MODEL FOR BIG DATA WAREHOUSING AND ANALYSIS IN HEALTHCARE

BY

SHAIKH SOHEL RANA

A project paper submitted in fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

Kulliyyah of Engineering
International Islamic University Malaysia

JANUARY 2022

ABSTRACT

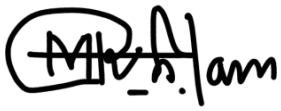
Health care providers, researcher, scientists and analysts are facing huge identified and unidentified problems with massive data those are produced from various heterogeneous systems. For a modern healthcare industry, massive amount of electronic data are produced everyday from various various healthcare systems in the form of structured and unstructured data which includes health records of patients, hospital or clinics records, records from laboratory tests and diagnosis result of the patients, relevant pharmaceuticals as well as many digital and analog signals, image processing data format, signal processed data from medical equipment. Research on bioinformatics produces bulk amount of data related public healthcare. But the problem is, these data are not integrated, united, well managed and well structured. Most of these data are from heterogeneous sources with different formats: some are structured, some are semi-structured and even some are totally unstructured. These data require integration storage, uniform structure, good management and analysis in order to achieve meaningful development of the healthcare industry. The government, national healthcare authorities, researchers, scientists, healthcare providers or patients cannot use their own health data which are stored in various data storage systems until those data are united and integrated from disparate sources to a central location. There should be a storage place where all unstructured data will be turned into a unified, meaningful and structured information and which data can be retrieved later on for various stakeholders for various purposes. An effective model of data warehousing can solve this problem. Data warehousing model performs data extraction, data cleansing, data transformation at first stage and then it ensures data staging design for data assessment and loading. Finally it performs architectural modelling, deployment and implementation. After deploying, improved data can be analysed and visualized in different dimensions. Therefore Scattered raw data or metadata from heterogeneous sources are stored centrally, then process for readiness and become available for real-time usage. From the one central DW, all analysts can retrieve the same data for their own analytical purpose. This is a nice platform where analysts, scientists, healthcare providers, controlling authorities, governments are working with same unique dataset with unidentified dimensions of data mining scopes. For improving public health and future research, multidimensional big data analysis is required. In order to achieve this, healthcare data and bioinformatics can be combined to a uniform system. Therefore, in this thesis, an effective model is proposed for health data warehousing and also for effective mining and analysing techniques for big data.

