Real-time analysis of healthcare using big data analytics

IOP Conf. Series: Materials Science and Engineering **263** (2017) 042056 doi:10.1088/1757-899X/263/4/042056

Real-time analysis of healthcare using big data analytics

Antony Basco J and Senthilkumar N C
School of Information Technology and Engineering, VIT University, Vellore-632014, Tamil Nadu, India.

E-mail: ncsenthilkumar@vit.ac.in

Abstract Big Data Analytics (BDA) provides a tremendous advantage where there is a need of revolutionary performance in handling large amount of data that covers 4 characteristics such as Volume Velocity Variety Veracity. BDA has the ability to handle such dynamic data providing functioning effectiveness and exceptionally beneficial output in several day to day applications for various organizations. Healthcare is one of the sectors which generate data constantly covering all four characteristics with outstanding growth. There are several challenges in processing patient records which deals with variety of structured and unstructured format. Inducing BDA in to Healthcare (HBDA) will deal with sensitive patient driven information mostly in unstructured format comprising of prescriptions, reports, data from imaging system, etc., the challenges will be overcome by big data with enhanced efficiency in fetching and storing of data. In this project, dataset alike Electronic Medical Records (EMR) produced from numerous medical devices and mobile applications will be induced into MongoDB using Hadoop framework with Improvised processing technique to improve outcome of processing patient records.

1. Introduction

In this era, Digitization plays its part in all activities. Digitization leads generation of large amount of data which is directly proportional to need of large storage spaces. These are some methods where the data is generated: Doctors: Health Data (From EMR or EHR or EPR), Payers: Through Claims and also from Cost Data, Developers: From Medical Devices via Research and Development and Pharmacy, Government: Public Health Data or through Population, Consumers or Marketers: Sentiment Data and Patient Behavior.

When there is a large amount of data, there emerges the concept of Big Data. Big Data has its own characteristics which defines its properties. In this project we focused on Health related data which is a never ending [1,2], dynamic cause for data generation and collected the datasets from various sources and inserted in to MongoDB database which uses NoSQL concept with Hadoop as framework to enhance the MapReduce. Hence Hadoop Map Reduce will be tuned with java and integrated with MongoDB and Healthcare Dataset will be imported in to MongoDB and tested for performance.

