Spring 4-10-2019

# Developing Virtual Data Warehouse for Rehabilitation Registry in Sabah, Borneo: Towards Big Data Analytics and Geomapping

Fatimah Ahmedy
*Universiti Malaysia Sabah*, fatimahmedy@ums.edu.my

Soo Fan
*Universiti Malaysia Sabah*, soofun@ums.edu.my

Nooralisa Mohd Tuah

Syed Nasirin
*Universiti Malaysia Sabah*, snasirin@ums.edu.my

Syahiskandar Sybil Shah
*Queen Elizabeth Hospital*, syahiskandar77@gmail.com

# DEVELOPING VIRTUAL DATA WAREHOUSE FOR REHABILITATION REGISTRY IN SABAH, BORNEO: TOWARDS BIG DATA ANALYTICS AND GEOMAPPING

***Research-in-progress***

**Fatimah Ahmedy**
Faculty of Medicine & Health Sciences, Universiti Malaysia Sabah,
Kota Kinabalu, Malaysia

**Tan Soo Fun**
Faculty of Computing & Informatics, Universiti Malaysia Sabah,
Kota Kinabalu, Malaysia

**Nooralisa Mohd Tuah**
Faculty of Computing & Informatics, Universiti Malaysia Sabah, Labuan
International Campus, Malaysia

**Syed Nasirin**
Faculty of Computing & Informatics, Universiti Malaysia Sabah, Labuan
International Campus, Malaysia

**Syahiskandar Sybil Shah**
Department of Rehabilitation Medicine, Queen Elizabeth Hospital,
Ministry of Health Malaysia

## Abstract

*Clinical registry, defined as an organised system for the collection, storage, retrieval, analysis, and dissemination of information on individuals with a condition that predisposes to the occurrence of a health-related event, are designed through data repository or data warehouse. Data repository is described as a real-time database that consolidates data from a variety of clinical sources that offers a comprehensive source for storage and retrieval of relevant clinical information needed. However, data warehouse is a data repository that concentrates on data queries and data analytics. Rehabilitation registry in Malaysia is still at its infancy with lack of data sharing and integration. As rehabilitation is a subspecialty concerned with the prevention, diagnosis, and rehabilitation of disabling conditions, a registry would allow identification of patients' demographics, clinical and functional outcomes improvement, benchmarking the delivery of rehabilitation services, and research purposes. The application of virtual data warehouse, cloud computing, big data analytics and geomapping for clinical registries have been implemented well in countries like China and United Kingdom. The main objectives of this research-in-progress paper are to demonstrate the feasibility of developing and designing virtual data warehouse framework based on cloud computing technology, in an attempt towards big data analytics and geomapping implementation for inpatient rehabilitation registry in Sabah, Malaysia.*

**Keywords**: Rehabilitation Registry, Virtual Data Warehouse, Big Data Analytics

## 1.0 Introduction

In the wake of evidence-based medicine, a significant rise of data repositories is observed globally leading to the development of thousands clinical registries. Clinical registries are defined as “an organised system for the collection, storage, retrieval, analysis, and dissemination of information on individual persons who have either a particular disease, a condition that predisposes to the occurrence of a health-related event, or prior exposure to substances (or circumstances) known or suspected to cause adverse health effects” (Gliklich, Dreyer & Leavy, 2014). Clinical registries are designed through the development of data repository or data warehouse. Abdelhak et al. (2014) described data repository as "a real-time database that consolidates data from a variety of clinical sources to present a unified view of a single patient that offers a comprehensive source for storage and retrieval of relevant clinical information needed". On the other hand, data warehouse is a data repository that concentrates on data queries (e.g. Online Analytical Processing (OLAP)) and data analytics (e.g. descriptive and prescriptive analytics) (Chandra, & Gupta, 2018).

The numbers of rehabilitation registries are increasing exponentially for the past decade, not only focusing on data repository but aptitude for big data analytics while serving as a health educational platform. Rehabilitation registry in Malaysia is still at its infancy with lack of data sharing and integration. Rehabilitation medicine is a subspecialty concerned with the prevention, diagnosis, treatment and rehabilitation of disabling conditions including stroke, spinal injury and traumatic brain injury. Rehabilitation registry allows accurate identification of patients’ demographics, clinical and functional outcomes improvement, benchmarking the delivery of rehabilitation care services (New, Simmonds & Stevermuer, 2011), and serves as the platform for research purposes.

In the absence of a clinical registry, data on rehabilitation population’s demographic, disease complications, and rehabilitation outcomes are not accurately visualised to justify the allocation of health resources and funding. In this research-in-progress paper, we proposed to design and develop virtual data warehouse framework based on cloud computing technology, in an attempt towards big data analytics and

geomapping feature for supporting the inpatient rehabilitation registry in Sabah, located in northern Borneo, Malaysia.