خلاصة البحث

يواجه مقدمو الرعاية الصحية والباحثون والعلماء والمحللون مشاكل ضخمة محددة وغير محددة مع البيانات الضخمة التي يتم إنتاجها من أنظمة غير متجانسة مختلفة. بالنسبة لصناعة الرعاية الصحية الحديثة ، يتم إنتاج كمية هائلة من البيانات الإلكترونية كل يوم من مختلف أنظمة الرعاية الصحية في شكل بيانات منظمة وغير منظمة والتي تشمل السجلات الصحية للمرضى أو سجلات المستشفيات أو العيادات ، وسجلات الاختبارات المعملية ونتائج التشخيص للمرضى ، الأدوية ذات الصلة بالإضافة إلى العديد من الإشارات الرقمية والتناظرية وتنسيق بيانات معالجة الصور والبيانات المعالجة بالإشارة من المعدات الطبية. البحث في المعلوماتية الحيوية ينتج كمية كبيرة من البيانات المتعلقة بالرعاية الصحية العامة. لكن المشكلة هي أن هذه البيانات ليست متكاملة وموحدة ومدارة بشكل جيد ومنظمة بشكل جيد. تأتي معظم هذه البيانات من مصادر غير متجانسة ذات تنسيقات مختلفة: بعضها منظم وبعضها شبه منظم وحتى البعض الآخر غير منظم تمامًا. تتطلب هذه البيانات تخزينًا تكامليًا وبنية موحدة وإدارة جيدة وتحليلًا من أجل تحقيق تنمية ذات مغزى لصناعة الرعاية الصحية. لا يمكن للحكومة أو سلطات الرعاية الصحية الوطنية أو الباحثين أو العلماء أو مقدمي الرعاية الصحية أو المرضى استخدام بياناتهم الصحية المخزنة في أنظمة تخزين البيانات المختلفة حتى يتم توحيد هذه البيانات ودمجها من مصادر متباينة إلى موقع مركزي. يجب أن يكون هناك مكان تخزين حيث سيتم تحويل جميع البيانات غير المهيكلة إلى معلومات موحدة وذات مغزى ومنظم وأي البيانات يمكن استردادها لاحقًا لمختلف أصحاب المصلحة لأغراض مختلفة. يمكن لنموذج فعال لتخزين البيانات أن يحل هذه المشكلة. يضمن نموذج تخزين البيانات بعض أعباء العمل الشائعة مثل: () تحليل المتطلبات () استخراج البيانات ، وتنقية البيانات ، وتحويل البيانات ، () تنظيم البيانات ، وتصميم تقييم البيانات وتحميلها ، () النمذجة المعمارية ، () النشر والتنفيذ. يتضمن نموذج تحليل البيانات استخراج البيانات والنمذجة والتحول والتصور. سيتم تخزين البيانات الأولية المتفرقة أو البيانات الوصفية من مصادر غير متجانسة مركزيًا ، ثم تتم معالجتها من أجل الاستعداد وتصبح متاحة للاستخدام في الوقت الفعلي. من مستودع البيانات المركزي الواحد ، يمكن لجميع المحللين استرداد نفس البيانات لأغراضهم التحليلية الخاصة. إنها ظاهرة رائعة أن المحللين والعلماء ومقدمي الرعاية الصحية والسلطات الرقابية والحكومات يعملون بنفس مجموعة البيانات الفريدة ذات الأبعاد غير المحددة لنطاقات التنقيب عن البيانات. لتحسين الصحة العامة والبحوث المستقبلية ، يلزم تحليل البيانات الضخمة متعدد الأبعاد. لتحقيق ذلك ، يمكن دمج بيانات الرعاية الصحية والمعلوماتية الحيوية في نظام موحد. لذلك ، في هذه الورقة ، تم اقتراح نموذج فعال لتخزين البيانات الصحية وأيضًا اقتراح تقنيات التعدين والتحليل الفعالة للبيانات الضخمة.

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a project paper for the degree of Master of Science (Computer and Information Engineering).



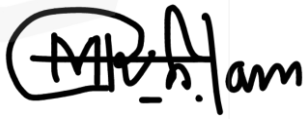
……
Md Rafiqul Islam
Supervisor

……
Mohamed Hadi Habaebi
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a project paper for the degree of Master of Science (Computer and Information Engineering).

……..…
Farah Diyana Bt. Abdul Rahman
Internal Examiner

This project paper was submitted to the Department of Electrical and Computer Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)



……
Md Rafiqul Islam
Head, Department of Electrical
and Computer Engineering

This project paper was submitted to the Kulliyyah of Engineering and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computer and Information Engineering)

……..
Sany Izan Ihsan
Dean, Kulliyyah of Engineering

DECLARATION

I hereby declare that this project paper is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Shaikh Sohel Rana

Signature ..... Date .....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

MODEL FOR BIG DATA WAREHOUSING AND ANALYSIS IN HEALTHCARE

I declare that the copyright holders of this project paper are jointly owned by the student and IIUM.

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Shaikh Sohel Rana

……..…….. ……..

Signature Date

ACKNOWLEDGMENTS

All the research work done in this study is done in the Research Laboratory under the supervision of Prof. Dr. Md Rafiqul Islam and my all respected teachers at the International Islamic University Malaysia. I would like to thank him for guiding me throughout my studies.

I would like to thank my family for being there as a source of moral and emotional support during my journey. I also appreciate my parents for enabling me to study for a master's degree and always urging me to never give up.

Finally, I am thankful to my all friends here at IIUM and outside who have supported me through my struggles and have been there every time I have been worried about academic or personal issues.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$g^{/dl}$	Gram per decilitre
^	And set element of Factor
Z_C	Z Score for normalized data
t_p	Test parameter of single patient
P_I	Total patient in class/group i
R_i	Number of report in class i
l_i	Number of laboratories in class i
Z_0	Coefficient zero
≤	Less than or equal to
≥	Greater than or equal to
∑	N-Ary summation

