Performance Evaluation of RDBMS with Several NoSQL Databases in aspect of Medical Images

Rupal Parekh[1], Dr. Achyut C. Patel[2]

[1]Department of Computer Science, Saurashtra University
[2]Smt. M. T. Dhamsania Commerce College, Saurashtra University
[1]rbparekd@gmail.com, [2]acp2809@gmai.com

Abstract

The use of digital medical imaging systems has greatly increased in healthcare institutions and they are currently valuable tools supporting medical decision and treatment procedures. NoSQL databases have been replacing relational databases in some scenarios, due to their horizontal scalability and to their flexibility to adapt to dynamic requirements. This research focus on DICOM images, as unstructured data as it is the standard used in healthcare industries. The fitness of NoSQL Databases in view of addressing all the challenges revolving around digital imaging is to be studied and analyzed. In this context, the performances of the Relational Data Model and the NoSQL Databases in storing and retrieving Medical Images are to be analyzed. We consider the many methods in handling digital images. For the experiment here in this research we consider MySQL as a RDBMS and MongoDB and Cassandra as NoSQL databases. We designed a set of experiments with a huge number of various aspects for storing and retrieving of digital images for the comparing two databases by the same data for RDBMS and NoSQL. The results show that NoSQL data model performs better for storing and retrieval of medical images.

Keywords*—NoSQL; SQL;RDBMS; Big data; Database; MongoDB; Cassandra; DICOM; PACS; Performance evaluation*

1. INTRODUCTION

In today's era Digital imaging laboratories are dealing with large amounts of data which tends to increase, since the full patient’s history might help medical doctors in future diagnosis. In order to support this development and to manage a huge amount of data, several technologies have been becoming more attractive and gaining higher acceptance [1,2]. One such example has been the emergence of NoSQL (Not only SQL) databases, such as Cassandra, CouchDB, BigTable, HBase, MongoDB, Redis. These technologies were developed to handle large amount of schema-less data and to provide high performance. At the same time, they have been increasingly adopted by industry and research applications [4].

Conventional databases are based on the relational model for storing data (RDBMS), and they were named the SQL databases after that the SQL query language was used for querying them [3]. However, in recent years non-relational databases that are known as NoSQL databases, have been highly in trend and research centric .

In today's era a major part of the data produced is vast and unstructured. The Data produced is stored into the cloud for the many reasons. RDBMS is insufficient to handle unstructured data and data management in the cloud having drawbacks . The researcher’s forecast in [5] regarding data challenges expressed by RDBMS has led to a new type of non-relational databases known as NoSQL. Many researchers have argued the data handling challenges in the cloud and specially the incompatibility of RDBMS in the cloud environment [6[7][8].

The data records in a relational model database are represented as a schema. The related data recorded in this structure are grouped in tables as rows and columns, and each row has the same number of columns. The relational database tables help the designer to avoid data redundancy when the tables are normalized. The normal result is a multi-table design. The queries usually require a combination of the tables and merging the information in the tables. For this reason, the join operation should be used. Join operations need to define foreign keys and this imposes a large overhead to the database.

The key advantage of a NoSQL Database is its distributed structure which is complete contrary to a Relational Model. There is no need schema definition to be predetermined for unstructured data. The ACID transaction properties of the traditional SQL databases are ignored or receive less attention in NoSQL databases. Database like MongoDB, the ACID properties are replaced by the BASE architecture. On the other hand, the concepts of joins and transactions are not supported in NoSQL databases due to their specific architectures.

In this research focus is in the comparison between well-known document-oriented type of NoSQL databases i.e. MongoDB and column oriented database i.e. Cassandra with the MySQL as a relational database individually. We examine the fundamental differences between these two databases management systems in terms of performance of storing and retrieving the digital medical images. Our selected queries are run on the same dataset with the same number of records.

2. RELATED WORKS

Medical images are stored in Picture Archiving and Communication Systems (PACS). In order to support common data and communication formats, when handling medical images between different devices and vendors, the DICOM standard was created. Currently, any equipment in medical institutes follows the DICOM standard to communicate, store, and visualize medical data. As a consequence, PACS require robust information and communication infrastructures to ensure that all these devices communicate in a secure and timely manner. The existing data storage capabilities are not able to satisfy the needs of this massive amount of medical imaging data. This is a huge challenge for healthcare organizations where it is a big struggle to share, manage and access this data in less cost [9].