## 2.0 Literature Review

The state of registries is highly advanced in developed countries as such the platforms for data repositories incorporated big data analytics, research databases and educational tools. For example, in the UK, data registry from Global Burden of Diseases Injuries and Risk Factors Study (GBD, 2010) was further utilised to scrutinise the patterns of health loss and the leading preventable risks (Murray et al., 2013).

Advancing Information System (IS) technology with simultaneous improvement of the healthcare system in developing region has resulted in increased development of clinical registries. The China National Stroke Registry is a nationwide prospective registry initiated in 2007 aiming to evaluate the delivery and quality of stroke care (Wang et al., 2011). Indians' National Cancer Registry Programme was developed in 1981 for epidemiological studies and developing human resources for planning and monitoring cancer control activities (Rath & Gandhi, 2014).

Clinical registries in the field of rehabilitation medicine have evolved significantly for the past decade since the introduction and commendation of using World Health Organisation International Classification of Functioning, Disability and Health (ICF) framework that recognised the needs to identify and quantify disabilities (Stucki, Reinhardt, Grimby & Melvin, 2008). Model Systems Knowledge Translation Center at American Institute for Research has developed comprehensive physical medicine and rehabilitation databases that not only include registries for traumatic brain injury (TBI), spinal cord injury (SCI) and burn, but also generate up-to-date big data analytics and concurrent health information resources for patients.

In Malaysia, rehabilitation registry is not fully established across the nation. Having a registry allows contextual identification of benefits and barriers encountered in the provision of rehabilitation services leading to an accurate number of disabilities, better allocation of rehabilitation resources and research potentials. Thus, initiating

rehabilitation registry would be one of the first initiatives to bring forward this field into industrial revolution within the country.

Data repository terminology is often used interchangeably with the data warehouse to describe a collection of clinical data from various sources. Payne (2011) has defined the data repository as “a database that is optimised for storing and viewing clinical information sent over interfaces from departmental systems". Abdelhak et al. (2014) subsequently described the data repository as "a real-time database that consolidates data from a variety of clinical sources to present a unified view of a single patient that offers a comprehensive source for storage and retrieval of relevant clinical information needed". On the other hand, the data warehouse is a data repository that concentrates on data queries (e.g. Online Analytical Processing (OLAP)) and data analytics (e.g. descriptive and predictive analytics) (Chandra, & Gupta, 2018; Golfarelli & Rizzi, 2018). One of the major roles of rehabilitation registry is focused on analysing the collected data as an attempt to produce patients’ health care profile and benchmarking rehabilitation care received through descriptive and inferential statistical analytics.

Generally, the data warehouse can be categorised into two main types: 1) virtual data warehouse and 2) distributed data warehouse. Virtual data warehouse refers to layers of databases that sit on top of existing databases to permit users in the query of the whole databases as if they are a single entity (although these databases are logically and physically separated) (Golfarelli & Rizzi, 2018; Maity et al., 2018). Distributed data warehouse, on the other hand, refers to the physical architecture of a single database. Distributed architecture usually includes clusters of 2 or more nodes and mostly enables efficient separation of computing resources to support concurrent operations (Chandra et al., 2018).

Theoretically, constructing virtual data warehouse is feasible but the main downside is the query performance of the virtual environment upon demand of queries across data from different layers of databases. In the advancement of IS technology, such downside is overcome by cloud computing technologies **(**Barkhordari & Niamanesh, 2018; Chandra et al., 2018). This research-in-progress paper seeks to design a virtual data warehouse framework based on cloud computing technology for supporting the

development of inpatient rehabilitation registry in Sabah, Malaysia. This virtual data warehouse framework serves as the platform towards implementation of big data analytics and visualisation using geomapping feature.

## 3.0 Virtual Data Warehouse Development Methodologies

The methodology of developing a virtual data warehouse for inpatient Rehabilitation Registry in Sabah consists of 4 main phases as illustrated in Figure 1.

### 3.1 Phase 1: Problem Formulation and Preliminary Data Requirement Investigation

Sabah is located in northern Borneo and has various population and geographical distributions. Inpatient rehabilitation facility in Sabah is situated at the largest tertiary hospital in its capital state, Kota Kinabalu. Ward admission is based upon non-pre-emptive priority with capacity to accommodate 26 beds. Patients' demographics, clinical information, outcomes assessments and inpatient progress are all recorded and date stamped in inpatient medical files. These files are kept in the hospital's medical record office.

In developing a virtual data warehouse, it is necessary to understand user requirements from database users (i.e. rehabilitation medicine team members). Main objective of this phase is to investigate data behaviours and access patterns, data queries and Online Analytical Processing (OLAP), as well as data analytics requirements of Sabah's inpatient rehabilitation services. Due to the heterogeneity of data sources, qualitative research methods such as interview, questionnaire and observation are applied in this phase.

