用上的表現比較。

第三節　論文架構

本論文將分為五章，各章節主要內容簡述如下：

第一章、 說明研究背景與動機並訂定本研究之目的。

第二章、 依據研究目的探討相關之文獻，主要介紹資料倉儲基本概念、多維數據模型構成及集中式與分散式資料庫架構及其優劣、特性。

第三章、 本章主要介紹研究流程、設定數據分析需求以及設計因應查詢分析資料之多維數據模型，最後說明本研究實驗設計之方式。

第四章、 說明各項實驗測試結果與比較。

第五章、 研究總結探討並說明研究限制及未來方向。

第二章 文獻探討

第一節　資料倉儲

資料倉儲一詞是由資料倉儲之父 William Inmon 於 1990 年提出，主要功能是將經年累月所累積的大量資料，透過資料倉儲理論所特有的資料儲存架構，作一有系統的分析整理，以利後續如線上分析處理(OLAP)、資料採礦(Data Mining)等分析方法之應用，並進而支援如決策支援系統(DSS)、主管資訊系統(EIS)之建立，幫助決策者能快速有效的自大量資料中，分析出有價值的資訊[11]。

一　資料倉儲定義

蕭凱文等人指出[9]：「資料倉儲可以說是專門為查詢和分析設計，用來複製交易資料的一種架構。資料倉儲是將企業內異質性的資料加以合併，將歷史資料取出，來輔助決策分析。」

沈兆陽(2001)認為[1]：「資料倉儲不僅是指所需的分析資料，它還包含處理資訊所需的應用程式，像是將資料由外部媒體轉入資料倉儲內的應用程式，還有將資料進行分析並呈現給使用者的應用程式。」

資料倉儲是一份為了查詢與分析為目的，而由交易系統中複製出來的資料 [8]。資料倉儲為資料的集合體，是一個具有主題導向(Subject-Oriented)、經過整合(Integrated)、時間變動性(Time-Variant)、以及非易變性(Non-Volatile)等四項特質之資料集合 [7]。針對 Inmon 提出的資料倉儲四種特質整理如表 1。

表 1 資料倉儲特質

特質	說明
主題導向	資料倉儲儲存的是與某一主題相關的資料，主題可以不只一個，但是不需要儲存與主題無關的資料。
經過整合	將不同來源的資料以統一的規範，經過整理後 儲存在一起，所謂統一的規範指的是相同的資料型態、格式或度量等。
時間變動性	資料倉儲的資料都是過去所發生的事實資料，但是並非儲存過去的所有資料，而是在某一時間範圍內的資料。
非易變性	當我們將過去的資料轉入資料倉儲後，就不應該加以變更，但是可以以固定且週期性的方式定期加入新的歷史資料。

資料來源：[3]

二　資料倉儲架構

資料倉儲的資料從同質或異質、外部或內部的作業性資料庫萃取出來，經過資料轉換服務(DTS, Data Transformation Services)進行資料轉換或是將資料匯入資料倉儲，而資料轉換服務(DTS)包含了資料驗證(Data Validation)、資料搬移(Data Migration)、資料淨化(Data Scrubbing)、資料轉換(Data Transformation)的四個程序[6]，如圖 1 所示。

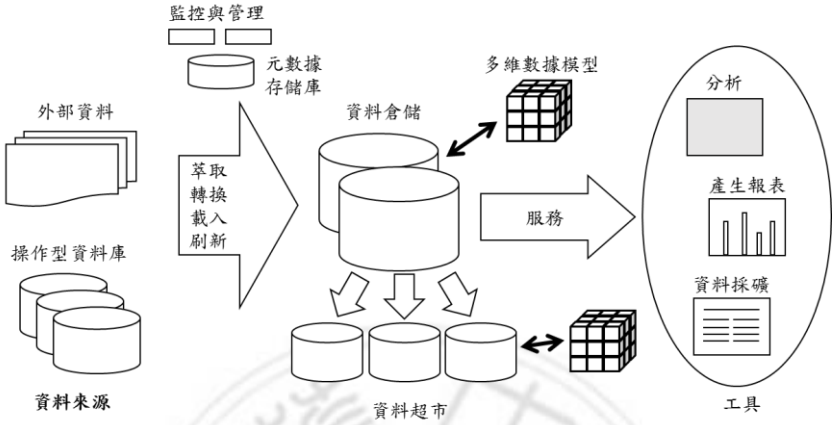

圖 1 資料倉儲架構

資料來源：[6]

三　交易型資料庫與資料倉儲比較

資料庫與資料倉儲最大差異點是資料庫進行資料的新增、刪除、修改、查詢等功能其主要目的於管理資料庫的存取；而資料倉儲重視的是資訊的獲得，以問題決策分析導向，強調多維度視野，視覺化的提供決策者資訊可用性[5]。

表 2 為資料倉儲與傳統資料庫差異比較。

表 2 資料倉儲與傳統資料庫差異比較

比較項目	交易型資料庫(OLTP)	資料倉儲(OLAP)
對象	針對工商企業現行業務的自動化設計	針對企業資料分析需求設計
功能	協助人員進行既有活動	協助人員進行決策分析
使用者	多使用者同時使用資訊系統	沒有太多使用者同時使用
資料	偏重細節	偏重高層級彙總資訊
來源	使用者日常生活輸入	OLTP 系統
資料庫內容	可以新增、刪除、修改	歷史性資料、不常更新
處理方式	以交易為單位	使用者的要求
設計方式	正規化	非正規化

資料來源：[5]

第二節　多維數據模型

一　多維數據模型構成

多維數據模型以欲探討的主題為主軸，由一至多張事實表(Fact Table)和維度表(Dimension Table)構成。構成成員資料整理如下表。

表 3 多維數據模型構成成員

構成成員	說明
事實表 (Fact Table)	存放數值性度量值(Measure)及連結維度表的外來鍵(Foreign Key)，一個事實表連接一個或多個維度表。一般而言存放到事實表的資料不會再更改，只能新增資料到表格中。
維度表 (Dimension Table)	維度表用以參照事實表中的資料，也欲分析資料的屬性，而屬性可能會有階層(Hierarchy)。例如銷售時間及銷售額等，其中銷售額的屬性可能會有年、季、月、周及天等不同階級。
度量值 (Measure)	度量值為將維度表屬性彙總計算出的項目，例如銷售數量、銷售金額等。分析者能透過度量值輕易地得到分析主題相關的資訊。

資料來源：[2]

二　多維數據模型設計架構

多維數據模型依照事實表及維度表結構及資料，以資料表的聚合關係分別設計為星狀模型(Star Schema)、雪花狀模型(Snow Schema)以及事實星座模型(Fact Constellation Schema)三種模型。說明如下。

1. 星狀模型

星狀模型是資料倉儲的最簡單樣式並且是最廣泛用於開發資料倉儲的方法。星狀模型中僅有一個紀錄分析主題的事實表，每一個維度僅用單一維度表表示。如圖 2 所示。

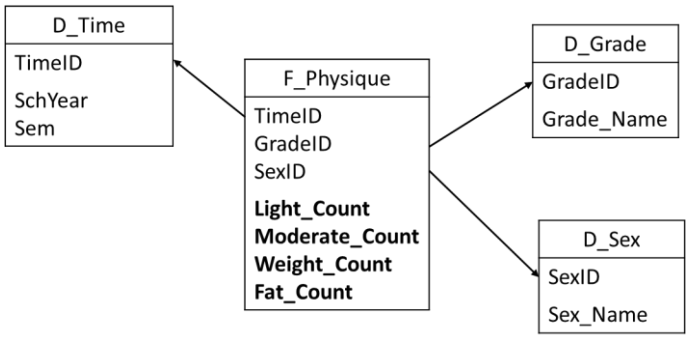

圖 2 星狀模型圖

2. 雪花狀模型

雪花狀模型將星狀模型中的某些維度表再進行正規化，並將維度表間建立關聯，減少資料多餘情形。比起星狀模型更能表示維度階層如圖 3 所示。

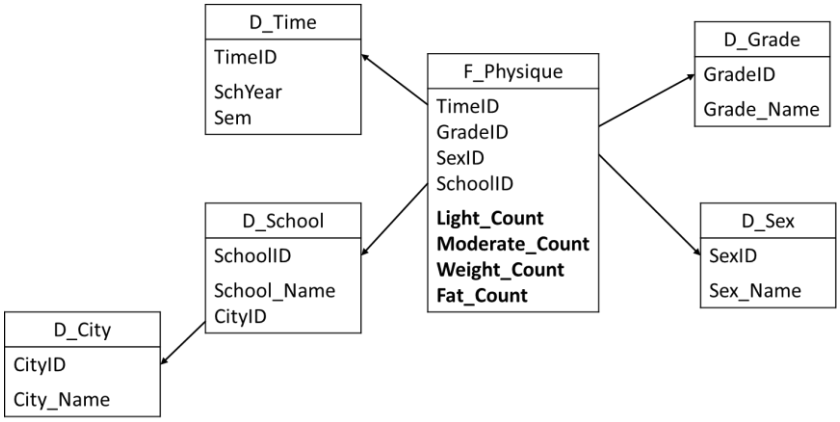

圖 3 雪花狀模型圖

3. 事實星座模型

事實星座模型內包含了不同的分析主題，相較於星狀模型及雪花狀模型僅有一個事實表，事實星座模型包含了多個事實表，且一個維度表能連結不同的事實表。如圖 4 所示。

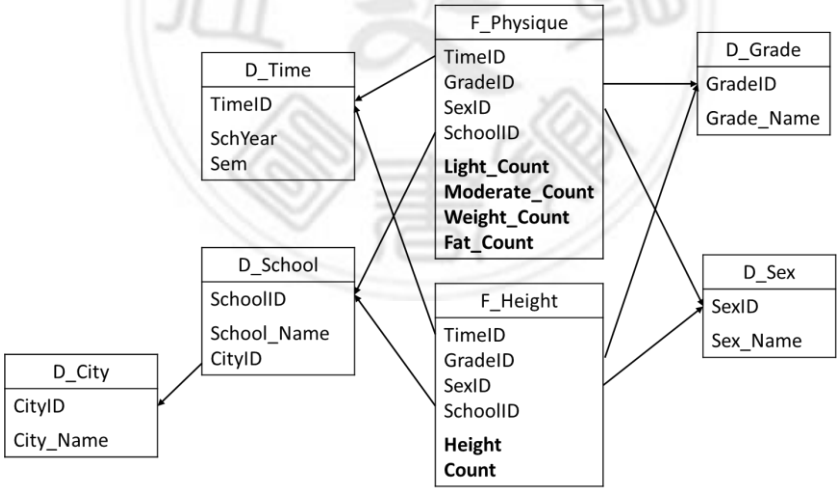

圖 4 事實星座模型圖

三　資料方體(Data Cube)

資料方體為多維度資料集之專有名詞，使用者可以對資料方體進行切片(Slice)、切塊(Dice)、上卷(Roll-up)、鑽取(Drill-down)以及旋轉(Pivot)等方法從不同角度取得多維度資料[2]。

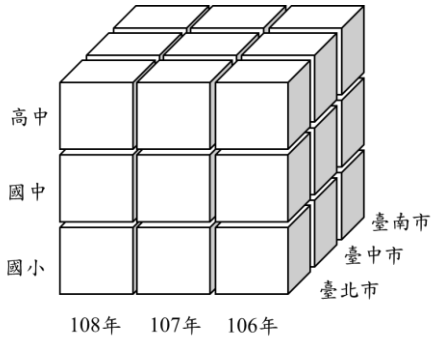

圖 5 資料方體

1. 切片 (Slice)

選擇某維度中的特定值進行分析，例如選擇臺北市、臺中市及臺南市 106 年到 108 年的國小身高數據。如圖 6 所示。

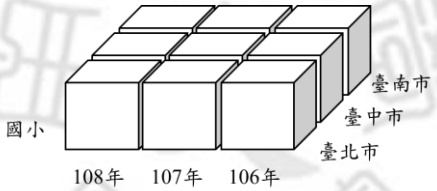

圖 6 資料方體切片

2. 切塊 (Dice)

