INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

IJARnD

(Volume2, Issue5)

Available online at www.ijarnd.com

Bigdata: A Survey On RDBMS And Various NOSQL Databases On Storing Medical Images

Divya .S , N. Shivaprasad

[1] *Student, Department of ECE, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, India.*

[2] *Asst Prof, Department of ECE, Sri Jayachamarajendra College of Engineering, JSS Science and Technology University, Mysuru, India.*

[1] *divyasathya05@gmail.com*, [2] *nagshivu@gmail.com*

ABSTRACT

Bigdata is a term that describes a large volume of data, which is gaining high importance in the medical IT industry. With the increased volume of medical data in due course, the effort of storing large medical data shoots up. According to the survey, about 70% of medical data consists of medical images. Every patient likes to know their medical history and storing that is a costly operation. Traditionally the medical industry stores their medical images in a system called Picture Archiving and Communication System (PACS) which uses the Relational database management system (RDBMS). With the advent of NoSQL databases, one can also look into this for storing and archiving their medical images. In this view, a survey on RDBMS and various NoSQL databases for storing the medical images is being documented by us in this paper.

Keywords: Bigdata, Picture Archiving and Communication System (PACS), Relational database management system (RDBMS), NoSQL, Digital Imaging and Communication in Medicine(DICOM).

I. INTRODUCTION

The IT industry has given solutions to various problems in different sectors in the society. Healthcare being the most important sector and the need for having collaboration with the IT has gone to the peak. Healthcare is one of the fields with the highest Big Data potential. The volume of data that is being generated through various processes in the healthcare Industry has become unmanageable. Despite the development in the database technology, the healthcare industry has seen many challenges with regards to storing data [8].

Experts predict a substantial increase in the volume of Medical images stored in the near future. It is estimated that in the future, 30% of world's storage will be related to health informatics, and mainly the medical images. Research forecasts that the market for medical imaging systems will grow to $49 billion in 2020. The volume of medical images stored has exceeded 1-exabyte mark today, which takes medical imaging into Big Data territory [6, 17].

There is a wide range of applications for Hospital Information Systems (HIS) and Radiology Information Systems (RIS). In both these cases, there is the medical image management activity, such as MRI, X-Rays, CT scans, among other types of examinations. These examinations generate as a practical result, one or a set of images that assist the physician in the diagnostic process about any potential disease. Still, in order to assist the management and communication of these images, the concept of PACS was inserted, which are a software component that focuses on managing medical images [1,3].

As we know PACS uses the traditional RDBMS strategy for storing and retrieving the medical images. So, with the advent of different NoSQL databases, a large scale distributed approach for storing medical data is possible. This paper gives a survey report on storing medical images on RDBMS and various NoSQL databases.

II. CLASSIFICATION OF DATA IN MEDICAL BIG DATA

The review through literature surveys on medical data identifies the patient care is destined for an electronic future, where more and more patients check in via electronic media like tablets, cell phones etc. and patients requests their medical information in an easily transportable format, resultant constitutes the 3V's of big data. Figure 1 shows the categories of data which fall under medical big data.

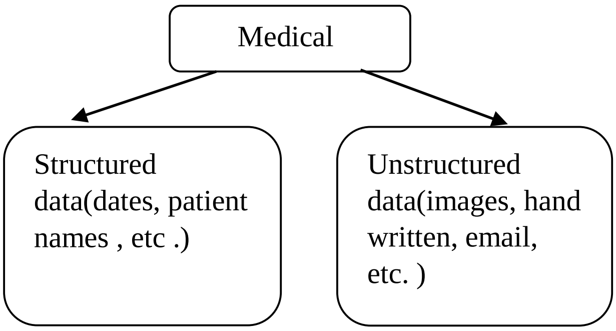
Medical
Structured data(dates, patient names , etc .)
Unstructured data(images, hand written, email, etc.)

Figure 1: Categories of data

A. Structured data

Data organized into specific fields as part of a scheme, with each field having a defined purpose. Data can be structured within each field(s) through data validation by enforcing the use of a standardized data format or allowing only a specific range of values entered in a field. Structured data information such as dates, patient names, identification numbers and diagnosis codes is easier to collect and exchange between systems, because it is standardized, pre-defined, computer-readable and typically quickly accessible from a database.

