Review of Big Data Tools for HealthCare System with case study on patient database storage methodology

Purva Grover
USICT
Guru Gobind Singh Indraprastha University
Sector 16-C , Dwarka , Delhi-110078
Email: purva.usict.00416@ipu.ac.in

Rahul Johari
USICT
Guru Gobind Singh Indraprastha University
Sector 16-C , Dwarka , Delhi-110078
Email: rahul@ipu.ac.in

***Abstract*—Over the years with automation more and more systems deployed in multiple industries are generating huge amount of data. In fact IT Industry itself has witnessed phenomenal growth of data in the recent years. The data generated in the last 5 years is much more then the data generated cumulatively by all the industries put together in the past 20 years. In the current work we focus on the ways and means to handle the data generated by PHIS(Personal Healthcare Information System). The big question which we have addressed in this paper is selection of the appropriate tool (Relational MySQL database or NoSQL MongoDB database) to store the patient data, its archival and storage, steps to mine it and concluded the work by depicting the comparative analysis in terms of space and time.**

***Keywords*—Patient data storage, Patient data schema, Diseases trends, MySQL, MongoDB, Relational database**

India's population in 2013 was 1.252 billion. If we store the health data of each citizen of India than the volume of data would be so huge, that it would be apt to call it big data. Big data storage, retrieval, transferring, searching and visualization is very much complex as compared to traditional databases. Mostly the relational database management software, desktop applications and visualization software face many difficulties in handling the big data. The Big data can be categorized on the basis of 5V's and 1C. The 5V's of big data are Volume, Variety, Velocity, Variability, and Veracity. The 1C of big data is complexity. Big data analysis helps in predicting, decision making, acquiring knowledge and knowing trends. Big data application works in various areas of scientific research, social administration, business, social media, commerce, government department, health sectors, banking sectors and many more areas. Some of the big data sets available in these areas are stock exchange rates dataset, supply chain management dataset, population distribution dataset, disaster dataset, social networking dataset, disease trends dataset and many more. This research paper lists the attributes needed to store for the patient database. Research paper also discusses the Pros and Cons of implementing patient database in. MongoDB and MySQL.

I. Problem Statement

As we know in 2013 India populations was 1.252 billion. If we need to store the health information of every citizen of India than lets calculate the memory space required to store the data in MySQL. Assume each record contains the Personal information and diseases he/she suffering from and each citizen of India is suffering from one or more diseases. The schema is defined in the section Patient database schema. The personal information takes 1122 bytes data for each citizens record. Assume each citizen can have at most ten diseases. The diseases take 523 bytes of storage space. In MySQL the attribute values is available or unavailable than also we need to reserve the memory space for the attribute value. Therefore each citizen record will take $1122 + 523 = 1645$ bytes. For all citizen of India we need at least $1645 * 1.252 billion = 2059.54$ billion bytes. Gigabyte GB is about a billion bytes. So for storing the data about it's own citizen we would be needing $2059.54GB$ of storage space. Thus MySQL is not efficient storage for the Health database and we have to look out for other options.

II. Related work

Chen et al. [1] had highlighted just some of the big data problems, challenges, Opportunities, applications, transmission, curation, analysis and visualization. Kambatla et al. [2] presented the scale and scope of the data analytics and the discusses the issues should keep in mind while designing the big data analytic applications. They also discusses about the impact on software and the hardware by the big data application. Hardware and Software consideration for data analytic includes memory storage, processing for data, network resources and energy consideration. Saeed et al. [3] had introduced new data collection architecture for storing monitoring data of ICU Patients. In their application they are using relational database Postgre for storing the data. Wang et al. [4] had introduced BioStar schema for storing the biomedical data. They had also discuss the challenges they face while storing the biomedical data in star or snowflake shema (Data warehousing Schemas). Grover et al. [5] had discuss the limitation of the relational databases that had given the rise to the Big Data. Bhardwaj et al. [6] discusses about the history and the emergence of big data, how traditional DBMS couldnt compete with large data set and what are the issues and challenges of big data and the tools currently being used to implement and analyze the big data. Purva et al. [11] had beautifully contrasted the open source databases by varying the number of records for diabetic database. In this work we are going to discuss the advantages