選擇某維度中特定區間的值進行分析，例如選擇臺北市、臺中市及臺南市 106 年到 107 年的各學制身高數據。如圖 7 所示。

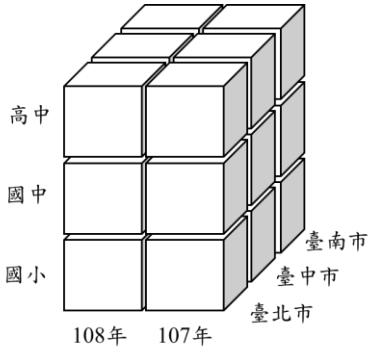

圖 7 資料方體切塊

3. 上卷 (Roll-up)

將某一維度資料由下層提升至上層，明細資料至高階匯總資料，例如以臺北市、臺中市及臺南市可依地區匯總成北部、中部及南部，看到各地區之資料。如圖 8 所示。

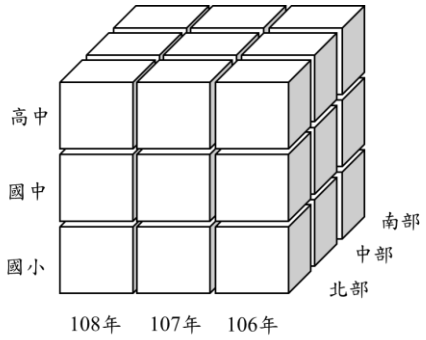

圖 8 資料方體上卷

4. 鑽取 (Drill-down)

鑽取是與上卷相反的動作，將維度階層由上層降至下層，使資料轉換至更細節的數據，例如由各學制查看到各年級之身高數據。如圖 9 所示。

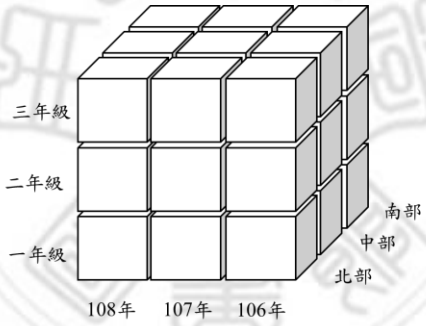

圖 9 資料方體鑽取

5. 旋轉 (Pivot)

改變資料立方體中的維度位置，如同二為陣列的行列轉換，例如通過旋轉將縣市與學制位置互換。如圖 10 所示。

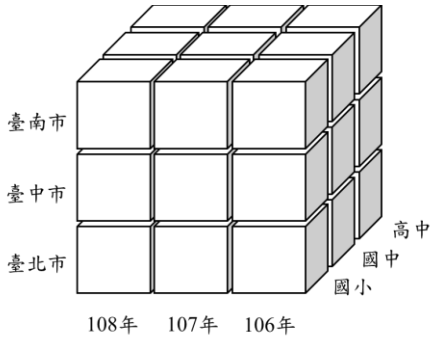

圖 10 資料方體旋轉

第三節　集中式與分散式資料庫架構

一　集中式資料庫

集中式資料庫(Centralized database)資料的儲存、維護及查詢全集中在一台資料庫伺服器。集中式資料庫架構如圖 11，所有資料存於單一資料庫，各個用戶端的請求統由單一資料庫回應。此架構優勢在於因所有資料存儲單一位置，有助於盡可能保持資料的準確和一致性並增強數據可靠性且資料安全性較高，在操作方面易於從同一位置同時訪問所有訊息，維護上也較為容易[9]。

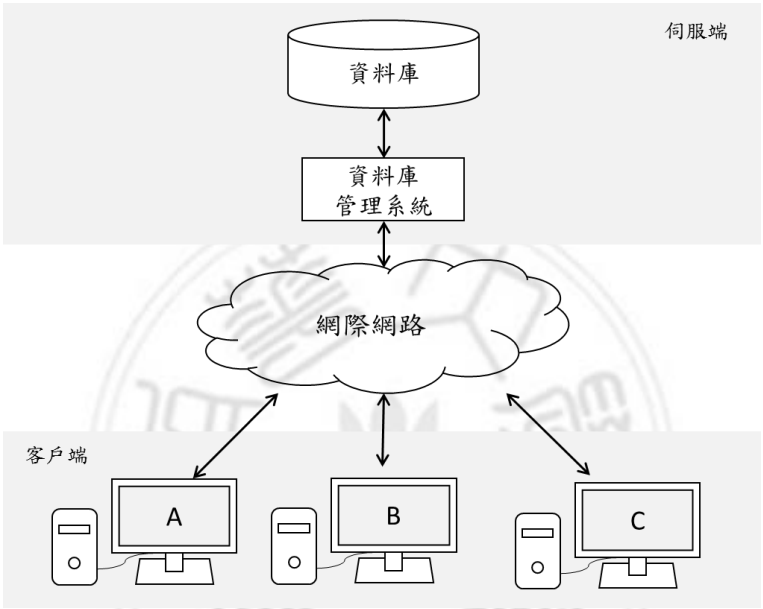

圖 11 集中式資料庫系統架構圖

資料來源：[10]

二　分散式資料庫

分散式資料庫(Distributed database)將資料依照特性分散儲存在不同的資料庫伺服器，再以網路將分散在電腦網路中的眾多邏輯資料庫關連在一起的集合如圖 12。分散式資料庫優勢整理如下[4]。

1. 增加可靠性與可用性

   由於資料分散存在各個資料庫，如某個資料庫故障，使用者僅會無法取得部分資料，但能可以存取其他部分資料。

2. 改善效能

   資料區域化(data localization)可減少爭奪 CPU 和 I/O 服務的情形，同時縮短在廣域網路中的存取延遲。將資料切分存至多個資料庫，由於各資料庫資料存量較少且藉由平行查詢處理，查詢資料的效能較佳。

3. 容易擴展

系統的擴展無論是增加資料、增加資料庫大小或是增加處理器，會比較容易。

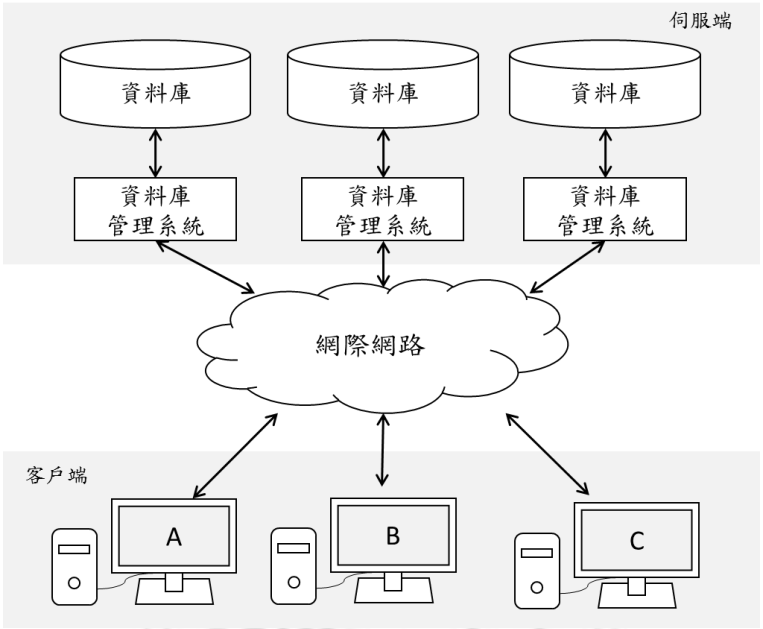

圖 12 分散式資料庫系統架構圖

資料來源：[10]

在分散式資料庫中，必須決定那些資料要存於哪一個資料庫中，因此需要運用資料分段的技術，一般而言資料分段有水平分段(horizontal fragmentation)及垂直分段(vertical fragmentation)兩種技術。水平分段是將資料列依照某種意義來分組，然後再將這些分段配置到分散式系統的不同資料庫中[4]。水平分段方式如圖 13 所示，將記錄身高體重資料表存放之資料，依性別資料分割存於不同資料表。垂直分段是將部分資料欄進行切分，垂直分段方式如圖 14，將體重資料分割存於不同資料表，兩張表都要保留相同主鍵，這樣才能確保資料完整性。

ID	SexID	Height	Weight
001	男	165	70
002	男	135	39
003	男	157	62
004	女	178	96
005	女	196	75
006	女	154	52

ID	SexID	Height	Weight
001	男	165	70
002	男	135	39
003	男	157	62

ID	SexID	Height	Weight
004	女	178	96
005	女	196	75
006	女	154	52

圖 13 水平分段

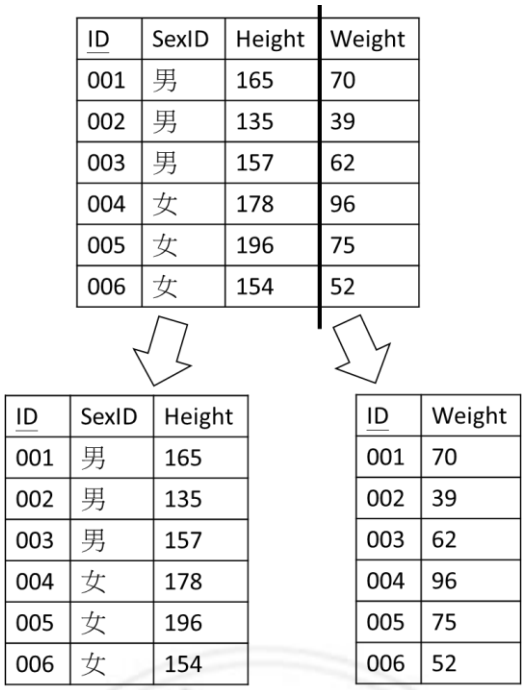
ID SexID Height Weight
001 男 165 70
002 男 135 39
003 男 157 62
004 女 178 96
005 女 196 75
006 女 154 52
ID SexID Height
001 男 165
002 男 135
003 男 157
004 女 178
005 女 196
006 女 154
ID Weight
001 70
002 39
003 62
004 96
005 75
006 52

圖 14 垂直分段

第三章 研究設計

本章第一節會先說明研究流程，介紹整體研究之脈絡；第二節會探討學生體格生長資料所需要產出的分析項目；第三節說明因應分析需求所設計之集中式與分散式的多維數據模型架構及其概念；最後第四節則是說明實驗所使用之測試資料集處理方式、所需的實驗工具、實驗方法以及實驗系統的配置方式。

第一節　研究流程

本研究之研究流程共有六步驟如圖 15，首先第一步會依據本研究設定之分析情境設計所需之數據分析需求，訂出要分析的數據及分析方式；第二步再以此分析需求下規劃、設計因應資料查詢及分析之多維數據模型；第三步則是針對設計出來的多維數據模型設計實驗測試方式；第四步則是對於前述設計之模型及實驗系統進行實作；再來第五步進行實驗測試；最後第六步將探討實驗結果及發現。

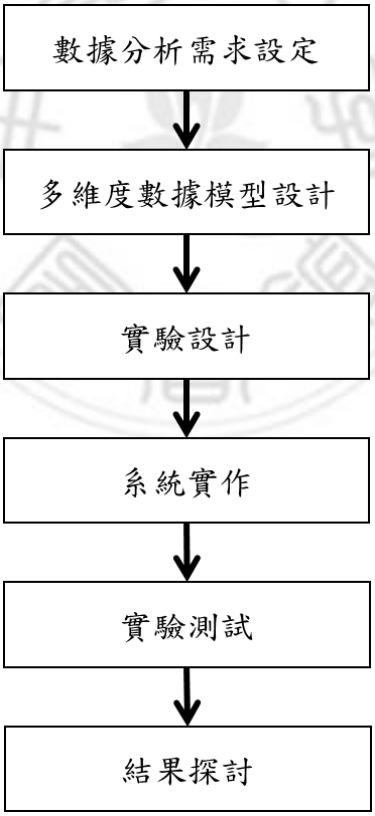

圖 15 研究流程圖

第二節　數據分析設定

本研究設計之資料分析情境，是以學生體格生長資料分析需求為基礎進行模擬，其資料量是以真實狀況為參照模擬而來，以每學期增長兩百多萬筆資料，共模擬 9 年之資料。在分析需求方面本研究以教育部健康資訊系統所提供之為分析項目作為參考，主要以敘述性統計分析為主，在數值型資料的分析方面，主要分析資料之集中趨勢(central tendency)以及離散程度(statistical dispersion)，而在類別型資料則須分析個數總計及百分比。分析面向要可以看出各縣市、鄉鎮區、學校、學校轄屬單位、學校設立別、學年度、學期、年級、性別、半歲齡、學生身分等不同面向的統計結果。分析的資料有包含身高、體重、身體質量指數(BMI)及健康體位判別結果，其中身高、體重及 BMI 資料為數值型資料，而健康體位判別資料屬於類別型資料，分為過輕、適中、過重及肥胖四類。分析項目需求整理於表 4。

