ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Challenges in Storage and Retrival of Healthcare Data:Review of various NoSQL Technologies

Rakesh Kumar [1], Florella Anna Fernandes [2]

Assistant Professor, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

M.Sc.ST, Scholar, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

ABSTRACT: In current digital Era, the volume of data that is being generated through various processes in the Health care industry has become unmanageable. More than structured data there are a lot of unstructured data that is being generated. Relevant research indicate that are many issues that have to be looked into.To manage this huge amount of unstructured data, SQL with its relational structure and strict schema design, is not efficient. The efficient way of managing this is through NoSQL, the database technology that does not follow property of ACID as we follow in SQL. Literature show that NoSQL databases have considerable Advantage in healthcare systems. In this paper an attempt has been made to identify an efficient technology for storage and retrieval of healthcare Industry Data.

KEYWORDS: Healthcare-systems, EHR, NoSQL Databases, ACID,CAP theorem, BASE.

I. INTRODUCTION

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 health care Industry has become unmanageable. Despite the development in the database technology the healthcare industry has seen many challenges with regards to storing and sharing of data.

Data intensive information systems require a strong database management systems in order to function properly. The volume and variety of data in modern distributed systems is increasing very quickly [1]. To manage Big Data in terms of Volume, Velocity, and Variety is the biggest challenge in the present day. SQL is a database computer language designed for the retrieval and management of data in relational database [2]. However, SQL is not an efficient way of managing Big Data. There are new systems to manage data and these new systems are referred to as "NoSQL" databases. It is another type of data storage other than databases (that were used earlier) that is used to store huge amount of data which keeps on increasing day by day. NoSQL is a non-relational database management system, it is a fast information retrieval database and is portable. NoSQL has many advantages such as high scalability, lower costs and availability.

II. BACKGROUND

The concept of Healthcare records or computerized Patient records is not a new concept. It was introduced in early 1990's.Healthcare records are identified by different names such as Electronic health records(EHR) ,Electronic Medical records(EMR),Electronic Patient records(EPR). An electronic health record (EHR) is a digital version of a patient's paper chart. EHRs are real-time, patient-centred records that make information available instantly and securely to authorized users. The international standard organization (ISO) defines EHR as "a repository of information regarding the status of a subject of care in a computer process able form and, transmitted securely, and accessible by multiple authorized users" [3] of Implementations of such systems was not possible cause of many reasons which include lack of technological standards, difficulty in using systems and system cost. These EHR systems have a great potential in the healthcare industry in India. The national health Portal India intends to introduce a uniform system for maintenance of Electronic Medical Records / Electronic Health Records (EMR / EHR) by the Hospitals and healthcare providers in the country.

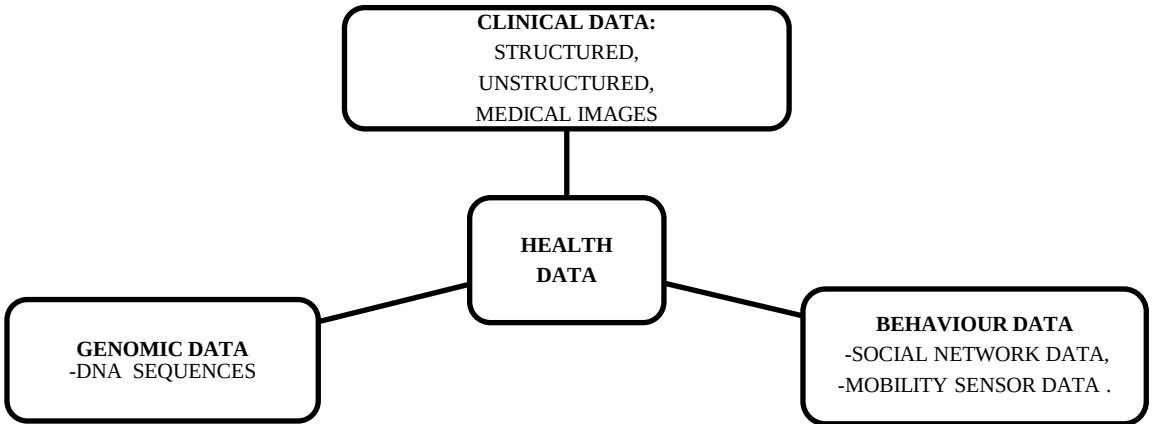
ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

DATA:



III. TECHNOLOGICAL ISSUES AFFECTING THE HEALTHCARE SYSTEM

There are three primary areas that need to be addressed with healthcare data i.e. storage issue, management issue and processing issue. There are a number of obstacles and challenges in relation to EMR systems mentioned in the literature, such as standardisation of vocabulary, security, privacy and data quality. The challenges include capturing, storing, searching, sharing and analysing.