LIST OF ABBREVIATIONS

AWS	Amazon Web Services
CDW	Central Data warehouse
CSV	Comma-separated values
DW	Data warehouse
DWH	Data warehouse for health
HDC	Health data cube
ISO	International Organization for Standardization
KDD	Knowledge Discovery from Data
MHC	Malaysia Healthcare
MoH	Ministry of Health, Malaysia
MoS	Ministry of Statistics Malaysia
OLAP	Online analytical processing
RDBMS	Relational Database Management Systems
SDW	Standard Data Warehouse
SISH	Sub Information System for health
SQL	Structured Query Language

CHAPTER ONE

INTRODUCTION

1.1 RESEARCH BACKGROUND

A Data Warehouse (DW) is a storage data and information of operational systems for a corporate organization as well as different data resources. Data Warehouse is designed as supportive role for decision making procedure via data collection from different sources, processing, analysis, and through research. DW has become one of the main parts of corporate business intelligence. The architecture of DW was first developed by IBM in 1980's to transform data from traditional operational system to decision making systems. From a design view, DW is generally a part of mainframe server for a business domain.

In DW, data come from various sources into a single storage. After that data are transformed into a valid format via a definite processing process for future storage and utilization.

In 1950, punch card was the main tools to handle computer generated data. In 1960 magnetic storage replaced punch card and become popular until mid of 1980. Hard disk and floppy disks were introduced, developed and enhanced by IBM from 1964 to mid 1980. These storage tools were widely used by personal as well as business domain. But after developing RDBMS and SQL, the whole scenario were changed and most of the giant business domains moved to RDBMS and SQL for storage, analysis and decision making purposes of their businesses.

But after 1990, huge change occurred in global trade and technology simultaneously. That was the revolutions for both business domains and information technology. Internet was increasingly become popular, free trade agreement globally taken business to a new height.

Networking of businesses, computerization and globalization made the demand of Data Warehouse. Client-server technology, improvement of online data bandwidth and it's capacity made possible for DW for receiving data from heterogeneous sources and processing those data to a desired format for applying business intelligence. Improvement and enhancement of DW technology was continuous during 1990 onward. For handing their massive data, In 2008, Facebook started to use NoSQL for Big Data in their data warehousing.

For today's modern civilization, the most valuable treasure is data or information. The domain who has got the most enriched data or information, it become one of the most powerful controller over technology. Google is now one of the most powerful game changer of this civilization. Without a day without Google, it is hard to think. Google has developed the widest and powerful data warehouse which belongs almost over 90% of world's data searching engine: direct or indirect. When we need any information we go to Google for help. From its DW server, Google provides us. Google DW is so powerful and structured that it can provide data or information within a nanosecond. Navigation, medical surgery, vehicle tracking, traffic management, astronomical operation, quantum analysis- in all sector we are taking help from Google data warehouse and management system. Microsoft Azure and Amazon (AWS) also are the two great examples of DW and analytical models of big data. They are successful because they have good DW and management capabilities.

In terms of tasks or job related definition we can say that, Data warehousing is the set of process for extracting, transforming, cleansing, organizing and storing data in such a way that data becomes more enriched, efficient and easy for retrieval, sighting, reporting, visualization and various types of analysis. This process of sets of tasks are referred to transformation of data into valuable information.

(DW): Data Warehouse specified for Medical data is similar with Information of enriched health related Data Warehouse. According to ISO: the Healthcare Data warehouse is different access points to a uniform data storage and management system from heterogeneous sources of a full system or sub-system which enables data analysis on the next secondary level. These transformed and loaded data should be clearly understandable for the health system and thus these data should be supportive tools for improvement of the existed health system and required maintenance.

Grouping of data accessible by a single data management system, possibly of diverse sources, pertaining to a health system or sub-system and enabling secondary data analysis for questions relevant to understanding the functioning of that health system, and hence supporting proper maintenance and improvement of that health system.

The 'Big data' is a significant relevant term in healthcare which handles huge volume of data collected and extracted via information technology. Big data for health in particular includes medical records of the patient, relevant test records from laboratories, research data and other medical information. It is a term used to describe massive volumes of information created by the adoption of digital technologies that collect patients' records, diagnosis reports, bio-information, pharmaceuticals data, medical research dataset and it also support for improving performance of healthcare. Basically big data refers to huge and bulk data with complex architecture (Sonia Ordoñez Salinas and Alba Consuelo Nieto Lemus, 2017).

