ISSN No: 2309-4893
International Journal of Advanced Engineering and Global Technology
Vol-03, Issue-07, July 2015

Big Data Electronic Health Records Data Management and Analysis on Cloud with MongoDB: A NoSQL Database

Sreekanth R
Department of Networks and Infrastructure Management
Faculty of computing,
Botho University,Gaborone, Botswana
rallapalli.sreekanth@bothouniversity.ac.bw

Golajapu Venu Madhava Rao
Department of Software Engineering,
Faculty of computing,
Botho University,
Gaborone, Botswana
venu.madhava@bothouniversity.ac.bw

Srinivas Nanduri
Department of Engineering,
Faculty of Engineering and Applied Science, Botho University,
Gaborone, Botswana
Srinivas.nanduri@bothouniversity.ac.bw

***Abstract*— The emergence of cloud computing architecture allows huge computations to run inexpensively and efficiently. Big Data systems like Hadoop are designed to run on commodity hardware and can process huge data of any data types. This makes operational Big Data workloads much easier to manage, cheaper and faster to implement. Traditional relational databases cannot scale horizontally when the data grows. A new database architecture which can handle the structured and unstructured data like NoSQL databases are designed for cloud computing environment to handle such data. NoSQL databases are natively able to handle load by spreading data among many servers, making them a natural fit for the cloud computing environment. The document data model used in NoSQL databases like MongoDB makes it natural fit for cloud computing environment. MongoDB is a database which is built specially for the cloud because it supports scale out architecture. In this paper we study how big data Electronic Health Records (EHR) systems data management and analysis on cloud can be achieved using MongoDB. We also compare how this NoSQL database performs well than SQL based EHR systems.**

Keywords—BigData; NoSQL; Cloud computing; EHR

I. INTRODUCTION

A systematic collection of health records and stored in a database systems for further aggregation and visualization is called as Electronic Health Records. With advancement in the information technology, new applications like openEMR provide the healthcare organizations to collect the patient records in well organized manner. Most of the Health systems developed earlier have SQL database for storage of the data as SQL provides more structured way of collecting, accessing of records. With the increase of EHR data and various types of data is collected from different sources (structured, semi-structured and unstructured) it is difficult for traditional databases like SQL cannot handle these data. SQL is not designed to handle big data. There is a need for a database architecture where it can handle the big data and various types of data arriving from different sources. A NoSQL database will handle a large data and can handle semi structured and unstructured data. MongoDB [1] is such a NoSQL database and an open source document database. It guarantees that it can handle the large data better than traditional SQL database. Traditional DBMS have limitations with regards to scalability and infrastructure cost issues. "Not only SQL" database systems can scale horizontally with no single point of failure bottleneck because of shared-nothing architecture [2].

II. LITERATURE REVIEW

In order to address certain issues related to licensing, scalability and low cost solutions an open source database is required. NoSQL database offers solutions for the issues that currently organizations are facing.

ISSN No: 2309-4893
International Journal of Advanced Engineering and Global Technology
Vol-03, Issue-07, July 2015

NoSQL systems have been influenced by Google's Bigtable and Amazon's Dynamo systems which can easily scale up to accomadate large data sets [3]. Many other NoSQL systems have been developed on related line such as HBase, MongoDB, CouchDB, Cassandra, etc. and currently there are around 150 databases belong to these systems [4]. In order to improve health care information sharing function plays a major role. An EHR system supports efficient and quality integrated health care systems by recording all the information throughout lifetime [5].

III. CLOUD AND MONGODB

MongoDB is an open source document database that provides high performance, high availability and auto scaling features which currently most of the organizations required. As lots of data being generated scaling up is an important factor for the organizations. MongoDB document database contains field and value pairs. For Electronic health record data it may contain the field and values as

```
{
        Id:"A1001"
        name: "sree"
        age:35,
        Blood group: "o+"
        LabReports:["ECG",X-RAY,SCAN]}
```

MongoDB provides high performance data persistent to support embedded data models which reduces I/O activity of database systems. It provides replica sets and redundant data for high availability. It provides horizontal scaling by automatic sharding[6] and distributes data across a cluster of machines.

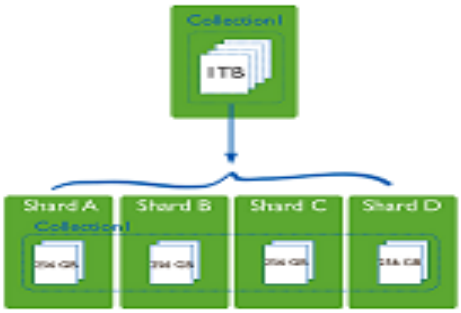

Fig1: 1TB data is split into 4 Shrads of 256 GB

Running MongoDB on cloud will significantly reduce the operational overhead. You can quickly spin up new shards to scale writes and resyncing replica nodes is trivial due to the exceptional networking performance across the platforms like Google Compute Engine Zones.

IV. AGGREGATION OF EHR VALUES

Let us take an example of Electronic Health Record fields for aggregation. MongoDB aggregation framework generates the aggregations values returned by the query. It is very similar to the concept of SQL GROUP BY statement. The documents in a collection will pass through an aggregation pipeline where they are processed by operators. Expressions produce the documents based o calculations by input documents. The accumulator expressions used in $group operator maintain state as documents progress through the pipeline.