Volume
Health care organizations are collecting more data. Analysing this huge collections of data can improve the accuracy, for better results which might give users to find unexpected patterns and insights [4].

Variety
To provide capital for this variety of data that is available, organizations need software solutions that can help them capture, integrate, and analyse this huge amount unstructured data. Data management solutions can help meet an organization's requirements in integrating data from multiple sources and help ensure the data is reliable.

Velocity
High-velocity data should be captured and analysed by Analytics software and the infrastructure on which it runs to deliver results in time. Patient monitoring systems such as those used in ICUs generate critical data at a rapid pace. If organizations can produce insights from that data in real time or near-real time, they can provide those insights to patient care teams at the moments when interventions will have the greatest benefits.

IV. NoSQL DATABASES

NOSQL stands for Not Only SQL. It is a class of database management system which is identified by its non-adherence to the widely used relational database management system (RDBMS) model with its structured query language (SQL) [5].

Features of NoSQL
- Massive scalability
- Flexible schema
- Quicker/cheaper to set up
- Higher performance and availability
- No declarative query language (i.e. SQL)
- Relaxed consistency

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

BASE

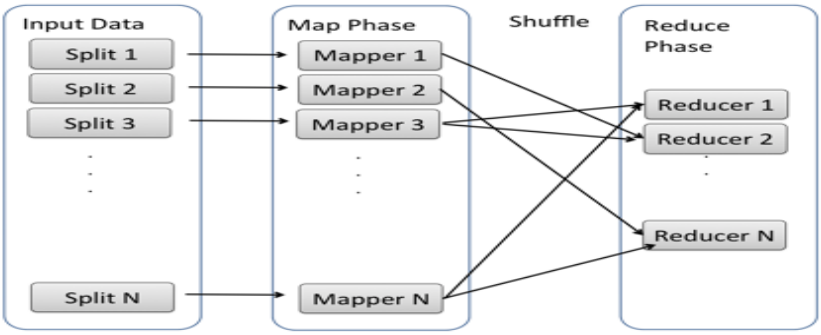
- **Basically Available**: The database system always seems to work!
- **Soft State**: It does not have to be consistent all the time.
- **Eventually Consistent**: The system will eventually become consistent when the updates propagate .
- Because of the distributed model, any server can answer any query
- Servers communicate amongst themselves at their own pace (behind the scenes)
- The server that answers your query might not have the latest data

V. MAP REDUCE[6]

It is a software frameworkfor easily writing applications which process vast amounts of data (multi-terabyte data-sets) in-parallel on large clusters (thousands of nodes) of hardware in a reliable, fault-tolerant manner.

A Map Reduce program is composed of:

- **Map() :** performs filtering and sorting (such as sorting students by first name into queues, one queue for each name)
- **Reduce ():** performs a summary operation (such as counting the number of students in each queue, yielding name frequencies).



VI. TYPES OF NoSQL DATABASES

There are about 150 different types of NoSQL databases which are grouped into four basic categories based on requirement of data for a specific application

A. KEY-VALUE PAIR (KVP) STORES

All data is stored in key value pairs. Keys are unique values that are used to access the information stored in values. In this type, there is no required format for the data i.e. data may have any format. It has an extremely simple interface. Records are distributed to nodes based on key.
The basic operations are:

Insert (key, value),Fetch (key), Update (key), Delete (key)

The "Value" is stored as a "BLOB". Here the application is responsible for understanding the data. In simple terms, a NoSQL Key-Value store is a single table with two columns: one being the (Primary) Key, and the other being the Value. Azure Table Storage, DynamoDB, Redis, are examples of key-value stores [7].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Example of unstructured data for user records

Key: 1	ID: sj	First Name: Sam

Key: 2	Email: jb@gmail.com	Location: London	Age: 37

Key: 3	Facebook ID: jkirk	Password: xxx	Name: James

B. DOCUMENT STORES

This is similar to Key-Value Stores, except that the value here is a "Document".It is in the form of "key" and "document" pair. The document format can beXML, JSON or any other semi-structured formats. The **Document Stores** store extensible structures as a "value" [8].Operation are based on document contents.