Healthcare applications will be dealing with large binary files which are results of various tests done on the patient from various medical imaging devices. It can be radiology test results, x-ray and MRI images, CT scans. This may also include medical records created by scanning paper documents. It is not uncommon for a large hospital or a medical insurance provider to have 50-100 TB of patient data, out of which 90% of the information are unstructured binary data. The healthcare providers look for new methodologies and paradigms to store these huge volumes of unstructured and semi structured data. Handling medical images poses the following challenges [10] :(a) Handling Different types images, (b) Handling huge sized images.

(a) Handling Different types images

The medical images are from different sources which may be of different formats [6,7,8].The medical image files sourced from different modalities may be either in DICOM format or in raw format. Picture Archiving and Communication Systems (PACS) is used to store Medical images which use a Relational Database Management System (RDBMS) in the background [5] The medical images are so far managed using RDBMS. Medical images are semi structured. The traditional Relational data model (RDBMS) handles structured information very effectively. RDBMS cannot handle semi-structured data efficiently So there is a need to look for an alternate solution to handle the medical images.

Typically, PACS use a Relational Database Management System (RDBMS) to support their archive systems [5]. Also, in these systems there are already few solutions intending to use NoSQL [5, 11], especially for index and retrieval. Nevertheless, they do not exploit the full possibility to store also files inside the NoSQL databases, as BLOB stores. This type of solutions can be used, not only by the traditional archive used by medical staff, but also by researchers that are dealing with big data issues.

(b) Handling huge sized images.

Medical images are very huge in size. Handling huge images poses a challenge in archiving, retrieving and sharing as well. The use of RDBMS in storing huge images is a challenge, as the size is limited in RDBMS. NoSQL databases can easily handle huge sized images at ease. Also there is a hype going around moving the health care information to the cloud. As RDBMS is a worst fit for the cloud applies to say storing medical images as well. So these considerations are to be placed before going for a NoSQL database.

3. SUITABLE DATA MODEL TO HANDLE MEDICAL IMAGES

Medical images are stored in Picture Archiving and Communication Systems (PACS) are using a RDBMS in the background. PACS is a RDBMS based structure that stores Medical images. RDBMS has many disadvantages when the medical images are moved to the cloud or to a distributed environment. Some of the challenges in RDBMS are:

i) It is difficult to store/handle unstructured/semi-structured data in tables of any RDBMS.
ii) RDBMSs don’t scale out.
iii) Integrated search functions are not available in RDBMS.
iv) Portioning the image data and distributing is inefficient.
v) Strictly follows ACID properties of databases.

NoSQL databases can solve the challenges described. A NoSQL Database has many advantages like

i) It’s easy to store/handle unstructured/semi-structured data.
ii) It supports Horizontal scaling.
iii) For the distributing of huge images NoSQL databases supports chunking of data, which helps in auto sharding.
iv) Integrated search functions are available which provide better search results.

NoSQL Databases have following advantages over RDBMS:

Table 1. Advantages of NoSQL Databases over RDBMS

Scalability:	NoSQL databases use a horizontal scale-out methodology that makes it easy to add or reduce capacity.
Performance:	By simply adding resources, enterprises can increase performance with NoSQL databases.
High Availability :	NoSQL databases are generally designed to ensure high availability and avoid the complexity like typical RDBMS architecture. NoSQL has very good write speed and low latency query speed
Global Availability:	By automatically replicating data across multiple servers, data centers, or cloud resources. It can run over multiple data centers and its cloud enabled.
Flexible Data Modeling:	NoSQL offers the ability to implement flexible and fluid data models.
Data Redundancy:	The data is available with redundancy across one or more locations.

4. PERFORMANCE COMPARISON

To find a suitable data model to store and Retrieve Medical Images, the research design started with a comparative study between NoSQL and RDBMS. It was necessary to prove experimentally to proceed with a better Data model. So it was decided to compare the performances of NoSQL and RDBMS. The

Data model with a better performance to be selected.