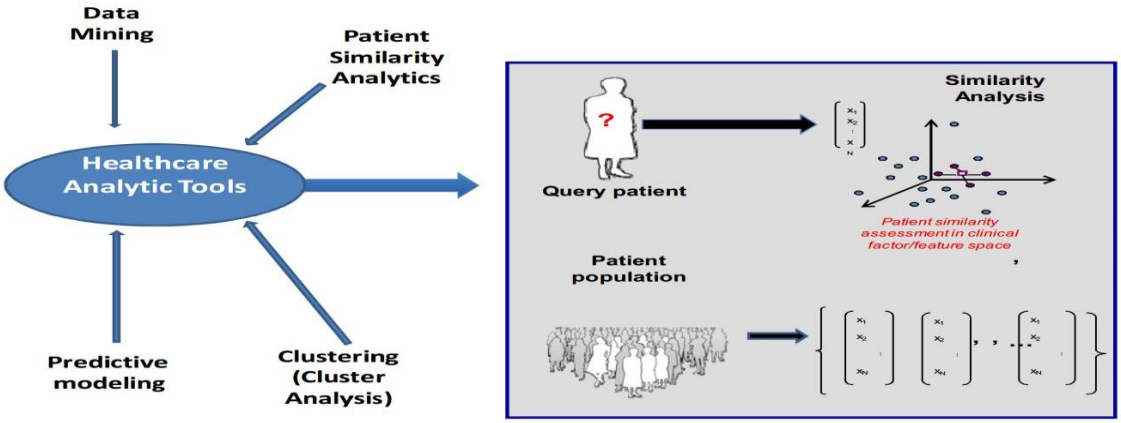

Figure 1 Big Data Analytics in Healthcare

2. Materials and Methods

2.1 Big data framework:

Hadoop is unique cloud calculating contexts to obtain verified to gauge and achieve sound on clouds. Presently, numerous recognized enterprises develop tenders constructed proceeding Hadoop, to say as Yahoo, Face book and etc. Hadoop , which is mainly instrument made of java for MapReduce(disseminated organization) industrialized by the Apache Software Foundation [3], so as is extremely error -prone and is considered near be installed on few expensive component, Hadoop customs hadoop disseminated file system(HDFS) towards sustenance the less - phase dispersed packing. HDFS affords from top to bottom material admittance to submission numbers and is appropriate meant for submissions that consume enormous records.

HBase significant Apache Hadoop-based development also offers a disseminated, column- focused on facts collection exhibited subsequently Google's Big table , and this one enhances a disseminated, burden - lenient accessible record on mists, constructed arranged maximum of the HDFS, per arbitrary immediate read/write contact in the direction of great numbers.

It alternated among NoSQL and RDBMS, it recovers info simply finished the key significant (row key) and vital array, in addition first funds sole racket contract. Similar to hadoop, the primary goalmouth of HBase stays to upsurge the totaling also packing volume via cumulating the squat - rate service servers.

2.2 Storage mechanism of framework

HDFS takes master/slave construction. HDFS collection comprises solitary NameNode which executes arranged the main attendant which succeeds case organization namespace too normalizes admission near records via consumers[10], too quantity of DataNodes executes taking place the Slave waitrons which succeed packing committed on the swellings which actually they lane proceeding, these are intended to run on product machines . HDFS discoveries a sleeve organization namespace besides permits consumer figures deposited in records. Within, sleeve will be divided addicted to lone or additional lumps besides

these hunks are stockpiled hip a established DN. NN implements file system namespace processes alike renaming files besides encyclopedia's. Similarly Her regulates the plotting of lumps near DN

The DN stand accountable intended for allocation of deliver and carve operations originated as of the heading arrangement of shoppers. The DN also accomplish lump replication in addition also its creation including deletion which acts upon order from the NameNode.

HDFS is envisioned to dependably pile enormous documents transversely tackles trendy a numerous gathering. Her provisions individual heading line of lumps, all chunks popular a heading excluding final block of similar magnitude Therefore, the vast HDFS heading be situated severed interested in Sixty four MB portions trendy the system of < Block_ID, MetaData> (Block_ID be situated the consecutive numeral portion, Metadata Binary data of portion) also deposited in the formula <key, value> .

Gamble probable, every portion determination be stored taking place a distinct Dn. Lumps dossier be situated imitation for error neglecting. The Nn marks conclusions concerning imitation of slabs. Her intermittently accepts a and a Block report individually every Dns popular the band. Block report comprises a slope very lumps taking place a Dn . Replication tactic be situated the principal of HDFS, also this one extreme guidance taking place delivered to carve routine HDFS, location terminated records tin accustomed be located prime distinguish amongst hadoop heading arrangement and further scattered case arrangements .

2.3 Programming prototype:

MapReduce program writing prototype made by Google, castoff to resolve difficulties transversely massive records on many lump gatherings atmosphere. two key occupations takes place popular a MapReduce background, Map plus Reduce occupation Map job receipts in a key/value duos plus produces transitional key/value duos. Next the Reduce job drive at that point profits standards accompanying for equal key and yield the finishing result.

The master node will be known as 10bit tracker which evenly separates the job and distributes the divided jobs also knows as replace careers to slave bumps which drive be known as TaskTrackers, these methodologies happen in an Authentic Map Phase Process.

Then Task Trackers will execute the sub jobs which will then permits the result rear on the way to her leading lump. I the Diminish period, leading swelling associations the rejoinder on or after slave swellings on the way to contract a clarification for central career.

The master node will be known as 10bit tracker which evenly separates the job and distributes the divided jobs also knows as replace careers to slave bumps which drive be known as TaskTrackers, these methodologies happen in an Authentic Map Phase Process. Then Task Trackers will execute the sub jobs which will then permits the result rear on the way to her leading lump. I the Diminish period, leading swelling associations the rejoinder on or after slave swellings on the way to contract a clarification for central career .

