Liliana BYCZKOWSKA-LIPIŃSKA, Agnieszka WOSIAK

Institute of Information Technology, Lodz University of Technology

Multimedia NoSQL database solutions in the medical imaging data analysis

Abstract. *Technological capabilities of medical imaging contributes to the increasing use of image analysis in diagnostic medical systems. Medical imaging is one of the key sources of information to medical staff. Therefore, the implementation of functions associated with the medical imaging in computer systems for medical diagnosis is desirable, or even necessary. Relational databases are the dominant technology for data storage. However, the substantial growth of multimedia data indicates other solutions, primarily to enable faster data access and scalability of the system, even if it means a temporary, partial lack of consistency. This approach has gained general name of NoSQL solutions. In the study a system for rheumatologic medical data analysis for the presence of children in juvenile idiopathic arthritis was designed. The proposed solution is based on non-relational data storage in conjunction with a relational database.*

Streszczenie*. Możliwości technologiczne w zakresie obrazowania medycznego przyczyniają się do coraz szerszego stosowania analizy obrazów w diagnostycznych systemach medycznych. Obrazowanie medyczne jest obecnie jednym z kluczowych źródeł informacji dla personelu medycznego. W związku z powyższym implementacja funkcji związanych z obrazowaniem medycznym w komputerowych systemach diagnostyki medycznej jest wskazana, a nawet konieczna. Rozwiązania relacyjne stanowią dominującą technologię w zakresie przechowywania danych. Jednakże stały i znaczny przyrost danych multimedialnych powoduje, że w wybranych zastosowaniach wskazane byłyby inne rozwiązania, przede wszystkim umożliwiające niezwykle szybki dostęp do danych i skalowalność systemu, nawet kosztem chwilowego, częściowego braku ich spójności. Takie podejście zyskało ogólną nazwę rozwiązań NoSQL. W ramach przeprowadzonych badań zaprojektowany został system analizy reumatologicznych danych medycznych pod kątem występowania u dzieci młodzieńczego idiopatycznego zapalenia stawów. Zaproponowane rozwiązanie uwzględnia zastosowanie nierelacyjnej bazy danych Oracle NoSQL Database w połączeniu z bazą relacyjną Oracle Database. (**Multimedialne bazy NoSQL w analizie medycznych danych obrazowych**).*

Słowa kluczowe: NoSQL, medyczne bazy danych, obrazowanie medyczne, informatyczne systemy medyczne, analiza danych
Keywords: NoSQL, medical databases, medical imaging, computer medical systems, data analysis

Introduction

Technological capabilities of medical imaging contributes to the increasing use of image analysis in diagnostic medical systems. Medical imaging is one of the key sources of information for medical staff, which is largely due to the fact that the accuracy of conclusions drawn by the physicians from this type of the data presentation is very large in comparison with other forms (verbal description, numerical data) [2, 15]. Accordingly, implementation of the function associated with the medical imaging in computer systems for medical diagnosis is desirable, or even necessary [9, 12].

The main issue to consider while designing and implementing a system for supporting a full analysis of medical data and taking into account the medical imaging, is to choose the right data storage technology. The solutions for collecting images have changed in time with the development of information technology. Initially, images were stored in files outside databases and inside databases only their paths were collected. Then a new type of data - BLOB (Binary Large Object) was developed and therefore a possibility for storing images in the database was introduced. This enabled accessing to images as part of a single transaction in the same manner as to other types of stored data. Currently, the possibility of processing multimedia data in relational database management systems are specifically defined in the SQL/MM standard (SQL Multimedia and Application Packages) defined by ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission).

Relational solutions are the dominant technology for data storage. It should be noted, however, that the BLOB type for binary data storage cannot fully benefit from the functions of database management systems because access to the binary content from the SQL is not possible. In addition, steady and substantial growth of multimedia data makes other non-relational solutions desirable in some applications, primarily to enable extremely fast data access and scalability of the system, even if it means a temporary, partial lack of consistency. This approach has gained general name of NoSQL approach [3, 4, 8].

