Big Data Analytics for Health Systems

Vikash Yadav, Monika Verma, Vandana Dixit Kaushik
Department of Computer Science & Engineering
Harcourt Butler Technological Institute, Kanpur, India
vikas.yadav.cs@gmail.com, mvmonikaverma261@gmail.com, vandanadixitk@yahoo.com

Abstract- The speedily growing field of big data analytics has started to play a pivot role in the advancement of healthcare practices and research. It has provided tools to mount up, manage, analyze, and incorporate large volumes of unrelated, structured, and unstructured data produced by current healthcare systems. Big data analytics is a useful technique that is useful to provide better analysis of disease. However, the acceptance rate and research development in this space is still delayed by some fundamental problems inbuilt within the big data standard. Current research which focuses on utilization of huge volume of medical data while combining multidimensional data from different sources is discussed. Some areas of research within this field which have the ability to provide significant impact on healthcare delivery are also examined.

***Key Words*-** Big Data, Hadoop, Repositories, Healthcare

1. Introduction

Big data analytics is a latest analytics standard which is used to examine a collection of data, which cannot be managed or processed with the presently accessible technologies. Big Data mining is used to extract significant and valuable information from the vast datasets [1]. Useful information such as hidden patterns, unidentified correlations and the likes are discovered from the big data. Big data analytics (BDA) has emerged from two distinct concepts big data and analytics. Together it represents a new information management technique that has been intended to obtain previously available acumen and insights from data to address numerous innovative and significant questions.

Big data has become the new leading edge of information managing given the amount of data today's systems are generating and consuming. It derives the need for technological infrastructure and tools that can capture, store, analyze and visualize huge amounts of structured and unstructured data [2]. There are many possibilities for using BDA in health care. BDA can be used to help researchers to find the cause, treatments for diseases and care so expensive resources associated with a treatment are not administered to a patient who cannot benefit from the intervention.

Big data in healthcare is the electronic health data sets that are so huge and complex and hard to manage with conventional software, hardware and general data management tools [4]. Big data in healthcare system is very huge not just because of its volume but also of the variety of data types and the rate at which it should be manage [4]. The total data related to patient healthcare is being made up of "big data". It contains clinical data from CPOE and clinical decision support systems (physician's written notes and prescriptions, medical imaging, pharmacy, laboratory and other data), patient data in electronic patient records (EPRs).

Big data is huge amount of collection of data. Discovering relations and understanding patterns within the data, big data analytics has the potential to perk up care, save lives at lesser costs. Thus, big data analytics applications in healthcare take advantage of the explosion in data to extract insight for making better informed decisions [5]. When big data is synthesized and analyzed, patterns and trends discovered healthcare providers and in the

978-1-4673-7910-6/15/$31.00 ©2015 IEEE

healthcare delivery system can build up more careful and understanding diagnoses and treatments, resulting, one can expect, in higher quality care at lesser costs and overall better result [6].

2. Related Work

At present there are many techniques employed for the health monitoring systems. On one side there is a vast commercial monitoring and managing solution systems such as HP System's Insight Manager, IBM's Tivoli, and VMware's vCenter which are used for data center environments. All the centralized data gathering and analysis is done at this end and it also provide few supports for script based triggering mechanisms. The hardware-level support is relied on definite physical subsystems, such as HP's iLO or IBM's Director solutions for blade centers. In case of network traffic business monitoring tools such as TCP-dump, CoralReef, Wireshark, and Cisco NetFlow are available. There are various open source tools for gathering and monitoring, which use a hierarchical approach to monitoring where attributes are replicated within clusters using multicast methods and stored via a tree structure [8]. Open source tools to monitor network data such as Snort [10], Bro [11] and Tstat [12]. None of these available solutions presently scale to the sizes needed in next generation data center systems. The major networks and system health monitoring tools at present run on a single server but they are not able to cope up with a huge quantity of traffic received at high-speed links of routers in a scalable manner.

It has been illustrated that the data mining can be extensive ahead of the traditional relational data to the real time structured and unstructured data [7]. For the network system monitoring the application of big data is quite less [9]. Together with the toughness and scale properties of the above said methods the current paper also includes data collection and integration tasks.

It also customize the data mining approaches to the big data analytics infrastructure, simultaneously providing the scalability all the way through cloud based distributed computing strategy.

3. Architectural framework

The traditional health analytic system is almost similar to the theoretical structure for a big data analytics. In normal health analytics, the analysis is done with an intelligence tool installed on a separate system, such as a desktop or a laptop. The large data sets currently use the distributed processing to tap into their large data repositories to gain imminent for making better informed health related results. The open source software like Hadoop/MapReduce is also very useful and used in the area of big data analytics in healthcare.