2.4 Storage database:

MongoDB, a NoSQL database, is a document-oriented database. It practices a bendable ideal to deliver dynamic schema. Data is deposited in BSON layout, a binary-encoded serialization of JSON-like forms{11}. The arrangement of MongoDB is diverse after traditional relational database systems. Ever since its core construction is a document base, a pool could be similar to a table although a file might be correspondent to a record in a relational some. MongoDB has on no account columns. Documents are similar to registers in a relational database. Documents comprises of field and value combinations. Pitches stay filaments, and principles can stay figures, twines, ranges or things. An item is that one a established of pitch and charge combinations. Therefore, its construction permits meant for uncontrolled dividing. A interrogation might stay entreated by distribution of JSON-like layout near associate per the pool trendy the organization.

MongoDB practises shredding to progress a horizontal scale-up. To support shading, we want to express MongoDB which data and collection that we wanted to do shading and which characteristic in the document will be castoff as a shard key. An attribute(s) aiding as a shard key must occur in all forms and will be indexed for future usage. MongoDB allow both shell method for implementation and accessing and also a diversity of drivers to provision furthermost current programming languages and development environments to enhance the usage protocols [7] and also user friendly commands.nMongoDB can also be integrated with any IDE to develop programs which may be in languages like PIG LATIN or JAVA and other programming languages. To integrate it to IDE's use libraries in to the Programming languages will suffice also several drivers are available to make it a default storage engine.

2.5 Test data preparation:

In order to make experimental result objectively showing the actual performance of spatial indexing, a specific console test program is developed for two products respectively. The test is conducted in the same machine and development environment (Tab.1), so that the result should be only driven by the product data driver and influenced by the actual performance.

Materials	**MongoDB**	**Hadoop**
Hardware Configuration	Intel(R) Core(TM) i3-5010U CPU @ 2.10GHz	
Operating System	Windows 10 (64bit)	
IDE	Eclipse Mars2	
Language	Java	
Tools	MongoDB(native)	Hadoop Objects
Data Driver	mongodb-driver-3.4.2	mongo-hadoop-core-2.0.2
Database	MongoDB	HBASE

Table 1: Features of Machine and Development Environment

3. Clinical document architecture:

Patient information which will be a clinical document will be in NoSQL format stored in the MongoDB database which will be stored and retrieved using the syntax proposed by common standard NoSQL

format. A more detailed view is displayed in table 2. The data conversion will be created with respect to the need of analytics. A common object model will be generated to understand the document and process involved in it.

A class diagram is drawn to explain the detail information about the process involved to get the result. It consists of list of patients along with their id and encountered problems. It also consists of classes required to initiate the process and also the reducer part. The reduced part consists of section code and collection of similar clinical statements. The rest of the classes consists of atomic object and environment which supports the system to execute.

3.1 Instrumentation

The proposed work will be carried out in Shell environment, where records will be manually inserted using MongoDB commands and MapReduce will be generated using java platform on eclipse IDE which will be executed in Hadoop environment using Hadoop basic commands of execution. The output will be generated automatically by the MapReduce program and stored in MongoDB.

Section	**Contains**
ID	Identification number
Chol	Cholesterol level
Stab.glue	Stable Glucose level
HDL	High density lipoproteins transport cholesterol
Ratio	Ration of HDL
Glyhb	Glycosylated hemoglobin for diabetes
Location	Location of Patient
Age	Stage of Enduring
Gender	Sex of Enduring
Height	Tallness of Enduring
Weight	Heaviness of Enduring
Bp.1s	Blood pressure stage 1
Bp.1d	Diastolic Blood pressure stage 1
Bp.2s	Blood pressure stage 2
Bp.2d	Diastolic Blood pressure stage 2
Waist	Midriff size of patient
Hip	Hip size of patient
Time	Time of readings taken