Health or healthcare information science can be considered as the mid CenterPoint and intersection point between Computer Information Engineering and modern healthcare science. Both are concentrating on procedures and data resources which is needed for extraction, storage and analysis health related data for healthcare reporting, various analysis, healthcare services, bio-informatics research, scientific research, statistical surveys, machine

learning analysis, pharmaceuticals data, nursing care, patient and drugs management and public health-oriented issues (Roddick JF, Fule P, Graco WJ, 2003) and (Cios K, 2002). KDD can be defined as a distinguished technology driven process that can identify useful and enriched pattern of data. In KDD definition, data mining is one of the very important part which includes data analysis and high level algorithms for producing enriched and on-demand data patterns and format (Fayyad UM, Shapiro GP, Smyth P, 1996), (Khosla R, Dillon T, 1997) and (Inmon WH, 1992). Generally health related data means any data or information containing medical records, diagnosis reports of the patient as well as data of disease, surveys, trends, medicine, research, pharmaceuticals, about healthcare services and relevant record: direct or indirect: research or non- research: flat or patterned. These all types of information can be gathered from clinics, hospitals, test laboratories, medical practitioners or related organizations or kind of direct, indirect scientific research result-set. The pattern of these data can be text, CSV, pdf, metadata, image or other database format. Sometimes the format can be patients' information, their ethnics, region, behaviour, data from laboratories, family or disease history. It also includes electronic signals (digital or analog) such as MR, ETT, Angiogram, ECG and also images like ultrasound, radiology, histology etc. Now a days video file also a part of diagnostic reports. Here some complication arises as patient personal information and sensitive identification details shouldn't be used publicly. These type of sensitive data can be omitted while storing like other parameters. Besides these challenges another difficulty is that different disease and it's behaviour can be described, stored and analysed in different data glossaries, elements and various vocabularies (Cios K, 2002), (Stolba N, Banek M, and Tjoa AM, 2006), (Sahama TR, Croll PR, 2007) and (Lyman JA, Scully K, Harrison JH, 2008). Maintaining confidentiality with not compromising security, risk and disaster management also are the challenges to be met for building a sustainable DW.

1.2 PROBLEM STATEMENT

Health data or bio-information is the one of the most valuable parameters for medical services and research which is still a challenge to be integrated as central data warehouse. The challenge come from the problem of how all data can be extracted from heterogeneous sources, different repositories and store those data into a single integrated storage to ensure those data for all types of users including scientists, analysts, healthcare providers, patients and all kind of general users. It is very general idea that, everyday massive data or huge bulk health data are generated from different clinics, hospitals, diagnosing centres, laboratories and other relevant health organizations. These medical data are generally stored in a various information storage and retrieval system owned by the healthcare organizations in the form of image processing formats, health meta file, pdf format, dbf, mdb, xlsx, radiology and imaging, picture archiving and others many formats. Various hospitals and clinics use various data storage system and different file format. These are huge drawback to improve central healthcare system. With existing hardware (memory chip and processor) efficiently handing massive data producing everyday and use them for analysis and business intelligence is also a big challenge.

1.3 RESEARCH OBJECTIVES

The objectives of this research are as follows:

1. To design a model for healthcare warehousing that can process and analyze massive data
2. To simulate the proposed warehouse model for healthcare systems in Malaysia
3. To evaluate the performance of warehouse model by analyzing the results

1.4 SCOPE OF RESEARCH

The focus of this thesis is to propose a model for healthcare data warehouse which can extract, assess and process massive data from different sources and then ensure availability for users for query retrieval, analysis, reporting and mining. The prime focus of this research

will be on architecture that can ensure multidimensional data mining technique, performance, less memory usage, optimum processor usage, enrich data quality, handling complex queries and various utilization of DW.

The purpose of this research is to find out the problems of integrating health data from various heterogeneous sources and to propose a modern and sustainable model for data warehouse suitable for heterogeneous data extraction, transformation, storage, loading and integration for future use of various purposes including improvement of healthcare service, scientific research, reporting, analysis or any.