表 4 分析項目需求表

分析資料	資料型態	分析項目	分析面向
身高	數值	集中趨勢：平均數、中位數、眾數 離散程度：標準差、百分位數	縣市、鄉鎮區、學校、學校轄屬單位、學校設立別、學年度、學期、年級、性別、半歲齡、學生身分
體重			
BMI			
健康體位判別	類別	個數總計、百分比	

第三節　設計多維數據模型

依照資料分析需求，本研究設計因應資料查詢之多維數據模型，並設計集中式及分散式兩種資料庫架構，其說明如下。

一　集中式多維數據模型

此架構將歷年學生體格生長資料透過萃取、轉置、載入之動作，依照資料分析需求，將資料整合於設計之多維數據模型。而此多維數據模型皆存放於單一資料庫之中。如圖 16 所示。

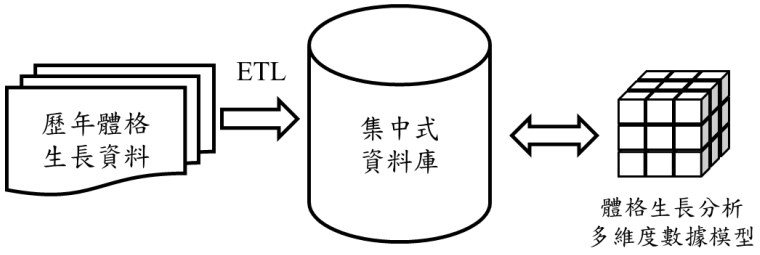

圖 16 集中式資料庫示意圖

體格生長資料分析之集中式多維數據模型，採用事實星座模型所設計，此模型記錄各個維度健康體位判別結果總數資料以及記錄各個維度底下身高、體重及 BMI 分布狀況，以提供後續分析各個維度體格生長數據之應用。此模型共有 10 張維度表及 4 張事實表所組成，如圖 17 集中式體格生長資料多維數據模型圖所示，維度表部分有歷年學校、學校設立別、學校轄屬單位、鄉鎮區、縣市、學期、年級、性別、學生身分及半歲齡,其中歷年學校維度表有與學校設立別、學校轄屬單位及鄉鎮區維度表進行外鍵關聯，鄉鎮區維度表再與縣市維度表進行外鍵關聯，以減少資料重複性；事實表部分首先介紹健康體位事實表，此表有 7 個維度外鍵欄位及 5 個度量值還有 1 個流水號欄位所構成，維度欄位有學年度、學校代號、學期代號、年級代號、性別代號、身分代號以及半歲齡代號，度量值欄位則有體位總數、體位過輕數、體位適中數、體位過重數及、體位肥胖數，而流水號欄位資料是代表各維度切分後的顆粒，讓身高、體重及 BMI 事實表進行外鍵關聯，目的是減少身高、體重及 BMI 三張事實表的維度資料重複，而身高、體重及 BMI 事實表則需再整理每一個顆粒底下對應數值分布狀況,如身高事實表須再對每一個顆粒底下再依相同身高進行分類加總，以供後續進行數值類型資料之分析。

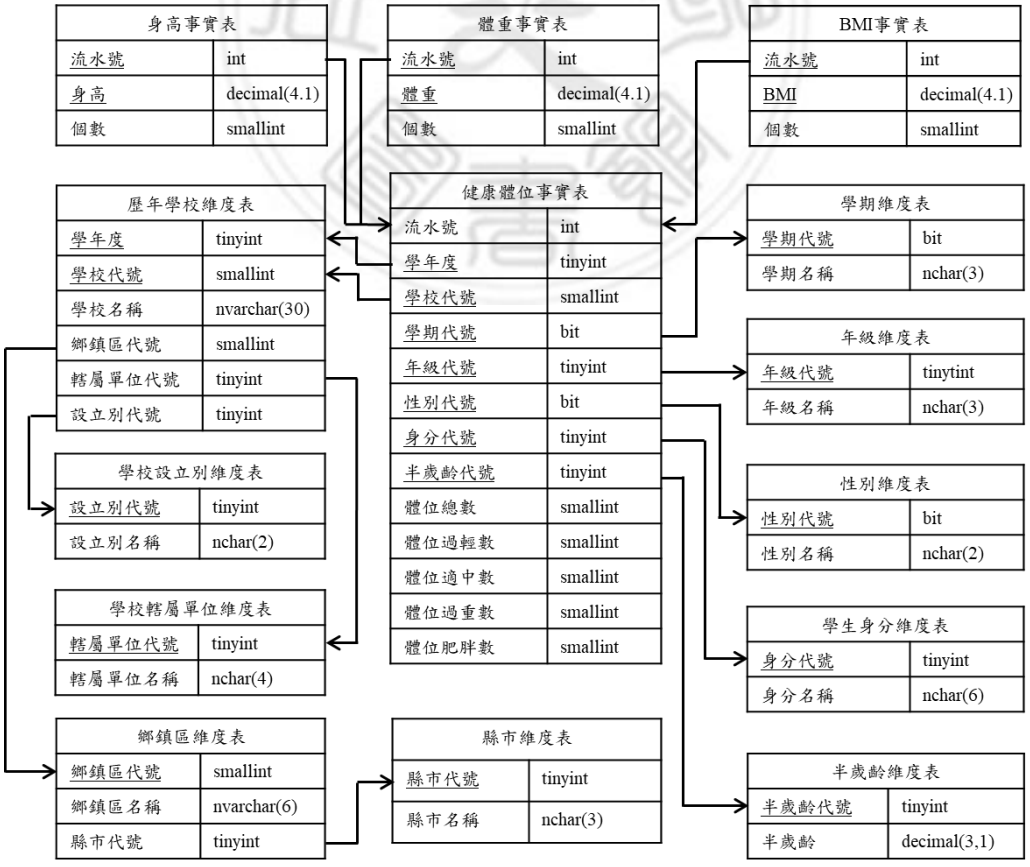

圖 17 集中式體格生長資料多維數據模型圖

二　分散式多維數據模型

本研究設計之分散式架構是以集中式架構為基礎進行拆分，依照資料量評估，設計以每一個學年度資料進行水平分段，將分段資料分散儲存於各資料庫，以防止資料大量累積於單一資料庫，導致查詢效能變差之狀況，資料分段方式如圖 18。此架構每一年即會產生一個資料庫，每一個資料庫存放對應當年的多維數據模型及資料，往後每年新增之資料就會以此方式建立新的資料庫加以儲存。如圖 19 所示

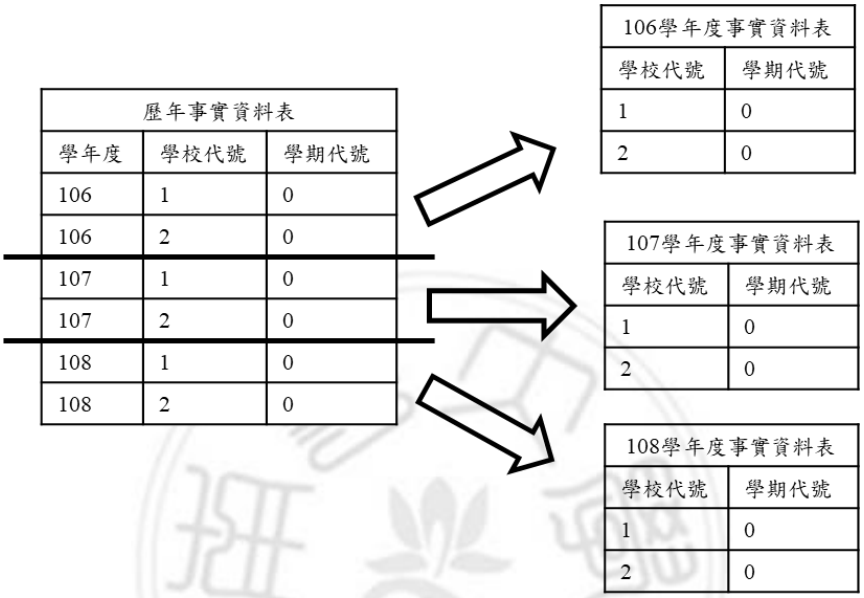

圖 18 分散式模型資料分段示意圖

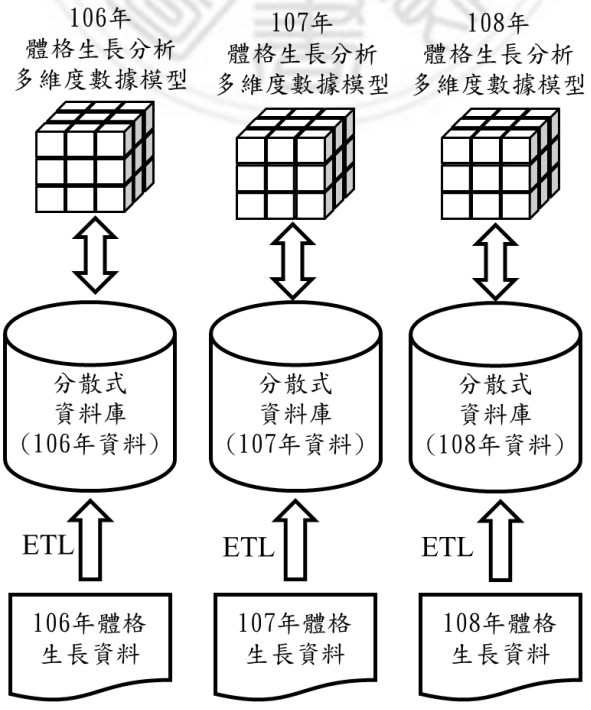

圖 19 分散式資料庫示意圖

在分散式多維數據模型部分，因每一學年度資料存於對應的資料庫，所以相較於集中式的多維數據模型，則不須紀錄學年度維度資料即可依照訪問的資料庫區別資料所屬之學年度，所以學校維度表及健康體位事實表相較於集中式多維數據模型則不需紀錄學年度資料。

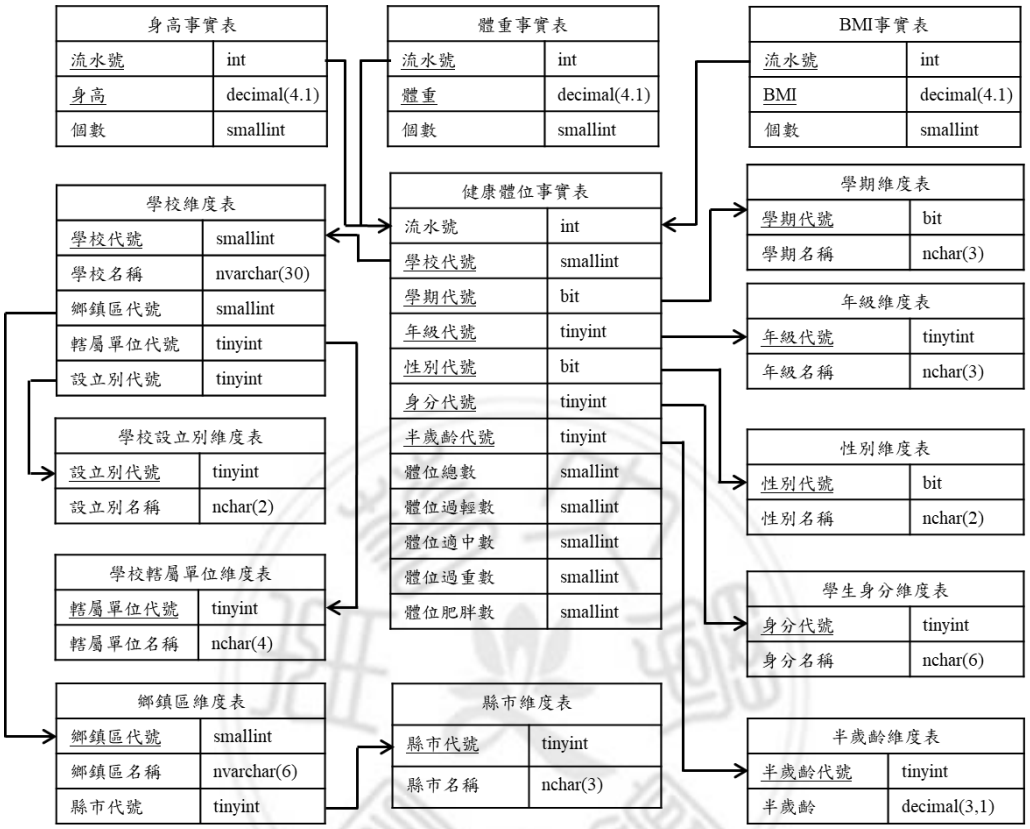

圖 20 分散式健康體位多維數據模型

第四節　實驗設計

本節將說明本研究使用之模擬資料之資料量及其資料模擬方式，後續介紹實驗工具及實驗方法，最後說明整套實驗系統的配置方式。

一　實驗資料集

本研究是採用高級中等以下學生體格生長之資料作為測試資料集，因真實之學生體格生長資料有個資問題，故本研究將以模擬之體格生長資料作為測試資料集。此測試資料集是以實際之體格生長資料及其資料量為基礎模擬而來，而資料模擬方式是先模擬產生某一學年度上學期學生體格生長資料，共 2,362,192 筆，再依產生資料複製轉存成各學年度上下學期之資料，共模擬 9 年 18 個學期的資料，其資料量達到 42,519,456 筆。