Table 2: Detailed description of diabetes patient's information

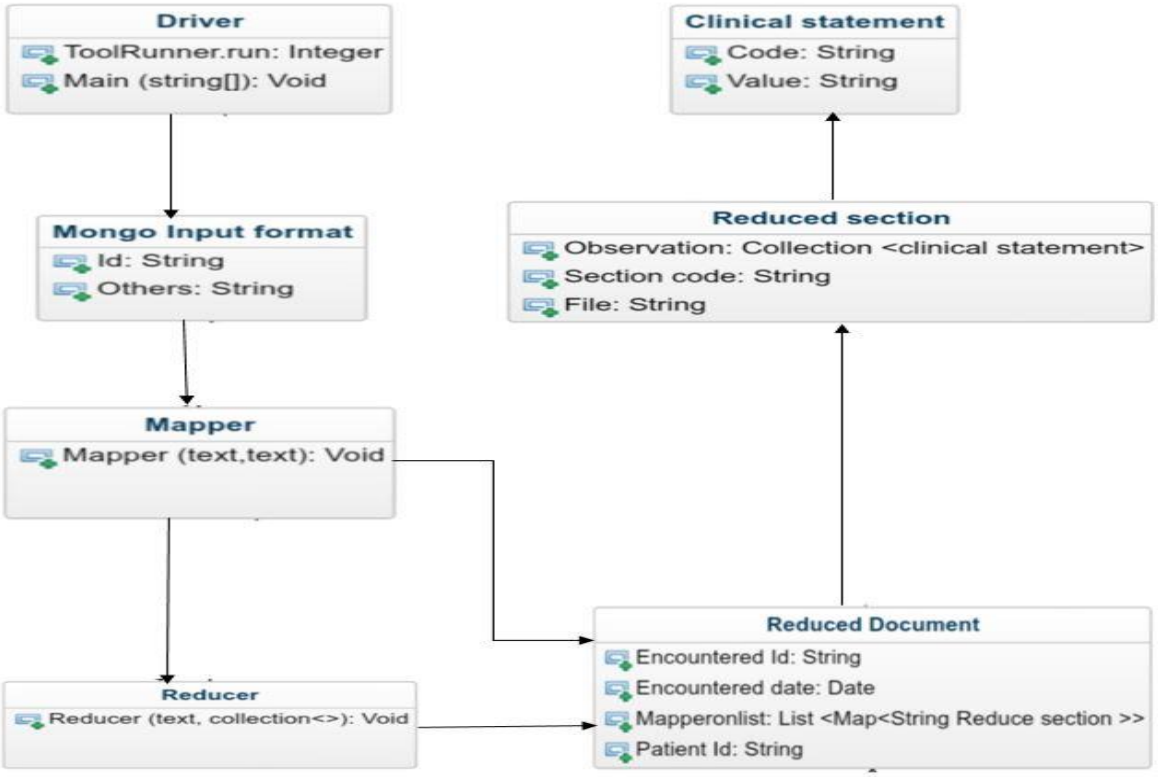

Figure 2 Class diagram

4. Proposed framework

The input of the proposed framework is health and patient related clinical standard files from hospitals and practices. This model will be having a Big data storage and processing level. This level considers the necessity of volume or size of the documents and the complexity of the files or records.

4.1 Big data container and processing layer

The layer comprises of repository known as HBase and Map Reduce scheduling algorithm which is inherited from installation of Hadoop environment. The detailed explanation of components are given below.

4.1.1 Semantic partitioner

HDFS is used on hadoop end storage which has a semantic partitioner to mine the entire set of healthcare related documents. The entire data must be extracted as the storage is based on document format and to ensure the integrity entire collection of data is must.

4.1.2 Standard schema validator:

The basic functionality is to process and parse the healthcare documents and validate the integrity of the healthcare documents.

4.1.3 Query formulator

It works on two phases, first extracts the data and builds for mapping. It helps to map the main concept of the queries to HDFS repository.

