東　　海　　大　　學
工業工程與經營資訊學系

碩士論文

應用 Hadoop 系統架構於生理大數據分析
-以心臟衰竭為例

研 究 生：夏可清
指導教授：張炳騰 教授

中 華 民 國 一 〇 六 年 八 月

Apply Hadoop System Structure in Physiology Big Data Analysis–With As Example of Heart Failure

By
Ko-Ching Hsia

Advisor：Prof. Ping-Teng Chang

A Thesis
Submitted to the Institute of Industrial Engineering and Enterprise Information at Tunghai University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in
Industrial Engineering and Enterprise Information

August 2017
Taichung, Taiwan

應用 Hadoop 系統架構於生理大數據分析-以心臟衰竭為例

學生：夏可清　　指導教授：張炳騰 教授

東海大學工業工程與經營資訊學系

摘　　要

本研究採用 OLAP 編碼方式，以及用於處理大數據資料的 Hadoop 為架構，針對心臟衰竭的病患資料，進行分維和預測病況的結點之架構整合。在編碼部分，根據病患的加入順序，結合以個人資訊維、地區維、時間維、感測維、心知識維五大維度去做分層。依照編碼的三大原則，推算出列舉之案例以兩千人為單位的醫療體系將不堪交易量的與日俱增。為解決不確定因子如:突如其來的環境因子造成的交易數據量過多，本研究提出資料蒐集應結合醫學知識為基礎的基於 Hadoop 的 OLAP 系統架構，進而降低數據量，從傳統的紀錄方式 164.79 GB 降低到 14GB，相比約少了 11.7 倍。

在基於 Hadoop 的 OLAP 系統架構裡面，本研究為強調 Hadoop 優異的 MapReduce 計算與分析方式，將以個案的形式描述一個患有心律不規律波型的患者的處理流程。用於分析上應具備的四大流程包括:計數、分類、過濾處理、相關計數。此外還要建立一個造訪資料的流程，以交代初始資料的建置、修改以及運用。本研究列出偽代碼，以及 Hadoop 在分布式文件系統中運用的 map reduce 組件之間的訊息傳遞流程和相關的外加功能 Hive 資料倉庫工具，提供強大的類 SQL 查詢功能。當中會敘述到基於 Hadoop 的 OLAP 系統架構中數個模塊間的使用，以及經過 MapReduce 運算處理後的新表單。

關鍵字詞：線上分析處理、Hadoop 應用程序、MapReduce 軟體架構、心臟衰竭、儲存、大數據

Apply Hadoop System Structure in Physiology Big Data Analysis–With As Example of Heart Failure

Student：Ko-Ching Hsia Advisor：Prof. Ping-Teng Chang

Department of Industrial Engineering and Enterprise Information
Tunghai University

ABSTRACT

In this study, we use OLAP coding method and Hadoop, which is used to process big data and construct the node framework integration of fractal dimension and condition prediction for patient data with heart failure. In the coding part, according to the joining sequence of the patients, it combines personal information dimension, regional dimension, time dimension, sensing dimension, and heart knowledge dimension, with the five dimensions for stratification. In accordance with the three principles of coding, it is projected that the amount of transactions for a medical system with listed cases of two thousand people will be unbearably increasing day by day.

In order to solve the problem of uncertain factors, such as sudden environmental factor causing excessive data, this study proposes that data collection should be combined with medical knowledge based on Hadoop OLAP system architecture, thereby to reduce the amount of data from the traditional recording method 164.79 GB to 14 GB, comparatively about 11.7 times less. This study lists pseudo-code, and Hadoop in the distributed file system used in the map reduce component between the message delivery process and related functions hive data warehouse tool to provide a powerful class SQL query function. And it describes the use of several modules in the Hadoop-based OLAP system architecture, as well as new forms that have been processed by MapReduce operations.

Keywords：OLAP, Hadoop, MapReduce, Heart Failure, Store, Big Data

致謝詞

感謝本論文得以完成，得感謝生命中許多的貴人。首先要謝謝指導老師張炳騰教授，老師除了視野廣闊，更重要的是在強大的思考碰撞下還能保持極為清晰的邏輯，讓我的論文寫作不至於太過發散。且除了信仰之外，老師也是唯一一個能活生生的展現對學術的熱情，在每週督導後，也不忘拍拍我的肩，甚至邀請我到車上聊聊做論文的心情。沒有這樣的「領導」，我是不可能完成論文，這項在學生階段裡最後也是重要的任務。

再來要感謝爸媽，讓我有機會在台中就讀大學，當我決定繼續留校讀碩士，他們也表示支持，在無後顧之憂的情況下，更能為學術做出貢獻。

其次是上帝，這個信仰是真正堅定我一切的力量，在碩二下一開始時曾經一度想放棄，是因為真的覺得好辛苦、好累，連看操作手冊也會覺得不堪其擾，一整個情緒是很低落的；然而對周身的小事卻又感受強烈，這是一個躁動和不安的時候，幸好有牧師、團契的家長和好的屬靈夥伴一同陪伴，我總能在聖經的話語裡，找到主要我做的事情，就是完成眼下份內的事。

再來要感謝研究室的同學，謝謝志豪，陪我度過這麼多個研究室過夜的夜晚，我所說的「牛肉」，將呈現在這份論文裡。感謝威鈞、哲偉，一同衝刺的階段，能有你們的笑聲，總讓我的不安能稍作化解，未來有好的咖啡店，記得還要一同品嘗，只是別在熬夜時約我就好。感謝瑞陽，就是這麼的硬漢，你的執著，只有在寫謝詞的時候，才會更有濃度。還有庭瑜，謝謝妳總是在我這麼緊急的關頭願意提供協助，真的是非常感動。最後就是柏健、緯綸、晉維，滿滿的嘴砲，也是滿滿的精力泉源，再次謝謝你們。

未來，依舊靠著神，持續走在祢的帶領下。

夏可清 謹誌於

東海大學工業工程與經營資訊學系

中華民國一〇六年八月

目錄

表目錄

圖目錄

第一章　研究背景

1.1 研究背景與動機

心血管疾病的人數眾多，若成功預測，將能對社會有極大貢獻心血管相關疾病已經連續 30 年（2016 年止）蟬聯十大死因第二或第三名。

目前，對於心血管疾病的防治，現有的醫療模式以疾病治療為核心，重視生命後期的治療和護理，而相對忽視疾病早期的診斷和幹預。然而研究結果表明，心血管疾病有一個突出特點：病程的自然不可逆性，而且往往存在一個緩慢的隱性發展時期。疾病一旦發展到晚期，不僅治療效率低、成本高、後遺症多，而且患者完全康復的可能性小，特別是死亡前數月，醫療成本會有非常顯著的增加。

因此，疾病的預防和早期幹預治療是目前公認的降低心血管死亡率和減輕醫療負擔的最有效手段。

除此之外，心血管疾病還具有另外一個特點就是臨床表現的個體差異性，不僅患者具有很大的個體差異，同時生活習慣、社會心理因素等都會對臨床表現產生影響。個體差異性反映到臨床治療上，需要充分考慮到患者的個體化病理狀態並製定合理的手術方案，以最大限度地提高預後和降低手術風險。正是個體的差異性決定了需要為患者提供個體化的心血管疾病臨床檢測和治療手段。同時，患者個體化的精確診斷與風險預測也成為了心血管疾病防治的關鍵。

1.2 研究目的

本研究以 Hadoop 這項技術，來探討生理數據之運用，其目的並不是建出一個完整的生理數據分析系統，而是以一個非醫學專業領域的人，做一項跨資訊、硬體、醫學的知識串聯。

1.3 研究貢獻

本研究透過基於 OLAP（John Von Neumann，1945）（Online Analytical Processing）的編碼，能使以每日兩千人次數據來往的醫院，將數據量的大小縮減為原本的 11.7 倍。並提出將感測器讀取數據的頻率降為原本的 1/3，在不影響生理數據判讀的前提下，再降低數據庫和維護人員的負荷。

透過此種 Hadoop 技術的成長，能賦予資料流動性，打破過去資料湖泊（Cold）資料庫（Warm）等設備的賠錢的裝置形象，利用感測器激發所有數據的可能，最終由 Hadoop 進行平行運算，讓困擾業主的「暗數據」比例得以下降，在提高每個裝置和節點的效益，將有助於人類用更低的成本，維護生理狀態。

第二章　文獻探討

2.1 大數據簡介

大數據是隨著電腦技術及網路技術高速地發展產生的必然資料現象。個體的價值也在這樣的現象中膨脹。現在的社會產生大量的數據，例如：手機通訊、FaceBook 或微博回覆的留言、商品產生、物流運送、電信網路中的帳單數據。

以上各種資料構成了我們對大數據相對廣泛的描述，大數據並非一個簡單能定義的名詞，但是能以幾項特徵對他做一個概括性的描述，有以下三種特徵：資料大小、資料類型、資料時效。

2.1.1 資料量大小

過去學者以 1 分鐘為單位，觀察並累積這個世界的數據，得到了以下的發現，下列五項為例（MIIT，2012）：

1. E-mail：全球所有電子郵件使用者發出 2.04 億封電子郵件。
2. 搜尋：Google 搜尋引擎處理了 200 萬次的搜尋請求。
3. 圖片在分享網站 Flickr 的使用者上傳了 3,125 張新照片，並有兩千萬張照片被瀏覽。
4. 通訊：以中國為例，產生的了總長為 531 萬分鐘的通話時間，並發出了 165 萬條簡訊。
5. 電子商務：以 eBay 為例，產生七萬次頁面造訪的記錄，以及增加的 35 GB 的資料。

根據國際數據資訊（International Data Corporation, IDC）的研究報告指出，若是拿 eBay 目前累積的資料量與人類開始記錄歷史的資料量相比，到 2006 年為止，人類共印刷了資料量約為 50PB 的書本。換言之，以 eBay 平臺累積的資料來說，只需要三年的時間，就超過全人類書本的資料量（IDC，2007）。

2.1.2 資料類型

從資料組織形式的角度，可以把資料類型簡單幾分為兩種，結構化資料和非結構化資料。結構化資料是可以用二維表結構來表達和實

現，並可儲存在資料庫中的資料，如銀行交易資料、航班資訊...等嚴謹的資料庫資料。而非機結構化資料是指那些無法透過預先定義的資料模型表述或無法存入關聯式資料庫中的資料，如公司文件、圖片、音訊和影片等。

在醫療衛生資訊化的建設過程中，數據可以分為 3 種：結構化數據、半結構化數據和非結構化數據。其中，大量的數據屬於業務過程中產生的文檔等非結構化數據。表 2.1.1 整理了三種資料的類型，並對其優劣做簡單的說明。

根據 IDC 的統計，在企業資料中目前已有超過 80%的資料是以非結構化資料的形式存在的，結構化資料僅佔 20%不到。而在這個網際網路領域，非結構化資料已占到整個資料量比例的 75%以上，並且非結構化資料超越結構化資料的速度仍在加速中，現在整個資料領域，非結構化資料的年成長速度大約為 63%，遠超過結構化資料 32%的成長速度。

2.1.3 資料時效性

隨著資料量的巨量增加和資料類型的多樣化，資料中所蘊藏價值的時效性特徵也隨之愈加凸顯。例如一間電子商務網站必須在當天分析用戶的購買行為，並預測第二天的貨物短缺狀況，如果不能達到這樣的處理速度，第二天的缺貨狀況勢必將引來用戶的流失和收入的損失。這樣的現象也適用在本研究的案例，若是一家醫療機構不能及早提供用戶當日的健康建議，使用戶因未獲得專業建議而產生原本可避免之病狀，將使此醫療機構的信用大為下降。

表 2.1.1 資料類型（資料來源:本研究整理）

解釋	資料的分類		
	結構式資料	半結構化資料	非結構式資料
描述	資料擺放得整齊，在放置進資料庫時就已經受到了準確定義，而且格式固定，沒有例外。舉例來說，每筆資料都有固定的欄位、固定的格式、固定的順序甚至是固定的佔用大小。	數據的結構和內容混在一起，沒有明顯的區分。	因為突破了：數據定長的限制，支持子、重複、變長欄位的存在，在處理連續資訊（包括全文資訊）和各種多媒體資訊有著傳統關係型資料庫所無法比擬的優勢。
數據模型	二維表（關係型）	樹狀、圖形	無
資料類型	行數據、存儲在資料庫裡，可以用二維表結構來邏輯表達實現的數據、關係型資料庫、面向對象資料庫中的數據	所謂半結構化數據，就是介於完全結構化數據和完全無結構的數據之間的數據，HTML 文檔就屬此類。	所有格式的辦公文檔、文本、圖片、XML、HTML、各類報表、圖像和音頻/視頻信息等。
優點	時間最久也使用最廣，大部分新的資料建置仍會採用這種成熟的技術。除了使用得久和多，因為結構簡明，因此查詢統計比較方便；連帶維護容易，可以降低成本。	容易收納變數極多或難以預測的資料集，並依然保有可查找性。 能夠靈活的進行擴展，資訊進行擴展式只要更改對應的 DTD 或者 XSD 就可以了。	容納全部的數據： 一般來說，企業會希望能完整記錄非結構化的資料，將其收納存放在一邊，作為日後各種未預期應用的參考。
缺點	不能適應數據的擴展，不能對擴展的資訊進行檢索，對項目設計階段沒有考慮到的同時又是系統關心的資訊的存儲不能很好的處理。 指令較繁雜，所需的記憶體空間較大，因為記憶體空間大，造成程式執行時間拉長。	因為每筆資料結構都不一致，不能共用欄位名稱，而必須每筆都單獨記錄一次。也因如此會降低查詢效率，要借助 XPATH 來完成查詢統計，隨著資料庫對 XML 的支援的提升性能問題有望能夠很好的解決。	建置成本過高： 在最常用，最需要效能的地方，還是會希望使用結構化資料庫。
簡單來說	優點：地圖清楚很好跑。 缺點：地圖太大要一直跑。	優點：地圖上標註了很多名字，容易定位。 缺點：在地圖內上上下下得跑容易累。	優點：地圖上什麼都有。 缺點：地圖會很貴。

2.2 大數據的處理

大數據在各種環境以及新的挑戰下截然不同，但從這些問題的本質出發，可歸納為三個問題：大數據存儲、高性能計算和系統容錯性。

2.2.1 大數據存儲

提升系統儲存容量有兩種方式，一種是提升單硬碟的容量，透過使用新的材質和新的讀寫技術，單個硬碟的容量已經從 MB、GB 跨入了 TB 時代。在這裡主要關注在多硬碟的環境下，如何提升系統的整體儲存容量。經過多年的發展，系統儲存技術已經由早期的直連式儲存（Direct-Attached Storage, DAS），發展出網路接入儲存（Network-attached storage, NAS）和儲存區域網路（Storage area network, SAN），並在近幾年進入到雲端儲存階段。

2.2.2 高性能計算

對於單個硬碟，提升輸送量的主要方法是提高硬碟轉速、改進硬碟介面形式或增加讀寫快取等。而提升資料儲存系統的整體輸送量，比較典型的技術是早期的專用資料庫機體系。

在 1970 年代，為了更好地支援企業營運和商業決策，一些大型企業需要對資料倉庫中累積的海量歷史資料進行深入分析，因此需要對這些大數據進行大量的關係性查詢。在當時的技術條件下，資料庫伺服器普遍採用基於馮·諾依曼（John Von Neumann，1945）架構實現的通用電腦，在這種架構及當時的硬體條件下，通用資料庫伺服器在處理當時的大數據時出現了極大的不足。資料庫領域的一些學者指出，在當時基於以下架構實現的資料庫伺服器是不適宜用於大數據處理的：採用通用計算單元以處理所有的資料操作；使用有限能力的 I/O 匯流排在分離的記憶體元件和硬碟元件間傳輸大量資料（Slotnick, 1970；Baum & Hsiao, 1976）。

其原因在於，基於通用電腦架構實現的資料庫伺服器將大量的計算能力用於解析軟體發出的資料庫操作請求，然後調用一系列軟體模組去處理這些請求並搜尋出相應的資料，並透過 I/O 操作將大量資料從次要儲存元件（如硬碟）複製到主要儲存元件（如記憶體），最終

經過大量運算得到最終結果返回給應用軟體。通用電腦設計面向操作更多的是計算，其特點是少量資料/大量運算， 關注的是計算與定址，實現方式是計算單元造訪高速儲存零組件（如記憶體）中的資料獲得計算結果。

而資料庫操作更多的是搜尋與更新，其特點是大量資料/少量運算，關注的是查詢與內容，實現方式則是計算單元造訪大型儲存區零組件（如硬碟）中的資料獲得處理結果。同時，由於通用電腦上的作業系統隔離了資料庫軟體模組與底層硬體，使得對資料儲存零組件和 I/O 快取的控制變得異常困難，因而導致資料造訪效率低下。這些就是在當時的條件下，大數據庫操作需求與通用電腦結構間存在的差距。