B. Unstructured data

Unstructured data is the information that typically requires a human touch to read, capture and interpret properly. Data that cannot be easily organized using pre-defined structures. It includes machine-written and handwritten information on unstructured paper forms, audio voice dictations, email messages and attachments, and typed transcriptions--to name a few. Sources of unstructured data in healthcare organizations Medical claims, Explanation of Benefits (EOB), Invoices and purchase orders, shipping documents, scanned medical reports, signed patient consent forms, handwritten notes, drawings, diagnostic images, voice dictation, email messages, email attachments, text messages, blog posts, tweets, online video, web pages, documents from other organizations.

III. MEDICAL IMAGES AND RDBMS

The following subsections discuss the medical images storage using traditional RDBMS technology.

A. Medical images

In the last 40 years we have seen an exponential growth in the use of digital systems in medicine, especially figured in the form of modern medical equipment and computer systems in general. In the context of medicine, currently, hospitals, medical centers, and clinics can utilize the support of medical equipment based on images to assist the diagnosis of their patients. Among these devices, one can cite as examples the type of CT, MRI, X-Ray, and others.

To support this enforcement, at the end of the 70s, the American College of Radiology (ACR) and the National Association of Electrical Manufacturers (NEMA) have teamed up to create standards to transfer medical images and information exchange between different equipment from different manufacturers, thus seeking, better interoperability. Thus, the pattern appeared in 1985 called DICOM Figure 2 shows the screenshot captured using the itksnap viewer, Figure 3 and 4 are the coronal and sagittal view respectively. All these are the examples of dicom images.

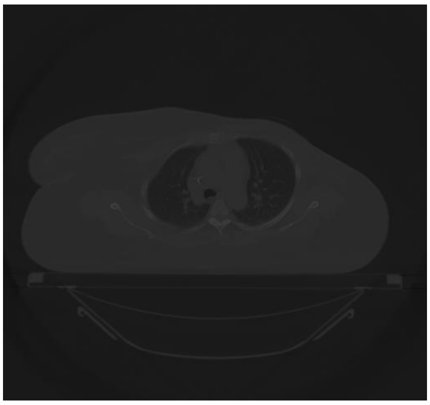

Figure 2: DICOM image

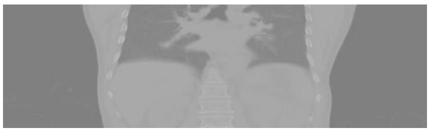

Figure 3: Coronal view

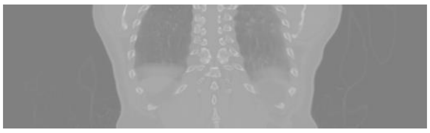

Figure 4: Sagittal view

B. Dicom File Format

The file format of the DICOM standard was described based on Part 10 of the standard specifications. The group of bytes of the Data Set, the image itself, is placed after the DICOM File Meta Information Header. This header is mandatory in all DICOM files and includes the identification of the Data Set. It is composed of a preamble file of 128 bytes, a prefix DICM of 4 bytes and the DICOM File Meta Information Header. The preamble can be used by a specific implementation and its content is not defined by the DICOM standard The DICOM File Meta Information Header is composed of a group of attributes. Each attribute has a label and an associated length of its content. The label is represented by two hexadecimal numbers, the first is the Group Number and the second is the Element Number inside of this Group. The length of the attribute content is represented by four hexadecimal numbers.

C. PostgresSQL

It is a Database Management System (DBMS) which is well known in the academic circuit. It is developed at the University of Berkeley where the project was led by Prof. Michael Stonebreaker. He is one of the most prominent scientists in the database research area. The project ended in the early 90's, two students decided to add support to the SQL language creating the PostgreSQL. It is fully Atomicity, Consistency, Isolation, and Durability (ACID) compliant.

PostgreSQL runs stored procedures in many programming languages, like Java, Python, C/C++ and much more. There are many interfaces available like JAVA (JDBC), ODBC.