The remainder of this paper is organized as follows. Section 2 presents database solutions for medical purposes. Next section concerns NoSQL approach - its characteristics and classification criteria. In Section 4 we describe the studies that were conducted. We introduce data collected for this medical application and discuss the selected database architecture. Finally, in Section 5 we draw the conclusions.

Medical database solutions

Medical centers involved in research using medical imaging techniques generate approximately from a few to tens of gigabytes of image data a day. These data must be efficiently collected, processed and made available to eligible persons. Furthermore, in order to improve medical diagnosis they must be related to other patient data obtained in other medical studies. The methods of organization and storing data are key issues related to the designed system because of the need to obtain achieved results quickly and effectively. Considering the logical storage model we can distinguish the following approaches:

- hierarchical,
- relational,
- object-oriented,
- entity-attribute-value.

Hierarchical databases are the simplest and yet the easiest solution for the future implementation [6]. They have a tree structure, in which the parent node may leave many children nodes, and the child node is assigned to only one parent. This type of database was introduced as the primary solutions of the medical sector. However some elements of the hierarchical model are implemented in further applications of object-oriented and relational databases, for example XML data stored in databases. One of the most popular medical implementation of the hierarchical database is the Massachusetts General Hospital Utility Multi-Programming System (MUMPS) [1].

Relational databases are currently the most common way for data storage. In this approach the collected data are interrelated and have a fixed, rigid structure resulting from

the designed pattern. The idea of relational databases derives from Codd's theorem [5], although today's implementations have deviated greatly from the original solutions and incorporate a variety of elements from object-oriented programming languages.
Object-oriented data model has been developed to enable tighter binding storage structures with object-oriented programming languages. In medical applications, this model has been implemented, inter alia, in [16, 17] researches.
Entity-attribute-value (EAV) databases have been developed to meet the strong diversity of medical data (heterogeneous data) and became a strong alternative to the relational model [7, 11]. The concept of EAV database also involved in those cases where some data was stored in a relational database, but the majority was stored in the form of EAV [10] . It can be assumed that the EAV approach is the basis of non-relational solutions described in this paper.

NoSQL storage approach

NoSQL solutions are becoming more an more popular. They are chosen mainly by network software developers, for whom the possibility of using a distributed database is a priority. An additional advantages of non-relational databases are a relatively simple structure and implementation in the commonly used programming languages. These elements are a good basis for large data storages [14]. Moreover the non-relational solutions can provide an excellent architecture for multimedia databases.
The main characteristics for non-relational storage includes:
- avoiding joints,
- enabling or facilitating scalability,
- building a flexible storage model.

The CAP theorem formulated by Eric Brewer in 2000 (proof of this theorem was given in 2002) is considered as the foundation of NoSQL technologies [3]. The foundations of CAP principle are three concepts of distributed databases:
- Consistency, meaning that at any time, all users have the same view of the data,
- Availability, which guarantees the ability to obtain response data to all the users' requests at any time,
- Partition tolerance, which means the continuous work of the system despite the physical failures of any part of the system.

According to Brewer's theorem, only two of the three properties are possible to obtain with a distributed data storage.
Relational database management systems primarily ensure availability and consistency. Their basic characteristics can be defined by a set of ACID properties:
- Atomicity - each transaction must be executed entirely or not at all,
- Consistency - ensuring the integrity of the database, bringing the database from one valid state to another valid state,
- Isolation - one particular transaction is independent of other concurrent transactions,
- Durability - after the transaction, the changes are saved permanently.

Non-relational storages put the emphasis on the capabilities of partitioning, even at the expense of integrity or availability. The requirements for non-relational databases are defined by the set of BASE properties:
- Basically Available - the system is generally available, but availability is not guaranteed,
- Soft state - the state of the system may change over time, even without the new data entries,
- Eventual Consistency - ultimate consistency: temporary database inconsistency is accepted, consistency is restored with a delay.