1.5 DISSERTATION OUTLINE

This thesis is divided up into seven chapters:

1. Chapter One: This chapter is the research introductive part. It provides a general overview of the study followed by the problem statement and objectives.
2. Chapter Two: Literature reviews on data warehouse, knowledge discovery from data and data warehouse for health data.
3. Chapter Three: This chapter describes several portions of National Healthcare data warehouse and its design issues.
4. Chapter Four: Data warehouse size calculation techniques, simulations results, pre processing techniques, KDD, data mining, data fabrication and summary of size reduction algorithm and recommendation of fragmentation mechanism.
5. Chapter Five: Chapter five concludes with the final remarks and suggestions for future research.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Data Warehouse integrates all various types of data of an organization into a centralized, uniform and single data structure and storage. The data can be structured, semi-structured or even unstructured. The purpose of centralizing these data is to analysis, mining, and business Intelligence. After storage, the next step is data quality assessment, cleansing, modelling and transformation. This part makes the data valuable for analysing. these data are then used for reporting, policy making, process mining, retrieving, digital marketing, biometric analysis, query analysis, scientific analysis, fundamental analysis and so on. Data source of the warehouse can be extracted from online and offline sources. A data warehouse for health is a very important data center where hospitals, clinics and other healthcare providers and stakeholders, medical and bio analysts can easily access for the analysing and caring part of the patients and virology and other depth analytical part as well as survey analysis. Extracting, Transforming and Loading (ETL) of healthcare domain data to a single and uniform data warehouse gives the provision of data storage efficiently, enrich quality analysis and decision making on real time basis will be facilitated.

Modern healthcare system is depends on several parameters like treatment facilities, patient care, accuracy of the diagnosis, rapid responds capabilities, adaptation of an emergency or facing new pandemic, outbreak, medical survey, regional and race oriented survey as well as medical and biometric analysis. But in reality health care institutes are far behind to fulfilled all of these criteria. A very recent example is COVID-19 pandemic. Because of not having a Global and centralized medical datacentre or Data warehouse, scientists were struggling to identify DNA structure of Coronavirus and regional and ethnic survey. As s result, it

became delayed which caused a massive causalities. Not only that but also the hospital oxygen capabilities, ICU bed related issues, relevant treatment, secondary infections of the patients, vaccination program- all are showing many question marks. An efficient DW can ensure fastest analysis (trillion data within a second) which can be the nucleus of medical analysis.

Different patients go different clinics and hospitals for treatment with different symptoms of disease and medical condition. All healthcare providers not using same database or data storage mechanism. This causes heterogeneity problem. To solve this problem a integrated data storage mechanism is needed which is called Data Warehouse. Data warehouse for Healthcare (DWH) is specialized type of data warehouse of having bulk data processing and storage mechanism.

2.2 REVIEW OF CURRENT HEALTH STRUCTURE, MALAYSIA

Before designing our proposed model for DW, we need to go through a quick review of Malaysian Health System (MHS). MHS has several entities where central Health Ministry (MoH) where refers the government control, monitor and operate overall total health system with collecting and providing statistical data. Medical personnel is another entity which refers to doctors, nurses, medical volunteers and others who can provide health service, care patients, do lab tests and produce health data- direct or indirect. The third entity of MHS is medical institutions which includes government and non-government hospitals, private clinics, diagnosis centers, community healthcare, blood banks, dental care providers, maternity homes, nursing homes, hospice, ambulatory care centre, combined facilities, community mental health centre. The next part is test laboratories those produce patient test data as per doctors' consultation. The final and most important entity for DWH is source of data which refers to unorganized, non-centralized scattered data from health institutes, health ministry (MoH), test laboratories, departments of statistics or other sources.

The above components of Malaysian healthcare system are producing huge data daily. While designing we need to focus on who are the users and who are the stakeholders of this warehousing system. Health data needs to be compiled and processed. Ministry of Health (MoH), Departments of Statistics and other controlling & monitoring authority currently may compile some health data. But central DW should be introduced soon for having greater views of multidimensional analysis, reporting, regional patient & disease trend analysis, identifying the limitations, bio-informatics analysis, virology analysis, finding out the cause of deaths, cause of infections spreading out, effectiveness of preventions measures taken and benefits of data mining of a total integrated system.

The current scenario and structure of healthcare system of Malaysia are illustrated in Figure 2.1. All components are still not integrated to a single uniform system. So data collection, processing and availability for analysing still not easy. So simulation and mining much depends on scattered data.