基於以上研究，並結合當時日趨成熟的資料關係模型（Codd, 1983）中可描述任意複雜資料操作的基本資料操作集理論，資料庫領域的學者們提出了一種在當時解決大數據處理的作法，即將一些基礎的資料操作功能（如搜尋、更新等）在單獨的專用硬體上實現，而將通用計算資源和 I/O 通道釋放出來用於其他複雜處理，進而實現高效的資料造訪。基於這樣的作法，並利用當時逐漸提高的硬體技術和不斷降低的硬體成本，逐步實現了用於支援大規模高速資料庫存取的專用電腦和硬體系統，即資料庫機（Database Machine）。

2.2.3 系統容錯性

大數據環境下的計算和處理，通常需要多個計算節點構成的叢集進行，即使如此，大數據的計算過程往往也需要耗費較長的時間。在這樣的情況下，大數據處理平臺的整體可用性也顯得尤為重要。衡量系統的可用性通常用系統的正常服務時間與總執行時間的百分比來計算，其計算方式為：

$$\text{Availability}=\text{MTTF}/(\text{MTTF}+\text{MTTR})\ \text{X}100\% \quad (1)$$

公式中 MTTF（Mean Time To Failure）為平均無故障時間代表系統的平均正常服務時間；MTTR（Mean Time To Recovery）為平均維修時間，代表系統從發生故障到恢復正常服務所需要的平均維修時

間。從公式(1)可以看出，在系統總執行時間一定的情況下，如果系統具有很高的可靠性，很少發生故障，即 MTTF 值很大的情況下，系統可用性也會很高。或者系統的自我恢復能力很強，在發生故障後能很快恢復正常工作，即 MTTR 值很小的情況下，系統可用性也會很高。通常而言，一個系統的 MTTF 是要大於 MTTR 的，因此提升 MTTE 值可以更加有效地提升系統可用性。而要保證系統正常運行，通常有兩種方式，一種方式是避免發生故障，即避免系統的硬體和軟體出現故障，這通常需要使用更好的硬體和軟體，而這也意味著更高的成本。因此從提這性價比的角度考慮，採用容忍錯誤的方式提高 MTTF 更為有效。容忍錯誤即，是指在系統中的部分硬體或軟體發生故障的情況下，整個系統仍然能夠正常運行。容錯的主要目標是降低故障帶來的影響，避免系統整體崩潰，確保計算任務正確完成，並且在故障部分恢復正常後，系統應能及時啟用恢復的計算節點，充分利用計算能力。容錯通常會涉及故障檢測、容錯與備份、故障恢復等多項技術。對於大數據環境下的處理平臺而言，容錯主要包括兩方面，資料儲存容錯和計算任務容錯。

2.3 大數據處理架構

自 Hadoop 技術的兩項核心技術 MapReduce 和 Hadoop Distributed File System（HDFS）提出並公佈其開放原始碼實現後，由於其具有在低成本的硬體平臺上實現高性能資料處理的強大優勢，迅速被企業和組織採用。隨著 Hadoop 技術的逐漸流行，更多圍繞 Hadoop 框架的拓展技術和工具也逐漸出現，例如 HBase、Pig、Hive 等。本研究參考多家大型公司基於 Hadoop 的雲端運算技術，找出一個基於 Hadoop 的運算架構圖，如圖 2.3.1 所示。

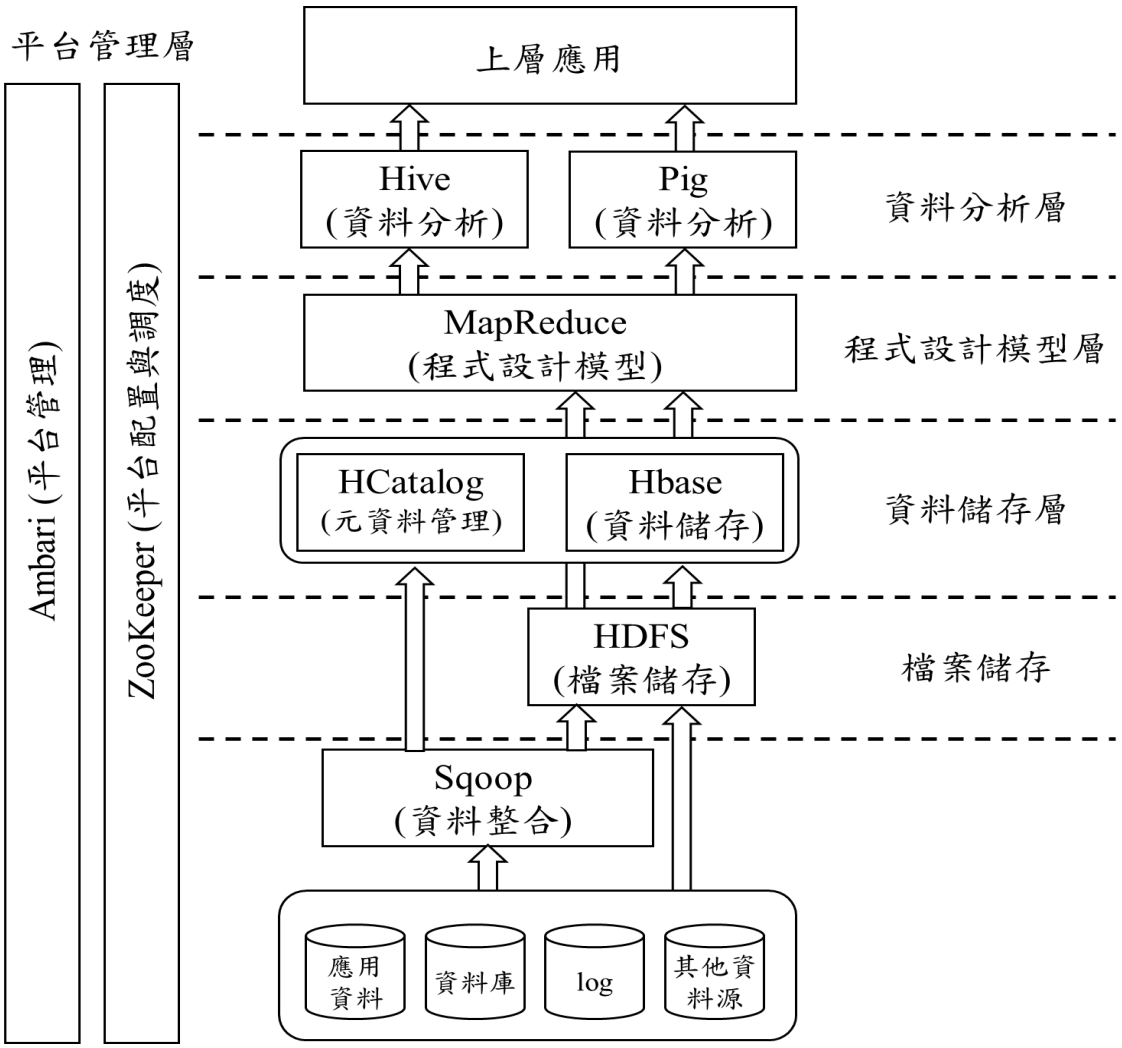

圖 2.3.1 基於 Hadoop 的運算架構圖

（資料來源：劉軍，2013）

下面由下而上的對此架構中的主要組成部分進行介紹：

1. 資料整合層：資料整合層在整個資料架構的最下方，是系統需要處理的資料來源，包括私有的應用資料、儲存在資料庫中的資料、被分析系統運行產生的 log 資料等。這些資料具有結構多樣、類型多變的特點，既有結構化的資料，也有非結構化、半結構化的資料；既有純文字格式的 log 資料，也有多媒體格式的網頁數據。在這些資料中，有些倉資可以直接儲存在 HDFS 中，例如格式化的 log 檔，還有些資料是可以被 MapReduce 程式解析後直接處理的。但是當要處理傳統企業應用保存在資料庫中的資料時，例如經營分析系統產生的歷史資料，由於這些資料不是儲存在 HDFS 中，因此 MapReduce 程式在處理時就需要透過外部 API 的形式造訪這些資料。這種方式既不靈活也不高效，因此 Hadoop 框架引入了一個資料整合層。資料整合層中的元件的作用，是在外部資料來源和檔案儲存層或資料儲存層之間進行連接，以實現雙向的資料高資料整合層元件的典型實例就是 Sqoop 工具。利用 Sqoop 工具，一方面可以將儲存於關聯式資料庫中的資料匯入 Hadoop 元件中以利於 MapReduce 程式或 Hive 工具進行後續處理，甚至直接匯入 HBase 中；另一方面還可以支援將處理後的結果匯出到關聯式資料庫中。

2. 檔案儲存層：檔案儲存層是利用分散式檔案系統技術，將底層數量眾多且分佈在不同位置、透過網路連接的各種儲存裝置組織在一起，透過統一的介面向上層應用提供物件級檔案存取服務能力。檔案儲存層為上層應用遮罩了儲存裝置類型、型號、介面協定、分佈位置等技術細節，並提供了資料備份、故障容忍、狀態監測、安全機制等多種保障可靠的檔案存取服務的管理性功能。同時，利用分散式並行技術，雲端運算大數據處理環境下的檔案儲存層還支援對海量大檔進行高效的並行造訪。在整體架構中，檔案儲存層向下與資料來源和資料整合層連接造訪具體的儲存

資源，向上為程式設計模型和資料儲存層提供檔案存取服務。HDFS 是檔案儲存層的一個典型元件。

3. 資料儲存層：資料儲存層的功能是提供分散式、可擴展的大量資料表的儲存和管理能力。與傳統的關聯式資料庫不同，基於雲端運算的大數據處理架構中的資料儲存層元件，並不要求具有完整的 SQL 支持能力，也不要求資料採用關聯式資料模型進行儲存。它更強調的是在較低成本的條件下實現大數據表的管理能力，可以支援在大規模資料量的情況下完成快速的資料讀寫操作，並且隨著資料量的快速增長，可以透過簡單的硬體擴充實現儲存能力的線性增長。目前 Hadoop 已為資料儲存層提供了兩項技術基礎，HBase 和 HCatalog。HBase 實現了一個欄位導向（Column-oriented）的分散式資料庫儲存系統。HCatalog 是一個資料表和儲存管理元件，可以支援 Pig、Hive、MapReduce 等上層應用間進行資料共用操作。

4. 程式設計模型層：程式設計模型層中的元件的作用，是為大規模資料處理提供一個抽象的平行計算程式設計模型，以及為此模型提供可實施的程式設計環境和執行環境。程式設計模型層是整個處理架構的核心部分，它的運行效率決定了整個資料處理過程的效率。目前在基於雲端運算的大數據處理領域，MapReduce 模型可以說是佔據了主流地位。雖然 MapReduce 的基礎思 想並不具有顛覆性，但是其簡潔高效的特性與電腦叢集及高速網路結合後，具備了處理大數據的高效生產力，因此得到了廣泛的應用，並成為 Hadoop 技術的核心。MapReduce 元件在整個架構中擔當了承先啟後的關鍵角色。一方面程式設計師可以使用 MapReduce 程式設計模型直接架構資料處理程或；另一方面，上層的拓展工具，例如，Hive 等也利用了 MapReduce 計算能力進行資料造訪和分析。

5. 資料分析層：對於大多數資料分析人員來說，掌握複雜的平行計算程式設計能力是一個成本很高的過程。他們應該更關注資料分

析的核心問題，例如建立資料模型、挖掘商業價值等。資料分析層中的元件，就是提供一些高級的分析工具給資料分析人員，以提高他們的生產效率。Hadoop 體系中的 Pig 和 Hive 是這一類的工具。Pig 提供了一個在 MapReduce 基礎之上抽象出的更高層次的資料處理能力，包括一個資料處理語言及其執行環境。而 Hive 則可以將結構化的資料對應為一張資料表，為資料分析人員提供完整的 SQL 查詢功能，並將查詢語言轉換為 MapReduce 任務執行。

6. 平臺管理層：平臺管理層中的元件是確保整個資料處理平臺平穩安全運行的保障。跟其他系統中的管理元件相同，平臺管理層中的元件提供了包括配置管理、運行監控、故障管理、性能優化、安全管理等在內的全套功能。由於基於 Hadoop 技術目前還處於發展期，因此對 Hadoop 平臺管理的成熟組件還不算太豐富，更多的是利用已有的一些開放原始碼元件進行有針對性的管理。對於 Hadoop 叢集的開放原始碼專案主要有 ZooKeeper 和 Ambari。ZooKeeper 主要提供配置管理及元件協調功能，Ambari 則提供了一個用於安裝、管理和監控 Hadoop 叢集的 Web 介面工具。

需特別說明，此架構並非建立一個運用 Hadoop 的技術標準，因此在本研究的第三章會提出一個類似的架構用於切合生理數據監測的 Hadoop 運算模型。

2.4 醫療大數據

2.4.1 移動健康服務

無處不在的健康計算設備、通信網絡和雲計算，將整個世界連接在一起，通過對健康資訊的感知、採集、傳輸、存儲、分析、決策、應用等，打造一個健康資訊服務的社會，將改善人們的生活質量，促進經濟的可持續發展以及和諧社會的建設。

表 2.4.1 訊息種類的階段特徵

種類	硬體	操作系統	特徵	結構
大型機	大型計算機	Unix	穿孔紙帶	主機多終端
微型機	微軟、蘋果	MS-DOS MAC OS	人-機	單機
互聯網	筆記型電腦	Window MAC OS LINX	多機人-人	客戶機和服務器
移動網絡	手機 平板電腦	Android IOS	多機人-物體	瀏覽器和服務器

（資料來源：姚志洪，2013）

2.4.2 醫療大數據的種類

隨著遠程醫療、移動醫療和健康物聯網等新技術的不斷湧現，出現了大量新型數據，與過去主要是結構化數據不同，如今主要的數據類型是半結構化或者非結構化的數據，如 XML 文檔、電子病歷（HL7 CDA）、電子健康檔案（OpenEHR）等。臨床資訊系統（Clinical Information System, CIS）中包含大量的非結構化的數據，例如心電圖、超音波、CT、MR、CR、DR 和 DSA 等，臨床的大量影像檔是醫生診斷的重要依據。醫院資訊系統中，醫學影像存檔與通信系統（Picture Archiving and Communication System, PACS）的數據量遠比醫院資訊管理系統（Hospital Information Management System, HIS）的數據量大得多。在「大數據」時代，傳統的數據庫分析系統正面臨著一次歷史性變革。

2.4.3 個性（別）化醫療

隨著科學家們解密人類基因組，已發現遺傳和非遺傳因素對人類健康，疾病和藥物反應的相互關聯。個性（別）化醫療是根據病人的

遺傳特徵以及所處環境的特點，應用基因組資訊來指導醫療決策，幫助醫生和病人診斷疾病、選擇最有效的疾病治療方法，更好地控制疾病，以及預防疾病的發生，從而實現最佳的診斷、治療和預防效果。由於個體間基因的差異，使得個體問的藥物反應自然存在著不同程度的差異。據文獻報導，美國每年約有 10 萬人因嚴重藥物不良反應而致死，已成為第 4-6 位的死亡原因。在中國每年 5000 多萬住院病人中，經統計至少有 250 萬人住院是因為藥物不良反應有關，引起死亡數達到 19 萬人之多。查明患者基因結構將有助於實現個性化醫療，既可節省醫療費用，又可以達到安全治療的目的。人類基因組學的新進展使得從健康到疾病的預測和積極的幹預成為可能。利用人體終身穩定的基因組或者 DNA 的資訊，在人體健康狀態，甚至在出生時，就能夠對疾病進行定性分析和預測，SNP 組合將為健康危險因素評估提供重要依據。

表 2.4.2 醫療資訊界面

	限制或問題	說明
醫療資訊面	資料取得	其資料多屬於病患個人資料，目前也將面臨個資法之問題，故對於資料的取得更增加許多障礙。 另外硬體設備的運用也受到多家不同廠牌間的競爭與相關法規之規定，使得具有專業評估病患的專業人員無法掌握確切數據。
	專業醫護人員輔助	從文獻上能取得的專業醫療資料有限，許多疾病之危險因數或是判別方法大多為專業醫護人員之內隱知識。
雲端平臺導入面	安全性	安全性為雲端備受眾人所質疑的問題，繼駭客攻擊成功地癱瘓 Google 的雲端基礎設施與 Sony 個資外洩等事件，雲端的安全性也備受爭議，若將病人之風險分析與相關資訊放置雲端中，將可能形成上述之問題產生。
	應用程式開發	雲端應程式之開發可用的 API 主要以 Google GAE、微軟 Windows Azure、Apache Hadoop 與趨勢科技 TCloud 等為主。各大廠皆以不同的技術架構與語法發開，皆有不同的進入門檻。
	資料庫設計	相較於傳統關聯式資料庫，在雲端平臺上許多廠商提供了較不同的資料存取方式，如本研究所使用的 GAE 雲端平臺，其資料庫設計方式以 BigTable 為主，與傳統關聯式大不相同，需重新學習與設計。