An important and main advantage of non-relational solutions is the speed of access to data which comes from the fact of storing the data with simple relationships between them and avoiding joins, which in the case of relational databases are the fundamentals for the schemas' designing.
The developers offer many non-relational storage solutions and the NoSQL market is extremely dynamic. There are new databases, and some of the already existing ones are withdrawn from commercial and non-commercial offers. Currently, users of NoSQL databases can choose from more than 150 solutions. This amount in comparison to the number of relational database systems (less than 10) is significant, and therefore the decision of choosing a particular solution is not easy. Each of them has a different approach to data storage. To help with choice of NoSQL database, the classification of all the non-relational databases has been performed taking into account certain criteria described and compared in table 1.
There are four basic types of NoSQL databases [8, 13]:
- key-value store,
- column family,
- document store,
- graph.

The architecture of all the mentioned above groups of non-relational databases affects their ability to complexity, performance, flexibility and scalability.

Tabel 1. The comparison of the basic properties for relational and non-relational databases.

Database type	Complexity	Efficiency	Flexibility	Scalability
Relational	high	medium	low	medium
Key-value	low	high	high	high
Column family	low	high	medium	high
Document store	low	high	high	high
Graph database	high	medium	high	high

The misleading term "NoSQL" is usually interpreted as "no to SQL". However, the idea underlying attitudes of non-relational databases indicates the development of that term as "Not Only SQL". This means that both methods of database management can and should complement each other, depending on the needs and applications [13]. Each of the methods has its advantages and disadvantages of data storage, which incorporation allows better meeting the chosen technology to the needs of the implemented system and its future users.

The basic elements to consider while selecting a relational solution or NoSQL, are the type of data to be stored and the application schema. NoSQL is a better solution when we deal with a large amount of data with a lower value. For the important (valuable) data it is better to choose RDBMS. This selection stems from the fact that relational databases offer a full integrity and availability of data and well-defined standards. , NoSQL solutions provide a very good basis for storing large amounts of data offering the distribution and partition tolerance. Considering application schema it can be concluded that the dynamic structure of the heterogeneous data and continuous changes indicate non-relational solution. For databases with fixed schema it is better to implement solution using relational database management system.

Experimental studies

The module of the system for rheumatologic medical data analysis for the presence of children in juvenile idiopathic arthritis was designed for the purpose of this research. The proposed solution includes the use of non-relational approach as Oracle NoSQL Database in conjunction with traditional relational database - Oracle Database 11g.

The data obtained in the diagnosis of patients include: personal information, laboratory tests, imaging studies as DICOM files and related descriptions and the results of additional studies gained by specialist consultation (usually rehabilitation and ophthalmology).

There are the following requirements for data acquisition and managing in the designed system:

- for each of the institutions all the treated patients' data should be stored in the system,
- the access to patient data is permitted to healthcare professionals (physicians, nurses, laboratory technicians), call-center staff and patients themselves,
- the system should include the ability to store sensitive data, but also the specific data required at the institution, depending on its policies,
- the conclusions should be drawn using all the data, including medical imaging metadata,
- the access to patient data and the authorization process should be as soon as possible.

All these requirements show that a relational database should be considered, because of the overall data consistency and availability. On the other hand NoSQL solutions offer greater opportunities for distributed storage and rapid access to selected data. Moreover they are a very good choice for multimedia storage. For these reasons it was decided to build a hybrid database consisting of a relational database management system Oracle Database 11g and Oracle NoSQL Database [15].

The process of data integration within the hybrid database can be achieved by:

- loading data from NoSQL database to a relational database,
- loading data from a relational database to NoSQL database,
- use separate software to manage each database and treat them as separate data storages.