D. Oracle

It is an RDBMS (Relational Database Management System) which is well known in commercial circuit. Oracle stores data logically in the form of tablespaces and physically in the form of data files. It is not fully ACID compliant. Oracle 12C has extended their support to unstructured data management [19].

E. PACS and RDBMS

PACS have been adapted incrementally over time to accommodate the evolution of medical imaging: for the increasing volume of medical data, storage capacity has been added accordingly by many database providers, and user-requested functionality has been implemented to allow simultaneous viewing at multiple radiology sites, both locally and remotely. However, the image archiving architecture has not been changed substantially over the years. PACS generally have relied on relational, schema based Structured Query Language (SQL) databases for medical images data management. RDBMS have a fixed design in which a schema of tables and relations between those tables (“joins”) is defined at the outset and any data inserted thereafter must conform to the predefined schema. To improve the speed of data retrieval operations, an “index” (a list of values extracted from a particular column in a table and stored in a format that allows faster access) is usually created during database design and implemented at the expense of storage space [19]. Although the most the RDBMS providers are robust which supports the PACS architecture they may be no longer enough for the distributed approach.

IV. NoSQL DATABASES TO STORE MEDICAL IMAGES

The relational approach is a dominant technology for data storage. However, the BLOB type for the binary data storage doesn't benefit the functions of database management system completely. Because the access to the binary content from the SQL is not possible. Above all, there is a substantial growth in medical images that makes the non-relational solutions desirable in some applications, primarily to enable fast data access, scalability, consistency etc. This non-relational approach evolved to be as a NoSQL solution.

A. NoSQL approach

NoSQL databases are non-relational databases which do not follow a strict schema. NoSQL Databases are advantageous over RDBMS

i) The data is available with redundancy across one or more locations.
ii) It can run over multiple data centers and its cloud-enabled.
iii) It has very good write speed and low latency query speed

iv) Supports scale-out architecture where it is possible to add more processing power and storage capacity can be increased. It is highly scalable [6].

And this NoSQL is formulated by CAP and BASE theorem. The CAP principle has three concepts as described in Figure 5

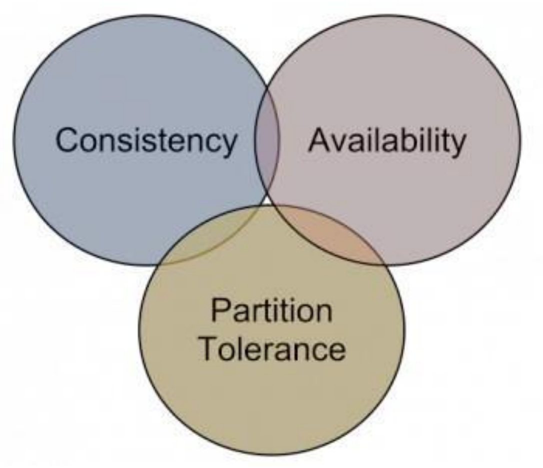

Figure 5: CAP Theorem

i) Consistency meaning that all the users get to read the most recent write anytime. Achieved by replicating the data across different machines.
ii) Availability meaning that guarantees to obtain the response to all the users anytime. Updating several nodes before allowing further reads.
iii) Partition-tolerance continuous work despite any physical failures in the system.

According to Brewer's theorem, only two of the three properties are possible to obtain with a distributed data storage. The requirement for the non-relational databases are emphasized by BASE properties which are defined as follows

i) Basic Availability meaning that the database appears to work most of the time, generally available but not guaranteed.
ii) Soft-state meaning that, stores don’t have to be write-consistent, nor do different replicas have to be mutually consistent all the time.
iii) Eventual consistency meaning that stores exhibit consistency at some later point (e.g., lazily at the reading time).

BASE properties are much looser than ACID guarantees, but there isn’t a direct one-for-one mapping between the two consistency models.