The interface of the traditional health analytic system varies with that of the big data sets while their algorithms and models may be similar. While the interface of the former are user friendly, the platform for the latter are very difficult, programming rigorous, and they require the application of a range of skills. They lack the support and user easiness that vendor driven proprietary tools have. As indicated in Figure 1, the complication is regarding the data. Big data in healthcare system can come from anywhere including clinical decision support systems, electronic health records, laboratories, pharmacies, insurance companies etc. The data come in multiple formats such as relational tables, ASCII/text, flat files, .csv etc. and also resides at multiple locations. The Sources and data types consist of following:

a) Social media sites and the websites like the data which is accessed from the facebook, blogs etc. It can also integrated the websites related to health plan etc. [13].

b) Machine to machine data. In this the data is found out from the readings from remote sensing devices, meters, and satellites.

c) Big transaction data: All the data that is available either in structured or unstructured formats related to health care or some billing information's can be one of the data types.

d) The biometric data as palm prints, retinal scans, finger prints, genetics, handwriting, x-ray including the medical images also.

e) The data that is human generated examples of which include unstructured and semi structured data such as physician's notes, EMRs, email, and hand written documents.

Once the data has been collected it has to be processed or transformed into a type that is suitable for further processing. A service oriented architectural approach together with web services is one example of transforming the data [14]. The data is in a raw state and the services are used to access, extract and transform the data. In data warehousing, data is taken from various sources and is made ready for processing. Through the various steps of extracting, transforming, and loading (ETL) the data from diverse sources is cleansed and readied. Several data formats can be input to the big data analytics platform, depending on whether it is structured semi structured or unstructured.

In the theoretical frame a number of decisions are made about the data input technique, distributed design, platform selection and models through which analysis are done. The four major areas of big data analytics in healthcare system contain queries, reports, OLAP, and data mining. Visualization embraces all the above mentioned applications. A broad range of techniques and technologies has been developed through the drawings from the fields such as computer science, statistics, economics and applied mathematics [18]. These technologies are used for the aggregation, manipulation, analysis, and visualization of big data in healthcare.

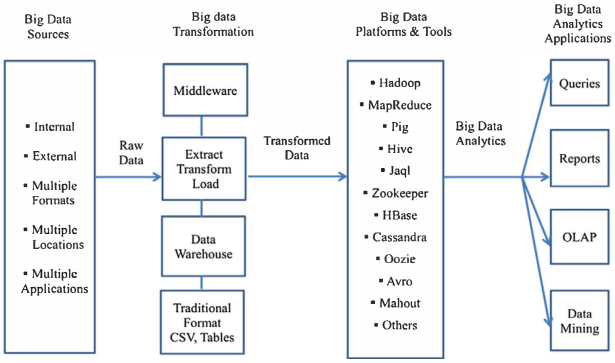

Figure 1. A theoretical model of big data analytics

For the aggregation of web search indices the very first platform for the big data analysis is the Hadoop. Hadoop is the open source of distributed data processing. This platform belongs to the class "NoSQL" technologies while others include CouchDB and MongoDB and many more that were developed to compile the big data in distinctive ways. Hadoop can handle an extremely large data set. It does so by distributing the task to several potential nodes each node solves different parts of the bigger program and then collects the results together.

[15].It serves the role of both a data controller and processing tool, offering a great deal of potential in enabling enterprises take care of data the has been up till now a challenge to handle. Either the data is structured or unstructured. Hadoop is used to process extremely huge volume of data. The adjacent ecosystem of extra platforms and tools supports the Hadoop distributed platform [17].

Several vendors including Hortonworks, AWS, Cloudera, and MapReduce Technologies distribute open source Hadoop platforms are distributed by various vendors as [16]. Many platforms such as Cassandra, HBase, and MongoDB are cloud versions which making them easily available. There are many trade-offs that the developers and the users of big data analytics should think about. Although the development costs are lower the technical support and minimal security have to be taken due considerations. In the healthcare industry, all these carry much more significance and therefore the trade-offs should be addressed. Moreover the tools/platforms require good programming skills. Governance issues which include possession, security, privacy and standards have yet to be addressed regarding the recent emergence of big data analytics in healthcare. In the coming section the analysis of big data is done whereby a method to develop and implement a big data scheme for healthcare system providers is given.

4. Methodology

There are various methodologies have being discovered in this promising field. Figure 2 represents one of them. In Step 1, the interdisciplinary big data analytics in healthcare team develops a 'Main statement'. This is a first stage. The concept is followed by the description. After acceptance of Main statement we can proceed to Second Step, the plan of action stage. Here, more details have provided. Based on the Main statement, several questions are addressed. We also have to provide necessary information for the problem as well as previous work and research done in this area.

Step 1 Main statement
- Set up need for big data analytics in healthcare based on the "4Vs", volume, variety, velocity, and veracity.