978-1-4673-8203-8/16/$31.00

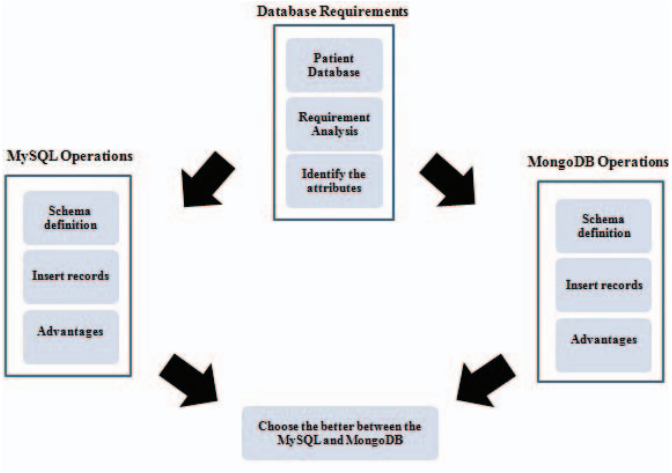

Fig. 1: Block Diagram indicating the Methodology followed

and disadvantages of storing patient data in Relational database MySQL and NoSQL database MongoDB.

III. Motivation for doing current work

Now days we can hardly meet any person who is not suffering from any disease. Day by day diseases are increasing in world. Most common disease [6] in the world are Gastroesophageal reflux disease, Acne, Allergy etc. Therefore to carefully analysis the increase in spread of diseases among the people so that proper and timely healthcare services can be provided to them is biggest motivation for undertaking current research work.

IV. Methodology Adopted

The flowchart in Fig. 1 details the methodology adopted for this research.

Algorithm 1 DatabaseActivities (AttributeList[])

1. Design the Schema for the database
2. Insert the records in the database
3. Observe the Space required for storing the schema and record. Execution time for schema creation and record insertion.

Algorithm 2 ChooseSuitableDatabase(AttributeList[])

Require: : AttributeList[n] where n is number of the attributes, and attributes are stored in the array AttributeList

1. Perform the DatabaseActivities(AttributeList[])) using MySQL
2. Perform the DatabaseActivities(AttributeList[])) using MongoDB
3. Compare the MySQL and MongoDB on Space and Time Complexity
4. Choose the best one between the MySQL and MongoDB

The Algorithm 1 and Algorithm 2 are used for the comparison between the MySQL and MongoDB on the basis of Space and Time complexity of the database.

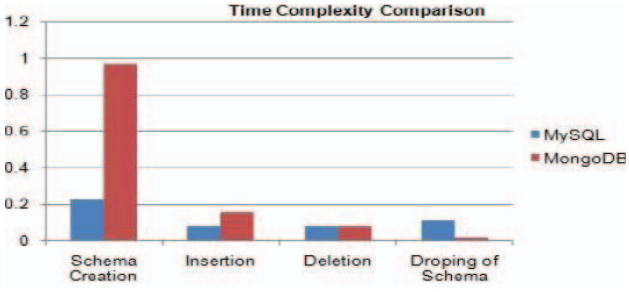

Fig. 2: Comparison between MongoDB and MySQL on basis of Time Complexity (Y-Axis represents time in seconds)

V. Patient Database Schema

The table I contains the attributes we need to store for patients data along with their reason to store it.

Attribute	Reason for storing
Patient ID	Auto increment and Primary key of the schema
Aadhar Card Number	To identify the each individual
Gender	Stores the Gender of the Patient
Age	Stores the Age of the Patient
Name	Stores the Name of the Patient
Highest level of Education	Stores the education level of the Patient
Marital Status	Stores the Marital Status of the Patient
Number of Children	Stores the number of children the Patient had
Household Income	Stores the household income of the Patient
State	Stores the state of the Patient
Pin code	Stores the pin code of the Patient
Employment Status	Stores the Employment status of the Patient
Diseases in patient	Stores the diseases the patient had
Last Updated	Stores the date when the record is last updated

TABLE I: Attributes for patient database schema

VI. Comparison between the MySQL and MongoDB

The queries for schema creation, insertion, deletion and schema drop for Personal Healthcare Information System was written for both MySQL and MongoDB and run on both the databases. Table II contains the execution time for these queries and Table II compares the two on the basis of time complexity, Fig. 2 shows the graphical representation of same and Table III compare the two on the basis of Space complexity.