而測試資料集是以 OLPT 資料庫之模式存放資料，因此須再將資料進行萃取、轉置

並載入本研究設計之兩種多維數據模型，經過整理後的資料在健康體位事實資料表單一學年度即存放 280,966 筆資料，9 年共 2,528,694 筆資料；而身高事實資料表單一學年度則是有 4,121,114 筆資料，9 年共 37,090,026 筆資料；至於體重事實資料表單一學年度則是有 4,229,782 筆資料，9 年共 38,068,038 筆資料；最後是 BMI 事實資料表單一學年度則是有 3,698,986 筆資料，9 年共 33,290,874 筆資料。

二　實驗工具

本研究進行實驗之工具分別有兩台資料庫伺服器以及一台網站伺服器，資料庫伺服器是佈署本研究設計之資料庫及多維數據模型所需，因要探討集中式及分散式之資料模型查詢效率，所以兩台資料庫伺服器的硬體及軟體規格設定上要一致，將兩者環境與效能等外部因素調整到最一致，這樣才能客觀評量兩架構查詢效率；而網站伺服器則是架設本研究針對多維數據模型分析效率測試所實作之網站。3 部伺服器之軟、硬體規格介紹如下。

1. 資料庫伺服器*2

   硬體規格

   - CPU：Intel(R)Xeon(R)CPU E3-1220 v6@ 3.00GHz
   - RAM：16 GB 2400 MHz
   - Disk：1TB

   軟體規格

   - OS：Windows Server 2012 R2 Standard
   - DBMS：SQL Server 2012 Express

2. 網站伺服器*1

   硬體規格

   - CPU：Intel(R)Xeon(R)CPU E5-1630 v3@ 3.70GHz
   - RAM：8 GB 2133 MHz
   - Disk：1TB

   軟體規格

   - OS：Windows 10 專業版
   - Web server：IIS 10

三　實驗方法

本研究欲探討動態大量資料之新增對於多維數據模型的查詢分析效率之影響，採用分析高級中等以下學生體格生長資料作為實驗情境，依據表 4 之分析需求所示，此分析情境有 11 個維度的分析面向，因各維度排列組合多達數萬種以上，所以本研究以實際較常分析之維度進行資料分析來測試資料查詢效能，常分析維度有學年度、縣市與年級，分析項目則是有身高、體重、BMI 之百分位數、平均數及標準差，健康體位則是分布個數及比率。

實驗方式是將模擬資料以學年度為單位分批匯入數據模型中，並且每次匯入資料後針對每一項分析項目進行 3 次的資料查詢，再以 3 次查詢的時間取平均數表示平均查詢時間，以此方式統計批次匯入 9 年資料對於查詢時間的變化差異。

在資料查詢方面，分散式多維數據模型欲查詢之資料若是分散存於兩台資料庫伺服器，本研究則會以平行查詢之方式，同時命令兩資料庫伺服器查詢資料；而集中式多維數據模型之資料僅存於單一資料庫伺服器，則未使用平行查詢之查詢方式。另外為排除可能因資料庫管理系統快取機制所產生的干擾，在每次查詢資料前本研究都會先將資料庫管理系統的快取資料先作清空的動作。

本次實驗共分為三個部分，列點說明如下。

1. 批次將部分資料匯入 OLTP 資料庫，測試其查詢分析資料的效率
2. 批次將資料匯入集中式多維數據模型，測試其查詢分析資料的效率
3. 批次將資料匯入分散式多維數據模型，測試其查詢分析資料的效率

四　實驗系統配置

依上述設計之研究方式，本研究將要測試之資料庫佈署在資料庫伺服器並透過自行開發之網站應用程式，使用者透過只要瀏覽器連線至此網站，即可操作網站應用程式對資料庫進行連線並查詢所需要分析的資料。

OLTP 資料庫及集中式多維數據模型在測試時分別會佈署在個別資料庫伺服器上，透過對網頁操作即可分別取的兩方查詢之資料及計算查詢時間。而在分散式多維數據模型佈署方面，100 學年度到 103 學年度四年份之學生體格生長資料會放置在第一部資料庫伺服器，另一部資料庫伺服器佈署了 104 學年度到 108 學年度五年份的學生體格生長資料。

第四章 效率分析與比較

第一節　OLTP 資料庫數據分析效率測試

本節將測試如若未以資料倉儲之模式預先將分析資料進行處理、轉換、儲存成多維數據模型架構，僅以 OLTP 資料庫直接查詢數據分析資料，所需要花費之時間成本。實驗將以批次匯入 5 年資料方式，且每匯入一年的資料就測試查詢歷年各縣市各年級體格生長分析資料，並記錄其查詢時間變化。測試結果由表 5 可看到每增加一年的原始資料，其查詢時間就會以等差級數上升，平均約上升 20 秒左右查詢時間，到資料量增加至 5 年份時，查詢時間已需要 100 秒左右。

表 5 OLTP 資料庫數據分析效率測試結果

資料查詢項目	查詢時間(ms)	累積資料量				
		1 年份	2 年份	3 年份	4 年份	5 年份
健康體位分析資料	第一次查詢時間	17,358	39,173	62,248	83,274	110,060
	第二次查詢時間	17,267	39,146	63,361	84,304	107,575
	第三次查詢時間	17,254	39,094	62,983	82,365	107,596
	平均查詢時間	17,293	39,138	62,864	83,315	108,410
身高分析資料	第一次查詢時間	9,525	28,684	50,171	70,153	95,557
	第二次查詢時間	9,597	28,207	50,984	69,819	95,915
	第三次查詢時間	9,630	27,946	50,523	69,804	95,052
	平均查詢時間	9,584	28,279	50,559	69,925	95,508
體重分析資料	第一次查詢時間	11,317	29,847	52,446	75,983	92,209
	第二次查詢時間	9,970	29,699	52,575	76,348	91,910
	第三次查詢時間	10,411	29,314	52,316	76,899	91,901
	平均查詢時間	10,566	29,620	52,446	76,410	92,007
BMI 分析資料	第一次查詢時間	9,698	28,200	55,420	80,867	109,923
	第二次查詢時間	9,590	28,273	55,059	81,047	110,258
	第三次查詢時間	9,381	28,618	55,032	81,268	109,673
	平均查詢時間	9,556	28,364	55,171	81,061	109,952
註：每一年份原始資料量為 4,724,384 筆						

第二節　多維數據模型分析效率測試

經由資料分析主題設計之多維數據模型查詢效率，為本研究主要實驗測試之重點，本節將分為兩個部分進行測試，第一部分測試長年資料累積於單一集中式多維數據模型對於資料查詢效率的變化；第二部分測試長年資料分散於個別分散式多維數據模型對於資料查詢效率的變化。兩者實驗測試方式皆是批次匯入 9 年之資料並測試查詢分析資料的時間變化。而查詢方式是以分析整個多維數據模型內的資料以及對其進行切片、切塊處理的分析兩種方式為基礎設計查詢主題，而查詢之主題本研究以教育部學生健康資訊系統主要提供的分析主題做為參考所設計，其查詢主題列表說明如表 6。

表 6 測試重點與查詢主題

測試查詢方式	測試重點	查詢主題
查詢整個多維數據模型內的資料分析	測試資料增長對於整體資料整合查詢能力	查詢歷年各縣市各年級體格生長分析資料
多維數據模型切片及切塊處理查詢	測試資料增長對於擷取部分資料整合查詢能力	查詢當年六都縣市國中小各年級體格生長分析資料
		查詢近三年六都縣市國中小各年級體格生長分析資料
		查詢當年、前 3 及前 6 年六都縣市國中小各年級體格生長分析資料

一 集中式多維數據模型分析效率測試

1. 測試查詢歷年各縣市各年級體格生長分析資料

在此部分測試結果如表 7，四項資料查詢項目皆會隨著資料增長對查詢時間也會成正比成長，而在查詢健康體位分析資料方面每增加一年份資料，平均查詢時間會上升約 300 毫秒左右，而在查詢身高、體重、BMI 資料方面每增加一年份資料，平均查詢時間會上升約 5,000 毫秒左右。

表 7 集中式多維數據模型查詢歷年各縣市各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量								
		1 年份	2 年份	3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	363	785	1,029	1,642	1,856	2,296	2,260	2,443	2,785
	第二次查詢時間	365	631	939	1,301	1,623	2,027	2,594	2,407	2,577
	第三次查詢時間	368	749	982	1,265	1,747	2,031	2,688	2,266	2,499
	平均查詢時間	365	722	983	1,403	1,742	2,118	2,514	2,372	2,620
身高分析資料	第一次查詢時間	4,619	10,607	15,728	20,110	25,612	30,537	35,719	36,907	39,636
	第二次查詢時間	4,531	9,701	15,352	20,407	25,832	30,439	36,759	37,983	41,722
	第三次查詢時間	4,860	10,568	15,386	20,723	25,707	30,601	36,248	39,288	40,430
	平均查詢時間	4,670	10,292	15,489	20,413	25,717	30,526	36,242	38,060	40,596
體重分析資料	第一次查詢時間	5,118	10,964	15,806	21,911	27,494	32,424	38,862	40,478	48,362
	第二次查詢時間	5,093	10,860	15,627	21,618	28,031	32,833	40,126	42,070	50,624
	第三次查詢時間	4,852	10,888	16,322	21,833	28,008	30,593	38,602	40,990	46,965
	平均查詢時間	5,021	10,904	15,918	21,787	27,844	31,950	39,197	41,179	48,650
BMI 分析資料	第一次查詢時間	4,856	8,964	14,295	18,680	25,902	27,636	34,231	39,910	44,338
	第二次查詢時間	4,198	8,923	12,653	19,250	25,236	27,016	35,175	36,010	44,871
	第三次查詢時間	3,794	8,892	13,693	19,131	24,318	26,696	34,472	35,598	43,967
	平均查詢時間	4,283	8,926	13,547	19,020	25,152	27,116	34,626	37,173	44,392

2. 測試查詢當年六都縣市國中小各年級體格生長分析資料

此部分測試結果可以出，在查詢健康體位分析資料方面每增加一年份資料，平均查詢時間都約莫 200 毫秒左右，對於資料增長沒有明顯影響；而在查詢身高、體重、BMI 資料方面可以看到資料量的累積增長會造成查詢新年度資料時間增加，每增加一年份資料其查詢時間約增長 1 秒左右，測試結果如表 8 所示。