4.1.4 Batch Scheduling:

Over a distributed environment the jobs are divided in to multiple small jobs and assigned towards the intermediate data processing to the data nodes present in the queue. It will be designed in such a way that same intermediate node will be directed towards a single data node with similar collection which is directly proportional to the execution time and performance.

4.1.5 Intermediate Processing Layer:

It has the functionalities of partitioning the data and parsing the data files which are descendent from big data storage layer. The following will explain individual processes

4.1.6 Mongo data reader:

This is a customized library function used in map reduce driver class to read the format of data. Distinct information's of the patients details., different header details are extracted and additional go through will help to identify required information about healthcare.

4.1.7 Speculative parsing

With the help of Schematics a speculative parse will parse in sections for healthcare records. This will help the schema validator present in the storage or data layer. It will ensure increase in efficiency as it follows predictions methods via prior knowledge about the documents with document schema. It will be default extracts the necessary fields and in a worst case it will scan the whole document.

4.1.8 Driver class

It is the main function for execution over a distributed system. It originates by invoking classes in sequence. It invokes mapper class and then reducer class and transfers the extracted information.

4.1.9 Analytics of healthcare

This layer will comprise of several scenarios and presently it will be disease related data or medication related data of diabetic patients.

5. Case study

Considering United States in to account, 29.1 Billion persons before 9.3% on U.S. Populace munch diabetes. And considering in to fact only 21 billion persons consume remained identified and 8.1 shedload persons not identified. These data are surveyed from a national diabetes statistics report, 2016 and the data is tremendous. The implementation is done through naïve Bayes which assists to predict huge volume of data and are parallel parsed.

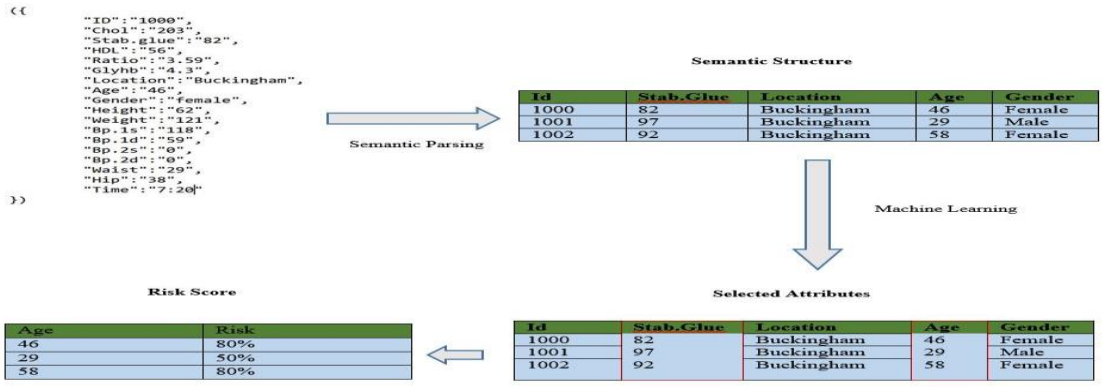

Figure 3 Case Study

The projected concept was implemented on real-time dataset collected from opinion of patients suffering from diabetes [5,6]. Human diabetic condition is related to disorders in produce and elimination of blood pressure. Treatment of diabetes needs to be recognized as it directly impacts the blood pressure.

6. Final remarks:

Even though usage of EHR, EMR has been underway for decades now, it's performance is still questionable and prone to inefficiency. In this paper the following main challenges has been overcome: Unstructured, Structured, Semi-structured data understanding, handling enormous volume of patient records, images, reports, etc., treatment in an efficient manner and quick decision taking, backtracking the efforts taken to cure the diseases.

6.1 Advantage of MongoDB and Hadoop in healthcare analytic data:

When it concerns with big data over healthcare sector by improvising the results by implementation of potential analytics in healthcare. Doctors will be able to take quick decisions based on the results, which are attained by smearing Big Data Analytics along with Hadoop and MongoDB[14]. The advances of healthcare standards will progress in recognizing and predicting diseases in initial stage and can be healed in minimum time. MongoDB helps to store the entire document in a single place which helps to store and retrieve in less time while Hadoop helps to reduce the complexity. The major advantage is early stage disease detection, analyzing and curing disease in an efficient manner.