### 3.2 Phase 2 – Data Collection and Requirement Analysis

Following an agreement with the studied hospital, the data collected is further analysed in order to identify the data requirements, business process and query practices, as well as desirable visualisation report condition. Several data analysis tools and feasibility studies are conducted to achieve the objective of this phase. Outcomes from requirement analysis are further documented as a narrative story.

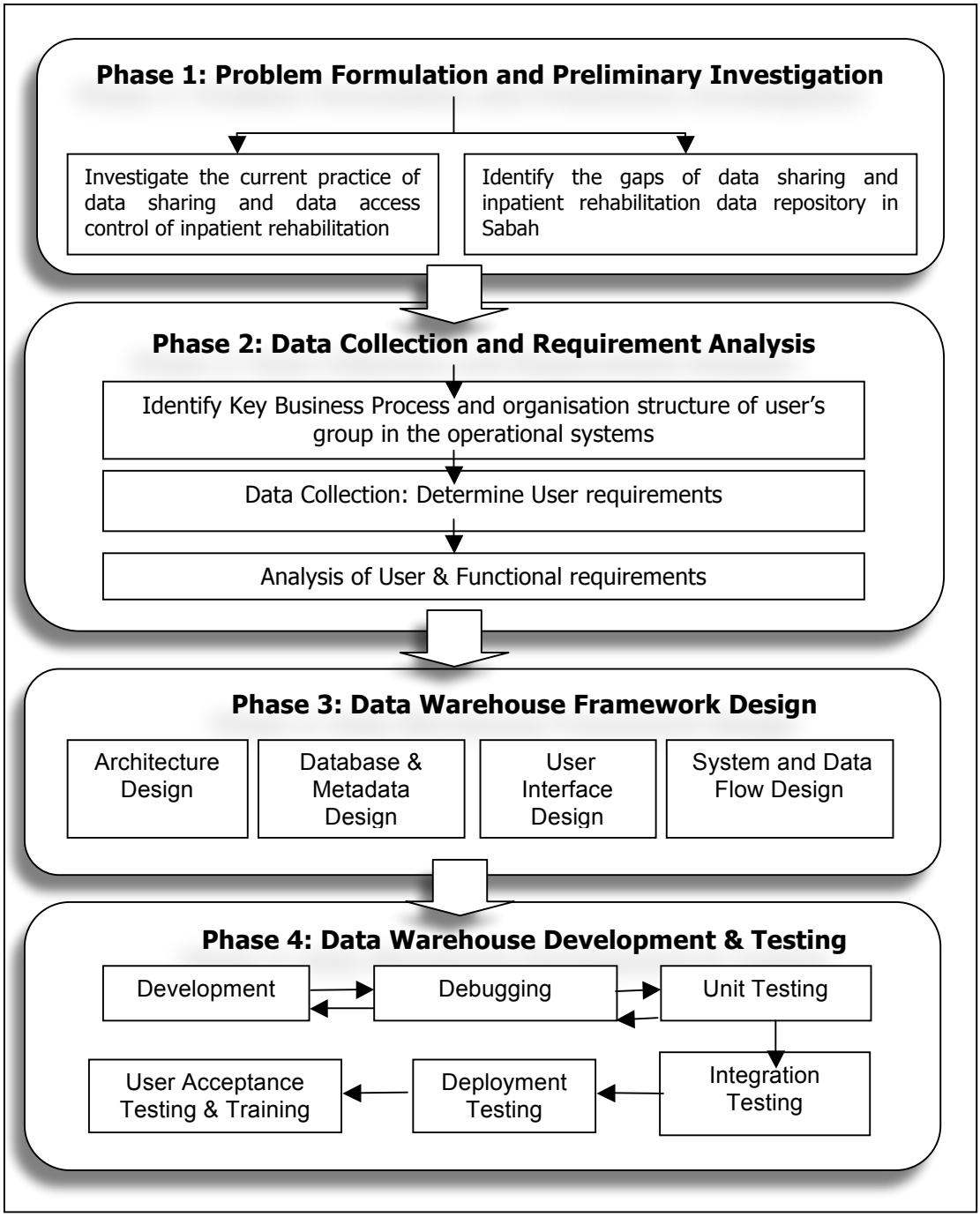


**Figure 1. Flowchart for developing a virtual data warehouse for inpatient Rehabilitation Registry in Sabah**

### 3.3 Phase 3 – Data Warehouse Framework Design

Based on the analysed data and business requirements, developing the virtual data warehouse involved designing cloud and virtual data warehouse architecture, data repository and metadata, OLAP, system user interface (UI) and report visualisation. The storyboard, data flow diagram (DFD) and flowchart diagram are used to communicate with the end-users.