（資料來源:簡聖哲 等人，2012）

2.5 疾病問題

老化與疾病經常並存，臨床醫師需熟悉老化對各器官功能之影響，方能正確判斷老年病患臨床資料與數據；對於老年人器官功能的衰退，凡是無法以正常老化來解釋，一定要追究其可能的病因，並設法治療之。

老人生理改變相關的研究常受許多因數之影響而導致不同的結論，例如：

1. 排除各種疾病干擾觀察變數之方法與周密程度
2. 人種、性別、生活型態之差異
3. 縱向性研究（Longitudinal Study）或橫斷面研究（Cross-sectional Study）之實驗架構
4. 基礎（Basal）或壓力（Stress）狀態下之功能觀測

因此，在參考相關文獻時，應檢討各個實驗觀察之設計、適用狀況及優缺點，以作最佳之判斷與運用。此外，老人有相當大的個人歧異性（Heterogeneity），臨床上面對之特定老年病患，即使其某一器官無明顯疾病，該器官功能也未必完全遵循正常老化的範圍，必須依個別狀況作適當之考量。以下簡述心臟血管系統，會隨著老化而改變與其臨床意義以及實驗室數據之變化。

2.5.1 心臟血管系統

心臟體積通常不會單純因老化而改變，但左心室壁的厚度可稍微增加。竇房結（SA Node）之細胞數目從 20 歲開始減少，至 75 歲時僅剩約 10%心臟瓣膜與傳導系統（Conduction System）會纖維化與鈣化。竇房結與傳導系統的退化，使老年人較易罹患病竇症候群（Sick Sinus Syndrome）與傳導異常（Conduction Disturbance）。另外動脈變長而呈現紅曲，其內膜變得不光滑，其厚度也增加。動脈壁中層之平滑肌層變厚，鈣化程度增加，彈性蛋白斷裂增多。老年人動脈硬化的高盛行率到底是老化或疾病所引起，目前仍有爭議。不過，動脈硬化使老年人容易發生高血壓、冠狀動脈心臟病與腦中風，卻是不爭的事

實。在休息狀態下，心輸出量（Cardiac Output）與心搏容量（Stroke Volume）不太受老化影響，心臟對交感神經或其介質鄰-苯二酚胺（Catecholamine）的刺激反應變差。運動時可達到的最快心跳速率會隨年齡增加而約略呈線性下降，吾人可以用 220 減去年齡來估算之。運動後，心臟恢復到休息狀態所需的時間會延長。心肌鬆弧（Relaxation）的速度減緩，使舒張早期由左心房流入左心室的血液量減少，心臟的前負荷（Preload）因而更依賴左心房收縮來維持。一旦罹患心房纖維顫動（Atrial Fibrillation），其心輸出量所受的不利影響將大於年輕人。另外，周邊血管的阻力上升，壓力反射（Baroreflex）的敏感度變差。老年人因壓力反射變差與血管變硬（Stiff），容易有姿勢性低血壓；因此，在老年人投與抗高血壓藥物或作用於中樞神經系統的藥物時，需注意是否加重姿勢性低血壓的現象。

血壓方面，多數流行病學的研究顯示收縮壓（Pulse Pressure）會隨年齡而上升；在美國的研究則顯示收縮壓隨著年齡增加而持續上升，舒張壓從 35 歲左右開始增加，至大約 60 歲便不再上升，甚至會稍微下降。在某些少數原始村落或未開發國家的研究卻發現收縮壓不一定隨老化而上升，顯示遺傳或環境因數會影響老化過程中血壓的表現。

另外，在心血管疾病危險因數之血脂方面，血中總膽固醇（Total Cholesterol）的濃度隨年齡而上升，其中低密度脂蛋白（Low-Density Lipoprotein, LDL）膽固醇的變化大致與總膽固醇相同，而高密度脂蛋白（High-Density Lipoprotein, HDL）膽固醇則逐漸增加。不過，有其它的研究顯示，女性的高密度脂蛋白膽固醇從 20 到 80 歲會下降 30%。至於三酸甘油酯（Triglyceride），其血中濃度會隨年齡而上升。不過，在某些原始部落內，沒有上述有關脂質變化的現象，顯示可能有其他因素的影響，例如遺傳或生活習性。

2.6 心臟建模

上世紀 40 年代末，美國的弗明漢心臟研究中心（Framingham Heart Study）就開始了針對人群的心血管疾病風險研究，不僅提出了危險因素的概念，並且預測不同危險水準的個體在一定時間內發生冠心病危險的概率。在此基礎上，許多國家也開展了各自的研究，陸續開發了基於各種心腦血管疾病預測模型的評估工具，如：SCORE、PROCAM、QRICK、Reynolds Risk Score 及 MUCA。雖然這些預測方法已經成為臨床中預防心腦血管疾病的主要工具，但是這些傳統的心腦血管疾病風險評估預測工具都是基於大規模樣本的長期隨訪，採用醫學統計的方法產生的。由於個體差異性，這些預測方法只能針對人群，無法實現個體風險的短期預測。多個國際研究機構在上世紀 90 年代，共同建立完整心臟的數學模型。2004 年，通過在「CardiacPhysiome Project」項目中的長期合作，兩所世界頂級大學的研究人員共同提出了一個多尺度的心臟電力耦合模型。歐盟在 FP7 框架協議下，於 2007 年啟動了虛擬生理人（Virtual Physiological Human, VPH）計劃。該計劃融合牛津大學、奧克蘭大學和帝國理工等 27 個高校和國家實驗室，成立專門的 VPH 網絡，為建立 VPH 所面臨的生物醫學、健康保健和資訊通訊技術挑戰開展系統的研究。

2.7 感測器

開發無擾式傳感器用於連續、實時的獲取心血管健康資訊無擾式傳感器可以記錄人體的生理信號和日常行為等資訊，如：心電信號（ Electrocardiograph, ECG ）、光電容積脈搏波（Photoplethysmogram ,PPG）、呼吸率、心率、體溫、血氧飽和度、血壓、姿態以及運動情況等，在採集信號的同時不影響人的正常生活，甚至可以不被覺察。這種無擾式的監測設備一般通過兩種方式來實現：一種是將傳感器嵌入到衣服或者飾物中，如耳環（Online: http://www.vphnoe.eu/.)、戒指（Paradiso et al., 2005)、手套（Poh，2010）等。另外一種是將傳感器嵌入到日常生活的背景環境中，比如說傢俱、家用電器、房屋建築（Duun et al., 2010； Rothmaier et al., 2008；Ishijima，1993）等。根據信號的來源是被動還是主動，我們可以將無擾式傳感器分成兩種類型。一種是無能源的傳感方法是被動對人體的信號進行無擾式的檢測，比如電容耦合方式的 ECG 測量。而有能源傳感方法，則是由傳感器主動向人體發送能量，並且檢測反射或者後向散射回來的能量，比如利用雷達遠程檢測心率或者利用紅外線進行溫度測量等。

無擾式設備測得的生理資訊可以通過無線傳輸技術傳送到遠程式控制制終端，這樣病人的心血管健康資訊就可以在院外實時獲得，不僅可以降低病人經常跑去醫院的醫療成本，同時可以在發生急性心血管疾病（Cardiovascular Disease, CVD）事件的時候及早採取救治行動。更重要的是，院外的測量結果往往比臨床的診斷更加有意義，因為臨床血壓測量並不一定能夠反映出真實的血壓資訊，甚至可能提供一些錯誤的臨床診斷資訊，比如白大衣高血壓。多方專家提出對於已知的或疑似高血壓的病人，更適合採用家庭式的血壓監測方法，並且應該使之成為一種常規的血壓測量方式（Lim，2006）。風險因數已成為當前醫療產業的重大需求。

最近幾年，國際上的研究團隊在無擾式設備監測心血管健康資訊方面開展了一系列的研究，並取得了突出的成績。對於 ECG 的測量，

基於電容耦合原理的非接觸式電子織物電極已經取代傳統的凝膠電極片，這種織物電極可以嵌入到床單或者衣物中進行無擾式生理參數監測（Chi，2010；Gu，2009）。Poh 等人（2018）提出了一種簡單的非接觸式的脈搏傳感技術，它採用攝像頭記錄人臉區域，然後採用獨立成分分析方法從數字化的人臉彩色圖像中提取出血容量脈衝，並且自動計算心率、呼吸率等生理資訊（Humphreys et al., 2007）。K.Humphreys 等人(2007)同樣採用非接觸傳感方法建立了一個基於攝像頭的雙波長光電容積脈搏波採集系統，可以遠程測量血氧飽和度（Chen et al., 2000）。基於脈搏波傳導速度的血壓估計演算法使無袖帶、連續的血壓測量成為可能。脈搏波傳導時間（Pulse Transit Time, PTT），也就是脈搏波從心臟傳輸到外周的時間，可以簡單的利用 ECG 和 PPG 信號進行測量。國內外的研究者積極的開展研究希望能夠證明 PTT 可以替代傳統血壓測量（Poon & Zhang，2005；McCombie et al., 2006；Foo et al., 2006；Gesche et al., 2012；Axisa et al., 2005）。

除了傳感技術，無擾式監測設備在發展過程中，人們還比較關心的就是它的用戶友好性設計問題。近些年中，智慧紡織技術不斷延伸與發展，提供了穿戴式設備一種有效而又實際的解決方案。我們可以將所有的穿戴式傳器都無縫整合在一件衣服上，而衣服是我們生活中的必備品，這樣我們就可以自然而然地在日常生活中進行穿戴式測量（Kim et al., 2011）。

也就是說，柔性電子織物這個新興領域的出現為穿戴式設備的設計提供了全新的解決方式（Science，2011）。這些柔性電子織物具有很多良好的特性，比如織物結構、重量輕、具有生物相容性等，相信它會進一步推動可穿戴式設備向完全無擾的監測方式前進。

2.8 Hadoop 原理與運行機制

Hadoop 的核心由 3 個子專案組成：Hadoop Common、HDFS 和 MapReduce。Hadoop Common 子專案在 Hadoop 0.20 版本前被稱為 Hadoop Core，從其名稱就可以看出來，Hadoop Common 專案是為 Hadoop 整體架構提供基礎支援性功能，主要包括了檔案系統（File System）、遠端程式呼叫協定（Remote Procedure Call , RPC）和資料序列化庫（Serialization Libraries）。HDFS （Hadoop Distributed File System） 由早期的 NDFS 演化而來，是一個分散式檔案系統，具有低成本、高可靠性、高輸送量的特點。MapReduce 是一個程式設計模型和軟體框架用於在大規模電腦叢集上編寫對大數據進行快速處理的並行化程式。在實際應用環境中，Hadoop Common 更多的是隱藏在幕後為架構提供基礎支援，而 HDFS 和 MapReduce 的邏輯元件相互配合完成使用者提出的大數據處理請求。圖 2.8.1 展示了一個典型的 Hadoop 部署環境圖及邏輯元件之間的互動，並結合此圖對 Hadoop 的主要邏輯元件進行說明。

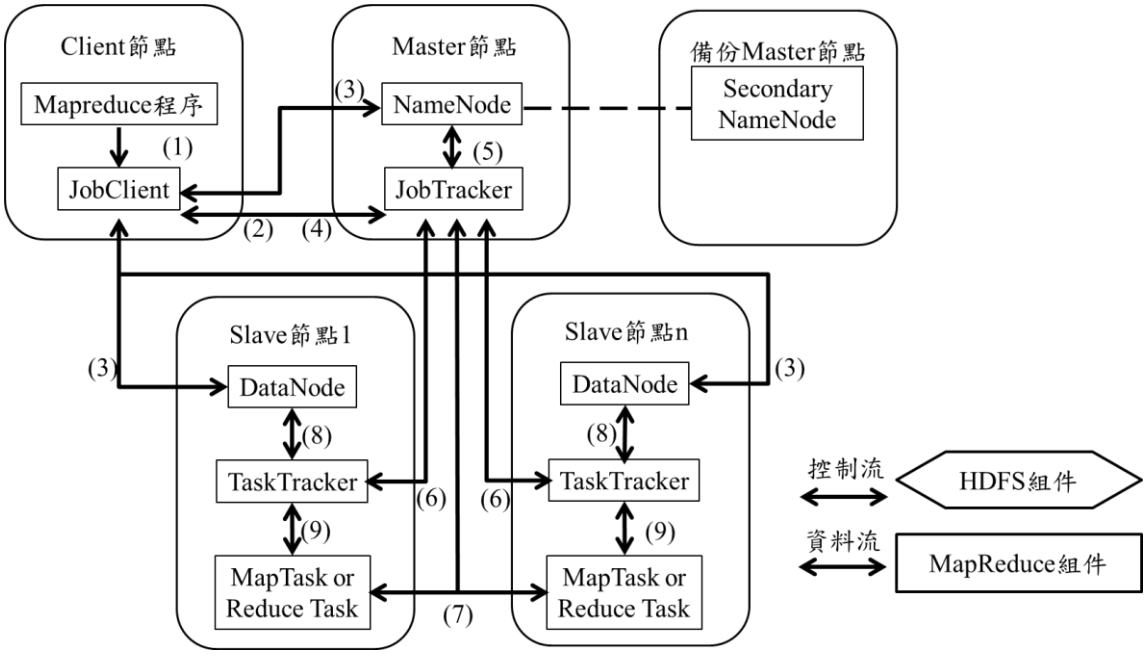

圖 2.8.1Hadoop 運行機制圖

（資料來源：劉軍，2014）

2.8.1 HDFS 組件

HDFS 是一個適合架構於廉價電腦叢集之上的分散式檔案系統，具有低成本、高可靠性、高輸送量的特點，由早期的 NDFS 演化而來。HDFS 組件包含：

1. NameNode：NameNode 是 HDFS 系統中的管理者，它負責管理檔案系統的命名空間，維護檔案系統的樹狀架構及所有的檔和目錄的中繼資料。這些資訊儲存在 NameNode 維護的兩個硬碟檔：命名空間鏡像檔和編輯 log 檔。同時，NameNode 中還保存了每個檔與資料區塊所在的 DataNode 的對應關係，這些資訊被用於其他功能元件查詢所需檔資源的資料伺服器。
2. Secondary NameNode：在一個 Hadoop 叢集環境中，只有一個 NameNode 節點，很顯然，NameNode 成了整個 HDFS 系統的關鍵故障點，一旦發生故障將影響整個系統的運行。為了避免這樣的問題出現，Hadoop 設計了 Secondary NameNode 節點，它一般在一台單獨的實體電腦上運行，與 NameNode 保持通訊，按照一定時間間隔保持檔案系統中繼資料的快照。當 NameNode 發生故障時，系統管理者可以透過手工配置的形式將保存的中繼資料快照恢復到重新啟動的 Namenode 中，降低資料丟失的風險。
3. DataNode：DataNode 是 HDFS 檔案系統中保存資料的節點。HDFS 中的檔通常被分割為多個資料區塊，以多備援備份的形式儲存在多個 DataNode 中。DataNode 定期向 NameNode 報告其儲存的資料區塊清單，以備使用者透過直接造訪 DataNode 獲得相應的資料。

2.8.2 MapReduce 組件

MapReduce 是一個程式設計模型和軟體框架，用於在大規模電腦叢集上編寫對大數據進行快速處理的並行化程式。MapReduce 組件包含：

1. JobClient：JobClient 是基於 MapReduce 介面庫編寫的用戶端程式，負責提出 MapReduce 作業。
2. JobTracker：JobTracker 是應用於 MapReduce 模組之間的控制協調者，它負責協調 MapReduce 作業的執行。當一個 MapReduce 作業提出到叢集中，JobTrack 負責確定後續執行計畫，包括需要處理哪些檔案、分配任務的 Map 和 Reduce 執行節點、監控任務的執行、重新分配失敗的任務等。每個 Hadoop 叢集中只有一個 JobTracker。
3. TaskTracker：TaskTracker 負責執行由 JobTracker 分配的任務，每個 TaskTracker 可以啟動一個或多個 Map 或 Reduce 任務。同時 TaskTracker 與 JobTracker 間透過心跳(HeartBeat)機制保持通訊，以維護整個叢集的運行狀態。
4. MapTask、ReduceTask：MapTask 和 ReduceTask 是由 TaskTracker 啟動的負責具體執行 Map 任務和 Reduce 任務的程式。