There are four basic types of NoSQL databases [15]:

i) Key-Value store
ii) Column based store
iii) Document based store
iv) Graph database

The classification of all non-relational databases has been performed taking into considerations of certain criteria described and compared in Table 1

Table 1: Comparison of basic properties between relational and non-relational databases

Database Type	Complexity	Efficiency	Flexibility	Scalability
Relational	high	medium	low	medium
key-value	low	high	high	high
Column	low	high	medium	high
Document	low	high	high	high
Graph	high	medium	high	high

Below briefs about the four basic types of NoSQL database

1) Key-Value Store: the name itself states that it is a combination of two things that is key and value. It is one of the low profile database systems. Key-Value databases are the mother of all databases of NoSQL. Key is a unique identifier to a particular data entry. Value is kind of data that is pointed by a key. Examples for Key-Value stores are Dynamo, Redis, Riak etc.

2) Column based store: This is based on Google's Big Table store. Here the data is stored as sections of columns of data and not as rows of data. It is not required to define the columns at the beginning. There can be a countless number of columns that can be grouped as super columns for a distributed approach. Cassandra and HBase are some of the examples for column based store.

3) Document based store: as the name suggests, designed to manage and store documents. These documents are encoded in a standard data exchange format such as Extensible Markup Language (XML), JavaScript Object Notation (JSON). Unlike the simple Key-Value store which is described above, the value column in document database contains semi-structured data which are attributes. A single column can accommodate hundreds of such attributes. Both keys and values are searchable in document based stores. MongoDB and CouchDB are examples for document based store.

4) Graph database: Graph databases replace relational tables with structured relational graphs of interconnected key-value pairings. They are similar to object-oriented databases as the graphs are represented as an object-oriented network of nodes (Conceptual objects), node relationships ("edges") and properties (object attributes expressed as key-value pairs). They are only of the four NoSQL types that concern themselves with relations, and their focus on the visual representation of information makes them more human-friendly than other NoSQL DMS. Neo4j is one such example.

B. Description on MongoDB and CouchDB

This section describes an example of the document based and column-based stores such as MongoDB and CouchDB Respectively.

1. MongoDB

The documents are grouped together as collections. Collections are similar to relational tables. MongoDB uses a BSON format to store documents. BSON is a binary way of JSON-type representation. Here the structure is similar to a nested set of key/value pairs. BSON supports more data types like regular expression, binary data, and date. BSON helps in storing and exchanging data; Also BSON helps in describing the contents in a given document. Due to this, it is not needed to specify the structure of the document in advance. JSON can be regarded schema-less as the documents can be updated individually or changed independently of any other documents. The performance of MongoDB is enhanced due to BSON. It even makes the processing and searching faster. BSON stores data in Binary and all objects in BSON is a set of key/value pairs. Each document in MongoDB is identified using a unique identifier called the _id key [6]. These characteristics of MongoDB makes it suitable for storing humongous multimedia data which refers to the medical images eventually.

2. CouchDB

It is an open source document based data store where JavaScript is used for querying and indexing. This makes appropriate for medical imaging data management by its characteristics like operational reliance on common web standards, managing attachments to the documents, data replication, data mining applications such as dose monitoring quality assurance and protocol optimization. CouchDB provides RESTful JSON Application programming interface (API) that can be accessed through any environment using HTTP requests. REST relies on HTTP protocols for basic create, read, update and delete (CRUD) operations. Since CouchDB uses standard HTTP requests for all its operations, it is instantly portable to HTTP implementations, which are available for every programming language like in JAVA spring framework supports for RESTful implementations. In addition, CouchDB acts as its own high-performance Web server, providing front-end functionality with direct coupling to the database, with no middleware like Nodejs.

V. RELATED WORKS

The survey papers have served as a basic source for gaining the knowledge regarding the various approaches with respect to storing medical images using RDBMS and different NoSQL databases.

Simón J. Rascovsky, Jorge A. Delgado, Alexander Sanz, Víctor D. Calvo, Gabriel Castrillón [19], have proposed how a document based datastore like CouchDB provided an efficient storage of all DICOM objects. They also conclude that "with the use of information retrieval algorithms such as map-reduce, all the DICOM metadata stored in the large database were searchable with only a minimal increase in retrieval time over that with the traditional database management system" and "Results also indicated possible uses for document-based databases in data mining applications such as dose monitoring, quality assurance, and protocol optimization" [19]. They also provided a schema and proved that "All data generated at any step in the fulfillment of a radiologic imaging order can be stored in the CouchDB database, regardless of whether the data are structured or unstructured, textual digital (binary)" [19].