Step 2 Plan of action
- What is the actual problem?
- Why is it important and motivating?
- Why we use big data analytics approach?
- Necessary stuff

Step 3 Methodology
- Propositions
- Indicators assortment
- Data gathering
- Data conversion
- Platform/tool assortment
- Theoretical model
- Analytic approaches
- Association, clustering, classification etc.
- Results

Step 4 Deployment
- Evaluation & validation
- Testing

Figure 2 Big data analytics in healthcare methodology

In third Step, the methodology is being implemented. The Main statement is divided into a sequence of propositions. Consecutively, the dependent and independent variables or indicators are recognized. The data sources, as mentioned in Figure 1, are also identified. The data is gathered, described, and converted for analytics preparation. The main step is platform/tool evaluation and selection. There are various options containing AWS Hadoop, IBM BigInsights and Cloudera are available. The subsequently step is to apply these techniques to the data. This process differs from regular analytics only in that the techniques are scaled up to large data sets. In last Step, the models and their results are tested and validated.

5. Future Scope

A. System Outcomes

a) Correct dimensions of environmental parameters.
b) User friendly interface for visualization.

c) Well-organized use of energy for remote site, mobile devices and vehicle based data courier component.

d) Data messenger and remote site should be able to operate for long periods without any shortcomings.

e) Data is efficiently spread across systems to prevent bottleneck problem.

f) Resources should be well allocated to devices.

g) System is able to learn from feedback received from usage and able to adjust its analysis of data.

h) Robust security implementation on WSBs and Android IOIOs.

i) Easy web front end for medical professionals.

B. Expected Impact

BDA mainly focused on providing better healthcare to the patients residing in rural areas. There will be a definite blow on the quality of healthcare which these patients can receive. This system will show some key features to show how the storage of medical data can be possible.

C. Security Outcome

Security is the main concern for BDA, we will use traditional method of evaluating a secure network.

E-health is the further step in the development of health services, the acceptance will depend on the quality, availability and the user experience. In respect with heath issues the impact of the environment pollution on the health of population.

6. Conclusion

For making up to date decisions the healthcare providers use some cumbersome ways to get the insight from either the clinical or the other data repositories. The Big data analytics has changed the ways of how to look at this scenario. In future we shall see the speedy, extensive execution and use of big data analytics across the healthcare industry and healthcare organization. As big data analytics becomes more conventional, issues such as safeguarding security, guarantee confidentiality, continually improving the tools and technologies will gather attention. Big data analytics and applications in healthcare are at an emerging stage of expansion, but quick advances in platforms and tools can speed up their growing process.

References

[1] Wu et al., "Data mining with big data", IEEE Transactions on Knowledge and Data Engineering, vol. 26 no. 1, pp. 97-107, 2014

[2] Content and Predictive analytics in Healthcare, IBM - February 2012.

[3] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/ wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[4] Frost & Sullivan: Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. http://www.emc.com/collateral/analyst-reports/frost-sullivan-reducing-information-technologycomplexities-ar.pdf.

[5] Ikanow: Data Analytics for Healthcare: Creating Understanding from Big Data. http://info.ikanow.com/Portals/163225/docs/data-analytics-for-healthcare.pdf.

[6] Knowledgent: Big Data and Healthcare Payers; 2013. http://knowledgent.com/mediapage/insights/whitepaper/482.

[7] Rabi et al., "Solving Big Data Challenges for Enterprise Application Performance Management", 38th Intl. Conference on Very Large Databases, pp. 1724-1735, 2012.

[8] M.L.Massie, B.N.Chun, and D.E.Culler. "The Ganglia Distributed Monitoring System: Design, Implementation and Experience". Parallel Computing, vol. 30, pp. 818-840, 2004.

[9] Yeonhee Lee, Youngseok Lee, Toward Scalable Internet Traffic Measurement and Analysis with Hadoop, ACM SIGCOMM Computer Communication Review Vol. 43, No. 1, pp. 5-13, 2013.

[10] M. Roesch, Snort - Lightweight Intrusion Detection for Networks, USENIX LISA, pp. 229-238, 1999.

[11] Bro, http://www.bro-ids.org. last accesses Nov 2014.

[12] A. Finamore, M. Mellia, M. Meo, M. M. Munafo, and D. Rossi, "Live traffic monitoring with tstat: Capabilities and experiences", 8th International Conference on Wired/Wireless Internet Communication, 2010.

[13] IHTT: Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry; 2013. http://ihealthtran.com/wordpress/2013/03/iht%C2%B2-releases-big-data-research-reportdownload-today/.

[14] Raghupathi W, Kesh S: Interoperable electronic health records design: towards a service-oriented architecture. e-Service Journal 2007, 5:39–57.

[15] Borkar VR, Carey MJ, Chen L: Big data platforms: what's next? ACM Crossroads 2012, 19(1):44–49.

[16] Ohlhorst F: Big Data Analytics: Turning Big Data into Big Money. USA: John Wiley & Sons; 2012.

[17] Zikopoulos PC, DeRoos D, Parasuraman K, Deutsch T, Corrigan D, Giles J:Harness the Power of Big Data. McGraw-Hill: The IBM Big Data Platform; 2013.

[18] Zhang, Jianguo. "Big data issues in medical imaging informatics", Medical Imaging 2015 PACS and Imaging Informatics Next Generation and Innovations, 2015.