介紹完這主要的 7 個元件後我們根據圖 2.8.1 上的流程編號做元件之間的關係說明：

1. MapReduce 程式啟動一個 JobClient 實例以開啟整個 MapReduce 作業（Job）。
2. JobClient 透過 getNewJobId()介面向 JobTracker 發出請求，以獲得一個新的作業 ID。
3. JobClient 根據作業請求指定的輸入檔計算資料區塊的劃分，並將完成作業需要的資源，包括 JAR 檔、設定檔、資料區塊，儲存到 HDFS 中屬於 JobTracker 的以作業 ID 命名的目錄下，一些檔案（如 JAR 檔）可能會以多備援備份的形式儲存在多個節點上。

4. 完成上述準備工作後，JobClient 透過調用 JobTracker 的 submitJob()介面提出此作業。
5. JobTrackerru 將提出的作業放入一個作業佇列中等待進行作業調度，以完成作業初始化工作。作業初始化主要是建立一個代表此作業的運行物件，作業運行物件中封裝了作業包含的任務和任務運行狀態紀錄資訊，用於後續追蹤相關任務的狀態和執行進度。
6. JobTracker 還需要從 HDFS 檔案系統中取出 JobClient 放好的輸入資料，並根據輸入資料建立對應數量的 Map 任務，同時根據 JobConf 設定檔中定義的數量產生 Reduce 任務。
7. TaskTracker 和 JobTracker 間透過心跳機制 TaskTracker 發送的心跳訊息中包含了當前是否可執行新的任務的任務資訊，根據這個資訊 JobTracker 將 Map 任務和 Reduce 任務分配到空閒的 TaskTracker 節點。
8. 被分配了任務的 TaskTracker 從 HDFS 檔案系統中取出所需的檔案，包括 JAR 程式檔和任務對應的資料檔案，並存入硬碟，並啟動一個 TaskRunner 程式實例準備運行任務。
9. TaskRunner 在一個新的 Java 虛擬機器中根據任務類別建立出 MapTask 或 ReduceTask 進行運算。在新的 Java 虛擬機器中運行 MapTask 和 ReduceTask 的原因，是避免這些任務的運行異常影響 TaskTracker 的正常運行。MapTask 和 ReduceTask 會定時與 TaskRunner 進行通訊報告進度，直到任務完成。

2.8.3 Hadoop 相關的外加功能簡介

MapReduce 和 HDFS 作為 Hadoop 技術髓系的兩個核心元件，已經廣為業界熟知。但實際上 Hadoop 技術體系不僅僅只有這兩部分，還包括了很多其他技術，才能構成一個完整的大規模分散式運算系統。下面再介紹三個與 Hadoop 體系相關的外加功能，它們是由 Apache 開放的原始碼，其整理如下：

1. Hadoop Database（HBase）：HBase 是一個分散式的、欄位導向（Column-oriented）的開放原始碼資料庫，不同於一般的關聯式資料庫，它是一個適合於非結構化大數據儲存的資料庫。
2. Hive：Hive 是一個基於 Hadoop 的資料倉庫工具，它可以結構化的資料檔案對應為一張資料庫表，並提供強大的類 SQL 查詢功能，可以將 SQL 語句轉換為 MapReduce 任務進行運行。
3. Pig：Pig 是一個用於大數據分析的工具，包括了一個資料分析語言和其執行環境。Pig 的特點是其結構設計支援真正的並行化處理因此適合應用於大數據處理環境。

第三章　研究方法

本研究從數據挖掘的視角開始做開端，因為其功能主要可以分為描述型和預測型，十分適合本研究的探討方式。此外， 透過多種感測器對生理數據收集，形成的大數據，本研究也將使用 OLAP 之後的更新科技 Hadoop 這項新的大數據處理辦法，針對此技術的核心 Mapreduce 以及 HDFS 描述在大數據下的生理數據之應用。

3.1 數據挖掘的功能選擇

數據挖掘的種類，按照功能可以劃分為兩類，即描述型的數據挖掘和預測型的數據挖掘（D'Agostino et al., 2008）。前者的功能主要是描述數據庫中所有數據的一般特性；後者的功能主要是在當前數據上進行推斷並以此推斷作出預測。因此，確定數據挖掘的任務時必須綜合考慮數據挖掘的功能、要挖掘的數據類型和用戶關心的主題與興趣。

常見的數據挖掘功能有基於數據特徵化和區分方法的概念描述、基於頻繁模式的關聯分析、基於有指導學習的分類分析、基於無指導學習的聚類分析、基於罕見事件出現概率預測的離群點分析以及基於規律和趨勢預測的時序演變分析等（Conroy et al., 2003；Assmann et al., 2002；Hippisley et al., 2007）。本文主要探討關聯分析、離群點分析和時序演變分析。

1. 關聯分析：應用於感測器蒐集到的各種數據，又或是醫院提供的資料。舉例來說，感測器蒐集到的使用者所在的緯度，就能當作一個資料作保留，以利於後續有更多的數據匯入後，能多出一個維層次進行分析。
2. 離群點分析：主要探討溫度的變化和空氣品質這兩類外部環境因素，對本身就有心血管疾病的患者的影響，會有多大的概率呈現暈眩昏厥，又或直接進入心肌梗塞的死亡階段。
3. 時序演變分析：心血管疾病本身竟是一種慢性病，符合此種分析法具有規律的趨勢特徵。它會隨著時間的推移逐漸地將本來不明

顯的徵狀逐步放大，在過去因為蒐集資料碰到一個最大的問題就是時間的不連續性，因此這類研究大多只能讓患者待在實驗室中進行，又或者建立數學模型做模擬，但現在有了方便隨身攜帶的感測器，我們將能在短時間內去找出發病狀的特徵或是變化當下的因素。

3.2 數據挖掘的過程

數據挖掘是從數據倉庫現有數據源開始，對其進行一系列的數據處理並形成相應的模型，對這些模型採取觀察者能夠理解和接受的方式進行展示和解釋，最終轉化為對企業管理決策有用的知識。數據挖掘是一個交互和反複的複雜過程，需要分析人員大量的決策分析型幹預行為。它包括 6 個步驟，在本研究中只使用 1 到第 4 步驟；而每個步驟之間又相互影響，可能會出現多次反復，因此整體上是一個螺旋式的上升過程。

3.2.1 確定系統挖掘的目標

主要任務是根據應用領域的經驗知識，從用戶角度考慮系統分析的需求從而確定數據挖掘的目標。需要綜合考慮系統工作的性質（如：哪些工作可由系統自動完成、哪些工作需要人工處理）、性能目標和挖掘模型的可理解性等多方面的因素。

3.2.2 數據選取

根據系統挖掘的目標，從數據源中選取與本次數據挖掘任務相關的數據集並將其集成，形成數據挖掘的目標數據。這個步驟中需要解決由於數據源數據類型、操作系統以及數據挖掘平臺的不同所造成的數據格式差異問題。

3.2.3 數據淨化

對目標數據中存在的不完整、不一致、不精確以及冗餘的“臟數據”採用基於規則的方法、神經網絡方法或者模糊匹配技術進行分析並實施相應的清洗後預處理操作，形成適合於數據挖掘的淨化數據。這個步驟中需要解決異質數據源數據之間的邏輯差異問題。

3.2.4 數據降維和轉換

對淨化數據進行降維和轉換操作，形成能夠進行數據挖掘的直接數據。降維是指在保證數據的不變表示或發現了數據的不變表示的前提下，減少變量的數目，將其轉換到一個更易找到解的空間上。數據轉換主要包括數據類型的轉換、數據組織方式的轉換、對數據的屬性進行算術運算元或邏輯運算元的轉換等，轉換時需要綜合考慮數據挖掘的目標、操作和技術等因素。

3.3 聯機分析處理（OLAP）

聯機分析處理（On-Line Analytical Processing, OLAP）是一種為決策者或者分析人員提供直觀的大量數據分析和深入理解的處理技術。此技術與數據倉庫技術相伴而發展起來，它彌補了數據倉庫在直接支持多維數據視圖方面的不足，下表 3.3.1 為 OLAP 相關的維度概念（Lloyd-Jones et al., 2010；Online: http://heart.physiomeproject.org/index.html；Online: http://www.vphnoe.eu/）。

表 3.3.1 OLAP 相關的維度概念

維度	描述	在 OLAP 技術中，一個實體的兩個或者多個的屬性定義為維度，通過「維度」這個概念的引入，使用 OLAP 技術的用戶可以從多維度對同樣的數據進行分析，從而得到不同程度的分析結果和效果，所以維度是 OLAP 技術的核心。
度量值	描述	是用戶查看 OLAP 分析結果所關注的面向某一個主題的一組數字數據，例如醫生數量、醫療器械的利用率、病床的使用率、藥品銷售額等。
維層次	描述	是用戶觀察數據分析結果時所關注的維度的細微性，比如「省份、城市、區、街道」就是本項目中醫療資源的地區維度的一個層次，用戶可以根據自己的實際需要調整維度的層次。
維成員	描述	是維度的一個具體的屬性值，例如「臺北市信義區孝東路四段 216 巷 8 弄 2 號」就是本項目中醫療資源的地區維度的維成員。

維單元	描述	是由度量值、維層次和維成員構成的集合，例如：(2009 年 11 月，信義區，助理醫師，1536)，就表示 2009 年 11 月臺北市信義區助理醫師的總人數為 1536 人。
聚集	描述	是為了減少對用戶的查詢回應時間提出了一種數據預處理，表現形式常為數據匯總。OLAP 通過數據的聚集有效的提高查詢效率，但是過多的數據聚集會增加磁盤存儲空間，使得存儲維護變得困難。

3.3.1 OLAP 的特點

OLAP 主要有以下的特點（Paradiso et al., 2005；Poh et al., 2010）：

1. 多維性

OLAP 提供對多維數據集的數據進行切片、切塊、旋轉和鑽取等多維數據動態分析操作，力爭為用戶提供滿足其需求的多角度多細微性的深度分析。

2. 聯機性

OLAP 的聯機性主要體現在支援用戶根據自身需要可以實時定制維度、維層次等，並能快速響應用戶的各種分析請求，保證在最大延時時間內對用戶的請求給出合理的響應。

3. 可擴展性

用戶不僅可以在 OLAP 系統中分析數據，也可以通過外部的 OLAP 輔助工具來完成數據的深度分析，具有較好的可擴展性。

4. 信息性

OLAP 系統實時在線管理和維護著海量的多模數據資訊，並能及時完成用戶對數據處理的各個響應級別的請求。

3.4 Hadoop 簡介

Hadoop 由 Apache Software Foundation 公司於 2005 年秋天作為 Lucene 的（Foo et al., 2006）子 Hadoop Logo 項目和 Nutch（Gesche et al., 2012）的一部分正式引入，它受到最先由 Google 實驗室開發的 MapReduce 和 GFS（Google File System）的啟發。Hadoop 是一個能夠對大量數據進行分佈式處理的軟件框架，它以一種可靠、高效、可伸縮的方式進行海量數據處理。用戶通過 Hadoop 可以在不瞭解分佈式底層細節的情況下，開發分佈式程式，充分利用集群的能力高速運算和存儲。簡單地說，Hadoop 是一個可以更容易開發和運行處理大規模數據的軟件平臺。

3.4.1 MapReduce 工作機制

MapReduce 運行框架中包含以下幾類獨立組件：

1. Client

   Client 節點上運行了 MapReduce 程式和 Job Client，負責提出 MapReduce 作業和為使用者顯示處理結果。

2. JobTracker

   JobTracker 負責協調 MapReduce 作業的執行，是 MapReduce 運行框架中的主控節點。JobTracker 的功能包括制定 MapReduce 作業的執行計畫、分配任務的 Map 和 Reduce 執行節點、監控任務的執行、重新非配失敗的任務等。每個 Hadoop 叢集中只有一個 JobTracker。

3. Map TaskTracker

   Map TaskTracker 負責執行由 JobTracker 分配的 Map 任務，系統中可以有多個 Map TaskTracker。

4. Reduce TaskTracker

   Reduce TaskTracker 負責執行由 JobTracker 分配的 Reduce 任務，系統中可以有多個 Reduce TaskTracker。

5. 分散式檔案儲存系統

分散式檔案儲存系統中儲存了應用運行所需要的資料檔案及其他相關設定檔。

6. 作業（Job）

作業是指 MapReduce 程式指定的一個完整計算過程，一個作業在執行過程中可以被拆解為若干 Map 和 Reduce 任務完成。

7. 任務（Task）

任務是 MapReduce 框架中進行平行計算的基本事務單元，分為 Map 和 Reduce 任務，一個作業通常包含多個任務。

3.4.2 MapReduce 作業運行流程

MapReduce 的作業主要按下面六大流程，21 步驟組成（Polo et al., 2010）：

1. 作業提出

(1) 使用者編寫 MapReduce 程式建立新的 JobClient 實例。

(2) JobClient 實例建立後，向 JobTracker 請求獲得一個新的 JobId，用於標識本次 MapReduce 作業。

(3) 然後 JobClient 檢查本次作業指定的輸入資料和輸出目錄是否正確。在檢查無誤後，JobClient 將運行作業需要的相關資源，包括本次作業相關的設定檔、輸入資料分片的數量，以及包含 Mapper 和 Reducer 類的 JAR 檔存入分散式檔案儲存系統中，其中 JAR 檔將以多個備份的形式存儲。

(4) 完成以上工作後，JobClient 向 JobTracker 發出作業提出請求。

2. 作業初始化

(1) 作為系統主控節點，JobTracker 會收到多個 JobClient 發出的作業請求，因此 JobTracker 實現了一個佇列機制處理多個請求。收到的請求會放入一個內部佇列，由作業調度器進行調度。JobTracker 為作業進行初始化工作。

(2) 初始化的內容是建立一個代表此作業的 JobInProgress 實例，用於後續追蹤和調度此作業。JobTracker 要從分散式檔案儲存系統中取出 JobClient 儲存的輸入資料分片資訊，以決定需要建立的 Map 任務數量，並建立對應的一批 TaskInProgress 實例用於監控和調度 Map 任務。而需要建立的 Reduce 任務數量和對應的 TaskInProgress 實例，則由設定檔中的參數決定。

3. 任務分配

(1) MapReduce 框架中的任務分配機制是採用拉的機制實現的。在任務分配之前，負責執行 Map 任務或 Reduce 任務的 TaskTracker 節點均已經啟動。TaskTracker 一直透過 RPC 向 JobTracker 發送心跳訊息詢問有沒有任務可以做。如果 JobTracker 的作業佇列不為空，則 TaskTracker 發送的心跳訊息將會獲得 JobTracker 給它派發的任務。由於 TaskTracker 節點的計算能力（由核心數量和記憶體大小決定）是有限的，因此每個 TaskTracker 節點可運行 Map 任務和 Reduce 任務的數量也是有限的，及每個 TaskTracker 有兩個固定數量的任務槽，分別對應 Map 任務和 Reduce 任務。在進行任務分配時，JobTracker 優先填滿 TaskTracker 的 Map 任務槽，及只要有空閑 Map 任務槽就分配一個 Map 任務，Map 任務槽滿了後才分配 Reduce 任務。

4. Map 任務執行

(1) Map TaskTracker 在節點收到 JobTracker 分配的 Map 任務後，將執行一系列操作以執行此任務。首先，建立一個 TaskInProgress 物件實例以調度和監控任務。然後將作業的 JAR 檔和作業的相關參數設定檔從分散式檔案儲存系統中取出，並複製到本地工作目錄下（JAR 檔案中的內容須經過解壓）。

(2) 完成這些準備工作之後，TaskTracker 新建一個 TaskRunner 實例來運行此 Map 任務。

(3) TaskRunner 將啟動一個單獨的 JVM，並在其中啟動 MapTask 執行使用者指定的 map( )函數。

(4) 使用單獨的 JVM 運行 MapTask 的原因是為了避免 MapTask 的異常影響 TaskTracker 的正常運行。MapTask 計算獲得的資料，定期存入快取中。

(5) 並在快取滿的情況下存入硬碟中。

(6) 在任務執行時，MapTask 定時與 TaskTracker 通訊報告任務進度。

(7) 直到任務全部完成，此時所有的計算結果會存入硬碟中。

5. Reduce 任務執行

(1) 在部分 Map 任務執行完成後，JobTracker 即將按照上面第 3 步同樣的機制開始分配 Reduce 任務到 Reduce TaskTracker 節點中。

(2) 與 Map 任務啟動過程類似，Reduce TaskTracker 同樣會產生在單獨 JVM 中的 ReduceTask 以執行使用者指定的 reduce( ) 函數

(3) 同時 Reduce Task 會開始從對應的 Map TaskTracker 節點中遠端下載中間結果的資料檔案。

(4) 直到此時，Reduce 任務還沒有真正開始執行，而僅僅是做好執行環境和資料的準備工作。只有當所有 Map 任務執行完成後，JobTracker 才會通知所有 Reduce TaskTracker 節點開始 Reduce 任務的執行。同樣，ReduceTask 定期與 TaskTracker 通訊報告任務進度，直到任務全部完成。