Luís A. Bastião Silva, Louis Beroud, Carlos Costa and José Luis Oliveira, compared several NoSQL solutions for medical imaging archiving [14].They compared the performance of the several solutions: Lucene (with the file system), and MongoDB and CouchDB, both with and without a file system. The main purpose was to compare the performance of different plugins, with different modalities [14]. They concluded that "MongoDB proved to be a very good solution to index the DICOM metadata, allowing quite fast storage and retrieval of information. Nonetheless, Lucene is the best solution if you want to query in every DICOM tag or if you are using free text query. CouchDB indexing, on the contrary, will be a great solution if you already know every kind of query that will be processed. And they both have some drawbacks when storing big files, which happens in a few modalities, such as XA or high-resolution mammography" [14].

Liliana, Agnieszka WOSIAK, provided a hybrid database for the medical imaging analysis [17]. The hybrid database consisting of a relational database management system Oracle Database 11g and Oracle NoSQL Database. The process of data integration with the hybrid database has three approaches "(1) loading data from NoSQL database to a relational database, (2) loading data from a relational database to NoSQL database, and (3) use separate software to manage each database and treat them as separate data storages." They concluded 3rd approach was more advantageous "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" [17].

D.Revina Rebecca, Dr.I.Elizabeth Shanthi analysed the suitability of storing medical images between MongoDB and Cassandra [6]. They used a concept called chunked storage in both MongoDB and Cassandra. They concluded that "the time complexity of Cassandra is less when compared with MongoDB for smaller files. But as the file size increases time complexity of MongoDB remains constant comparatively, so for a larger file, MongoDB seems to be a better candidate. In Cassandra, the time increases proportionally with the size of the file. Both MongoDB and Cassandra may be suitable to store Large Medical images. But MongoDB will be a better candidate" [6].

D.Revina Rebecca, I.Elizabeth Shanthi , proposed a NoSQL solution for efficient storage and retrieval of medical images [7]. They said that "RDBMS would be the worst fit to store medical images. NoSQL databases may be a better solution. The solutions given to store medical images are basically based on RDBMS. In the search for a better alternative to store medical images a comparative study of the performances with respect to storage and retrieval was done for both MYSQL and MONGODB. The time complexity was studied in i3 and i5 processors". They concluded that "the time for storing and retrieval in MONGODB was consistently lesser, even when the size of the images increased" [7].

N. D. Evangelista, J. F. Camapum, and E. Amemiya proposed the development of communication and storage of medical images in the database [9]. Proposed solution has different databases like PostgresSQL, Firebird, and Oracle. The major advantages of digital storage are online availability, random access to images and study data, department-wide analysis facilities and inter-departmental interchange by media or by wire [9].

CONCLUSION

The increased volume in medical images over the past decade makes the medical IT industry to analyze the suitability of databases to store them. This survey emphasizes more on the performance, time complexity of RDBMS and various NoSQL databases in storing the medical images. This Paper documents on the adaptability of various databases for storing multimedia according to the requirements.

REFERENCES

[1] Douglas D. J. de Macedo, Douglas D. J. de Macedo (2015): A Data Storage Approach for Large-Scale Distributed Medical Systems, 2015 Ninth International Conference on Complex, Intelligent, and Software Intensive Systems.

[2] D. Revina Rebecca, I. Elizabeth Shanthi (2017): A Hybrid Data Model to Share Medical Images, International Journal of Computer Applications (0975 – 8887) Volume 161 – No 9, March 2017.

[3] Leigh S. Shuman (2011): A Revolution In Radiology: PACS 101 (Part 2), The Journal of Lancaster General Hospital, Fall 2011, 78 Vol.6 – No. 3.

[4] Leigh S. Shuman (2011): A Revolution In Radiology: PACS 101 (Part 1), The Journal of Lancaster General Hospital, Summer 2011, 78 Vol.6 – No. 2.