The basic operations are:
Insert (key, document), Fetch (key), Update (key), Delete (key)**and** ***Fetch ()***.
Some of the examples of this system are CouchDB, MongoDB, SimpleDB, etc. The rrecordscan have different structures within a single table.An example record from Mongo, using JSON format, might look like:

Though records are called documents, they are not documents in the sense of a word processing document, you can store binary data (using BSON format) in any of the fields in the document. You can modify the structure of any document while it is still in progress, by adding and removing members from the document, you can do this by reading the document into your program or modifying it and re-saving it, or by using update commands.

C. COLUMN-BASED STORES

This is based on Google's Big Table store. In column based stores the data tables are stored as sections of columns of data and not as rows of data. It is not required to define columns at the beginning. There can be countless number of columns which can be grouped as Super Columns.Cassandra and HBase are some of the examples of column based stores [7] .The below figures explains the column-based stores.An example of column based databases.

Record #	Name	Address	City	State
0003623	ABC	125 N Way	Cityville	PA
0003626	Newburg	1300 Forest Dr.	Troy	VT
0003647	Flotsam	5 Industrial Pkwy	Springfield	MT
0003705	Jolly	529 S 5th St.	Anywhere	NY

D. GRAPH DATABASE SYSTEMS

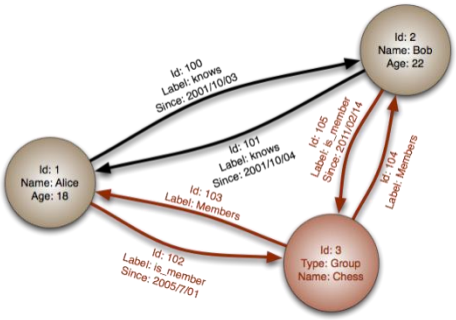
In the Graph Database Systems, the model is in the form of nodes and edges. The nodes can consist of properties and edges may have labels or roles. The graph theory is applied in the storage of information about the relationship between entries. To represent and store data, a graph database uses graph structures with nodes, edges, and properties. It provides index-free adjacency i.e. every element contains a direct pointer to its adjacent element and it is not necessary for index lookups [7].

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015



VII. RELATIONAL DATABASE THEORY

The transactions in the RDBMS are managed using ACID properties. It was developed by E.F. Codd. RDBMS has been widely used in most of the industry since a long time [9]. In the Relational database theory data are stored in rows and columns [10].

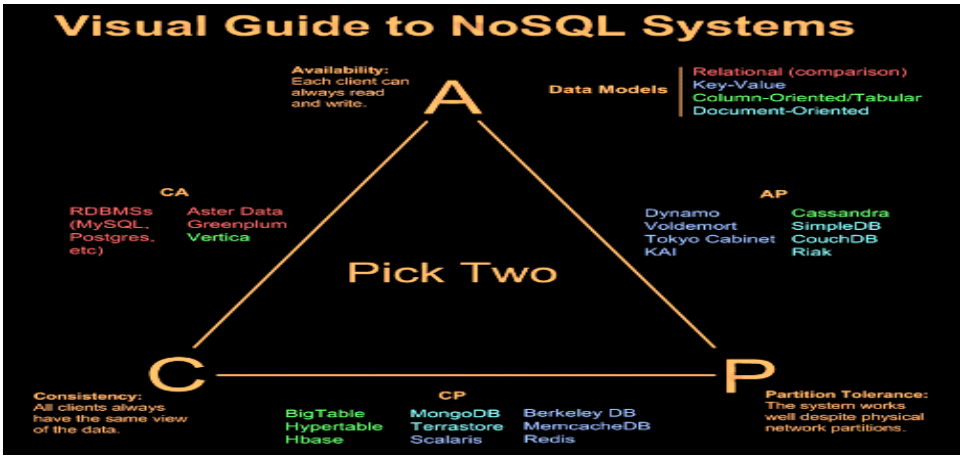
ACID PROPERTIES

- **Atomicity** – All of the work in a transaction completes (commit) or none of it completes
- **Consistent** – A transaction transforms the database from one consistent state to another consistent state.
- **Isolated** – The results of any changes made during a transaction are not visible until the transaction has committed.
- **Durable** – The results of a committed transaction survive failures.

VIII. CAP THEOREM

The CAP theorem states that there are three basic requirements which exist in a special relation when designing applications for a distributed architecture.