Work Sketch: Relational Model Vs NoSQL

Methodology:

- ➢ To identify the right data model to store and retrieve medical images
- ➢ A throughput and latency based comparison to be done between
  - ✓ MySQL(RDBMS) and Cassandra(Column database)
  - ✓ MySQL(RDBMS) and MongoDB(Document Database)

- **Medical images with MySQL**

In the RDBMS initially images were not an primary part of the image files were stored separately. A new type of data type (BLOB) or storage method was introduced to store images into the RDBMS. Using this images can be accessed as part of a single transaction[11]. Another type LBLOB(Long BLOB) can be used to store medical images which are huge in size. This data type LBLOB can hold data up to 4GB. As medical images size may exceed this limit and in such a case the user is expected to write coding to chunk it to handle it, this poses an extra overhead for the user and the application developer.

- **Medical Images and MongoDB**

The documents are grouped together as collections and MongoDB is a document Database. Collections are similar to relational tables. MongoDB can easily handle different image file formats and huge images also. MongoDB is having features like BSON and Chunked storage so that Storage and retrieval of Medical Images is made simpler.

MongoDB uses a an open data format called BSON which is short for Binary-JSON(JavaScript Object Notation). BISON is a great way to exchange data and better way to store data.

The Image will be chunked and stored in MongoDB using GridFS(Grid File System), which can handle large binary files. Binary files including videos, images, and PDFs. It allows large binary files to be chunked and stores in MongoDB. Each chunk can be handled independently. GridFS uses two collections to save a file to a database. One collection stores the file chunks, and the other stores file metadata.

- **Medical Images and Cassandra**

For managing huge amount of distributed data Cassandra is one of the popular NoSQL databases. The data model of Cassandra is column oriented, columns together form column family. Column family is nothing but collection of columns associated with the key.

Storing large images in Cassandra with single set operation is difficult. To store a file under a single key and column creates performance blockage as the streaming potential is designed around smaller objects. Cassandra permits large object to be divided into the small part (chunks) and then store them across the multiple columns. The chunk size can be specified in bytes. This can be done by using the utility Astyanax, which splits up large objects into multiple keys and can fetch them parallel.

5. THE EVALUTION METHOD

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 i) MySQL and MongoDB ii) MySQL and Cassandra.

The time complexity was studied using standard processor. The images was stored and retrieved in MySQL, MongoDB and Cassandra. The application program was written in JAVA and the time was recorded on single system. The experiments were done to find out the performances of RDBMS v/s NoSQL.

6. EXPERIMENT RESULTS

The time complexity for storing and retrieving medical images in MongoDB and MySQL and Cassandra and MySQL was recorded and the results are shown below.

- **MongoDB v/s MySQL**

Figure 1 shows the time based performance to Store Medical images in MongoDB and MySQL.

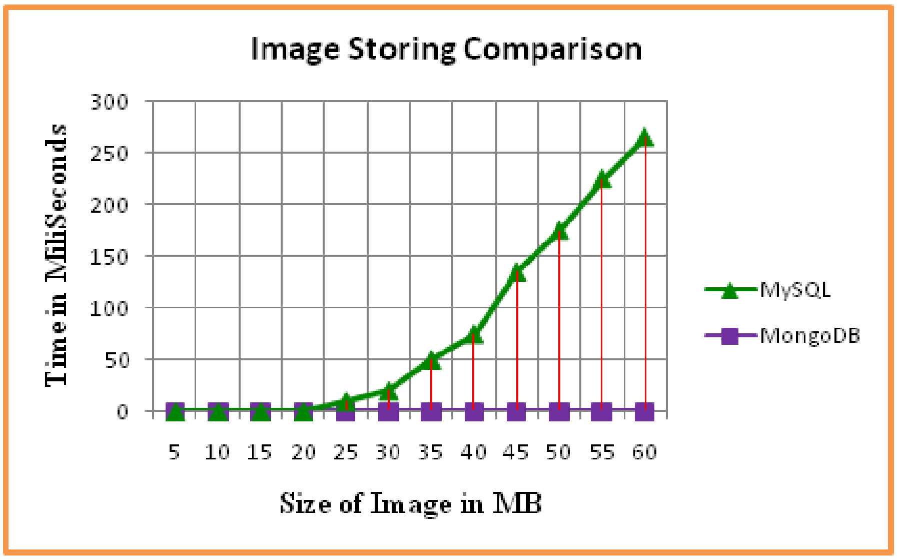

Figure 1. Image Storing Comparison of MySQL and MongoDB

Figure 2 shows the time based performance to Retrieve Medical images in MongoDB and MySQL.