6. 作業完成

(1) 在 Reduce 階段執行過程中，每個 Reduce task 會將計算結果輸出到分散式檔案儲存系統中的站存檔案中。

(2) 當全部 ReduceTask 完成時，這些暫存檔案會合併為一個最終輸出結果。JobTracker 在收到作業包含的全部任務的完成通知

後（透過每個 TaskTracker 與 JobTracker 間的心跳訊息），會將此作業的狀態設置為「完成」。當此候的 JobClient 的第一個狀態輪詢請求到達時，將會獲知此作業已經完成

(3) 於是 JobClient 會通知使用者程式整個作業完成並顯示必要的資訊。

第四章　基於 Hadoop 的 OLAP 系統架構

本文以建立六個人的生理數據為例，在這六個人的角色設定上為三對夫妻。之所以取三對夫妻而非以一個人當案例的原因是增加維度編碼上的示範；又因為示範不宜過度複雜，所以也僅選擇做三對。下圖 3.4.1 為本研究系統之簡圖。

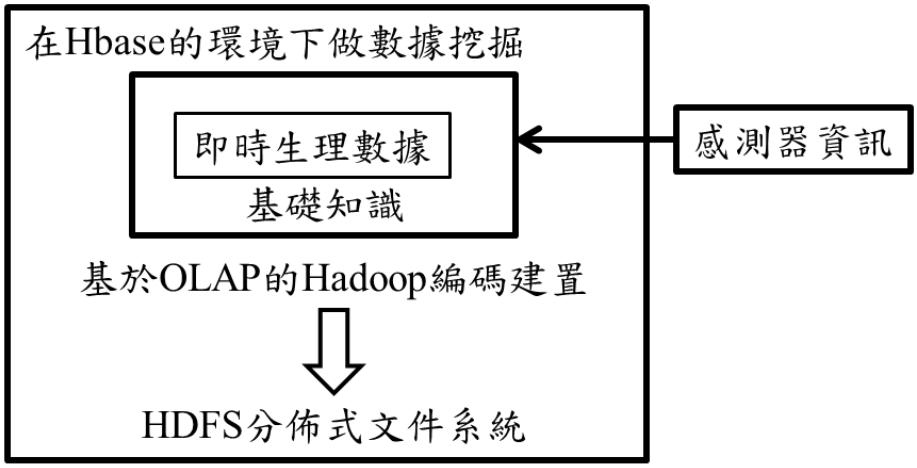

圖 3.4.1 本研究系統簡圖

4.1 確定系統挖掘的目標

在過去的研究，我們可以看到在各種特定病況下的死亡人數或是比例，其中還會夾雜大量的資訊，例如：發病的原因、受損的面積、醫治時用藥物...等，而這些都是在完成一個個病例後，經過完整的統計後，我們進行的一些歸納，例如心臟衰竭是一種慢性病，就是我們透過幾種相似的病歷紀錄，最終歸類出的一種說法。過去的知識，是藉由大樣本的採集，經過統計的檢定方式，得出一種因果關係，或是說明一個主效果的顯著性或是交互作用的現象。然而當資訊科技的進步，大數據的廣博使用，遠距醫療的興起，以個人作為核心的生理數據顯得特別重要，這將會使資料具有最高的純度，同時也能做到即時的應用，以上種種將會反饋到每個人特別需要受到的照護，進而滿足病患真正的需求。

大數據的應用，主要用於描述和預測，本研究依照這兩個應用進行資料存儲結構。首先本研究基於 Hadoop 的 0LAP 大數據按維存儲的模型建立。其中又分成兩大類：計算用的模型、各類數據的維度模

型。計算模型主要能體現大數據的預測能力，但需要先有各類數值維度的支持，才能完備預測的準確性。

本研究將列舉出幾個與心血管疾病相關的數據，透過數據類型的定義和分析，呈現在 OLAP 就運用的維度概念，並在此基礎上，作出 Hadoop 的兩項核心 MapReduce 和 HDFS。以圖 4.1.1 說明基於 Hadoop 的 OLAP 系統架構。

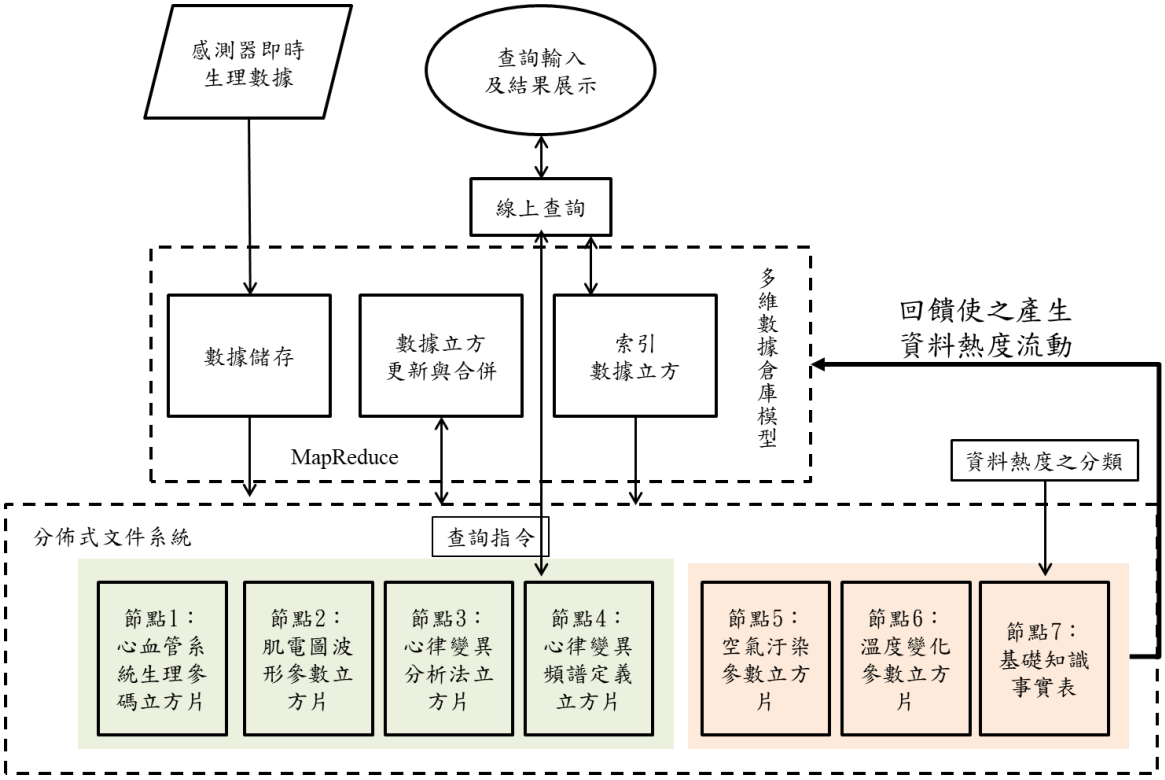

圖 4.1.1 基於 Hadoop 的 OLAP 系統架構

4.2 數據選取依照 Hadoop 的 Hbase 技術集成資料

HBase 雖然在 NoSQL 資料庫在依些方面也存在一些缺陷，並且還不能完全取代傳統的關聯是資料庫，但他們已經成為大數據處理領域的主要支援技術，並在開放原始碼的前提下快速進化，下面將會用這種技術，將這幾種數據加以整合與儲存。

value = Map(TableName, RowKey, Column Key, Version)

其中：

1. TableName（表名）是一個字串，為一張資料表的標識。
2. RowKey（行關鍵字）可以是最大長度 64KB 的任意字串，是用來搜尋紀錄的主鍵。
3. ColumnKey（列關鍵字）是由列族（Column Family）和限定詞（Qualifier）構成的。列族是 HBase 中很重要的概念， 因為資料是以列族為依據進行儲存的。在定義表結構時，列族需要提前定義好，但列的限定詞不需要，可以在使用時產生，且可以為空。HBase 透過這種方式實現了靈活的資料結構。
4. Version（版本）的存在是為了適應同一資料在不同時間的變化，尤其是網路上的網頁數據，在 URL 相同時， 可能會在多個時間存在多個版本。因此 HBase 中的版本就直接採用了時間截記來表示。
5. 由<RowKey, ColumnKey, version>三個元素確定的一個單元為 HBase 中的資料元（Cell），資料元中的資料以二進位形式儲存，由使用者進行格式轉換。

瞭解以上幾個關鍵概念後，可以將圖中的實例轉為通常習慣的資料表的形式，見表 4.2.1。

表 4.2.1 以心血管疾病為例的 Hbase 邏輯視圖表

行關鍵字	版本	列族：contents	列族：anchor
搜尋的議題	時間	內容	位置
溫度	t1	+4/溫度	cwb.gov.tw
溫度	t2	+3/溫度	weather.gov
空汙	t1	-2/PM	cwb.gov.tw
空汙	t2	-2/PM	weather.gov
基礎知識	t1	+8/血壓 mmhg	www.ntuh.gov.tw
基礎知識	t2	+7/血壓 mmhg	http://www.hopkinsmedicine.org
24h 監控	t1	+6/血壓 mmhg	cwb.gov.tw
24h 監控	t2	+4/血壓 mmhg	weather.gov

雖然 Hbase 的邏輯仕途可以採用與傳統關聯視資料庫類似的資料行表的形式表達，但實際上這些資料在進行實體儲存時是以列族為單位的進行儲存。這種邏輯視圖的對應可以透過圖 4.2.1 來理解

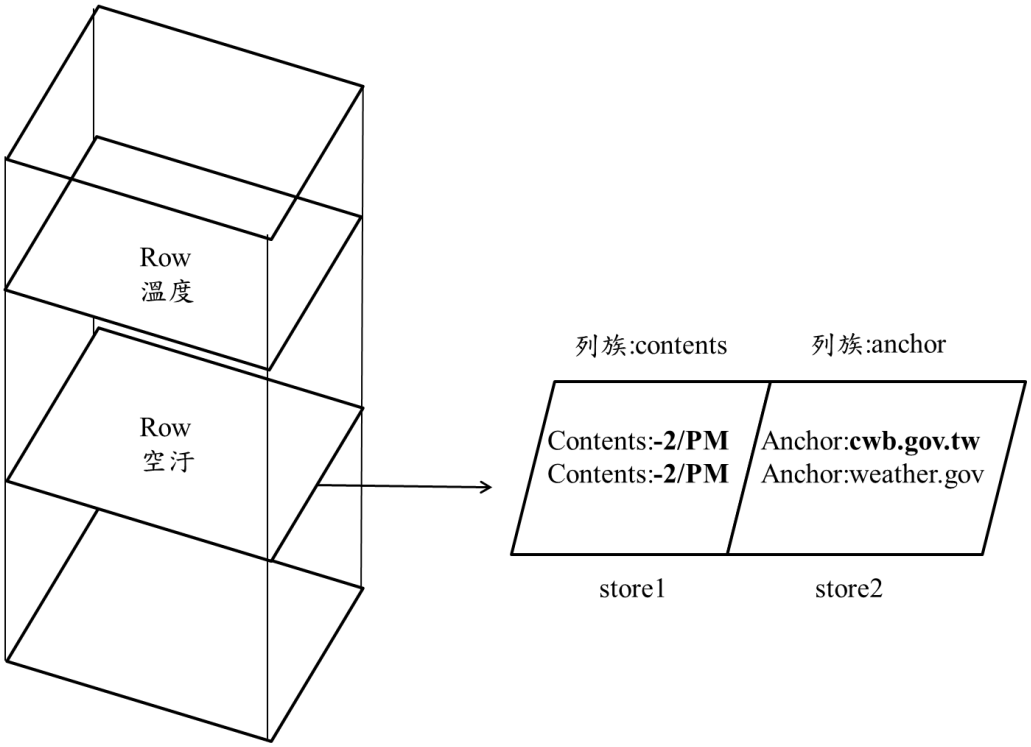

圖 4.2.1 以心血管疾病為例的 Hbase 邏輯視圖到實體試圖的對應

從圖 4.2.1 中可以看出，如果將邏輯視圖中的一行資料看做一個面，則這些面試由若干個 Store（Hbase 的核心儲存部件）構成的，每個 Store 儲存了同屬於一個列族的資料。經過這樣的對應，表 4.2.2 中的空汙行對應的資料就需要轉換為下面兩張實體儲存表。

表 4.2.2 以心血管疾病為例的 Hbase 實體視圖表 1-列族：contents

行關鍵字	**版本**	**列族：contents**
空汙	t1	-2/PM

表 4.2.3 以心血管疾病為例的 Hbase 實體視圖表 2-列族：anchor

行關鍵字	**版本**	**列族：anchor**
空汙	t1	-2/PM

此種做法是基於能將大量現有的分析資料，快速的導入以 HDFS 為基礎的 Hbase 中。

4.3 基於 OLAP 的 HDFS 編碼

為了有效避免 OLAP 查詢中檢索不必要的列數據，本章首先按維提取源數據中每一維對應的度量資訊；為了有效利用 OLAP 查詢中維層次特性提高查詢分析效率，本章構建了基於維層次特性的編碼；針對現有的 OLAP 海量數據處理平臺以行記錄為單位組織數據，不能有效滿足 OLAP 頻繁訪問某些列的讀請求，本章設計了基於 HDFS 的維存儲方案。

本文源數據的形式如表 4.3.1 下所示，包含 TID 列，維層次屬性和度量行。TID 表示該維層次屬性值在原始數據表中出現的位置，件單來說就是收到訊號的順序，感測器接收到的訊號是度量行內的數值，TID 和度量行之間是維層次屬性列。

表 4.3.1 初始 TID 表_1

維層次屬性										
TID	姓名	性別	年齡	BMI	階段一	階段二	階段三	階段四	心律不穩	高血壓
1	小豪	男	57	25	1	0	0	0	0	0
2	小英	男	63	27	0	0	1	0	1	0
3	小美	女	78	26	0	0	1	0	0	1
4	小元	男	74	22	0	1	0	0	1	0
5	小白	女	60	20	0	1	0	0	0	0
6	小麗	女	62	21	1	0	0	0	0	0

表 4.3.2 初始 TID 表_2

維層次屬性											度量行			
TID	城市	區	路	段	號	年	月	時	分	秒	mm Hg (收縮壓)	次 (心律)	L/min (心輸出量)	秒 (QRS 波)
1	台中	西屯	台灣大道	三	568	106	1	0	0	0	/	/	/	/
2	台北	大同	長安西路	0	39	106	1	0	0	0	/	/	/	/
3	台中	西屯	台灣大道	三	568	106	1	0	0	0	/	/	/	/
4	台南	東	大學路	西	89	106	1	0	0	0	/	/	/	/
5	台北	大同	長安西路	0	39	106	1	0	0	0	/	/	/	/
6	台南	東	大學路	西	89	106	1	0	0	0	/	/	/	/

如表 4.3.2 所示，首先按原始數據表中對應的 TID 和度量列的訊息，同時對每一維的成員值基於維層次特性進行分類編碼，然後將各維成員對應的 TID 和度量到 HDFS 上的 HDFile（Hadoop Dimension File）中。

4.3.1 按維分割

針對 OLAP 分析通常以維為單位進行聚集計算的特點，本文以維為單位提取每一維對應的度量資訊。按各維獨立的想法組織行數據，區別於傳統關係數據庫中以行記錄組織數據，避免了數據檢索過程中不必要的行掃描時間。具體的維分割思路如下：

首先，根據用戶的 OLAP 分析模式定義原數據中維的個數，提取建造共六個維度，分別是將姓名獨立成一個的「姓名維」；將性別、年齡、BMI、心衰竭階段、相關併發症集合的「個人資訊維」，此外這個維度還有再分層，在後續各維的解釋會提到；將城市、路、段、號集合的「地區維」；將年、月、時、分集合的「時間維」；將 mm Hg（收縮壓）、次（心律）、L/min（心輸出量）、秒（QRS 波）集合的「感測維」；以及將心血管系統、肌電圖波型、HRV（心律變異）分析法、HRV（心律變異）之頻譜定義集合的「心知識維」。

次之，本研究對維之下的層次做說明。通常每個維底下會包含多個層次，每個層次對應一列，並且維的層次之間有一定的語義關係，比如地區維中，城市的範疇就比鄉來得大；又或者時間維裡面，年的範圍也比月、時、分來得大。基於維數據的這種層次特性，本文把原始數據通過 OLAP 的維概念後重組的數據抽取出來，作為一個存儲單位。

