journal homepage: www.intl.elsevierhealth.com/journals/cmpb

Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage

Ken Ka-Yin Lee[a,*], *Wai-Choi Tang*[b,1], *Kup-Sze Choi*[a,2]

[a] *Centre for Integrative Digital Health, School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong*
[b] *Information Technology Committee, Hong Kong Doctors Union, Hong Kong*

ARTICLE INFO

Article history:
Received 13 April 2012
Received in revised form
24 October 2012
Accepted 31 October 2012

Keywords:
Clinical data
Relational database
NoSQL
XML
Information storage and retrieval

ABSTRACT

Clinical data are dynamic in nature, often arranged hierarchically and stored as free text and numbers. Effective management of clinical data and the transformation of the data into structured format for data analysis are therefore challenging issues in electronic health records development. Despite the popularity of relational databases, the scalability of the NoSQL database model and the document-centric data structure of XML databases appear to be promising features for effective clinical data management. In this paper, three database approaches – NoSQL, XML-enabled and native XML – are investigated to evaluate their suitability for structured clinical data. The database query performance is reported, together with our experience in the databases development. The results show that NoSQL database is the best choice for query speed, whereas XML databases are advantageous in terms of scalability, flexibility and extensibility, which are essential to cope with the characteristics of clinical data. While NoSQL and XML technologies are relatively new compared to the conventional relational database, both of them demonstrate potential to become a key database technology for clinical data management as the technology further advances.

1. Introduction

Clinical data is dynamic, sporadic, and heterogeneous in nature [1]. While they share some of the characteristics of the data managed by conventional data warehouse, special attention is required in the design of database schema because of the unique features they possess. Currently, storage of clinical data largely relies on relational database management systems. The relational database model is the most common and a proven approach to store and query data in various forms [2]. However, the major drawback is the need to pre-design the exact field structures of the data, which is required in the process of database normalization to ensure data consistency [3]. In addition, the relational database model is not practical for certain forms of data that require a lot of fields to handle different types of data involved, where most of the data fields are indeed left unused due to the nature of the data. A relational database storing these kinds of data will contain many empty fields, resulting in inefficient storage and poor performance. Medical data, especially clinical notes, are such an example. To deal with these issues, we attempt to use a class of database known as NoSQL and extensible markup language (XML) to develop databases that can cope with the special features of clinical data more effectively.

* *Corresponding author at*: Centre for Integrative Digital Health, School of Nursing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. Tel.: +852 2766 6399.
E-mail addresses: kenleeky@gmail.com (K.K.-Y. Lee), tang.eric233@gmail.com (W.-C. Tang), kschoi@ieee.org (K.-S. Choi).
[1] Tel.: +852 2423 8293.
[2] Tel.: +852 3400 3214.
0169-2607/$ – see front matter © 2012 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.cmpb.2012.10.018

To determine the suitable database approaches for the storage of structured clinical data, the performance of databases developed using NoSQL and XML are compared in this paper. The database approaches concerned are NoSQL database, XML-enable database and native XML database. The comparisons are made from the following aspects: query performance, scalability, flexibility and extensibility. The databases are populated with de-identified real data obtained from the clinic of a private general practitioner (GP) in Hong Kong. The rest of the paper is organized as follows. Section 2 gives an overview on the application of relational databases, the NoSQL database movement and XML databases for storing medical data. Section 3 describes the data structure used for clinical data management and the types of database approaches to be investigated in the paper. Section 4 discusses the three database approaches concerned based on the experimental results and our experience in the development. Finally, a conclusion is given in Section 5.

2. Backgrounds

It is generally agreed by many researchers that the storage and management of medical data is a difficult task. As pointed out by Cios and Moore [4], medical data are voluminous and complex in nature. They are the "most rewarding and difficult" data to analyze. The wide variety of medical data is indeed one of the major obstacles to the implementation of electronic health records (EHR) in hospitals [5]. Medical data are generated from multiple sources, e.g. laboratory or pharmacy, which resides in different databases of different data structures. They also exist in different forms, both semi-structured and structured formats. The heterogeneity makes integrated access of medical data a challenging issue [6]. Therefore, an effective approach for storing medical data and facilitating data analysis is needed.

A number of models have been proposed for medical data storage. Los [7] modified the conventional row-modelling database to accommodate the heterogeneous nature of medical data. Prather et al. [8] proposed a new knowledge discovery method for medical data in order to find the relationships between medical concepts and their related properties in a large clinical database. Rector et al. [9] emphasized the importance of the context of medical data and the relationships between different data sets. Nevertheless, most medical data storage systems are built based on relational database approach, which is not efficient and flexible enough to handle the data.

There is a constant demand for alternatives to the relational database. Recently, the movement towards the "NoSQL" is gaining attention. It refers to a class of data storage systems that is significantly different from the conventional relational database. For example, NoSQL does not require pre-defined schema, relationships and keys. Database table join queries are also not supported [10]. The NoSQL technology is inspired by the Web 2.0 developers and communities who realize that relational database is not suitable for the management of real-time social networking websites where the data are voluminous and heterogeneous. NoSQL databases are low-cost, schema-free, and horizontally scalable to accompany new computing resources when the need arises. In general, NoSQL uses key-value stores, BigTable implementation, document store, and graph databases which are uncommon in traditional relational database design [11]. As the nature of the data in Web 2.0 applications shares similarities with that of clinical data, a few attempts to apply the NoSQL approaches, e.g. using the key-value stores, have been made for clinical data storage [12,13], although the research effort is at an early stage. Despite a relatively new technology, NoSQL is significant in that it is a common database approach for cloud computing and thus a promising solution for cloud-based clinical systems.

On the other hand, XML is a well-known and robust markup language designed to encode documents electronically and represent the data in a structured way. It has been demonstrated widely that XML can be used for storing structured data with a high degree of flexibility. For example, a database architecture is proposed for medical data by exploiting XML's strength in data interoperability and application integration [14]. An XML-based record system is also developed to provide customizable storage of medical data based on the needs of patients [15]. Medical markup language (MML), a variation of XML, is used to design hospital information system for medical data exchange [16]. In these applications, medical data are represented using XML and stored as XML files. It is worth noting that popular standards for clinical records (e.g. EN-ISO13606, OpenEHR, HL7 RIM-CDA) also use XML to codify information.

Furthermore, XML databases are proposed and developed for efficient processing of structured XML data. Jagadish et al. proposed an XML database to optimize XML-based queries and processing [17]. Meier also proposed an XML database to store, index and query a large collection of XML data [18]. Both of these XML databases are considered "native" which refers to the characteristic that the internal model of the databases depends on XML and uses XML documents as the fundamental unit of storage. Alternatively, traditional relational databases can also be exploited to take advantage of the benefits of XML, where XML data is mapped to a conventional relational database [19]. Databases of this kind are known as XML-enabled databases. Few XML databases have been developed for medical data storage [20–22], and they model the data at the document level rather than the content level inside an XML document. Note that XML databases are often also classified as a type of NoSQL database [10].

Standardized and normalized data reference models are needed to represent clinical information. One of the benefits of clinical data models is that they support the level of complexity required to correctly interpret the information with context. The data models also facilitate the processing of semantic meaning within the data, which is of growing interest in EHR development. To achieve normalized clinical data models, a strategy known as the two-level methodology is proposed to decouple knowledge models from system design [23]. The methodology provides flexibility for the integration of different knowledge models with the clinical systems while the integration can be independent from the system design. This strategy is being adopted by an increasing number of systems. Examples of normalized reference models to represent the clinical information includes EN-ISO13606, OpenEHR, and HL7 RIM-CDA [24].