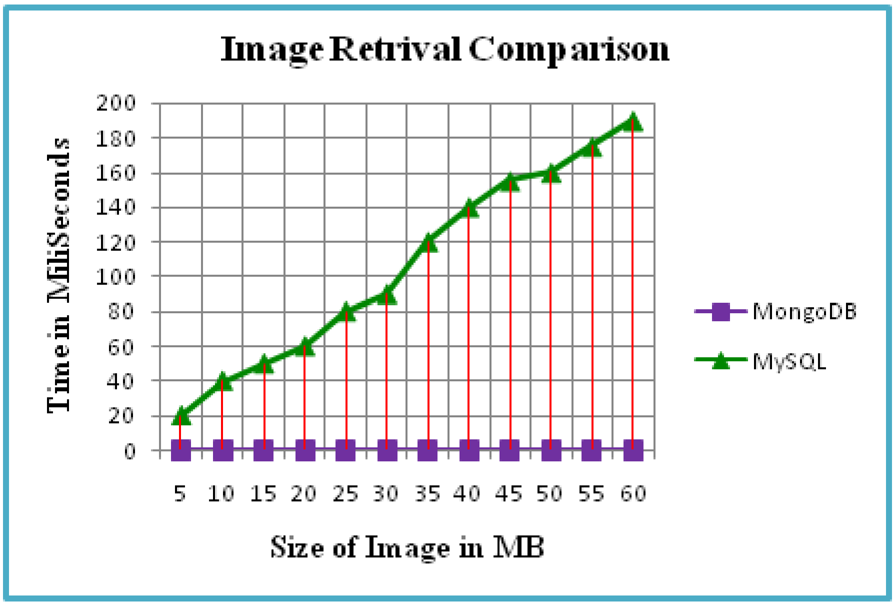

Figure 2. Image Retrival Comparison of MySQL and MongoDB

- **Cassandra v/s MySQL**

Figure 3 and 4 shows the results of the time taken for storage and retrieval Using Cassandra and MySQL respectively. The input Data set remains the same.

