A Comparative Evaluation of NoSQL Databases for Philippine Cancer-Phenome Biobanking System

Philip John C. Sales[a], Alvin B. Marcelo[a], Ariel S. Betan[a,c], Michael C. Velarde[b,1]

[a] *Medical Informatics Unit, College of Medicine, University of the Philippines, Ermita, Metro Manila*
[b] *Regenerative Biology Research Laboratory, Institute of Biology, University of the Philippines, Quezon City, Diliman*
[c] *College of Arts and Sciences, University of the Philippines, Ermita, Metro Manila*

Abstract

The field of healthcare is rapidly accumulating data of complex types and formats. The current methods of storing these diverse data relies heavily on traditional relational database management system (RDBMS). While RDBMS offers many advantages, they also have notable limitations particularly in adapting to the increasing volume and variations of clinical and biomedical data; thus, a constant demand for alternative modeling approaches are in place. NoSQL databases had gained market tractions and had been cited as a viable solution due to its numerous benefits; however, there are few published studies regarding the evaluation of NoSQL for health information systems (HIS).This paper aimed to adapt a framework with application-specific and context-based parameters for evaluating three (3) types of different NoSQL databasesThe results showed that document-based NoSQL database is the best choice for extensibility, flexibility, and query readability whereas key-value pair is the most efficient for performance and scalability. Moreover, columnar-wide type is advantageous in terms of storage capacity. Among the shortlisted NoSQL databases, MongoDB

*Corresponding author at Medical Informatics Unit, UP College of Medicine, Tel. +63(2) 536 1396

Email addresses: pcsales@up.edu.ph (Philip John C. Sales), admarcelo@up.edu.ph (Alvin B. Marcelo), asbetan@up.edu.ph (Ariel S. Betan), mcvelarde@up.edu.ph (Michael C. Velarde)

[1] https://regenlab.weebly.com

was found to be the recommended choice for the current implementation and immediate need of the biobanking HIS.

Keywords: Clinical data, Relational database, NoSQL, HL7 - Fast Health Interoprable Resource (FHIR), Information storage and retrieval

1. Introduction

The increasing size and complexity of data being generated in healthcare industry is reaching to the point of being unmanageable (Kumar et.al, 2015; Ercan et.al, 2014). This was brought about by the technological advances in medical and biomedical fields which have highlighted the importance of collecting and storing data to benefit clinical research and medical services.

Currently, storage of clinical data largely relies on relational database management systems (Goli-Malekabadi et.al, 2016; Lee et.al, 2012) and has since been the most common approach of data storage since the 90's (Atzeni et.al 1993). From the time of its origins in the 1970's, the relational model had been adapted widely in many database management system (Berg et.al, 2016; Suciu 2001), and since then had remained unchanged for decades (Ercan et.al, 2014). This modeling approach had become the de facto standard for most industries, but remains ineffective in the field of healthcare due to the characteristics of clinical and biomedical data.

Healthcare data are dynamic, hierarchal, sporadic, and heterogeneous in nature. It is stored in various types including free text, coded data elements, images, signals, logs, or notes that is formatted in either structured, semi-structured, or unstructured presentation (Kruse et.al, 2016; White S, 2014; Ercan et.al, 2014; Lee et.al, 2012; Wasan et.al, 2006). The heterogeneity of the data, increasing volume, and ever-changing structure makes it difficult to model and manage health data using the traditional relational database (Al-Fatlawi et.al, 2015; Schmitt and Majchrzak 2012; Lee et.al, 2012; Jin et al. 2011).

In addition, one of the major drawback of relational databases is the need for pre-designed schema. The problem of pre-defining data structure is the fact

that it is unrealistic to pre-determine all the variety of data fields to be collected in a medical record (Lee et.al, 2012).

The database model of any health information systems must be flexible to accommodate constant changes in data requirements. Consequences of not updating database schema may cause failure to capture the additional information, whilst an attempt to restructure may mean additional maintenance upkeep which in turn is time consuming and expensive that eventually leads to operational disturbance, dissatisfied staff, loss of business, or even impractical for some organization (Al-Fatlawi et.al, 2015; Manoj et.al 2014; Gilchrist et.al, 2010).

