**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](mailto:vikas.yadav.cs@gmail.com), [mvmonikaverma261@gmail.com](mailto:mvmonikaverma261@gmail.com), [vandanadixitk@yahoo.com](mailto: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 relatively new analytics paradigm that is used to analyze datasets, which cannot be managed or processed with the currently available technologies. Big Data mining is used to extract meaningful and useful 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 derive 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 refers to electronic health data sets so large and complex that they are difficult to manage with conventional and traditional software, hardware and common data management tools and methods [4]. Big data in healthcare is very huge not because of its volume but also of the diversity of data types and the speed at which it should be manage [4]. The whole of data related to patient healthcare and well being made up “big data” in the healthcare industry. It includes clinical data from CPOE and clinical decision support systems (physician’s written notes and prescriptions, medical imaging, laboratory, pharmacy and other data), patient data in electronic patient records (EPRs).

Big data is vast amount of array of data. By discovering associations and understanding patterns and trends within the data, big data analytics has the potential to improve care, save lives at lower 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 revealed healthcare providers and in the healthcare delivery system can develop more thorough and perceptive diagnoses and treatments, resulting, one would expect, in higher quality care at lower costs and in better outcomes overall [6].

1. **Related Work**

At present there are many techniques employed for the health monitoring systems. On one side there is a vast commercial monitoring and management 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 collection and analysis is done at this end and it also provide some support for script based triggering mechanisms. The hardware-level support is focused on certain physical subsystems, such as HP’s iLO or IBM’s Director solutions for blade centers. In case of network traffic commercial monitoring tools such as TCP-dump, Wireshark, CoralReef and Cisco NetFlow are available. There exist several open source tools for collecting monitoring data and for cluster-level monitoring, which use a hierarchical approach to monitoring where attributes are replicated within clusters using multicast methods and aggregated via a tree structure [8]. Some of the open source tools to monitor network data include Snort [10], Bro [11] and Tstat [12]. None of these solutions currently 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 large amount of traffic received at high-speed links of routers in a scalable manner.

It has been demonstrated that the data mining can be extended beyond 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 robustness and scale properties of the above methods the current paper also includes data collection and aggregation tasks.

It also customize the data mining algorithms to the big data analytics infrastructure, simultaneously providing the scalability through cloud based distributed computing approach .

1. **Architectural framework**

The traditional health analytic system is almost similar to the theoretical framework for a big data analytics. In normal health analytics, the analysis is performed with a business intelligence tool installed on a separate system, such as a desktop or laptop. The large data sets nowadays use the distributed processing to tap into their large data repositories to gain insight for making better-informed health related decisions. The open source platforms such as Hadoop/MapReduce are also very useful and used in the application 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 extremely complex, programming intensive, and they require the application of a variety of skills. They lack the support and user easiness that vendor driven proprietary tools possess. As indicated in Figure 1, the complexity is regarding the data. Big data in healthcare can come from anywhere including electronic health records, clinical decision support systems, CPOE, laboratories, pharmacies, insurance companies etc. The data come in multiple formats such as flat files, .csv, relational tables, ASCII/text, etc. and also are residing 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 and the likes. It can also include health plan websites 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, finger prints, genetics, retinal scans, 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 EMRs, physician’s notes, email, and paper 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 from various sources is taken 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 the data is structured or unstructured.

In the conceptual framework several decisions are made regarding the data input technique, distributed design, platform selection and models through which analysis is done. The four major applications of big data analytics in healthcare include queries, reports, OLAP, and data mining. Visualization embraces all the above mentioned applications. A wide variety of techniques and technologies has been developed through the drawings from the fields such as statistics, computer science, applied mathematics and economics. These technologies are used for the aggregation, manipulation, analysis, and visualization of big data in healthcare.

**Figure 1.** An applied conceptual architecture 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 unique 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 larger program and then collects the results together.

[15].It serves the role of both a data organizer and analytics 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 makes it possible to process extremely large volumes of data. The surrounding ecosystem of additional platforms and tools supports the Hadoop distributed platform [17].

Numerous vendors including AWS, Cloudera, Hortonworks, and MapR 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 must consider. 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 must be addressed. Moreover the tools/platforms require great programming skills. Governance issues which include ownership, privacy, security, 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 methodology to develop and implement a big data project for healthcare providers is given.

1. **Methodology**

There are various methodologies are being developed in this promising field, here we present one of them. Figure 2 shows the main stages of the methodology. In Step 1, the interdisciplinary big data analytics in healthcare team develops a ‘concept statement’. This is a first stage. The concept is followed by the description. Once the concept statement is accepted, then we can proceed to Step 2, the proposal development stage. Here, more details are filled in. Based on the concept statement, several questions are addressed. We also have to provide background information on the problem domain as well as prior projects and research done in this domain.

**Step 1** Concept statement

• Establish need for big data analytics project in healthcare based on the “4Vs”.

**Step 2** Proposal

• What is the problem being addressed?

• Why is it important and interesting?

• Why big data analytics approach?

• Background material

**Step 3** Methodology

• Propositions

• Variable selection

• Data collection

• ETL and data transformation

• Platform/tool selection

• Conceptual model

• Analytic techniques

* Association, clustering, classification, etc.

• Results & insight

**Step 4** Deployment

• Evaluation & validation

• Testing

**Figure 2** Outline of big data analytics in healthcare methodology

In Step 3, the methodology is being implemented. The concept statement is divided into a series of propositions. Simultaneously, the independent and dependent variables or indicators are identified. The data sources, as mentioned in Figure 1, are also identified; the data is collected, described, and transformed in preparation for analytics. The important step is platform/tool evaluation and selection. There are several options available, as mentioned before, including AWS Hadoop, Cloudera, and IBM BigInsights. The next step is to apply the several big data analytics techniques to the data. This process differs from routine analytics only in that the techniques are scaled up to large data sets. In Step 4, the models and their findings are tested and validated and presented and tested.

1. **Future Scope**

***A. System Outcomes***

1. Correct dimensions of environmental parameters.
2. User friendly interface for visualization.
3. Well-organized use of energy for remote site, mobile devices and vehicle based data courier component.
4. Data messenger and remote site should be able to operate for long periods without any shortcomings.
5. Data is efficiently spread across systems to prevent bottleneck problem.
6. Resources should be well allocated to devices.
7. System is able to learn from feedback received from usage and able to adjust its analysis of data.
8. Robust security implementation on WSBs and Android IOIOs.
9. 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.

1. **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 the future we’ll see the speedy, extensive implementation and use of big data analytics across the healthcare organization and the healthcare industry. As big data analytics becomes more conventional, issues such as guarantee privacy, safeguarding security, establishing standards and governance, and 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] [7] 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] Yang et al., “Data mining based fault isolation with FMEA rank: A case study of APU fault identification”, 2013 IEEE Conference on Prognostics and Health Management, pp. 1-6, 2013