COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 132 (2016) 75–82

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

An effective model for store and retrieve big health data in cloud computing

ARTICLE INFO

Article history:
Received 4 September 2015
Received in revised form
8 April 2016
Accepted 11 April 2016

Keywords:
Health data
Relational database
NoSQL
Big data
Information storage and retrieval
Cloud computing

ABSTRACT

Background and objective: The volume of healthcare data including different and variable text types, sounds, and images is increasing day to day. Therefore, the storage and processing of these data is a necessary and challenging issue. Generally, relational databases are used for storing health data which are not able to handle the massive and diverse nature of them.
Methods: This study aimed at presenting the model based on NoSQL databases for the storage of healthcare data. Despite different types of NoSQL databases, document-based DBs were selected by a survey on the nature of health data. The presented model was implemented in the Cloud environment for accessing to the distribution properties. Then, the data were distributed on the database by applying the Shard property.
Results: The efficiency of the model was evaluated in comparison with the previous data model, Relational Database, considering query time, data preparation, flexibility, and extensibility parameters. The results showed that the presented model approximately performed the same as SQL Server for "read" query while it acted more efficiently than SQL Server for "write" query. Also, the performance of the presented model was better than SQL Server in the case of flexibility, data preparation and extensibility.
Conclusions: Based on these observations, the proposed model was more effective than Relational Databases for handling health data.

1. Introduction

By development of technology, the request for storing and processing large amounts of data is increasing excessively. The volume of data is enormous right now; and it is predicted to reach 35 zettabytes by 2020 [1]. Huge and increasing amount of data are seen in all branches of science, specifically in healthcare. The volume of worldwide healthcare data was equal to 500 petabytes in 2012; and it is expected to be 25,000 petabytes in 2020. In addition, there are various types of data in healthcare including personal medical records, radiology images, clinical trial data, FDA submissions, human genetics and population data, 3D imaging, sensor readings, genomics, etc. [2,3].

These complex data are collected from different sources with several structures. It is a difficult task to store and analyze medical data. However, cumulative analysis of healthcare is necessary for discovering new useful patterns, recognizing unknown relationships, making universal decisions, identifying effective treatments and selecting best practices for a group of patients. Therefore, an effective method is needed for health data management in order to fulfill the requirements in this branch. Moreover, the model should be available and consistent in voluminous data [4,5].

In recent years, maintenance and processing of various and high volume data have created the "Big Data" challenge. As Gartner said: *"Big Data is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"* [6].

Generally, most organizations store their data in relational database management systems (RDBMS) [7]. There are some restrictions on using relational databases (DBs) for Big Data including efficient processing, effective parallelization, scalability and costs [7–9]. For example, users must convert Big Data into tables with pre-design fields; and it creates a complex and difficult structure of database that works slowly. In addition, data cannot fit into a table simply [7]. As another restriction, attaching tables to the distributed system is not easy. Therefore, these databases do not easily act in a distributed system [7,10]. On the other hand, due to scalability and availability requirements, it is inefficient to store massive amount of data locally. Another restriction is inability of RDBMS for storing huge data due to its limitation on database size. Moreover, companies must focus on providing a lot of management resources

to store massive data rather than focusing on their business innovation; and it consumes a lot of time and cost.

Recently, some new technologies such as a new class of databases, known as NoSQL, and a computing model, called Cloud Computing, have been developed to resolve these problems. NoSQL databases provide better performance for storing and maintaining large scale data regardless of their formats. These databases can also operate without predefined schema and relationship [5,7,11–13]. It provides high throughput for voluminous and heterogeneous data in a distributed environment [5]. Moreover, users do not need to be familiar with SQL language.

Due to their simple data models, NoSQL databases are used easily. These databases can function in a distributed manner. So, users can scale a single database on additional inexpensive machines instead of a more powerful and costly single machine. Data processing in NoSQL databases is generally faster than relational databases. Unstructured, semi structured data and data with various form and size are effectively used in these databases [7,14].