Suppose if we want the aggregate operation to return above 10 million patients who have undergone the ECG test.

```
db.ehr.aggregate( [{ $group: { _id: "$name", patients:
{ $sum: "$ { $match : { LabReport : "ECG" } }" } } },
{ $match: { total Patients: { $gte: 10*1000*1000 } }
}] )
```

V. BENEFITS OF MONGODB FOR EHR

There are lot benefits that NoSQL databases like MongoDB can provide for Electronic Health Records. As the size of healthcare data will be increased over a period of time and the big data of health records will become bottleneck for the EHR systems. As NoSQL database provides horizontal scalability the records can be automatically scaled up for storage. As health records are mostly unstructured the data model should able to handle all the forms of data. MongoDB offers data model which can handle the unstructured data easily. As MongoDB offers high availability due to its distributed nature and replication of data the healthcare data is always accessible for continuous services. As healthcare data is shared across locations for EHRs it requires high performance systems to responds to queries in a timely manner. MongoDB provides high performance for all these issues.

ISSN No: 2309-4893
International Journal of Advanced Engineering and Global Technology
Vol-03, Issue-07, July 2015

VI. EHR SYSTEM WITH MONGODB

In this section we discuss on Electronic Health Record system with NoSQL database that is built using MongoDB. We study the architecture, database components for this system. Then finally we compare the SQL systems with the NoSQL systems to conclude how best NoSQL systems for EHR provide benefits. MongoDB stores the data as Binary JSON (BSON). This extends the JSON (Java script object notation) which will include the additional type notation such as int, long and floating point. For EHR systems all our data will be stored in ehr.json.

In relational model for an EHR application [7] the database will contain multiple tables. Let us simplify this by providing few tables with fields and relation between each table. The following Fig 2 show the relational data model for an EHR application.

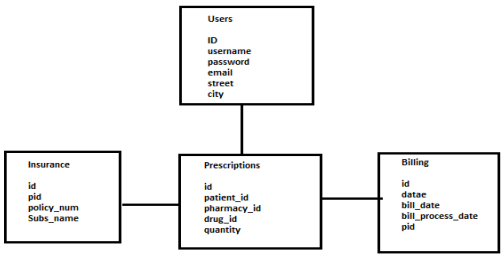

Fig2: Relational model for an EHR application

In relational database the information for a single record is spread across multiple tables. With MongoDB document model data is more localized with reduces the need to join separate tables. The final queried results shows higher performance and scalability across commodity hardware as a single read to database can access the entire document which contains all the related data. Fig 3 shows the data as documents for an application which use MongoDB database.

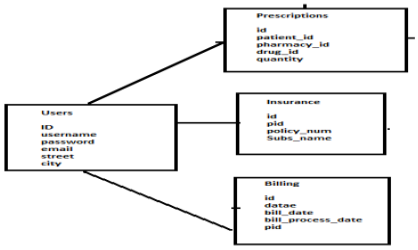

Fig 3 Data as documents in MongoDB for EHR application

VII. DATA MANAGEMENT

As discussed in section III MongoDB provides horizontal scale out for the database at low cost commodity hardware or cloud infrastructure with a technique called sharding. This technique provides transparency to the applications and distributes data to multiple partitions. It is an automatic built in the database. If Electronic health records are increasing due to increase in patient records, the data can be sharded across multiple partitions. Fig 4 shows automatic sharding provides horizontal scalability.

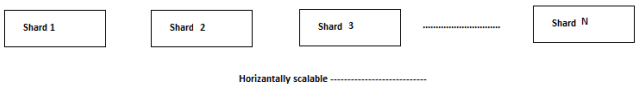

Fig 4 Automatic sharding by MongoDB

The relational data management of the data is more expensive as it requires licensing and other issues. MongoDB can be one tenth of the cost than RDBMS to build and run applications using these databases. Data management can be cost effective with MongoDB database.

VIII. COMPARISION WITH NOSQL AND SQL BASED EHR SYSTEMS

As relational database systems requires a well defined structure with fixed number of attributes that hold the data, the Electronic health records systems are built in the same way to hold the records. But NoSQL databases like MongoDB allows document based records to be stored and allows free flow of operations. When querying the data in relational database systems we use SQL standards. NoSQL databases implement a unique way to work with data. All data is stored in

documents like JSON which can be imported to the database. As the data with reference to electronic medical records are increasing horizontally it is required to scale them horizontally. SQL and NoSQL are designed to scale vertically but SQL cannot scale horizontally i.e to create clusters of multiple machines. Electronic health records need NoSQL based systems so that it can be efficiently utilized with lower costs.

CONCLUSION

In this paper we have seen how MongoDB a NoSQL database is better to build an EHR system. As SQL have many advantages like transaction security, but NoSQL systems can be build with cost less than 10 times of SQL systems. Also NoSQL systems are designed to scale horizontally. As electronic health care records are expected to scale out, it is highly required to build a system using NoSQL based system. Further we have seen how MongoDB can be scaled using Sharding. Electronic health care records systems can be built using document based JSON files. NoSQL based systems performs better than SQL based systems.

References

[1] Abramova, V., and Bernardino, J. 2013. "Nosql Databases: Mongodb Vs Cassandra," in: Proceedings of the International C* Conference on Computer Science and Software Engineering. Porto, Portugal: ACM, pp. 14-22.
[2] Cattell, R. 2011. "Scalable Sql and Nosql Data Stores," SIGMOD Rec. (39:4), pp. 12-27.
[3]Borkar, V.R. Carey, M.J., and Li, C. 2012 “Big Data platforms: what’s next? XRDS (19:1), pp.44-49.
[4]www.nosql-database.org
[5] Narayan S., Gagne M., and Safavi-Naini, R. 2010 “Privacy preserving EHR Systems using attribute-based infrastructure in 2010 ACM workshop on cloud computing security workshop, ACM pp 47-52.
[6]http://docs.mongodb.org/manual/core/sharding-introduction/#sharding-introduction
[7] Z. Yaowei, X. Yabin; “Design of Electronic Medical Record Data Integration Model Based on OGSA-DAI,” IEEE 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), on page(s): V3-278 - V3-281, 2010.