Time (in seconds)	MySQL	MongoDB
Schema Creation	0.22	0.967
Insertion	0.08	0.156
Deletion	0.08	0.078
Droping of Schema	0.11	0.016

TABLE II: Attributes for patient database schema

Database Tool	Space Required
MySQL	0.02 MB
MongoDB	8192 Bytes

TABLE III: Comparison between the MySQL and MongoDB on basis of Space Complexity for a single record

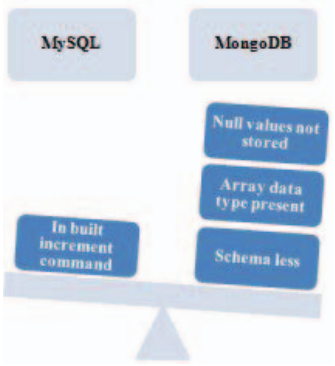

Fig. 3: Comparison between MongoDB and MySQL

VII. Conclusion

We conclude by commenting that for the current research work MongoDB performance appears much better for the patient database. Therefore on the facts presented in this paper we select MongoDB database for storage and querying in the patient database. The fig. 2 shows the comparison between the MySQL and MongoDB. **Some of the MongoDB advantages over MySQL :**

1) Schema less
2) Supports Array data type
3) The attributes having NULL values are not stored in the collection and memory space is not wasted

MySQL advantages over MongoDB : Inbuilt command AUTO-INCREMENT is available to make the attribute automatically increase its values whereas in MongoDB to auto increment the value we have to take the help of the function getNextSequenceValue(sequenceName) and collection counter.

VIII. Future work

The Patient database store in MongoDB can be used for analysis of the disease trends, disease occurrence, identifying the group people mostly get affected etc. In future we would design the Online Analytical Disease Processing (OADP). The Fig. 3 given the four application of patient database. This application can be designed using the patient database.

References

[1] Chen, CL Philip, and Chun-Yang Zhang. *"Data-intensive applications, challenges, techniques and technologies: A survey on Big Data."* Information Sciences 275 (2014): 314-347.

[2] Kambatla, Karthik, et al. *"Trends in big data analytics."* Journal of Parallel and Distributed Computing 74.7 (2014): 2561-2573.

[3] M Saeed, C Lieu1, G Raber, RG Mark. *"MIMIC II: A Massive Temporal ICU Patient Database to Support Research in Intelligent Patient Monitoring"*, Computers in Cardiology 2002;29:641-644.

[4] Liangjiang Wang and Aidong Zhang. *"BioStar models of clinical and genomic data for biomedical data warehouse design"*, Int. J. Bioinformatics Research and Applications, Inderscience.

[5] Purva Grover, Rahul Johari. *"BCD : Big Data, Cloud Computing and Distributed Computing"*. IEEE Global Conference on Communication Technologies (GCCT -2015) Kanyakumari TamilNadu, April 2015.

[6] Vibha Bhardwaj, Rahul Johari *"Big Data Analysis: Issues and Challenges"*, IEEE International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), VIIT, Visakhapatnam, Andhra Pradesh, January 2015.

[7] http://www.ranker.com/list/list-of-common-diseases-most-commonillnesses/ diseases-and-medications-info

[8] http://www.tutorialspoint.com/mongodb

[9] http://dev.mysql.com/downloads/mysql

[10] https://www.mongodb.org/

[11] Purva Grover, Rahul Johari. "MVM : MySQL Vs MongoDB". Springer Soft Computing for Problem Solving (SocProS -2015), Roorkee 2015.