Another new technology mentioned above is Cloud Computing. This computing model is a new paradigm that reduces the costs for management of resources such as hardware and software by offering resources on the network. So, users may access to demand services everywhere in the world. Cloud is a kind of parallel and distributed systems with a collection of computers. It provides computing resources based on service-level agreements established between the service provider and users. Cloud environments have some advantages such as availability, scalability, performance, multi-tenancy, elasticity, fault tolerance and load balancing. This computing model enables users to consume computing resources (e.g., networks, servers, storage, applications and services) as a utility over the Internet. Instead of local servers or personal devices, resources are shared with minimal management effort to handle applications [11,15].

In this paper, based on NoSQL technology, an efficient model has been proposed by considering the essential requirements of health data. The model is compatible with Cloud environment and utilizes capabilities of Cloud. In the following, its performance has been compared with former model considering query time in some workload, data preparation, flexibility and extensibility.

The rest of the paper is ordered as follows. Section 2 offers an overview on different approaches to handle and maintain big data such as health data. Section 3 describes the nature of health data and presents appropriate model for it. The performances of former and new model are discussed in section 4. Subsequently, a conclusion is given in section 5.

2. Background

In the big data revolution, health sector is a notable aspect. The volume of health data is huge and increases explosively. On the other hand, these data are complex in the nature and very difficult to analyze. Medical data are generated from multiple sources in different forms. Healthcare organizations must convert the information into useful knowledge through operational studies, research, and tools. Comprehensive analysis of healthcare-related data provides desirable results that can improve the quality of decision making process. For example, side effects of some drugs are indicated via analyzing patient histories [5,16].

Most medical storage systems are based on relational databases, which are not efficient and sufficiently flexible, according to the nature of these data [5]. Recently, the use of Cloud Computing has become epidemic in healthcare. Wang et al. [17] introduced some important Big Data applications in healthcare domain including large datasets for health information systems (HIS) and clinical decision support systems (CDSS), Medical Body Area Networks (MBANs). They offered Cloud Computing infrastructures for designing and developing Big Data Analytics due to its conjunction with fast communication networks, programming models, semantic web, and machine learning algorithms [17]. In a study, Xbase, a hybrid approach based on RDBMS and Hadoop was proposed, which resulted in higher performance and lower costs. The system was implemented in a Cloud Computing infrastructure that provided infinite distributed storage and computation capability [18].

Microsoft Health Vault, Google Health, Dossia, and Mphrx are some public health management systems based on Cloud Computing. These systems are capable to store and maintain personal data of patients through an online access. So, users can retrieve their data anytime and anywhere from every device. Also, they can take care of themselves via these systems according to their health status. These cloud services are a kind of Software as a Service (SaaS) platform. Despite, Google Health has been permanently stopped from 2013, while Microsoft Health Vault, Dossia, and Mphrx are still available [19–22].

In order to process and analyze large genomic datasets, Cloud Computing and Hadoop have been introduced in the study done by O'Driscoll as proper technologies. Genome data are significant in healthcare branch that can add much more value in this sector. This model provides a distributed and parallelized infrastructure for data sets with petabyte scale [23]. Nguyen et al. proposed an approach based on the Apache HBase (a type of NoSQL DBs) and the Map-Reduce programming to store and process clinical data. This system is integrated with a web-based layer to parallelize the computation process [24].

Doukas et al. [25] presented the implementation of a mobile system based on Cloud Computing that enables electronic healthcare data storage, update and retrieval. Amazon's S3 cloud service was established in this project which provided online management of patient's data. In other study [26], the authors addressed a gap between potential and actual data usage by focusing on open, visual environments. A framework was developed for efficient use of healthcare data by integrating the MIMIC database in a RapidMiner environment. Also, Hadoop and some analytic algorithms were used for data analysis.

In a study [27], the architecture of healthcare SaaS Platform was analyzed for Decision Support Service. Microsoft's Azure was introduced as a Cloud service in this model. Vukićević et al. [28] propose a cloud-based system for the analysis of biomedical data. This system integrated meta-learning framework in order to select the best predictive algorithms and open source big data technologies for analysis. In another research, a communicational framework was proposed which related key