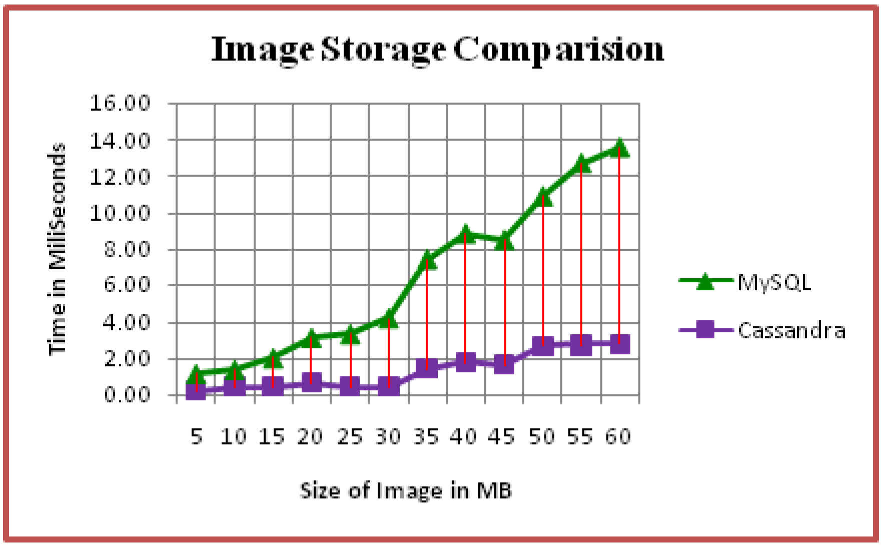

Figure 3. Image storage Comparison of MySQL and Cassandra

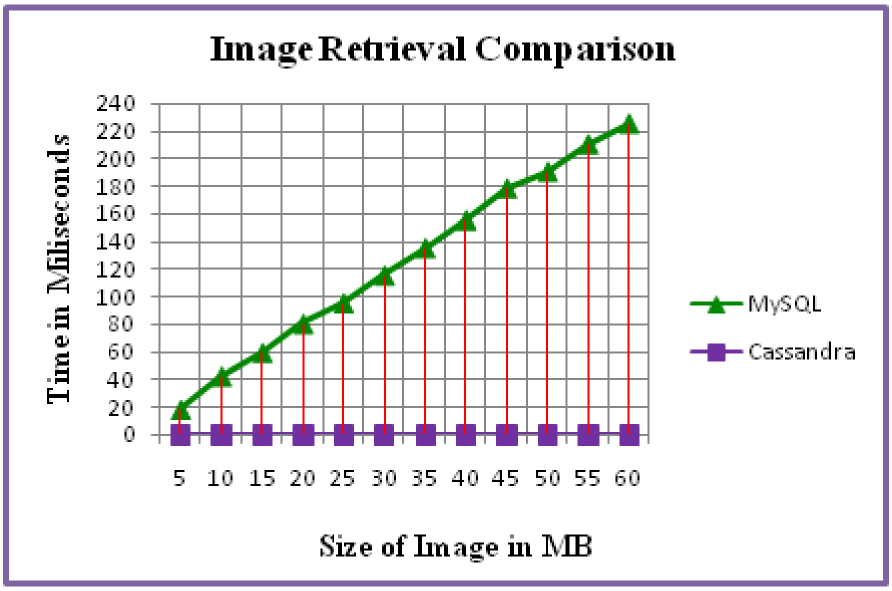

Figure 4. Image Retrival Comparison of MySQL and Cassandra

7. CONCLUSION

The experimental result of this study clearly indicates the performance of the NoSQL databases is better than MySQL. This indicates that NoSQL data model is a better option to store medical images. The challenges faced in RDBMS can be overcome using the NoSQL data model. The time for storing and retrieval in MongoDB and Cassandra was constantly lesser, even when the size of the images increased.

8. Future Work

The need for a better storage alternative for handling medical images may be done through any of NoSQL databases in the future. Also this takes us to the next level, where we can combine this NoSQL database with cloud environment for effective storing and retrieval of medical images.

So far archiving and sharing of medical images in the cloud was done only through relational databases, which has lot of drawbacks. Future work will be based on moving these medical images to the cloud with a better performance using NoSQL data model.

9. REFERENCES

1) *N. V. Chawla and D. A. Davis, "Bringing big data to personalized healthcare: A patient-centered framework," Journal of general internal medicine, vol. 28, pp. 660-665, 2013.*
2) *J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," 2011.*
3) *Z. Parker, S. Poe, and S.V. Vrbsky, "Comparing NoSQL MongoDB to an SQL db," In Proceedings of the 51st ACM Southeast Conference, p. 5. ACM, 2013.*
4) *N. Leavitt, "Will NoSQL databases live up to their promise?," Computer, vol. 43, pp. 12-14, 2010.*
5) *S. J. Rascovsky, J. A. Delgado, A. Sanz, V. D. Calvo, and G. Castrillón, "Informatics in Radiology: Use of CouchDB for Document-based Storage of DICOM Objects," Radiographics, vol. 32, pp. 913-927, 2012*
6) *Gang Chen.et.al, Federation in Cloud Data Management: Challenges and Opportunities, IEEE Explore, 2014.*
7) *Daniel J. Abadi, Data Management in the Cloud: Limitations and opportunities, Bulletin of the IEEE CST Committee on Data Engineering.*
8) *Divyakant Agrawal, Amr El Abbadi, Shyam Antony, Data Management Challenges in Cloud Computing Infrastructures.*
9) *D.Revina Rebecca et al, Impact of adapting Cloud Computing in health care industry for storing medical Images,National conference on emerging innovative technologies, (2014)-ISBN:978-81-923796-5-4.*
10) *Rick Cattell, Scalable SQL and NoSQL Data Stores , SIGMOD Record, December 2010 (Vol. 39, No. 4)*
11) *C. Costa, C. Ferreira, L. Bastião, L. Ribeiro, A. Silva, and J. L.Oliveira, "Dicoogle-an open source peer-to-peer PACS," Journal of Digital Imaging, vol. 24, pp. 848-856, 2011.*
12) *John Klein et.al, Application Specific NoSQL Databases, IEEE Explore.*