7. Conclusions

This papers briefs the role of Big Data in day to day healthcare sector. The conversion of consuming classy machineries by healthcare consumers to achieve insights from clinical datasets and make knowledgeable judgments had transformed by Big Data Analytics. With the help of Hadoop and MongoDB, the concept of achieving effective data driven services to patients by means of predications has been made possible. In future, all healthcare organizations will be benefitted by achievements of health care analytics. The tools that are used by this paper for managing big data has been discussed and also hoped to provide better solution to face difficulties in future needs. Diabetes issue is chosen as our application due to its complex standard volume and data.

7.1 Future Research Recommendations

Latest Hadoop method has only been verified on a trivial cluster using a moderately small medical dataset dimension. Additional investigation is compulsory to see if this method would endure to scale as cluster size is amplified. Added work would also be mandatory to measure if the tactic would endure operative on dataset sizes past one billion triples. Added research work could be achieved as to the optimum size of triple collections more to the distributed cache. In this effort the commonly joined to triple groups are under single gb in size so it would be stimulating to study how presentation is affected as more triples are added to the distributed cache. In accumulation to other features might be examined to achieve bottomless insight into the routine.

References

[1] Krishnan S 2016 Application of Analytics to Big Data in Healthcare *Southern Biomedical Engineering Conference*
[2] Koppad S H and Kumar A 2016 Application of Big Data Analytics in Healthcare System to Predict COPD *International Conference on Circuit, Power and Computing Technologies*
[3] Adil A, Kar H A, Jangir R and Sofi S A 2015 Analysis of Multi-diseases using Big Data for improvement in Healthcare *Electrical Computer and Electronics IEEE*
[4] Reddy A R and Kumar S P 2016 Predictive Big Data Analytics in Healthcare *Second International Conference on Computational Intelligence & Communication Technology*
[5] Yadav V, Verma M and Kaushik V D 2015 Big Data Analytics for Health Systems *Green Computing and Internet of Things*
[6] Asri H, Mousannif H, Moatassime H A and Noel T 2015 Big Data in healthcare: Challenges and Opportunities *International Conference on Cloud Technologies and Applications* (CloudTech)
[7] Jangade R and Chauhan R 2016 Big Data with Integrated Cloud Computing For Healthcare Analytics *International Conference on Computing for Sustainable Global Development*
[8] Hussain S and Lee S 2015 Semantic transformation model for clinical documents in big data to support healthcare analytics *The Tenth International Conference on Digital Information Management*
[9] Vaishali G and Kalivani V 2016 Big Data Analysis for Heart Disease Detection System Using Map reduce Technique *International Conference on Computing Technologies and Intelligent Data Engineering*
[10] Ni J, Chen Y, Sha J and Zhang M 2015 Hadoop-based Distributed Computing Algorithms for Healthcare and Clinic Data Processing *Eighth International Conference on Internet Computing for Science and Engineering*

[11] Jin Y, Deyu T and Yi Z 2011 A Distributed Storage Model for HER Based on HBase *International Conference on Information Management, Innovation Management and Industrial Engineering*
[12] Kookarinrat P and Temtanapat Y 2015 Analysis of range-based key properties for sharded cluster of Mongo DB *International Conference on Information Science and Security*
[13] Mathew P S and Pillai A S 2015 Big Data Solutions in Healthcare: Problems and Perspectives *International Conference on Innovations in Information Embedded and Communication Systems*
[14] Hou B, Li K Q L, Shi Y, Tao L and Liu J 2016 MongoDB NoSQL Injection Analysis and Detection *International Conference on Cyber Security and Cloud Computing*
[15] Chickerur S, Goudar A and Kinnerkar A 2015 Comparison of Relational Database with Document-Oriented Database (MongoDB) for Big Data Applications *International Conference on Advanced Software Engineering and Its Applications*