[5] Mehmet Zahid Ercan, Michael Lane (2014): An Evaluation of NoSQL Databases for Electronic Health Record Systems, 25th Australasian Conference on Information Systems 8th -10th Dec 2014, Auckland, New Zealand.

[6] D. Revina Rebecca, I. Elizabeth Shanthi (2017): Analysing the suitability of storing Medical Images in NoSQL Databases, International Journal of Scientific & Engineering Research, Volume 7, Issue 8, and August-2016.

[7] D. Revina Rebecca, I. Elizabeth Shanthi (2017): A NoSQL Solution to efficient storage and retrieval of Medical Images International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016.

[8] Rakesh Kumar, Florella, Anna Fernandes (2015): Challenges in Storage and Retrieval of Healthcare Data: Review of various NoSQL Technologies, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Special Issue 7, October 2015.

[9] N. D. Evangelista, J. F. Camapum, and E. Amemiya (2005): Communication and Storage of Digital Medical Images in Database, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China, September 1-4, 2005.

[10] Am Suk Oh1, Oh Hyun Kwon2 and Gwan Hyung Kim2 (2014): Design and Implementation of Standard DICOM Interface Module, International Journal of Bio-Science and Bio-Technology Vol.6, No.2 (2014), pp.141-146.

[11] Ahmad Fadzil M Hani, Irving Vitra Paputungan, Mohd Fadzil Hassan, Vijanth S Asirvadam, Megat Daharus (2014): Development of Private Cloud Storage for Medical Image Research Data,2014 IEEE.

[12] Gunjanbhai Patel (2012): DICOM Medical Image Management the Challenges and Solutions: Cloud as a Service (CaaS), ICCCNT'12 26th _28th July 2012, Coimbatore, India, IEEE-20180.

[13] Mehul Nalin Vora (2011): Hadoop-HBase for Large-Scale Data, 2011 International Conference on Computer Science and Network Technology, 2011 IEEE.

[14] Luís A. Bastião Silva, Louis Beroud, Carlos Costa and José Luis Oliveira (2011): Medical imaging archiving: a comparison between several NoSQL solutions, Conference Paper · June 2014, 2014 IEEE.

[15] Liliana BYCZKOWSKA-LIPIŃSKA, Agnieszka WOSIAK (2013): Multimedia NoSQL database solutions in the medical imaging data analysis, PRZEGLĄD ELEKTROTECHNICZNY, ISSN 0033-2097, 2012/2013.

[16] Olaniyi, Olayemi Mikail, Omotosho Adebayo, Robert Jane, Oke Alice.O(2013): Development of an Electronic Medical Image Archiving System for Health Care in Nigeria International Journal of Computer and Information Technology (ISSN: 2279 – 0764) Volume 02– Issue04, July 2013.

[17] Yang Jin, Tang Deyu, Zheng Xianrong(2011): Research on the Distributed Electronic Medical Records Storage Model, 2011 IEEE.

[18] Sarmad Istephan, Mohammad-Reza Siadat (2016): Unstructured medical image query using big data – An epilepsy case study, Journal of Biomedical Informatics 59 (2016) 218–226.

[19] Simón J. Rascovsky, MD, MSc, Jorge A. Delgado, MD, Alexander Sanz,

BS, Víctor D. Calvo, BS, Gabriel Castrillón, BS (2011): Use of CouchDB for Document-based Storage of DICOM Objects, 1From the Department of Research, Instituto de Alta Tecnología Médica de Antioquia, Cra 50 #63-95, Medellín, Colombia. Presented as an education exhibit at the 2010 RSNA Annual Meeting. Received March 11, 2011.

[20] HIMSS HIE inPractice Task Force (2016): Blending Structured and Unstructured Data to Develop Healthcare Insights.

[21] Datamark(2013): Unstructured Data in Electronic Health Record (EHR) Systems: Challenges and Solutions, Healthcare Content Management White PaperOctober 2013.

[22] Oracle (2016): Unstructured Data Management with Oracle Database 12c ORACLE WHITE PAPER | NOVEMBER 2016.