1. 姓名維

表 4.3.3 姓名維

TID	
維	姓名維
層一	姓名
1	小豪
2	小英
3	小美
4	小元
5	小白
6	小麗

2. 個人資訊維

表 4.3.4 個人資訊維

維	個人資訊維								
層一	性別	年齡	BMI	心衰竭階段				相關併發症	
層二				一	二	三	四	心律不穩	高血壓
1	男	57	25	1	0	0	0	0	0
2	男	63	27	0	0	1	0	1	0
3	女	78	26	0	0	1	0	0	1
4	男	74	22	0	1	0	0	1	0
5	女	60	20	0	1	0	0	0	0
6	女	62	21	1	0	0	0	0	0

3. 地區維

表 4.3.5 地區維

維	地區維				
層一	城市	區	路	段	號
1	台中	西屯	台灣大道	三	568
2	台北	大同	長安西路	0	39
3	台中	西屯	台灣大道	三	568
4	台南	東	大學路	西	89
5	台北	大同	長安西路	0	39
6	台南	東	大學路	西	89

4. 時間維

表 4.3.6 時間維

維	時間維				
層一	年	月	時	分	秒
1	106	1	0	0	0
2	106	1	0	0	0
3	106	1	0	0	0
4	106	1	0	0	0
5	106	1	0	0	0
6	106	1	0	0	0

4.3.2 建立編碼

將這些具有維概念的數據儲存單位進行數值淨化，是根據維數據的層次特性建立的編碼，其編碼的規範有以下三項原則，如下所述：

4.3.3 維層次編碼

定義 l：

維層次樹 DTree=(v，E)，其中節點 V 是維中各個層次所有取值的集合；根節點是一個抽象節點，不具有實際含義。邊 E 是各個取值之間的層次關係。若兩個取值具有層次關係，則在 DTree 中，層次較高的值成為層次較低的值的父節點。

地區維層次樹中的節點由（城市、區、路、段、號）的所有取值組成，其維層次樹如圖 4.3.1 所示。延續地區的例子，城市對區、路、段、號具有層次關係，因此在維層次樹中，「城市」成為區、路、段、號的父節點。

編碼的維層次樹：

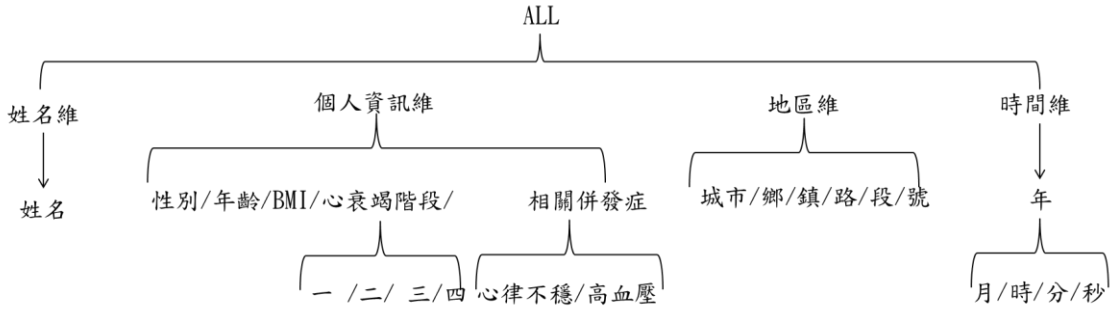

圖 4.3.1 完整的維層次樹

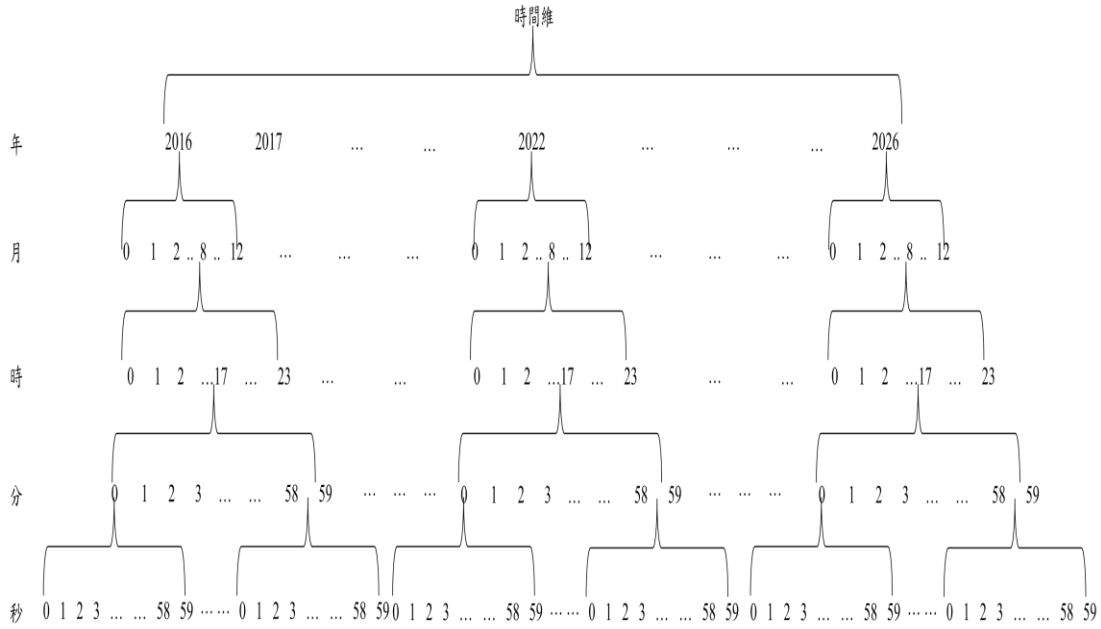

圖 4.3.2 時間維的維層次樹

定義 2：

層次屬性編碼長度說-Code-Len(L)=I-log：m-I。m 是 L 層中不同成員的最大個數。需要說明的是，文中不會使用全部為零的編碼，所以在確定編碼位數的時候要多加一位避免出錯。假設一個維層次屬性包含 25 個不同的屬性值，則基本編碼長度至少設為 5 位。並且為了適應一定程度的數據更新，本文會假設每個維度和層之間的編碼互不幹擾，能獨立進行編碼而不會有辨識上的衝突；下表 4.3.7 以時間維為例：

表 4.3.7 時間維編碼量

	年	月	時	分	秒
所需數量	每年編碼重新定義	固定 12 種編碼	固定 24 種編碼	固定 60 種編碼	固定 60 種編碼
編碼最低所需數量	暫以 10 年討論因此需 4 碼	至少 4 碼	至少 5 碼	至少 6 碼	至少 6 碼

定義 3：

層次屬性編碼 DL_Code(Ln)={((DL_Code(L1)<<Code_Len(L2)| DL_Code(L2))…)<< DL_Code_Len(Ln)|< b_{k-1}…b_i…b_0… > b_i ∈{o ,l}。其中 k 表示維 D 的所有維層次二進制編碼的位數之和，L_n 為維 D 中的第 n 層次屬性，其值域為 dom(L_n)={d_1, …,d_n,…,d_m}。對於分層的維，為維層次屬性的每一個不同屬性值指定一個唯一的編碼（編碼的分配基於屬性值在原數據中的先後次序）。

維層次樹中，根節點無維層次編碼。非根節點的維層次編碼為自根節點至該節點的路徑上所有編碼的串聯，除根節點外，所有節點的值編碼由定義 3 生成。

4.3.4 操作 I／O 的計算

在構建 HDFile 的同時將其存儲於 HDFS 上，基於 HDFS 中對檔的基本操作，可以對 HDFile 進行新建、讀取、寫入、追加、刪除等操作。在選擇 OLAP 聚集計算中涉及的維時，不檢索無關的列數據，可以有效降低聚集操作的 I／O 開銷。一天的數據量為 14GB，相較之下是經濟和可行的辦法。

本研究建議，由於心律週期的完整波型，時間為 0.8 秒，為了使整個資料量盡可能減少，我們會每 2.4 秒做一次數據的讀取（三個波型），之所以會取三是因為若是發生異常現象，最低判定標準為每三次週期裡面出現一次異常（楊麗&張囡囡，2016），若真的出現，再藉由感測器的預先設定，開始變為每 0.8 秒讀取一次的完全監控，並在搭配羅吉斯回歸判定下的危險因素，以及實際狀況，調整是否將判斷時間再次調降為每 0.8 秒一次。

表 4.3.8 範例 TID 表的完整編碼數

項目	性別	年齡	BMI	心衰竭階段	相關併發症
最低編碼數	3	7	6	3	3
項目	城市	區	路	段	號
最低編碼數	2	2	2	2	2
項目	年	月	時	分	秒
最低編碼數	4	4	5	6	6
總編碼數	22+10+25=57				

若以每天的實時紀錄來做計算，一個人每日的紀錄為 86,400 條紀錄，若每條，每條記錄佔用 lK 字節的表查詢為例，若對某維對應的度量值進行 SUM 操作，需要讀取的交易記錄數據量為 86,400*1024=0.082GB（未建索引的情況下）。但這還是一個人的情況下，假如要把這項系統推向以大型醫院時，若曾經造醫院並使用此系統的人數為兩千人，則一天就會產生 164.79 GB 的數據量。讀取並處理如此大的數據量會耗費較長的磁盤 i/O 時間，導致數據處理效率低、時間長。在 HDFile 模式下，僅僅是對該維所對應的 TID 和度量數據進行掃描，同樣以有 86,400 條記錄的一個單表為例，對特定的維進行查詢。假設該維對應的 TID 佔 5 個字節，度量屬性佔有 25 字節，讀取到的數據量只有 86,400×(5+22+10+25+25)=0.007GB。這相對於個人的全表掃描讀取 0.082GB 的數據量減少了約 11.7 倍。在建立大群體的系統時，以上文提到的兩千人為例，一天產生的數據不過 14GB，相較之下是經濟和可行的辦法。

本研究建議，由於心律週期的完整波型，時間為 0.8 秒，為了使整個資料量盡可能減少，我們會每 2.4 秒做一次數據的讀取（三個波型），之所以會取三是因為若是發生異常現象，最低判定標準為每三次週期裡面出現一次異常（楊麗&張囡囡 2016），若真的出現，再藉由感測器的預先設定，開始變為每 0.8 秒讀取一次的完全監控，以及實際狀況，調整是否將判斷時間再次調降為每 0.8 秒一次。

4.4 基於 mapreduce 設計模型

為了開發完整的心衰竭數據應用在 Hadoop 上，Mapreduce 的處理模式也需要建立完整的架構。其完整的架構，除了前一節提及的 OLAP 編碼，在本章還會加入多項資料用於實際病例的描述。

透過下圖 4.4.1 我們可以見到一個 Mapreduce 處理模式，在用於心臟衰竭的大數據分析上應具備的四大流程計數、分類、過濾處理、相關計數。此外還要建立一個造訪資料的流程，以交代初始資料的建置、修改以及運用。本章將以個案的形式描述一個患有心律不規律波型的患者的處理流程。

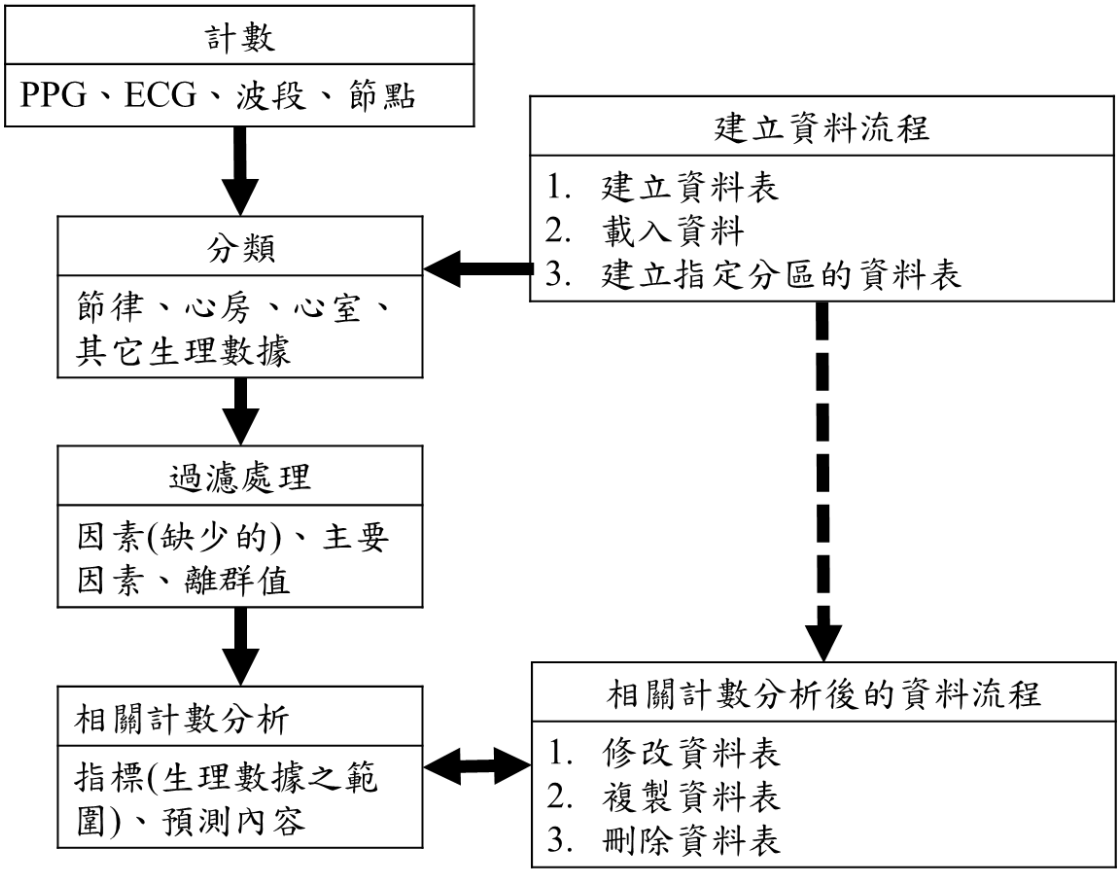

圖 4.4.1 Mapreduce 四大流程與資料流向

4.4.1 計數

在心臟疾病的分類上，我們多會利用波行來做判定，測定波型的工具有 PPG、ECG 兩大類，雖然 PPG 是較能做到即時監控的穿戴裝置，但鑑於本章節的模型是以病患初次接觸本系統的情境下做的描述，因此預做假設病患是來到醫院做的檢查，因而使用 ECG 來做紀錄，如下表 4.4.1。

表 4.4.1 兩種不規律的 ECG 圖形示範

疾病種類:不規律的 ECG
描述 1：心律不整、ST 段下降，倒 T 波

描述 2：ST 段下降，倒 T 波


先建立這位患有心律不整的患者的資料，如下表 4.4.2，這些資料會是病患一來到醫療單位時，就會做的詢問調查，以便建立完整的個人數據庫。

表 4.4.2 案例初診調查資料

變數項目	變數	範圍區間
年齡	中年	40-65 歲
是否有糖尿病	否	0
是否吸菸	是	1
血壓(收縮壓)	稍高	135-159 mm/hg
總膽固醇	稍高	235-279 mg/dl
低密度脂蛋白膽固醇	中	>159 mm/hg
風險值(輸出)	高	45-60%