表 8 集中式多維數據模型查詢當年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量								
		1 年份	2 年份	3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	245	235	199	184	221	231	203	251	298
	第二次查詢時間	233	186	181	182	186	200	182	239	216
	第三次查詢時間	183	184	190	182	186	186	198	184	185
	平均查詢時間	220	202	190	183	198	206	194	224	233
身高分析資料	第一次查詢時間	1,941	3,591	4,400	5,008	6,592	7,318	8,553	9,657	10,808
	第二次查詢時間	1,500	3,601	4,380	4,845	6,445	7,316	8,640	9,542	10,567
	第三次查詢時間	1,852	3,295	4,016	4,884	6,415	7,948	8,719	9,746	10,762
	平均查詢時間	1,764	3,496	4,265	4,912	6,484	7,527	8,637	9,648	10,712
體重分析資料	第一次查詢時間	1,520	3,592	5,080	5,921	6,865	8,076	8,995	10,074	11,210
	第二次查詢時間	2,006	3,593	4,864	5,984	7,009	7,987	9,015	10,041	10,953
	第三次查詢時間	2,013	3,654	4,881	6,015	6,844	7,821	8,904	9,939	10,844
	平均查詢時間	1,847	3,613	4,941	5,973	6,906	7,961	8,971	10,018	11,002
BMI分析資料	第一次查詢時間	2,096	3,072	3,730	4,531	5,183	6,015	7,119	8,806	9,677
	第二次查詢時間	1,361	3,025	3,771	4,658	5,185	5,901	7,175	8,877	9,497
	第三次查詢時間	1,412	3,139	3,458	4,521	4,831	6,099	7,174	8,916	9,712
	平均查詢時間	1,623	3,079	3,653	4,570	5,066	6,005	7,156	8,866	9,628

3. 測試查詢近三年六都縣市國中小各年級體格生長分析資料

在查詢健康體位資料每增加一年份資料，查詢時間都約莫 800 毫秒左右，最高查詢時間在匯入第五年份資料時測出，平均查詢時間是 1.087 秒，整體看來資料增長沒有明顯影響健康體位資料查詢時間；而三項數值型資料查詢結果每增加一年份資料其查詢時間約增幅約 1 秒，測試結果如表 9 所示。

4. 測試查詢當年、前 3 及前 6 年六都縣市國中小各年級體格生長分析資料

而此部分測試結果如表 10，在查詢健康體位資料每增加一年份資料，查詢時間都約莫 900 毫秒左右；而三項數值型資料查詢結果每增加一年份資料其查詢時間約增幅約 1 秒。

表 9 集中式多維數據模型查詢近三年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量						
		3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	620	758	1,077	1,035	713	985	980
	第二次查詢時間	548	745	1,029	738	728	954	984
	第三次查詢時間	564	742	1,154	773	710	937	987
	平均查詢時間	577	748	1,087	849	717	959	984
身高分析資料	第一次查詢時間	9,719	10,027	11,421	12,695	14,032	15,920	16,474
	第二次查詢時間	9,524	10,175	11,848	12,647	14,180	15,384	16,461
	第三次查詢時間	8,986	9,962	11,815	12,861	13,897	15,050	16,552
	平均查詢時間	9,410	10,055	11,695	12,734	14,036	15,451	16,495
體重分析資料	第一次查詢時間	9,423	10,823	11,968	13,417	14,394	16,313	16,940
	第二次查詢時間	9,800	11,099	12,370	13,152	13,980	16,258	17,194
	第三次查詢時間	9,647	10,703	12,119	12,692	14,702	15,591	17,038
	平均查詢時間	9,624	10,875	12,152	13,087	14,359	16,054	17,057
BMI分析資料	第一次查詢時間	8,038	9,710	10,449	10,590	11,972	14,095	14,998
	第二次查詢時間	8,011	9,213	10,316	11,162	12,327	14,462	15,182
	第三次查詢時間	8,187	8,744	10,138	11,173	11,540	14,143	15,161
	平均查詢時間	8,079	9,222	10,301	10,975	11,947	14,233	15,114

表 10 集中式多維數據模型查詢當年、前 3 及前 6 年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量		
		7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	859	950	962
	第二次查詢時間	801	933	961
	第三次查詢時間	859	943	939
	平均查詢時間	840	942	954
身高分析資料	第一次查詢時間	13,964	15,203	16,719
	第二次查詢時間	13,973	15,588	16,521
	第三次查詢時間	13,958	15,169	16,555
	平均查詢時間	13,965	15,320	16,598
體重分析資料	第一次查詢時間	14,718	15,690	17,374
	第二次查詢時間	14,364	15,828	17,140
	第三次查詢時間	14,965	15,150	16,576
	平均查詢時間	14,682	15,556	17,030
BMI分析資料	第一次查詢時間	12,182	13,853	15,380
	第二次查詢時間	12,317	13,937	14,965
	第三次查詢時間	12,067	13,654	14,991
	平均查詢時間	12,189	13,815	15,112

從上述實驗結果看來，資料大量累積於集中式多維數據模型中，對於資料查詢時間會有所影響，尤其在查詢身高、體重、BMI 的最新年度資料實驗時發現，雖僅查詢單一年度之資料，但隨著資料增長耗費查詢時間則是越來愈高。

二　分散式多維數據模型分析效率測試

1. 測試查詢歷年各縣市各年級體格生長分析資料

查詢測試結果如表 11，累積資料量截至 4 年份以前，隨著資料每年增長對查詢時間亦是隨著增加，之後從第 5 年開始至的 8 年，資料量增長對於查詢時間則無明顯影響，直到資料累積增加至第 9 年，查詢時間才開始增加。

表 11 分散式多維數據模型查詢歷年各縣市各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量								
		1 年份	2 年份	3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	445	943	1,400	1,857	1,867	1,938	1,922	1,896	2,437
	第二次查詢時間	389	887	1,144	1,927	1,952	1,874	1,966	1,913	2,441
	第三次查詢時間	341	708	1,230	1,821	1,918	1,892	1,917	1,977	2,373
	平均查詢時間	391	846	1,258	1,868	1,912	1,902	1,935	1,929	2,417
身高分析資料	第一次查詢時間	5,158	10,632	15,719	20,740	20,742	20,442	20,438	20,760	23,924
	第二次查詢時間	5,189	10,535	15,757	20,445	20,295	20,908	20,727	20,374	23,716
	第三次查詢時間	4,858	10,519	15,456	20,222	20,256	20,606	20,624	20,672	23,145
	平均查詢時間	5,068	10,562	15,644	20,469	20,431	20,652	20,596	20,602	23,595
體重分析資料	第一次查詢時間	4,989	10,971	16,778	22,262	22,144	21,733	22,214	21,735	25,272
	第二次查詢時間	5,303	10,875	16,559	21,924	21,818	21,837	21,874	22,017	24,579
	第三次查詢時間	5,114	10,930	16,582	21,987	22,110	21,856	21,995	21,986	25,096
	平均查詢時間	5,135	10,925	16,640	22,058	22,024	21,809	22,028	21,913	24,982
BMI分析資料	第一次查詢時間	3,957	8,400	12,466	17,234	16,572	17,128	16,757	17,399	20,577
	第二次查詢時間	3,116	8,737	11,980	16,602	18,012	17,825	17,704	17,004	20,362
	第三次查詢時間	3,710	9,073	13,935	17,538	17,001	17,637	17,408	16,952	19,835
	平均查詢時間	3,594	8,736	12,794	17,125	17,195	17,530	17,290	17,119	20,258

2. 測試查詢當年六都縣市國中小各年級體格生長分析資料

此部分實驗結果四項資料查詢時間在經由 9 年資料累積測試查詢結果，各項查詢時間變化不大，資料增長對於查詢時間沒有明顯影響，實驗結果如表 12。

3. 測試查詢近三年六都縣市國中小各年級體格生長分析資料

實驗結果如表 13，整體來看資料累積增長對於查詢時間無明顯增加，且在資料累積至第 5 年份及第 6 年份，由於欲查詢之資料分散存於兩台伺服器，利用平行查詢之方式，查詢效率表更佳。

4. 測試查詢當年、前 3 及前 6 年六都縣市國中小各年級體格生長分析資料

在此項查詢測試，健康體位分析資料部分的測試結果，資料增長對於查詢時間無明顯影響現象，查詢平均時間約在 400 毫秒左右；而查詢身高、體重及 BMI 平均時間也都在 5 秒左右，實驗結果如表 14。

表 12 分散式多維數據模型查詢當年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量								
		1 年份	2 年份	3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	218	247	188	210	202	209	215	210	206
	第二次查詢時間	168	220	273	203	203	227	219	228	218
	第三次查詢時間	167	202	220	238	182	203	220	202	217
	平均查詢時間	184	223	227	217	196	213	218	213	214
身高分析資料	第一次查詢時間	2,280	2,469	2,338	2,311	2,400	2,046	2,173	2,006	1,981
	第二次查詢時間	2,269	2,407	2,487	2,377	2,123	1,683	1,738	1,619	1,970
	第三次查詢時間	2,274	2,549	2,447	2,325	2,015	2,049	1,905	1,606	2,108
	平均查詢時間	2,274	2,475	2,424	2,338	2,179	1,926	1,939	1,744	2,019
體重分析資料	第一次查詢時間	2,311	2,513	2,640	2,373	1,771	2,361	1,796	1,900	1,869
	第二次查詢時間	2,321	2,573	2,385	2,382	1,765	1,775	1,765	2,022	1,884
	第三次查詢時間	2,283	2,626	2,420	2,329	2,267	2,424	2,061	1,723	1,859
	平均查詢時間	2,305	2,571	2,482	2,361	1,934	2,187	1,874	1,882	1,871
BMI分析資料	第一次查詢時間	2,277	2,496	2,537	2,377	2,060	2,367	2,358	1,831	2,342
	第二次查詢時間	2,019	2,132	2,206	2,081	1,923	1,497	2,241	1,390	1,830
	第三次查詢時間	1,969	2,277	2,229	2,326	1,434	1,786	1,622	1,381	1,864
	平均查詢時間	2,088	2,302	2,324	2,261	1,806	1,883	2,074	1,534	2,012

表 13 分散式多維數據模型查詢近三年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量						
		3 年份	4 年份	5 年份	6 年份	7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	649	699	507	391	583	545	655
	第二次查詢時間	605	640	424	392	604	642	731
	第三次查詢時間	570	693	440	375	565	613	622
	平均查詢時間	608	678	457	386	584	600	669
身高分析資料	第一次查詢時間	8,086	8,314	5,516	4,225	7,823	7,938	7,894
	第二次查詢時間	8,061	8,131	5,387	4,954	7,750	7,628	7,783
	第三次查詢時間	8,132	8,175	5,456	4,950	7,724	7,970	7,610
	平均查詢時間	8,093	8,206	5,453	4,710	7,766	7,846	7,762
體重分析資料	第一次查詢時間	8,349	8,489	5,623	4,647	7,957	7,952	8,169
	第二次查詢時間	8,052	8,398	5,519	4,802	7,989	8,252	8,180
	第三次查詢時間	8,184	8,261	5,608	5,249	7,443	7,802	7,965
	平均查詢時間	8,195	8,383	5,583	4,899	7,796	8,002	8,105
BMI分析資料	第一次查詢時間	7,404	7,390	5,013	3,817	6,907	7,232	7,069
	第二次查詢時間	7,252	7,390	4,884	4,298	6,915	6,790	6,987
	第三次查詢時間	7,074	7,285	4,887	4,347	6,824	7,037	7,021
	平均查詢時間	7,243	7,355	4,928	4,154	6,882	7,020	7,025

表 14 分散式多維數據模型查詢當年、前 3 及前 6 年六都縣市國中小各年級體格生長分析資料耗時測試結果

資料查詢項目	查詢時間(ms)	累積資料量		
		7 年份	8 年份	9 年份
健康體位分析資料	第一次查詢時間	384	374	406
	第二次查詢時間	368	400	439
	第三次查詢時間	363	360	387
	平均查詢時間	372	378	411
身高分析資料	第一次查詢時間	5,349	4,926	5,013
	第二次查詢時間	5,227	4,934	5,119
	第三次查詢時間	5,318	4,978	5,046
	平均查詢時間	5,298	4,946	5,059
體重分析資料	第一次查詢時間	5,368	5,281	5,250
	第二次查詢時間	5,462	4,928	5,387
	第三次查詢時間	5,380	4,948	5,461
	平均查詢時間	5,403	5,053	5,366
BMI 分析資料	第一次查詢時間	4,751	4,454	4,709
	第二次查詢時間	4,755	4,473	4,490
	第三次查詢時間	4,655	4,459	4,542
	平均查詢時間	4,720	4,462	4,580