In summary, the conventional relational database management system is not flexible, scalable, and extensible enough to overcome the problem of ever-changing requirements, heterogeneity, increasing volume, and the continuous need for updates of healthcare data (Goli-Malekabadi et.al, 2016; Al-Fatlawi et.al, 2015; Schmitt and Majchrzak 2012; Lee et.al, 2012; Jin et al. 2011).

That being said, there is a constant demand (Lee et.al, 2012) and practical need for alternatives outside the conventional relational data modeling approach. Additionally, there are increasing researches and industry projects on alternative databases that are driven by the practical experience of the conflicting static nature of relational database and dynamic characteristics of healthcare data (Ercan, et.al 2010).

2. Background

Out of the wide spectrum of viable alternatives, this study identified NoSQL database as supported by literature to be suitable in addressing the challenges of healthcare data. NoSQL approach had attracted the attention of non-clinical industries (Kruse et.al, 2016; Parker et.al, 2013; Ferreira et al. 2013) and researchers due to its flexibility, scalability, velocity, and availability which addresses the limitations of relational databases (Ercan et.al, 2014). While there are significant advantages of using NoSQL database, there is limited research

which has evaluated the use of NoSQL databases in the healthcare domain particularly with the use of Fast Health Interoperability Resource (FHIR) formatted data.

Additionally, it is not always straightforward to declare which database type is better than the other (Fowler, 2013). Concerns are raise in selecting which among the list NoSQL databases can best address the healthcare data challenge. Literature gap exist in determining what database evaluation framework can be used as reference when developing a health information systems (HIS) or more particularly a biobank information system.

Technology selection or evaluation framework are constructs predominantly used in the field of business sectors with focus only upon management and financial decisions (Chan et.al 2010), but dearth of source exists when discussing technology or database evaluation for healthcare sector. Moreover, most selection practices and evaluation models in digital health technologies are informed by traditional means: word of mouth, internet search, consultant's advice, expert's opinion (Ostrovsky et.al 2014), emerging systems, and technology trend. There are limited models, empirical evidences, and guidelines that facilitate methodological evaluation of database for HIS.

In light of the foregoing, this study attempts to implement a systematic way of evaluating the databases. To date, there is a paucity of literature that provides guiding principle, standard protocol, or framework on how to evaluate of database for healthcare domain. Therefore, this study will adapt a framework that will evaluate how different databases are at par with each other and against application-specific and context-based evaluation criteria.

3. Method

4. Results and Discussion

4.1 Performance.

4.2 Scalability.

4.3 Storage.

4.4 Query Readability.

4.5 Extensibility.

4.6 Flexibility.

5. Conclusion

Acknowledgements

Research Funding. The development of the biobanking health information system was funded by the Commission on Higher Education (CHED) under the Philippine California Advanced Research Institute (PCARI) program. The project name is "IHITM 2016-13: Establishment of a Philippine Cancer Phenome-Biobanking System and Biomonitoring Program" under Prof. Michael C. Velarde (MCV). Moreover, the research itself is also funded as stand-alone graduate thesis funding under the same program.

Author's Contributors. Designed the workflow: Philip C. Sales (PCS). Performed the experiment: PCS, Programmed the scripts: PCS, Miko C. Chu (MCC). Analyzed the results: PCS. Reviewed the paper: Prof. Michael C. Velarde (MCV), Dr. Alvin B. Marcelo (ABM), Prof. Ariel S. Betan (ASB), Prof. Bryann Chua, Dr. Jun Inciong. Wrote the paper: PCS, ABM, ASB, MCV.

6. Bibliography styles

There are various bibliograph styles availabe. You can select the styl of your choice in the preamble of this document. These styles are Elsevier styles based on standard styles like Harvard and Vancouver. Please use BibTEX to generate your bibliography and include DOIs whenever available. Here are two sample references: [1, 2]. [3]

References

[1] R. Feynman, F. Vernon Jr., The theory of a general quantum system interacting with a linear dissipative system, Annals of Physics 24 (1963) 118–173. doi:10.1016/0003-4916(63)90068-X.

[2] P. Dirac, The lorentz transformation and absolute time, Physica 19 (1-–12) (1953) 888–896. doi:10.1016/S0031-8914(53)80099-6.

[3] V. Abramova, J. Bernardino, Nosql databases: Mongodb vs cassandra, in: Proceedings International C* Conference on Computer Science and Software Engineering, ACM, Portu, Portugal, July 2013, pp. 14–23. doi:10.1145/2494444.2494447.