建立這個個人資料的基礎，是基於的架構中的節點 7：基礎知識事實表。本研究義整理出可置於此節點的資料列於下方圖 4.4.2 中。

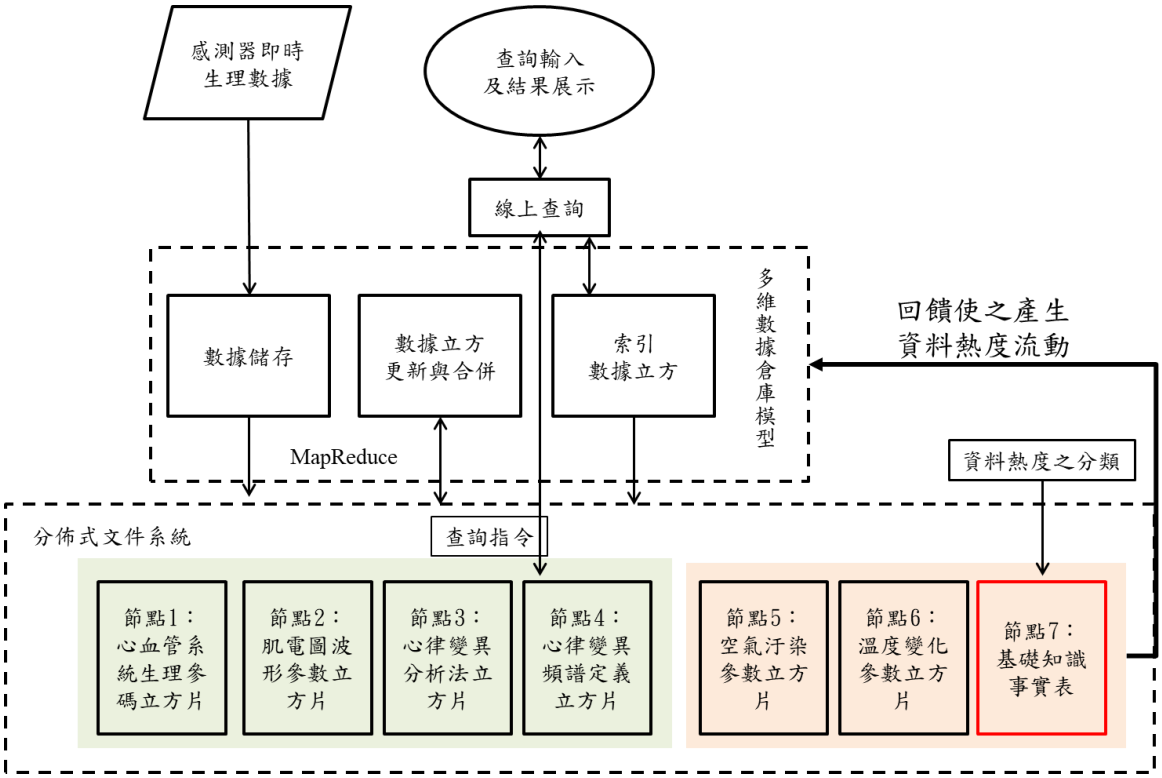

圖 4.4.2 基於 Hadoop 的 OLAP 系統架構_節點 7

4.4.2 分類

在進入第二流程分類前，需先介紹造訪資料的流程；我們必須先將一定的資料匯入 Hadoop 系統，才能做後續的分析。

1. 建立資料表

Hive 建立的表一般儲存在 HDFS 中，但也可以儲存在其他任何 Hadoop 檔案系統中。建立這個資料表，即可建立在原生檔案系統，也可以建立在外部資料表。內部資料表和外部資料表在使用上最大的區別，主要是在載入資料和刪除表這兩類操作上。

像是和表 4.4.4，都是屬於節點 7，這一類的資料僅需要關聯式資料庫即可儲存，因此可以藉由 Hive 內的驅動程式，將這兩筆資料透過中繼資料庫服務，進入到 MySQL 裡面。

表 4.4.3 生理變數與範圍

變數項目	變數	範圍區間
年齡	青年	<30歲
	壯年	25-45 歲
	中年	40-65 歲
	老年	>60歲
是否有糖尿病	是	1
	否	0
是否吸菸	是	1
	否	0
血壓(收縮壓)	低	<129 ㎜/hg
	中	125-139 mm/hg
	稍高	135-159 mm/hg
	高	>155 mm/hg
總膽固醇	中	<239 mg/dl
	稍高	235-279 mg/dl
	高	>275 mg/dl
低密度脂蛋白膽固醇	中	>159 mm/hg
	稍高	155-190 mm/hg
	高	>185 mm/hg
風險值(輸出)	非常低	<20%
	低	15-30%
	稍低	25-40%
	中	35-50%
	高	45-60%
	稍高	55-70%
	非常高	>65%

表 4.4.4 理數據與肌電圖波型

種類	標準區間
生理數據與肌電圖波型	安靜狀態時的收縮壓為 13.3~16.0kPa(100~120mmHg) 舒張壓為 8.0~10.6kPa(60~80mmHg) 脈壓為 4.0~5.3kPa(30~40mmHg) 平均動脈壓為 13.3kPa(93mmHg)左右
	心率(Heart rate, HR):是指每分鐘內心臟搏動的次數。成年人心率為每分鐘 60~100 次
	每搏輸出量(Stroke Volume, SV)：一側心室搏動一次所射出的血量，稱為每搏輸出量，簡稱搏出量。成年人搏出量約為 60~80
	心輸出量(Cardiac Output, CO)：一心是每分鐘射出的響亮稱為每分輸出量，及心輸出量，等於搏出量乘以心率。成年人安靜時的心輸出量為 4.5~6L/min，心輸出量與機體新陳代謝水平相適應，在肌肉活動、情緒激動、懷孕等情況下增加；此外，女子比同體重男子的心輸出量低約 10%
	血流阻力(Resistance Ofblood Flow)：血液在血管內流動時所遇到的阻力，等於血管兩端的壓力差與該段血管的血流量的比值
	血流量(Blood Flow)：是指單位時間內流過血管某一截面的血量，也稱容積速度，其單位為 ml/min 或 L/min
	血容量(Blood Volume):人體內的血液總量稱為些容量，是血漿量和血細胞量的總和，正常成年人的血容量約相當於體重的 7%~8%
	血管順應性(Compliance)：血液壓力改變一個單位時，所對應脈體積的變化量，他等於血管的可擴張度與血管段體積的乘積
	前負荷(Preload)：心肌收縮前所承受的負荷，可用心室舒張期術血液的充盈程度(容積)來表示
	後負荷(Preload)：心肌設血所面對的阻力，心室射血過程中，大動脈血壓起著後負荷的作用
	心肌收縮能力(Cardiac Contractility)：試紙通過新機本身收縮活動的強度和速度的改變而不依賴於前、後負荷的改變來影響每搏輸出量的能力
	非壓力容積(Unstressed Volume)：在一個彈性腔內不引起壓力變化的那部分容積，即該彈性槍壓力為零時的容積
	心率增加 10 次/分,將增加 20% 的全病因死亡和 14%的心血管死亡。有研究表明,對於心力衰竭患者心率減慢 15 次

種類	標準區間
	/分病死率可降低 30% 以上
	入院時心率> 90 次/分的心力衰竭患者死亡率要比心率< 70 次/分心力衰竭患者高 2 ~ 3 倍
	在 HRV 時域指標中,SDNN 和 SDANN 可以反映交感神經及迷走神經活動，PNN50 及 RMSSD 反映迷走神經張力大小，PNN50、RMSSD 降低提示迷走神經損害
	P 波：心房去極化，正常小於 0.12 秒
	QRS 波：心室去極化，正常不超過 0.11 秒
	ST segment： 1. 心臟早期的再極化 2. ST segment 位置(高低)較長短來得重要 3. 正常在±1mm 之間
	PR interval： 1. 評估心房至心室間的傳導速度 2. 正常值約 0.12-0.2 秒(3-5 小格)
	QT interva： 1. 代表整個心縮期的電位變化 2. 與心跳速率有關 3. 臨床上為藥物及離子對心肌影響的一個指標 4. 正常為 0.35~0.43 秒

2. 載入資料

Hive 可以從原生檔案系統或 HDFS 中匯入資料檔案，這個過程稱為載入資料。對於原生和外部表的匯入資料命令都是相同的，差別在於，對於原生資料表來說，載入資料相當於一個移動操作，也就是原本的資料檔案移動到 Hive 管理的檔目錄下。而對於外部資料表，卻是會放進中繼資料庫的虛擬機內，因此在使用外部資料時相對有風險，因為 Hive 不會檢查這個資料內是否有檔案，而是要等到真正進行資料操作時才會告知。

3. 建立指定分區的資料表

區分可以將資料表按照某個列進行切分然後儲存在不同的目錄下，提高對相應列的限制條件下，進行查詢效率的提升。本研究用一種常見的區分方式是將 log 檔中的大量記錄按照日期進行分區儲存，這樣對按照時間遞移而變化的心臟衰竭病況不僅是最有效的方式，更加惠於醫療機構查詢病患紀錄。

4.4.3 過濾處理

在實際的感測器運用上，常會發生雜訊的產生，主要來自於手指與電極間的接觸不良。身體的晃動、環境的干擾、手指過於乾燥或油漬、電極上的汙漬...等，均會造成雜訊。這類的雜訊，我們要分析上分成三種類：模糊因素（缺少的、微弱的）、主要因素、離群值。

根據既有的醫學知識以及個人生理數據的建立，通常主要因素是不會變動的，因此需要去注意的往往是模糊因素以及離群值。經過分類資料後，大部分要分析的項目已經都歸入所屬的檔案中，這時就可以使用 Mapreduce 的過濾功能，透過將波行傳遞出的訊息，包括振幅、波長以及訊號的強度進行過濾。

過濾處理也可以運用在將符合的紀錄取出，這麼做的好處是減少了一個管道需要處理的數據量。在 Hadoop 中減少處理的數據量是至關重要的，尤其當需要通過網絡和原生磁碟進行處理的時候，這個過程會將數據通過網絡寫入到磁碟中，所以擁有更少的數據就意味著 Job 和 MapReduce 框架的工作量也就越少，這樣 Job 的數據傳輸也會更快，CPU、磁碟、網絡設備的壓力也會減少。

以下代碼顯示一個多人的初檢結果，本研究選擇過濾排除 30 歲以下以及 65 歲以上的病患，並且只顯示是否有糖尿病與是否吸菸這兩種狀態，我們用表 4.4.5 來表示原始資料，並用表 4.4.6 來表示經過過濾後的資料。

public static class JoinMap extends Mapper<LongWritable, Text, Text, Text> { @Override protected void map(LongWritable offset, Text value, Context context) throws IOException, InterruptedException { User user = User.fromText(value); if (user.getAge >= 30；<=65；是否有糖尿病；是否吸菸) { context.write (newText (user.getName), new Text (user.getState));}}}

表 4.4.5 多人初檢結果表

	病患數	**1**	**2**	**3**	**4**	**5**	**…**	**629**	**630**
變數項目	範圍區間								
年齡	<30 歲 25-45 歲 40-65 歲 >60 歲	25	43	56	67	21	…	58	37
是否有糖尿病	1 0	1	0	1	0	0	…	1	1
是否吸菸	1 0	0	0	1	0	0	…	1	1
血壓（收縮壓）	<129 mm/hg 125-139 mm/hg 135-159 mm/hg >155 mm/hg	125	150	168	166	135	…	148	137
總膽固醇	<239 mg/dl 235-279 mg/dl >275 mg/dl	240	242	259	270	221	…	263	267
低密度脂蛋白膽固醇	>159 mm/hg 155-190 mm/hg >185 mm/hg	163	159	168	163	195	…	156	174
風險值（輸出）	<20% 15-30% 25-40% 35-50% 45-60%	10%	26%	30%	44%	84%	…	86%	89%

病患數	1	2	3	4	5	…	629	630
55-70%								
>65%								

表 4.4.6 過濾處理之多人初檢結果表

病患編號	3	4	5	…	629	630
年齡	56	67	21	…	58	37
是否有糖尿病	1	0	0	…	1	1
是否吸菸	1	0	0	…	1	1

4.4.4 相關計數分析

相關計數分析是用在指標以及預測內容的顯示。經過病患與醫療單位長時間的合作所累積的資料，在每次的回診或是調閱資料的時候才會運作的功能。其相關的更新處理，會接續第 4 節的資料流程繼續補充。

1. 修改資料表

Hive 中會支援已經建立好的表進行修改，包括修改表名、列名、列的欄位型、增加或替換列。例如一群例行來做健康檢查的病患，能夠讓醫療單位一次性的對新增的數據做加入。這樣做除了能整合數據的匯　入的時間，還能確保過去資料的持續追蹤，是相當具有彈性的設計。

2. 複製資料表

複製資料表的作用是將一個已有資料的表複製到另一個表中。HiveQL 支援多種複製的方法，包括單表、多表和建立表時複製。值得一提的是多表複製，這種做法是為了適應大數據處理環境下更高效的產生下一階段所需的資料表。

3. 刪除資料表

會將表對應的中繼資料和資料檔案一併刪除。

4.4.5 相關計數分析實例

本研究主要是凸顯多表單的同時運用，因此使用了表 4.4.7 和表 4.4.9 與表 4.4.3，做修改、複製與刪除的相關計數分析。本研究先將要彙整的資料以紅色作為表示，並先列出偽代碼，再把經代碼轉換後的表格貼上，與原本的作比較。

表 4.4.7 轉成表 4.4.8 用的偽代碼

複製類：
INSERT OVERWRITE TABLE 初始 TID 表; //複製出始初始 TID 表
修改類：
ALTER TABLE USER RENAME TO new _ TID 表; //重新命名複製的表
ALTER TABLE USER Swap [ROW] TO [COLUMN]; //欄列對調
ALTER TABLE USER Swap [COLUMN] TO [ROW]; //欄列對調
刪除類：
DROP TABLE CLONE FROM [COLUMN2] TO [COLUMN10] ;//刪除欄
DROP TABLE CLONE FROM [ROW6] TO [ROW8] ; //刪除列

表 4.4.9 轉成表 4.4.10 用的偽代碼

複製類：
INSERT OVERWRITE TABLE 生理數據與肌電圖波型; //複製生理數據與肌電圖波型
修改類：
ALTER TABLE USER RENAME TO new _ 生理數據與肌電圖波型; //重新命名複製的表
刪除類：
DROP TABLE CLONE FROM [ROW1] TO [ROW11] AND FROM [ROW14] TO [ROW16] AND [ROW18] ; //刪除列

表 4.4.3 的使用
複製類:
INSERT OVERWRITE TABLE 生理變數與範圍; //複製
修改類:
ALTER TABLE USER RENAME TO new _ 生理變數與範圍; //重新命名複製的表

表 4.4.7 初始 TID 表

	維層次屬性										度量行			
TID	城市	區	路	段	號	年	月	時	分	秒	mm Hg（收縮壓）	次（心律）	L/min（心輸出量）	秒（QRS 波）
1	台中	西屯	台灣大道	三	568	106	1	0	0	0	/	/	/	/
2	台北	大同	長安西路	0	39	106	1	0	0	0	/	/	/	/
3	台中	西屯	台灣大道	三	568	106	1	0	0	0	/	/	/	/
4	台南	東	大學路	西	89	106	1	0	0	0	/	/	/	/
5	台北	大同	長安西路	0	39	106	1	0	0	0	/	/	/	/
6	台南	東	大學路	西	89	106	1	0	0	0	/	/	/	/

表 4.4.8 新的 TID 表

	度量行			
TID	mm Hg（收縮壓）	次（心律）	L/min（心輸出量）	秒（QRS 波）
1	/	/	/	/
2	/	/	/	/
3	/	/	/	/

表 4.4.9 節點 7 基礎知識_生理數據與肌電圖波型

種類		標準區間
生理數據與肌電圖波型	1	安靜狀態時的收縮壓為 13.3~16.0kPa(100~120mmHg) 舒張壓為 8.0~10.6kPa(60~80mmHg) 脈壓為 4.0~5.3kPa(30~40mmHg) 平均動脈壓為 13.3kPa(93mmHg)左右
	2	心率（Heart Rate, HR）：是指每分鐘內心臟搏動的次數。成年人心率為每分鐘 60~100 次
	3	每搏輸出量（Stroke Volume, SV）：一側心室搏動一次所射出的血量，稱為每搏輸出量，簡稱搏出量。成年人搏出量約為 60~80
	4	心輸出量（Cardiac Output, CO）：一心是每分鐘射出的響亮稱為每分輸出量，及心輸出量，等於搏出量乘以心率。成年人安靜時的心輸出量為 4.5~6L/min，心輸出量與機體新陳代謝水平相適應，在肌肉活動、情緒激動、懷孕等情況下增加；此外，女子比同體重男子的心輸出量低約 10%
	5	血流阻力（Resistance Ofblood Flow）：血液在血管內流動時所遇到的阻力，等於血管兩端的壓力差與該段血管的血流量的比值
	6	血流量（Blood Flow）：是指單位時間內流過血管某一截面的血量，也稱容積速度，其單位為 ml/min 或 L/min
	7	血容量（Blood Volume）：人體內的血液總量稱為些容量，是血漿量和血細胞量的總和，正常成年人的血容量約相當於體重的 7%~8%
	8	血管順應性（Compliance）：血液壓力改變一個單位時，所對應脈體積的變化量，他等於血管的可擴張度與血管段體積的乘積
	9	前負荷（Preload）：心肌收縮前所承受的負荷，可用心室舒張期術血液的充盈程度（容積）來表示
	10	後負荷（Preload）：心肌設血所面對的阻力，心室射血過程中，大動脈血壓起著後負荷的作用
	11	心肌收縮能力（Cardiac Contractility）：試紙通過新機本身收縮活動的強度和速度的改變而不依賴於前、後負荷的改變來影響每搏輸出量的能力
	12	非壓力容積（Unstressed Volume）：在一個彈性腔內不引起壓力變化的那部分容積，即該彈性槍壓力為零時的容積