從上述實驗結果看來，分散式多維數據模型，在資料大量累積於單一資料庫伺服器中，對於查詢整個數據模型時間會有正向影響，但若是將資料分散存於多部伺服器中並利用平行查詢之方式，可有效提升查詢效率；而在切片、切塊分析處理查詢方面，資料增長對於查詢時間並無明顯影響，查詢時間相對平穩，且若是利用平行查詢之方式，可再有效減少查詢資料所需時間。

第三節　綜合比較

本節將對上兩節實驗結果進行比較，首先比較 OLTP 資料庫與兩種多維數據模型在查詢時間上的差異，接著再來比較兩種不同多維數據模型在查詢上的時間差異。比較方式以類別型分析與數值型分析項目各挑一項分析主題進行比較，類別型分析項目為健康體位分析，而數值型則是選擇身高分析。

一　OLTP 資料庫與多維數據模型分析效率比較

在查詢健康體位分析資料部分，OLTP 資料庫相較於兩種多維數據模型，查詢時間相當高，且隨著資料增長差距又更為明顯，相較之下多維數據模型時間則是少很多，且兩種架構查詢所耗時間非常接近，比較查詢結果如圖 21；在身高分析資料查詢方面，從圖 22 可以看出多維數據模型雖查詢時間一樣會隨著資料增長而有所增加，但可以看出相較於 OLTP 資料庫的查詢時間，多維數據模型的查詢表現好上許多，在加上隨著資料

累積，多維數據模型查詢所耗的時間相對於 OLTP 資料庫的查詢時間增幅較為平緩。

從上述兩種比較看來，原始資料經過預處理並轉存成多維數據模型對於資料查詢效率有明顯提升，其中又以健康體位分析項目查詢表現最為優異，該項目之資料經彙整後每年僅有 280,966 筆資料存於多維數據模型及可滿足資料查詢的需求，因此在資料少的情況下，查詢時間相較於其他三項相對少很多。而身高、體重、BMI 資料因在每一個顆粒下還要再對其數值進行分類，資料經統計壓縮的量有限，所以對於分析長年累積於整個資料方體內的資料，仍需要花費相當時間。

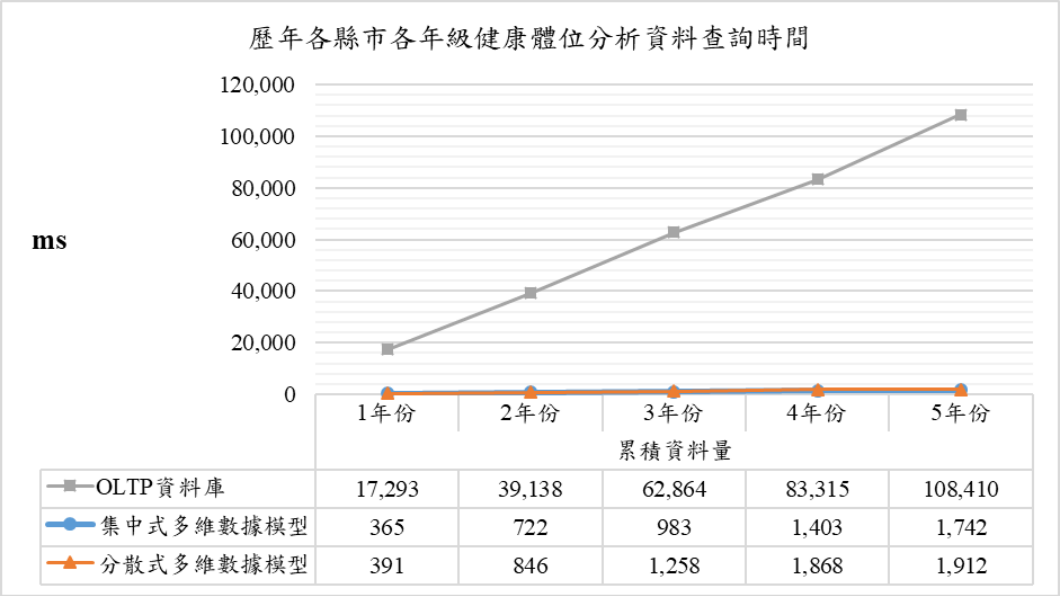

圖 21 OLTP 資料庫與多維數據模型查詢所有健康體位分析資料時間比較

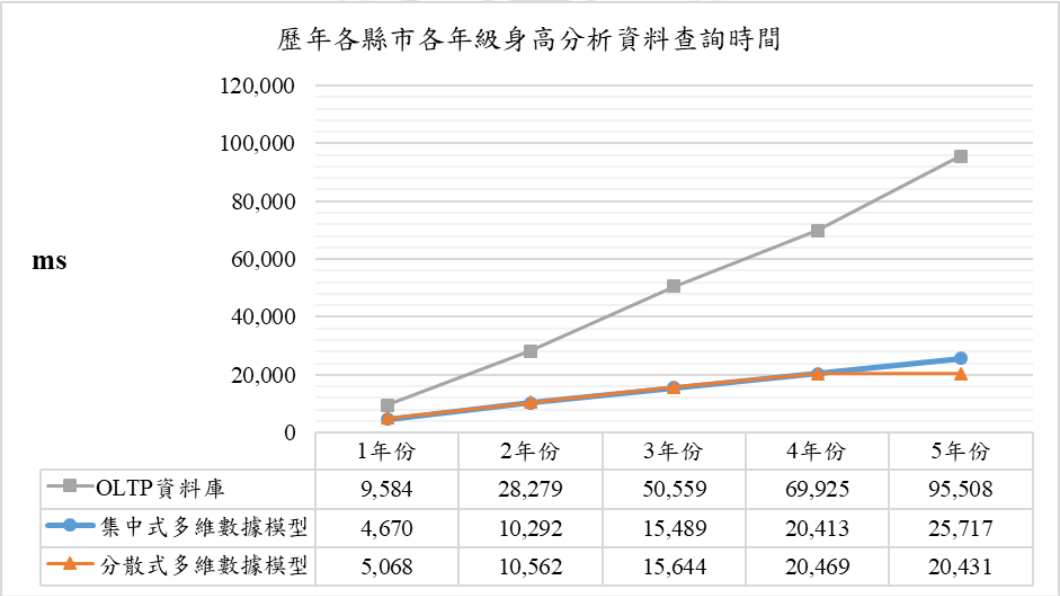

圖 22 OLTP 資料庫與多維數據模型查詢所有身高分析資料時間比較

二　集中式與分散式多維數據模型分析效率比較

1.　查詢整個多維數據模型資料耗時比較

從比較結果來看，集中式多維數據模型都有因資料增長而查詢時間也隨之增長的現象；分散式多維數據模型則藉由平行查詢方式，累積資料量從第 4 年以後到第 8 年之間，資料累積增長對於查詢時間則無明顯影響，到資料累積第 9 年查詢時間才開始上升。在健康體位資料查詢部分兩者架構查詢時間相近；而身高資料查詢部分，兩者架構結至累積資料量 4 年份以前查詢時間幾乎一致，之後資料繼續累積分散式架構的查詢時間則明顯優於集中式架構。比較結果如圖 23 及圖 24。

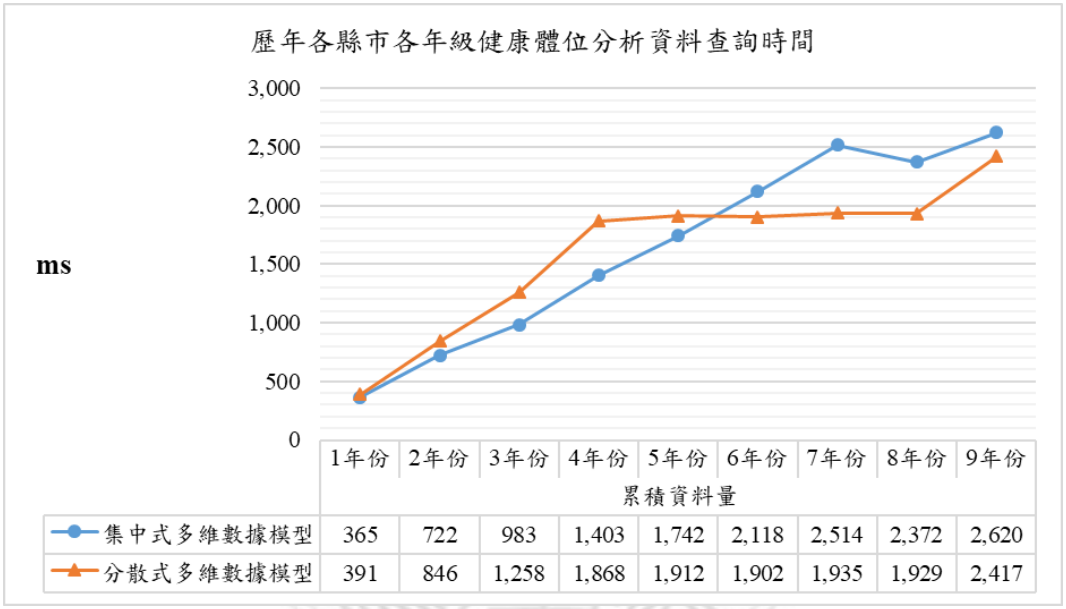

圖 23 集中式與分散式多維數據模型查詢所有健康體位分析資料時間比較

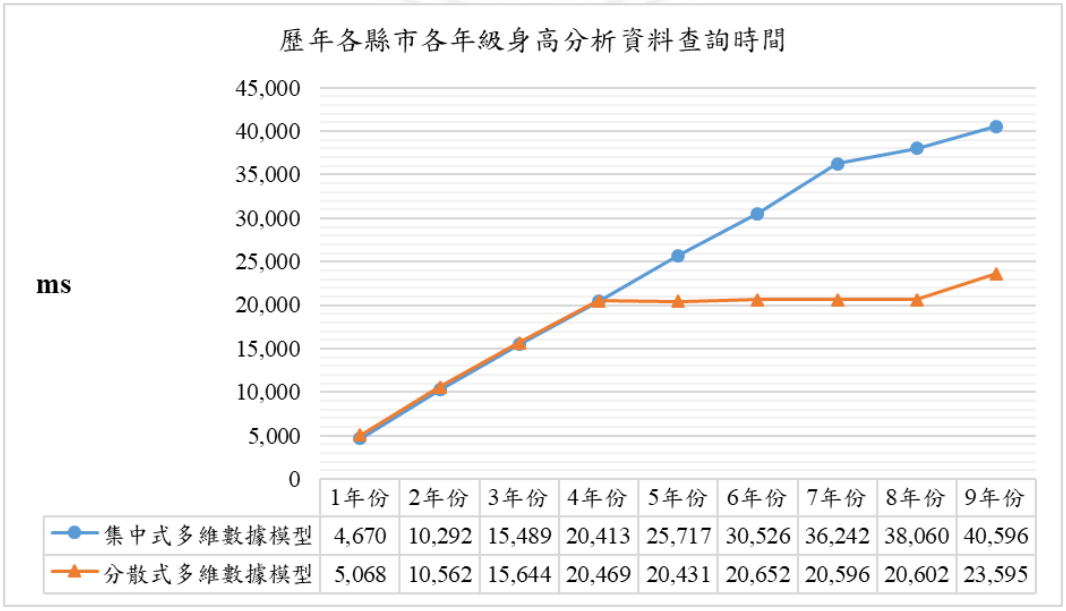

圖 24 集中式與分散式多維數據模型查詢所有身高分析資料時間比較

2. 多維數據模型切片及切塊處理查詢耗時比較

(1). 查詢健康體位分析資料耗時比較

從比較結果來看，對於 9 年資料累積，兩者多維數據模型在切片、切塊查詢健康體位資料都相當快速，其查詢當年度資料耗時都低於 250 毫秒，而查詢近三年及當年、前 3 及前 6 年的六都縣市國中小各年級資料，兩模型平均查詢所耗時間也在 1 秒以下，且分散式多為數據摩型查詢時間略優於集中式多維數據模型，比較結果如圖 25、圖 26 及圖 27。