### 3.4 Phase 4 – Data Warehouse Development & Testing

#### 3.4.1 Prototype and Incremental Development

Developing a virtual data warehouse prototype for the inpatient rehabilitation registry is implemented with the incremental development approach. Regular scrum meeting is conducted to obtain user feedback on the developed prototype for incremental improvement until an acceptable prototype is developed according to user requirements.

#### 3.4.2 Data Queries Implementation and Report Virtualisation

Following deployment of the virtual data warehouse on the cloud computing platform, OLAP data queries and report visualisation are further implemented and tested for big data analytics and geomapping. Concurrent regular scrum meeting with user is conducted for feedback followed by several testing including integration testing and deployment testing. Performance efficiency of queries implementation and report visualisation such as processing time and storage capacity are further recorded and benchmarked.

#### 3.4.3 Performance Evaluations and User Acceptance

The proposed prototype is evaluated using the System Usability Scale (SUS) approach for measuring the performance and user acceptance.

## 4.0 Summary

The preliminary result of this research-in-progress paper has shown that rehabilitation registry in Sabah is still in the infant stage, which lack of data sharing and integration. The utilisation of the advanced Information System (IS) technologies including data warehouse based on cloud computing and implementation of big data analytics and geomapping visualisation are proven to be advantageous in supporting the management of data repositories in several countries such as China and United Kingdom. Subsequently, this study investigated the feasibility of developing a virtual data warehouse for inpatient rehabilitation registry in Sabah, Malaysia.

The result has shown that cloud computing can be further adapted to construct a virtual centralised rehabilitation registry. The tradition on-premises rehabilitation registry systems might not be applicable well in Malaysia as the country is divided into two main regions: the Malay Peninsular and the island of Borneo, whereby Sabah

is located in the latter region. Future work of this study involves developing a virtual data warehouse for integrating the heterogeneity of data repositories from several inpatient rehabilitation facilities in the country, subsequently expanded to support the big data analytics and report visualisation with geomapping feature.

## References

Abdelhak, M., Grostick, S., & Hanken, M. A. (2014). *Health Information-E-Book: Management of a Strategic Resource*. Elsevier Health Sciences.

Barkhordari, M., & Niamanesh, M. (2018). Hengam a MapReduce-Based Distributed Data Warehouse for Big Data: A MapReduce-Based Distributed Data Warehouse for Big Data. *International Journal of Artificial Life Research (IJALR)*, *8*(1), 16-35.

Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. *International Journal of Information Technology*, *10*(2), 217-224.

Gliklich, R., Dreyer, N., & Leavy, M. (2014). Registries for Evaluating Patient Outcomes: A User's Guide. Two volumes.(Prepared by the Outcome DEcIDE Center [Outcome Sciences, Inc., a Quintiles company] under Contract No. 290 2005 00351 TO7.) AHRQ Publication No. 13 (14)-EHC111. Rockville, MD: Agency for Healthcare Research and Quality. April 2014. *Rockville, MD: Agency for Healthcare Research and Quality*.

Golfarelli, M., & Rizzi, S. (2018). From Star Schemas to Big Data: 20+ Years of Data Warehouse Research. In *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years* (pp. 93-107). Springer, Cham.

Maity, B., Sen, S., & Debnath, N. C. (2018). Challenges of implementing data warehouse in MongoDB environment. *Journal of Fundamental and Applied Sciences*, *10*(4S), 222-228.

Mastura, I., Zanariah, H., Fatanah, I., Idzwan, M. F., Shaariah, M. W., Jamaiyah, H., & Geeta, A. (2008). An audit of diabetes control and management (ADCM). *Med J Malaysia*, *63*(suppl C), 76-7.

Murray, C. J., Richards, M. A., Newton, J. N., Fenton, K. A., Anderson, H. R., Atkinson, C., & Braithwaite, T. (2013). UK health performance: findings of the Global Burden of Disease Study 2010. *The lancet*, *381*(9871), 997-1020.

Payne, T. (Ed.). (2014). *Practical guide to clinical computing systems: design, operations, and infrastructure*. Academic Press.

Rath, G. K., & Gandhi, A. K. (2014). National cancer control and registration program in India. *Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology*, *35*(4), 288.

Stucki, G., Reinhardt, J. D., Grimby, G., & Melvin, J. (2008). Developing research capacity in human functioning and rehabilitation research from the comprehensive perspective based on the ICF-model. *European journal of physical and rehabilitation medicine*, *44*(3), 343-351.

Wang, Y., Cui, L., Ji, X., Dong, Q., Zeng, J., Wang, Y., & Nguyen-Huynh, M. N. (2011). The China National Stroke Registry for patients with acute cerebrovascular events: design, rationale, and baseline patient characteristics. *International Journal of Stroke*, *6*(4), 355-361.