種類		標準區間
	13	心率增加 10 次/分,將增加 20%的全病因死亡和 14% 的心血管死亡。有研究表明,對於心力衰竭患者心率減慢 15 次/分病死率可降低 30% 以上
	14	入院時心率> 90 次/ 分的心力衰竭患者死亡率要比心率< 70 次/ 分心力衰竭患者高 2～3 倍
	15	在 Hrv 時域指標中，SDNN 和 SDANN 可以反映交感神經及迷走神經活動，Pnn50 及 RMSSD 反映迷走神經張力大小，Pnn50、RMSSD 降低提示迷走神經損害
	16	P 波：心房去極化，正常小於 0.12 秒
	17	Qrs 波：心室去極化，正常不超過 0.11 秒
	18	ST Segment： 1. 心臟早期的再極化 2. ST Segment 位置（高低）：較長短來得重要 3. 正常在±1mm 之間
	19.	PR Interval： 1. 評估心房至心室間的傳導速度 2. 正常值約 0.12-0.2 秒（3-5 小格）
	20.	QT Interval： 1. 代表整個心縮期的電位變化 2. 與心跳速率有關 3. 臨床上為藥物及離子對心肌影響的一個指標 4. 正常為 0.35~0.43 秒

表 4.4.10 新生理數據與肌電圖波型

種類	標準區間
13	心率增加 10 次/分，將增加 20%的全病因死亡和 14%的心血管死亡。有研究表明，對於心力衰竭患者心率減慢 15 次/分病死率可降低 30%以上
14	入院時心率> 90 次/分的心力衰竭患者死亡率要比心率< 70 次/分心力衰竭患者高 2～3 倍
18	ST segment： 1. 心臟早期的再極化 2. ST segment 位置（高低）較長短來得重要 3. 正常在±1mm 之間
20	QT interval： 1. 代表整個心縮期的電位變化 2. 與心跳速率有關 3. 臨床上為藥物及離子對心肌影響的一個指標 4. 正常為 0.35~0.43 秒

表 4.4.3 生理數據範圍

變數項目	變數	範圍區間	變數項目	變數	範圍區間
年齡	青年 壯年 中年 老年	<30歲 25-45 歲 40-65 歲 >60歲	總膽固醇	中 稍高 高	<239 mg/dl 235-279 mg/dl >275 mg/dl
是否有糖尿病	是 否	1 0	低密度脂蛋白膽固醇	中 稍高 高	>159 mg/hg 155-190 mg/hg >185 mg/hg
是否吸菸	是 否	1 0	風險值(輸出)	非常低 低 稍低 中 高 稍高 非常高	<20% 15-30% 25-40% 35-50% 45-60% 55-70% >65%
血壓(收縮壓)	低 中 稍高 高	<129 mg/hg 125-139 mg/hg 135-159 mg/hg >155 mg/hg			

最後將三張新的表單做合併產出新的表單

表 4.4.11 初診與後續追蹤表

TID	1	2	3
mm Hg(收縮壓)	/	/	/
次(心律)	/	/	/
L/min(心輸出量)	/	/	/
秒(QRS 波)	/	/	/
年齡	25	43	56
是否有糖尿病	1	0	1
是否吸菸	0	0	1
血壓(收縮壓)	125	150	168
總膽固醇	240	242	259
低密度脂蛋白膽固醇	163	159	168
風險值(輸出)	10%	26%	30%
備註	備 3、備 4	備 3、備 4	備 3、備 4

備註內容	
備 1	心率增加 10 次/分,將增加 20% 的全病因死亡和 14% 的心血管死亡。有研究表明,對於心力衰竭患者心率減慢 15 次/分病死率可降低 30% 以上
備 2	入院時心率>90 次/分的心力衰竭患者死亡率要比心率<70 次/分心力衰竭患者高 2-3 倍
備 3	ST segment： 1. 心臟早期的再極化 2. ST segment 位置(高低)較長短來得重要 3. 正常在±1mm 之間
備 4	QT interval： 1. 代表整個心縮期的電位變化 2. 與心跳速率有關 3. 臨床上為藥物及離子對心肌影響的一個指標 4. 正常為 0.35~0.43 秒

4.4.6 相關計數分析在 MapReduce 運行框架中的流程

1. JobClient 是基於 MapReduce 介面的用戶端程式，負責解決 MapReduce 作業，此時有一項作業是建立多人的初診與後續追蹤表。
2. JobClient 發出 getNewJobID( ) 的請求給 JobTracker，使之獲得新的作業 ID，其名稱即為_初診與後續追蹤表。
3. JobClient 根據此新 ID 將檔案做劃分並將 JAR 檔案（壓縮檔），資料區塊（即時、過去）；當有越多的穿戴裝備或是越長的時間，以及接近危險範圍的病患時，這些檔案就會越多。分別放入已被劃分的檔案內，此時就有多個 JobTracker，這裡有三個 JobTracker，分別處理：初始 TID 表、生理數據與肌電圖波型、生理數據範圍。
4. 多個 TaskTracker 將被 JobTracker 以心跳（Heart 機制）傳送資訊，並送給更多空閑的 TaskTracker 節點共同協作這三項運算。
5. 被分配任務的 TaskTracker 從 HDFS 中的 DataNode 的取出檔案，存入硬碟，並在 TaskRunner 程式準備運算，而真正的運算在虛擬機中完成的運算 Map & Reduce 的運算，這能夠避免任務運行異常時，影響到 TaskTracker 之間的溝通或運算。
6. 這些被使用的 TaskTracker 會記錄使用次數，也就是根據這些節點間的訊號紀錄，作為往後資料熱度的分類。至於訊號較弱或是在 DataNode 中深層的資料，也能透過相關指令去產生報表提供醫療單位做研究與提供意見的根據。

第五章　結論

過去前人的研究，都是希望不會產生心衰竭的越期的現象，甚至直接進入死亡階段，這部分本研究仍承襲過去學者們的論點，要以更先進的感測技術，實時的監控患者。此外，有賴於行動裝置的普及，對已經有前在發病的病患和已經證實患病的患者，都能及早知道危險環境，諸如：氣溫變化、空氣汙染...等，以利做迴避或預先的預防。

也因為即時的監控，院方能追蹤吃藥後對病情的穩定是否有正向的幫助，並能在用藥出現異常的當下，藉由 Hadoop 的雲端計算的監控，做到即時回報，告知病人回診。

另外，早期的預先評估方法，諸如 Score 評分方式、Framingham... 等十年發病預測，都是在已經發生疾病後就不再具有功效，然而我所建立的系統觀（硬體、傳輸、預測），卻不受這項限制，例如：持續追蹤惡化情況、再發病的可能性追蹤，都還會持續有用。這都是因為從過去的大樣本蒐集資料，轉成以個別資料為主的時代來臨，依據資料探勘的分類，心臟衰竭適用在時序演變分析，再利用感測器接收到的外部資料，更能將關聯分析與離群點分析，這兩個過去難以用於分析的探勘技術加入近來，形成更完整的資料蒐集和分析。

接著從數據的容納來看，本研究透過基於 OLAP 的編碼，能使以每日兩千人次數據來往的醫院，將數據量的大小縮減為一天僅有 14GB，與傳統的紀錄方式相比少了約 11.7 倍。並提出將感測資料從每 0.8 秒，記錄一次，增長為 2.4 秒做一次數據的讀取，這也將能再降低數據庫和維護人員的負荷。

透過案例描述相關計數分析，羅列出資料流程與偽代碼，並使用文字將抽象的 MapReduce 運行框架中的流程描述出來。

最後，透過此種 Hadoop 技術的成長，還能使資料具有流動性，打破過去資料湖泊（俗稱 Cold 資料）資料庫（俗稱 Warm 資料）似乎只是浪費金錢而沒意義的數據，利用感測器激發所有數據的可能，最終由 Hadoop 進行平行運算，讓困擾所有人的「暗數據」比例得以

下降，提高每個裝置和節點的效益，將有助於人類用更低的成本，維護生理狀態。

參考文獻

中文部分

[1] 姚志洪（2010）。健康資訊化從互聯網走向物聯網。**中國衛生資訊管理雜誌**，**4**，18-21。

[2] 姚志洪（2013）。醫療衛生資訊化 10 大視點。**醫學資訊學雜誌**，**1**，2-9。

[3] 張旭峰、姚志洪。(2011)。基於物聯網技術的慢病管理系統。**醫谷雜誌**，**17(8)**，28-34。

[4] 楊麗、張囡囡（2016）。心率減速率和連續心率減速率在心肌梗死後猝死患者中的變化及其意義。**中國循證心血管醫學雜誌**，**8(8)**，971-973。

[5] 簡聖哲、馬漢光、邱瑞科（2012）。以雲端運算為基礎建立醫療輔助決策系統-以冠狀動脈心臟病為例。**第二十三屆國際資訊管理學術研討會 (ICIM)**。

英文部分

[1] Anisimov, V. N., Birnbaum, L., Butenko, G., Cooper, R., Fabris, N. (1993). Principles for evaluating chemical effects on the aged population. *Environmental Health Criteria*, *144*.

[2] Assmann, G., Cullen, P., Schulte, H. (2002). Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the Prospective Cardiovascular Munster (PROCAM) study. *Circulation*, *105*(3), 310-315.

[3] Axisa, F., Schmitt, P. M., Gehin, C., Delhomme, G., McAdams, E., Dittmar, A. (2005). Flexible technologies and smart clothing for citizen medicine, home healthcare, and disease prevention. *IEEE Transactions on Information Technology in Biomedicine*, *9*(3), 325-336.

[4] Baum, R. I., Hsiao, D. K. (1976). Database Computers - A Step Towards Data Utilities. *IEEE Trans. Computers*, *25*(12), 1254-1259.

[5] Beck, J. C. (2002). *Geriatrics review syllabus: a core curriculum in geriatric medicine*. John Wiley & Sons.

[6] Cassel, C. K. (2003). *Geriatric medicine: an evidence-based approach*. Springer Science & Business Media.

[7] Chen, W., Kobayashi, T., Ichikawa, S., Takeuchi, Y., Togawa, T. (2000). Continuous estimation of systolic blood pressure using the pulse arrival time and intermittent calibration. *Medical and Biological Engineering and Computing*, *38*(5), 569-574.

[8] Chi, Y. M., Jung, T. P., Cauwenberghs, G. (2010). Dry-contact and noncontact biopotential electrodes: methodological review. *IEEE Reviews in Biomedical Engineering*, *3*, 106-119.

[9] Codd, E. F. (1983). A relational model of data for large shared data banks. *Communications of the ACM, 26*(1), 64-69.

[10] Conroy, R. M., Pyorala, K., Fitzgerald, A. P., Sans, S., Menotti, A., Bacquer, D. De, Backer, G. De, Ducimetiere, P., Jousilahti, P., Keil, U., Njolstad, I., Oganov, R. G., Thomsen, T., Tunstall-Pedoe, H., Tverdal, A., Wedel, H., Whincup, P., Wilhelmsen, L., Graham, I.M. (2003). Estimation of ten-year risk of fatal cardiovascular disease in Europe: the SCORE project. *European Heart Journal*, *24*(11), 987-1003.

[11] D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care - The Framingham Heart Study. *Circulation*, *117*(6), 743-753.

[12] Duun, S. B., Haahr, R. G., Birkelund, K., Thomsen, E. V. (2010). A Ring-Shaped Photodiode Designed for Use in a Reflectance Pulse Oximetry Sensor in Wireless Health Monitoring Applications. *IEEE Sensors Journal*, *10*(2), 261-268.

[13] Foo, J. Y. A., Lim, C. S., Wang, P. (2006). Evaluation of blood pressure changes using vascular transit time. *Physiological Measurement*, *27*(8), 685-694.

[14] Gesche, H., Grosskurth, D., Kuchler, G., Patzak, A. (2012). Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method. *European Journal of Applied Physiology*, *112*(1), 309-315.

[15] Gu, W. B., Poon, C. C. Y., Leung, H. K., Sy, M. Y., Wong, M. Y., Zhang, Y. T. (2009, September). A novel method for the contactless and continuous measurement of arterial blood pressure on a sleeping bed. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society IEEE Engineering in Medicine and Biology Society Conference*. The Hilton Minneapolis Minneapolis, MN, USA

[16] Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., Brindle, P. (2007). Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *British Medical Journal*, *335*(7611), 136-141.

[17] Humphreys, K., Ward, T., Markham, C. (2007). Noncontact simultaneous dual wavelength photo plethysmography: A further step toward noncontact pulse oximetry. *Review of Scientific Instruments*, *78*(4).

[18] Hunter, P. J., Borg, T. K. (2003). Integration from proteins to organs: the Physiome Project. *Nature Reviews Molecular Cell Biology*, *4*(3), 237-243.

[19] Ishijima, M. (1993). Monitoring of Electrocardiograms in Bed without Utilizing Body-Surface Electrodes. *Transactions on Biomedical Engineering*, *40*(6), 593-594.

[20] Kim, D. H., Lu, N. S., Ma, R., Kim, Y. S., Kim, R. H., Wang, S. D., Wu, J., Won, S. M., Tao, H., Islam, A., Yu, K. J., Kim, T. I., Chowdhury, R., Ying, M., Xu, L. Z., Li, M., Chung, H. J., Keum, H., McCormick, M., Liu, P., Zhang, Y. W., Omenetto, F. G., Huang, Y. G., Coleman, T., Rogers, J. A. (2011). Epidermal Electronics. *Science*, *333*(6044), 838-843.

[21] Lim, Y. G., Kim, K. K., Park, K. S. ECG measurement on a chair without conductive contact. *IEEE Transactions on Biomedical Engineering*, *53*(5), 956-959.

[22] Lloyd-Jones, D. M. (2010). Cardiovascular Risk Prediction Basic Concepts, Current Status, and Future Directions. *Circulation*, *121*(15), 1768-1777.

[23] McCombie, D. B., Reisner, A. T., Asada, H. H. (2006, September). Adaptive blood pressure estimation from wearable PPG sensors using peripheral artery pulse wave velocity measurements and multi-channel blind identification of local arterial dynamics. *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. New York, USA.

[24] Paradiso, R., Belloc, C., Loriga, G., Taccini, N. (2005). Wearable healthcare systems, new frontiers of e-textile. *Studies in Health Technology and Informatics*, *117*, 9-16.

[25] Park, S.B., Noh, Y. S., Park, S. J., Yoon, H. R. (2008). An improved algorithm for respiration signal extraction from electrocardiogram measured by conductive textile electrodes using instantaneous frequency estimation. *Medical & Biological Engineering & Computing*, *46*(2), 147-158.

[26] Pickering, T. G., Miller, N. H., Ogedegbe, G., Krakoff, L. R., Artinian, N. T., Goff, D. (2008). Call to Action on Use and Reimbursement for Home Blood Pressure Monitoring: Executive Summary a Joint Scientific Statement From the American Heart Association, American Society of Hypertension, and Preventive Cardiovascular Nurses Association. *Hypertension*, *52*(1), 1-9.

[27] Poh, M. Z., McDuff, D. J., Picard, R. W. (2011). Advancements in Noncontact, Multiparameter Physiological Measurements Using a Webcam. *IEEE Transactions on Biomedical Engineering*, *58*(1), 7-11.

[28] Poh, M. Z., Swenson, N. C., Picard, R. W. (2010). Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photo plethysmography. *IEEE Trans Information Technology Biomed*, *14*(3), 786-794.

[29] Polo, J., Carrera, D., Becerra, Y., Steinder, M., & Whalley, I. (2010, April). Performance-driven task co-scheduling for mapreduce environments. In Network

Operations and Management Symposium (NOMS), 2010 *IEEE* (pp. 373-380). IEEE.

[30] Poon, C. C. Y., Zhang, Y. T. (2005, September). Cuff-less and noninvasive measurements of arterial blood pressure by pulse transit time. *27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Shanghai, China,

[31] Ridker, P. M., Buring, J. E., Rifai, N., Cook, N. R. (2007). Development and validation of improved algorithms for the assessment of global cardiovascular risk in women - The Reynolds Risk Score. *Journal of the American Medical Association*, *297*(6), 611-619.

[32] Rothmaier, M., Selm, B., Spichtig, S., Haensse, D., Wolf, M. (2008). Photonic textiles for pulse oximetry. *Optics Express*, *16*(17), 12973-12986.

[33] Slotnick, D. L. (1970). Logic per track devices. *Advances in Computers*, *10*, 291-296.

[34] Tietz, N., W., Shuey, D. F., Wekstein, D. R. (1997). Laboratory values in fit aging individuals-sexagenarians through centenarians. *Clinical Chemistry*, *38*(6), 1167-1185.

[35] Timiras, P. S. (2007). *Physiological Basis of Aging and Geriatrics*. CRC Press.

[36] Virtual Physiological Human Network of Excellence [Online]. Available: http://www.vphnoe.eu/.

[37] Wellcome Trust Heart Physiome Project [Online]. Available: http://heart.physiomeproject.org/index.html.

[38] Wu, Y. F., Liu, X. Q., Li, X., Li, Y., Zhao, L. C., Chen, Z., Li, Y. H., Rao, X. X., Zhou, B. F., Detrano, R., Liu, K. (2006). Estimation of 10-year risk of fatal and nonfatal ischemic cardiovascular diseases in Chinese adults. *Circulation*, *114*(21), 2217-2225.