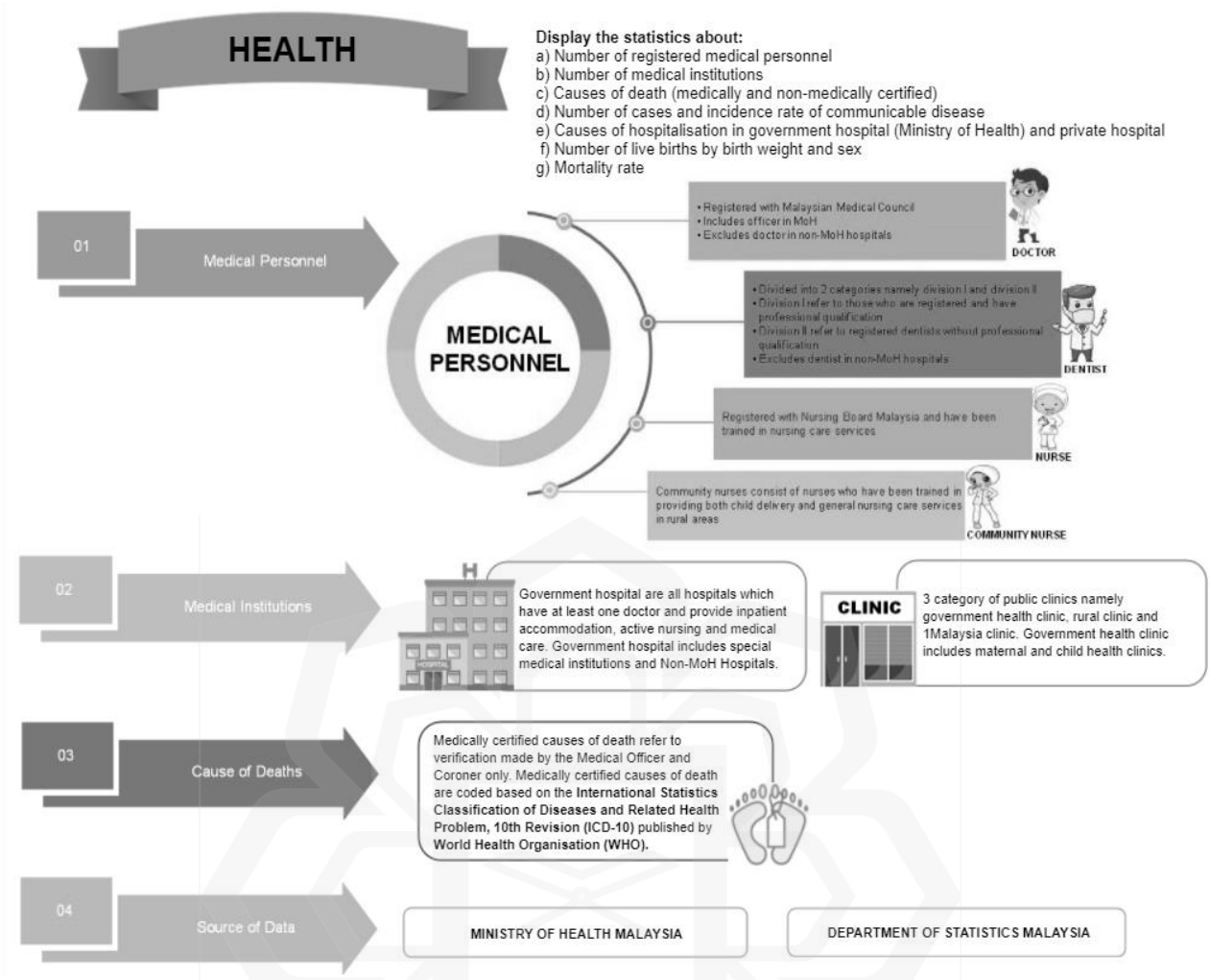


Figure 2.1 Healthcare Structure of Malaysia

2.3 HEALTHCARE DATA WAREHOUSE

This thesis is focusing to draw a model for DW to convert health data into meaningful information so that these valuable information can be used for research purpose, bioinformatics survey, virology, pattern of disease and forecasting the overall improvement of the healthcare system.

To develop a modern and sustainable healthcare system, the first step is to planning and decision making. For designing a modern healthcare system all types of previous and current data is required in a uniform file format with a central data storage mechanism. The data