The first approach offers the possibility of obtaining data from non-relational storage to relational database. Such an integration may be carried in the materialized way or just virtually. Using the virtual data availability of NoSQL database, the extraction of data is possible only after the the following query: a connection to a relational database is opened and it allows you to read data from the non-relational database. The greatest difficulty in this approach is the representation of non-relational unstructured data to gain the opportunity to read them through the SQL database. However, it is very useful solution due to the fact of the standardization for relational database management systems, and therefore provides a good fundamentals for hybrid solutions. In addition, the relational systems due to their maturity, usually provide additional mechanisms for the integration of data from different sources. Undoubtedly for software developers already working with relational systems, it is the most appropriate solution.

The second method indicates a non-relational database as the primary, which means that this is the source of the data to another application layer. Data from relational databases are loaded (usually virtually) to NoSQL database and are selected by the appropriate built-in query languages. Typically, such an approach is simpler than the first method, since the NoSQL databases allow for storage of any type of data: a structural or not. As the result, the problem of data transformation which may occur in the first method, here does not exist. The only question to be considered is the mechanism for loading data from a relational database to non-relational database, which must be carried out with higher layer NoSQL software.

The third way to obtain data from hybrid databases does not require the process of synchronization of operations, which is a main advantage. In this approach there is one abstraction layer for both the databases: relational and non-relational. The task for this layer is:

- to parse the query,
- to state from which database (relational or non-relational) the data should be retrieved,
- to select the appropriate data, and
- to combine them into one result presented to the user.

As part of the solution, an abstraction layer is defined. It connects the SQL and NoSQL databases, allows you to extract the data from both sources and generate consolidated results, depending on the given request. The diagram of search process in the hybrid database is shown in Figure 1.

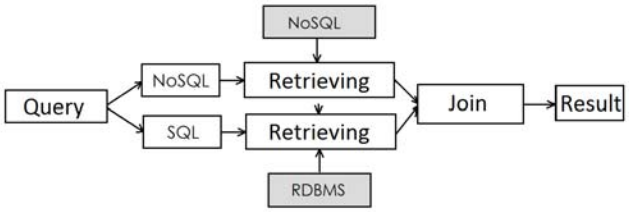

Fig 1. Search process for a hybrid database solution.

In the proposed system, most of data are stored in a centralized database, Oracle Database 11g. Each node of NoSQL storage is mainly used for storing multimedia content. Besides it enables loading data into the database, and for a quick search on the basic attributes of the data, especially using the patient's identifier. The advanced search is performed using a relational database, that offers mature indexing techniques to speed up this search process. Therefore, the query processor depending on the attributes specified in the QBF query (Query By Feature) uses the appropriate data source.

The process of loading data into both databases is performed in a single transaction. Moreover, there is a possibility to copy or move data for NoSQL database node to a relational database (and vice versa) in stated periods of time, depending on the predefined parameters. These processes of synchronization and data replication are preformed automatically, which is a characteristic feature for the active databases and can be implemented using triggers and jobs defined in the database.

Conclusions

NoSQL solutions are becoming more and more popular due to a relatively simple structure and implementation of the widely used object-orientd programming languages. The problem of complete integrity and availability of data remains unsolved, however taking into account the specific uses of these databases, it can be considered acceptable.

Considering the choice of non-relational techniques, it is necessary to take into account the fact that the market of non-relational solutions is very dynamic and it makes some of the solutions not to be developed any longer, which in turn can generate a nuisance for the system working without any technical support. An additional complication in many cases is negligible technical documentation, lack of specialized literature and experienced professionals.

However, there is a chance that the access of high experienced database providers like Oracle Corporation to the non-relational database developers will make NoSQL future users being able to choose long-term solutions. In addition, non-relational databases offer new opportunities, as discussed in this paper. The advantage of the NoSQL solutions is their full integration with the programming languages, which greatly simplifies the process of defining the logic layer of the application.