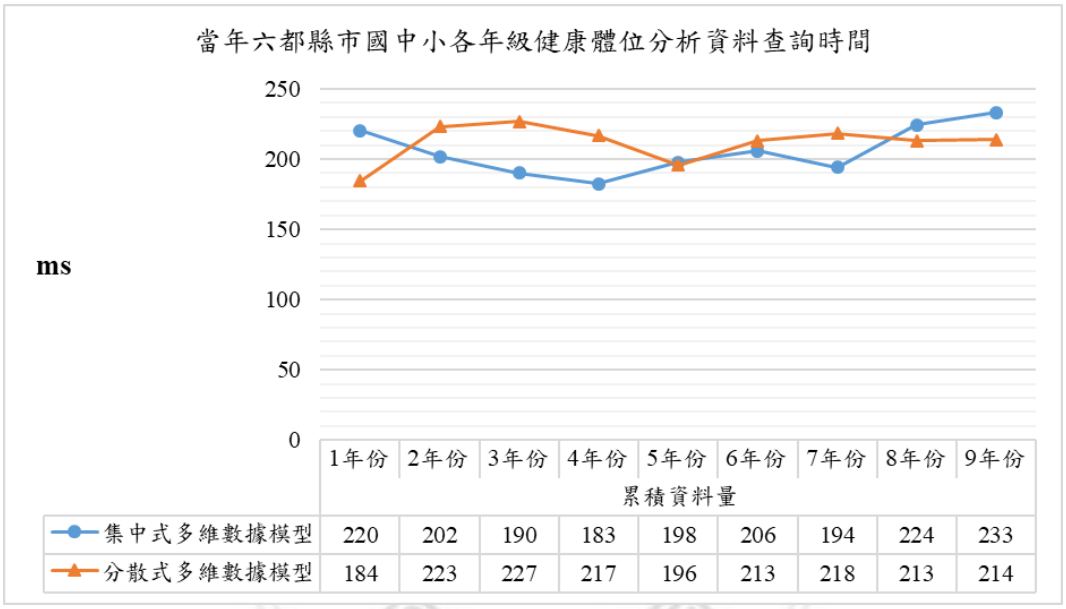

圖 25 集中式與分散式多維數據模型查詢當年健康體位分析資料時間比較

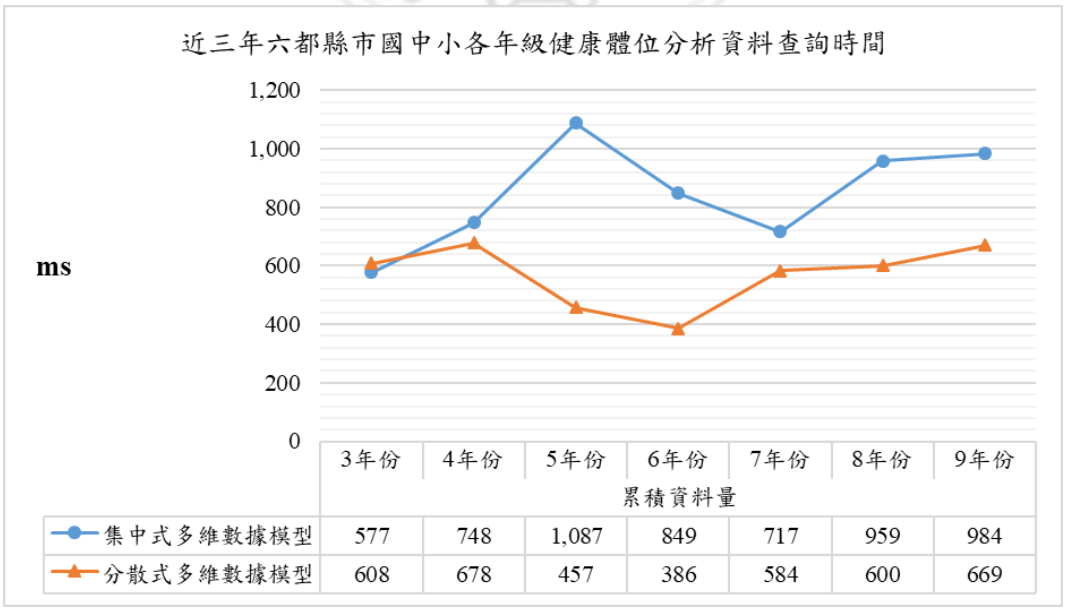

圖 26 集中式與分散式多維數據模型查詢近三年健康體位分析資料時間比較

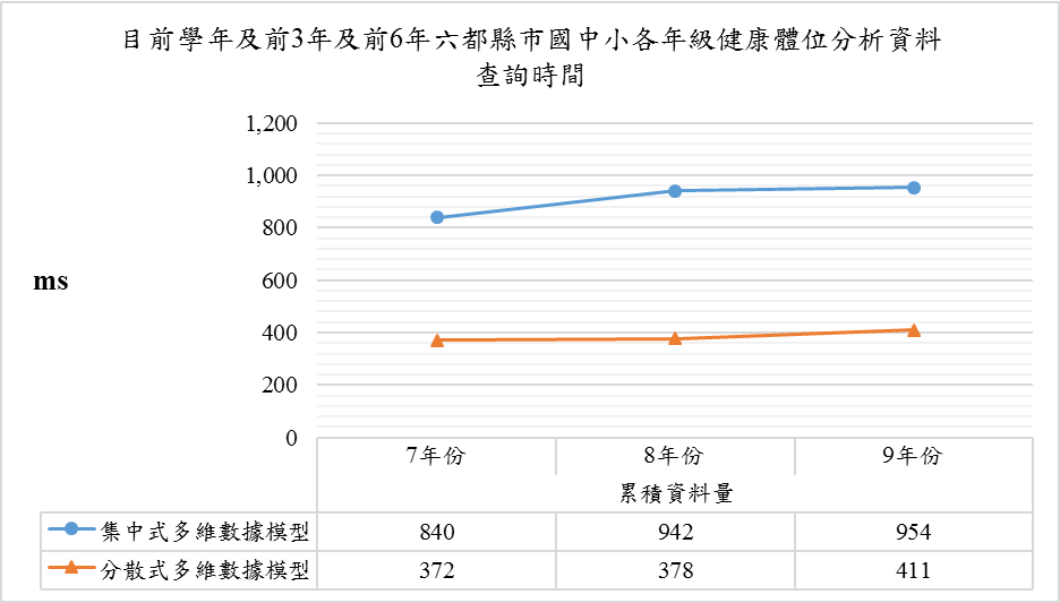

圖 27 集中式與分散式多維數據模型查詢當年及前 3 年及前 6 年健康體位分析資料時間比較

(2). 查詢身高分析資料耗時比較

而從查詢身高分析資料結果來看，在查詢當年六都縣市國中小各年級資料，可以明顯看到集中式模型對於資料增長，所需耗費的查詢時間越長，相較之下分散式模型查詢時間則沒因資料增長而有太大變化；而在查詢近三年六都縣市國中小各年級資料以及查詢目前學年及前 3 年及前 6 年六都縣市國中小各年級資料也有相同狀況，大量資料累積對於集中式模型在切片、切塊的查詢表現會造成影響，相較之下分散式模型則無明顯影響且利用平行查詢方式查詢效率表現更加提升，比較結果如圖 28、圖 29 及圖 30。