Consistency - This means that the data in the database remains consistent after the execution of an operation.
Availability - This means that the system is always on (Service guarantee availability), no downtime.
Partition Tolerance - This means that the system continues to function even if the communication among the servers is unreliable, i.e. the servers may be partitioned into multiple groups that cannot communicate with one another.



In a distributed system it is not possible to fulfill all the three requirements. However the CAP provides the basic requirements for a distributed system to follow two of the three requirements. Distributed systems should be partition tolerant (P), and then a choice is made between Consistency and Availability. Current NoSQL databases follow the different combinations of C and A from the CAP theorem as below:

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

- **CA** - Single site cluster, therefore all nodes are always in contact. When a partition occurs, the system blocks.
- **CP** - Some data may not be accessible, but the rest is still consistent / accurate.
- **AP** - System is still available under partitioning, but some of the data returned may be inaccurate.
-

IX. BENEFITS THAT NoSQL DATABASES PROVIDE FOR HEALTHCARE SYSTEMS

- **Scalability and Performance**: As current systems are based on RDBM systems, the increasing amount of data in the healthcare sector and the need for scalability are considered as an obstruction for the implementation of EMR Systems. The advantage of the NoSQl system is that it allows scaling up of large data sets .Cost effective scaling up is made possible with such NoSQL systems [12].

- **High Availability**: Replication is used to guarantee data availability among these systems. The nature of healthcare data need higher availability [13].

- **Consistency**: Healthcare data is usually added and not updated hence weaker consistency data can be applied using NoSQL Databases. Eventual consistency which is offered by NoSQL is sufficient.[12]

- **Open Source Availability**:NoSQL database have many open source alternatives, which may help in reducing the implementation cost by enabling access to the source code.

X. CONCLUSION AND FUTURE WORK

The main aim of this paper is to give an overview of the challenges faced in storing and retrieval of healthcare data. It describes the details that form the base of NoSQL databases like ACID, BASE and CAP theorem. Healthcare data cannot be managed with traditional Relational Database management systems. To satisfy the 3V's and to manage this huge amount of data, ACID properties are overtaken by BASE. Maintaining consistency all the time also is one of the big challenge along with the security.

REFERENCES

1) Borkar et al. 2012; Helland 2011; Konstantinou et al. 2011)
2) VatikaSharma,Meenu Dave, "SQL and NOSQL Databases", International Journal Of Advanced Research in Computer Science & Software Engineering. Vol. 2, Issue -8. ISSN: 2277 128X, October 2012.
3) ISO. 2004. "Ts 18308 Health Informatics-Requirements for an Electronic Health Record Architecture."
4) "Big Data: Issues and Challenges Moving Forward" 2013 46th Hawaii International Conference on System Sciences
5) "NOSQL" http://en.wikipedia.org/wiki/NoSQL,https://www.owasp.org/index.php/Testing_for_NoSQL_injection
6)"MAPREDUCE: Simpliefied data processwing on large clusters" By Jeffery Dean and Sanjay Ghemavat
7) Abramova, V., and Bernardino, J. 2013. "Nosql Databases: MongodbVs Cassandra," in: Proceedings of the International C* Conference on Computer Science and Software Engineering. Porto, Portugal: ACM, pp. 14-22.
8) Schmitt, O., and Majchrzak, T.A. 2012. "Using Document-Based Databases for Medical in-Formation Systems in Unreliable Environments," 9th International ISCRAM Conference, Vancouver, Canada.
9) Bailis, P., and Ghodsi, A. 2013. "Eventual Consistency Today: Limitations, Extensions, and Beyond," Commun. ACM
10) "SQL TUTORIAL", https://tutorialspoint.com
11) [Inc. 10gen. The MongoDBNoSQL Database Blog – BSON, May (2009). http:// blog.mongodb.org/post/114440717/bson
12) "Advantages and disadvantages of NoSQL over SQL" http://www.itworld.com/answers/topic/data-centerservers/question/what-are-advantages-or-disadvantages-nosql-over-sql
13) Dede, E., Govindaraju, M., Gunter, D., Canon, R.S., and Ramakrishnan, L. 2013. "Performance Evaluation of a Mongodb and Hadoop Platform for Scientific Data Analysis," in: *Proceedings of the 4th ACM workshop on Scientific cloud computing*. New York, New York, USA: ACM, pp. 13-20.
14)https://www.owasp.org/index.php/Testing_for_NoSQL_injection