Experimental studies conducted as a part of this work have shown the ability to implement a relational and a non-relational databases in the medical system. As a result it enabled acquiring the best properties of both technologies in one application. The choice of solutions from the same provider was intentional, but the hybrid solutions can virtually connect any relational database management system data with a NoSQL database.

The benefits that can be achieved using of hybrid solutions are currently under study, due to the short time of existing the non-relational databases. However, it can be already concluded that this approach deserves the attention of relational and non-relational researchers seeking new and faster solutions, as well as customers of future database systems.

REFERENCES

[1] BARNETT G.O.: Massachusetts general hospital computer system, chapter in Hospital computer systems (ed. Collen M.F.), 1974 Wiley

[2] BERNER E.S.(ed.): Clinical Decision Support Systems, Theory and Practice, 2007 Springer Science + Business Media, LLC

[3] BREWER E.A.: Towards Robust Distributed Systems, Keynote at the ACM Symposium on Principles of Distributed Computing Portland, Oregon, July 2000

[4] CATTELL R.: Scalable SQL and no-SQL Data Stores, SIGMOD Record - web edition, December 2010

[5] CODD E.F.: A relational model of data for large shared data banks, pp: 377-387, Comm ACM 1970:13

[6] COLTRI A.: Databases in health care, chapter in Aspects of electronic health record systems (eds. Lehman H.P., Abbott P.A., Roderer N.K., et al.), pp: 225-251, 2006 Springer

[7] DINU V., NADKARNI P.: Guidelines for the effective use of entity-attribute-value modeling for biomedicine databases, International Journal of Medical Informatics, pp: 769-779, Elsevier 2007:76

[8] JING HAN: Survey on NoSQL database, , 2011 6th International Conference on Pervasive Computing and Applications (ICPCA), 2011, pp: 363-366

[9] MARCOSA E., ACUNA C.J., VELA B., CAVEROA J. M., HERMANDEZ J.A.: A database for medical image management, Computer methods and programs in biomedicine, vol. 86, pp: 255-269, 2007 Elsevier Ireland Ltd

[10] NADKARNI P.M., MARENCO L.: Easing the transition between attribute-value databases and conventional databases for scientific data, Proceedings of American Medical Informatics Association, 2001, pp: 483-487

[11] NADKARNI P.M., MARENCO L., CHEN R., et al.: Organization of heterogenous scientific data using the EVA/CR representation,Journal of the American Medical Informatics Association 200:7 pp: 343-356,

[12] National Electrical Manufacturers Association: Digital Imaging and Communications in Medicine (DICOM), 2009

[13] SCOFIELD B.: NoSQL – Death to Relational Databases, CodeMash Presentation, January 2010, USA

[14] SKURZOK D., ZIÓŁKO B.: Special Key-Value Store - Header Only Database for N-gram Models, Journal of Applied Computer Science, 2012, vol. 20 no. 2, pp: 119-129

[15] TADEUSIEWICZ R.: Informatyka medyczna, Uniwersytet Marii Curie-Skłodowskiej w Lublinie, Instytut Informatyki, Lublin 2011

[16] WIEDERHOLD G.: Database technology in health care, Journal of Medical Systems, pp: 175-196, 1981 Springer

[17] WIEDERHOLD G., WALKER M.G., BLUM R.L., et al.: Acquisition of medical knowledge from medical record, Proceedings of Benutzergruppenseminar Medizine Systems, Munich 1987

***Authors**: prof. dr hab. inż. Liliana Byczkowska-Lipińska, Politechnika Łódzka, Instytut Informatyki, ul. Wólczańska 215, 90-924 Łódź, e-mail: liliana.byczkowska-lipinska@p.lodz.pl, dr inż. Agnieszka Wosiak, Politechnika Łódzka, Instytut Informatyki, ul. Wólczańska 215, 90-924 Łódź, e-mail: agnieszka.wosiak@p.lodz.pl*