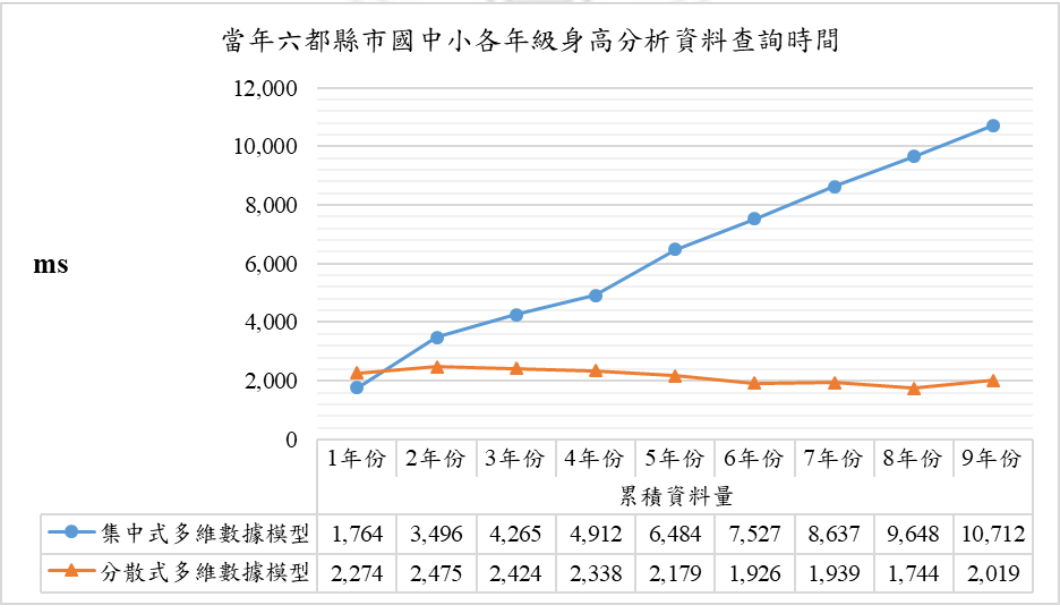

圖 28 集中式與分散式多維數據模型查詢當年身高分析資料時間比較

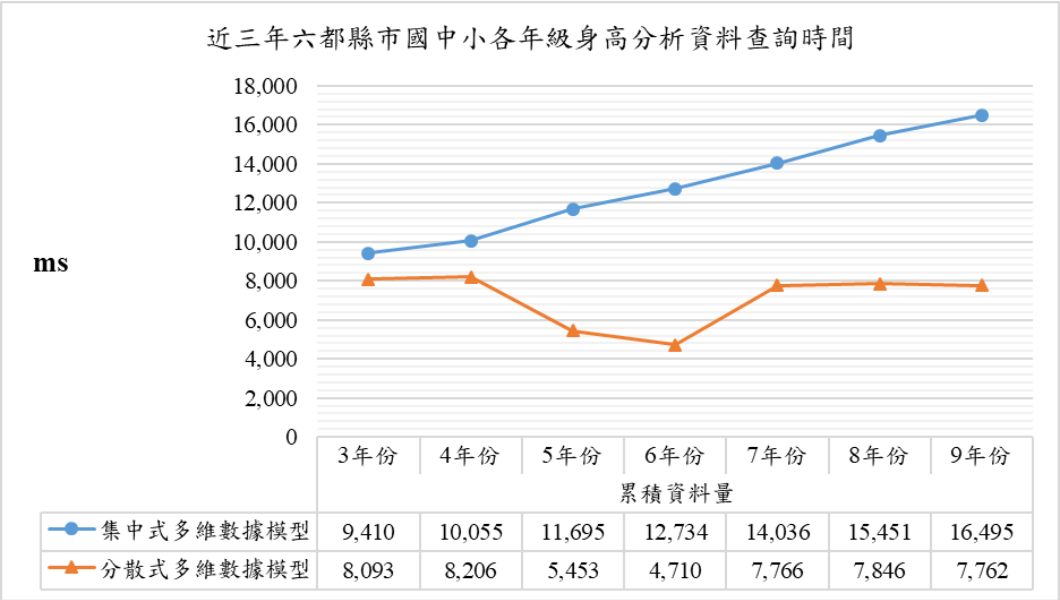

圖 29 集中式與分散式多維數據模型查詢近三年身高分析資料時間比較

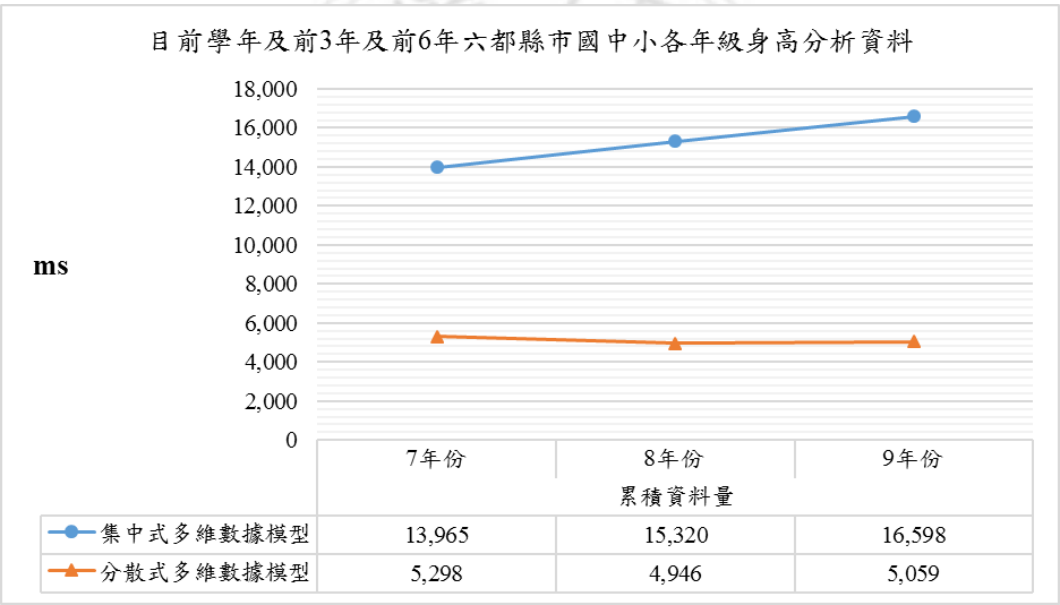

圖 30 集中式與分散式多維數據模型查詢當年及前 3 年及前 6 年身高分析資料時間比較

綜合本節實驗比較結果，其列點說明如下

1. 查詢整個多維數據模型資料

   從累積前 4 年份資料查詢比較可看出，分散式多維數據模型若是將資料僅存放於單一伺服器中，無使用平行查詢方式查詢資料，不論是資料量累積多寡，只要資料增長兩種架構的多維數據模型查詢時間也會隨之增長，兩架構在此部分實驗結果無太大差異，但若是將資料存放在多伺服器中，利用平行查詢方式查詢資料，分散式架構查詢時間則會優於集中式架構，此部分由累積資料第 5 年以後查詢時間比較可看出。

2. 多維數據模型切片及切塊處理查詢

   累積資料量較少的情況下對於兩架構多維數據模型查詢時間無太大差異，皆可在短時間內完成查詢；但在累積資料量大的情形下，集中式模型明顯會對查詢所耗時間造成影響，資料累積越多查詢時間越長，而分散式模型相較之下則無太大影響，且欲查詢之資料若是存於不同伺服器，透過平行查詢之方式，查詢效率則更加提升。

三　集中式與分散式多維數據模型應用上比較

本研究在實作兩種多維數據模型架構及開發測試查詢之網站，比較兩者架構在系統開發應用狀況，列點說明如下。

1. 系統設定比較

   本研究實驗測試時發現，分散式多維數據模型架構設計佈署於不同資料庫伺服器中，對於伺服器操作設定在實驗上需要花較多時間，而集中式模型因只存於單一伺服器中，在機器設定上所花時間相對較低。

2. 資料維護比較

   因歷年原始資料皆屬於同質資料，在進行資料清理、轉換並載入資料倉儲之處理兩者並無太大不同。

3. 資料查詢應用比較

   集中式的資料庫因資料都存於同一伺服器及同一資料庫管理工具，所以查詢資料之 SQL 指令相對容易撰寫，且在測試 SQL 指令上可搭配資料庫維護工具即可在查詢所有資料，而分散式模型架構因本研究將資料庫拆分到兩部伺服器上，因此若要查詢所有資料或是整合查詢的資料被切分到不同伺服器上，需要透過自行開發程式輔助查詢兩邊資料及整合，在開發上就會比較不易。集中式與分散式模型 SQL 查詢指令範例如附錄 1 及附錄 2 所示。

第五章 結論

第一節　結果探討

一　大量資料分析處理方式

從實驗測試結果看來 OLTP 資料庫對於大量資料之分析之處理表現與過去文獻結果一致，在查詢分析資料上需要花費許多時間，因此針對分析主題另外設計、建立多維數據模型來處理資料分析工作，不僅可以大大降低查詢時間，也可以明確與 OLTP 資料庫劃分工作，讓兩者在各自工作領域不互相干擾，在系統維護上相對單純，所以面對大量資料分析建立多維數據模型是必要的選擇。

二　集中式與分散式多維數據模型適用狀況

本研究設計集中式及分散式兩架構之多維數據模型，測試對於資料動態增長兩者表現為何，從實驗結果發現如下。

1. 累積資料量小的情況下，兩者模型資料查詢表現皆可快速完成查詢。
2. 對於查詢所有資料，資料量不論多寡，只要資料量有所成長，集中式模型查詢時間皆會增長；而分散式模型利用平行查詢方式，可有效縮減查詢所需時間。
3. 對於切片及切塊處理查詢，資料大量累積會對於集中式架構模型造成查詢時間上的增加，相較之下分散式架構模型則無明顯影響且查詢速度較集中式架構快。
4. 集中式架構模型在系統開發及維護上相較於分散式架構模型較為容易，開發以及維護的時間成本也會相對較低。

綜合上述發現，本研究認為集中式多維數據模型較適合資料量較少的狀況，一方面查詢效率與分散式架構皆可在短時間內完成，再者在維護及開發上也較為容易；而分散式架構則適合於資料量大的情況，雖然在維護及開發上也較為不易，但在切片及切塊處理查詢時間則明顯優於集中式架構，也可利用平行查詢優勢，再提升查詢效率，因此考量整體查詢效率及使用者查詢體驗，分散式架構還是優於集中式架構。

第二節　研究限制與未來方向

一　研究限制

1. 時間限制

本研究設計之實驗因時間上限制，僅針對部分項目進行查詢測試，無法將所有維度組合查詢一一探討，且分散式架構在設計上可以有多種模式，本研究未設計更多分散式架構，僅設計以學年度資料將切分進行資料，作為分散式多維數據模型作為探討，屬本研究限制之一。

2. 軟硬體環境限制

本研究實驗之系統環境雖資料庫伺服器軟硬規格相同，但實驗時仍可能受到網路、機器及軟體等其他因素影響，且本研究使用之資料庫管理系統 SQL Server 2012 Express 在查詢上僅限制以單個 CPU 核心來運算，無法發揮多核心運算能力，因此實驗查詢耗時結果偏高。

3. 測試資料集限制

本研究選擇使用高級中等以下學生體格生長資料作為測試資料集，然而真實的資料因有個資問題，因此本研究則以模擬資料當作測試資料集，雖模擬資料集之原始資料量與真實資料近似，但仍有可能因為模擬資料與真實資料差異造成彙整於多維數據模型之資料量有所差異，進而可能造成分析結果差異落差。

二　未來方向

1. 壓力測試

本研究實驗僅測試分析效率，且實驗系統也只有單一使用者在操作使用，而在真實情境下，同時間可能會有多位使用者進行查詢，因此探討兩架構模型承受大量訪問能力也是相當重要。

參考文獻

1. 沈兆陽(2001)，資料倉儲與 Analysis Services：SQL Server 2000 的 OLAP 解決方案，臺北市：文魁資訊有限公司。
2. 汪瑞勛(2019)，以多維度倉儲系統支援校務研究決策分析用的因果規則發掘，碩士論文，國立中央大學資訊工程學系。
3. 邱瑞科、詹前隆、翁頌舜、張啟明、顏哲傑(2002)，建立符合 HIPAA 標準規範之全國預防接種資料庫(倉儲)與管理決策輔助先導性系統之研究，行政院衛生署疾病管制局成果報告書。.
4. 陳玄玲(譯)(2008)，資料庫系統原理(原作者：Ramez Elmasri ,& Shamkant B Navathe)，臺北市：台灣培生教育出版股份有限公司.
5. 張玉婷(2014)，資料倉儲之資料品質改善程序研究-以個案公司為例，碩士論文，國立交通大學管理學院資訊管理學程。
6. 蕭凱文、薛志達、李政輝(1999)，Microsoft SQL Server 7.0 資料倉儲整合應用，臺北市：華彩。
7. William. H. Inmon, Building the Data Warehouse, John Wiley & Sons, Inc., 1996.
8. Ralph Kimball, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, John Wiley & Sons, Inc., 2003.
9. Wikipedia (2020)，Centralized database，擷取自：Wikipedia :https://en.wikipedia.org/wiki/Centralized_database。
10. Hightech(2019)，什麼是分散式資料庫？有哪些好處？，擷取自 STOCKFEEL: https://www.stockfeel.com.tw/什麼是分散式資料庫？有哪些好處？。
11. 維基百科(2020)，資料倉儲，擷取自: https://zh.wikipedia.org/wiki/資料倉儲

附錄一：SQL 指令範例

1. 集中式多維數據模型查詢近三年六都縣市國中小各年級學生健康體位分析資料 SQL 指令

```
SELECT  SchoolYear,
          a.CityID,
          b.GradeID,
          sum(Total)Total,
          sum(Light)Light,
          sum(Moderate)Moderate,
          sum(Heavy)Heavy,
          sum(Obesity)Obesity
FROM      (select   b.CityID,a.SchoolID,SchoolYear
          from      Concentrated_DW.dbo.School a
                    inner join
                    Concentrated_DW.dbo.Area b on a.AreaID=b.AreaID
          where     b.CityID in(1,2,3,4,5,7) and a.SchoolYear in(108,107,106))a
          inner join
          (select   SeqID,GradeID,SchoolID,SchYear,Total,Light,Moderate,Heavy,Obesity
          from      Concentrated_DW.dbo.HealthyBodyType
          where     GradeID in(1,2,3,4,5,6,7,8,9)) b on a.SchoolID=b.SchoolID and
a.SchoolYear=b.SchYear
group by SchoolYear,a.CityID,b.GradeID
order by SchoolYear,a.CityID,b.GradeID
```

2. 分散式多維數據模型查詢近三年六都縣市國中小各年級學生健康體位分析資料 SQL 指令

```
SELECT    108 SchoolYear,a.CityID,b.GradeID,sum(Total)Total,sum(Light)Light,
          sum(Moderate)Moderate,sum(Heavy)Heavy,sum(Obesity)Obesity
FROM      (select   b.CityID,a.SchoolID
          from      Dispersion_DW_108.dbo.School a
                    inner join
                    Dispersion_DW_108.dbo.Area b on a.AreaID=b.AreaID
          where     b.CityID in(1,2,3,4,5,7))a
          inner join
          (select   SeqID,GradeID,SchoolID,Total,Light,Moderate,Heavy,Obesity
          from      Dispersion_DW_108.dbo.HealthyBodyType
          where     GradeID in(1,2,3,4,5,6,7,8,9)) b on a.SchoolID=b.SchoolID
group by  a.CityID,b.GradeID
union all
SELECT    107 SchoolYear,a.CityID,b.GradeID,sum(Total)Total,sum(Light)Light,
          sum(Moderate)Moderate,sum(Heavy)Heavy,sum(Obesity)Obesity
FROM      (select   b.CityID,a.SchoolID
          from      Dispersion_DW_107.dbo.School a
                    inner join
                    Dispersion_DW_107.dbo.Area b on a.AreaID=b.AreaID
          where     b.CityID in(1,2,3,4,5,7))a
          inner join
          (select   SeqID,GradeID,SchoolID,Total,Light,Moderate,Heavy,Obesity
          from      Dispersion_DW_107.dbo.HealthyBodyType
          where     GradeID in(1,2,3,4,5,6,7,8,9)) b on a.SchoolID=b.SchoolID
group by a.CityID,b.GradeID
union all
SELECT    106 SchoolYear,a.CityID,b.GradeID,sum(Total)Total,sum(Light)Light,
          sum(Moderate)Moderate,sum(Heavy)Heavy,sum(Obesity)Obesity
FROM      (select   b.CityID,a.SchoolID
          from      Dispersion_DW_106.dbo.School a
                    inner join
                    Dispersion_DW_106.dbo.Area b on a.AreaID=b.AreaID
          where     b.CityID in(1,2,3,4,5,7))a
          inner join
          (select   SeqID,GradeID,SchoolID,Total,Light,Moderate,Heavy,Obesity
          from      Dispersion_DW_106.dbo.HealthyBodyType
          where     GradeID in(1,2,3,4,5,6,7,8,9)) b on a.SchoolID=b.SchoolID
group by a.CityID,b.GradeID
order by SchoolYear,a.CityID,b.GradeID
```