# A Survey Of Cost-Effective Big Data In Healthcare Applications

Vishnu S Basuthkar
Dept. of Computer Science and Engg.
East West Institute of Technology
Bangalore, India

Chetana Srinivas
Asst.Prof.: Dept. of Computer Science and Engg.
East West Institute of Technology
Bangalore, India

## ABSTRACT
The digitization of processes involved in the healthcare industry (HI) produces the massive amount of data, which is characterized by all the attributes of big data definitions. The analytics of these data may provide multifold benefits in both clinical practices as well as into management perspective in HI. There are requirements of supporting tools and methodologies to acquire, store, process, warehousing and analytics processes. This paper aims to collect initially generic tools for handling these issues of big data and later inferences the existing research attributes and practices, which are closely related to big data analytics. The methodology adopted is data collection mechanism from different sources such as specific blocks, textbooks, academic research papers. The outcome of this survey paper is beneficial for academician, researchers, and industries, who have interest into big data analytics and very specific to healthcare industry.

## Keywords
Analytics, Big data, Healthcare Industry, Warehousing.

## 1. INTRODUCTION
The healthcare industry is broadly composed of hospitals, druggist, Pharmacist, pathologist, radiologist as well as any another web services based applications which are related to health cure and management [1, 2]. In every country, the healthcare authorities are making it mandatory to digitize the process involved in healthcare industry [3]. In hospitals or at private clinics every new patient registration is supposed to be recorded in electronic registration system (ERS), and they need to be issued a secure chip-based data card so that their records can be updated in various departments. The advancement of a pathological process, digital clinical observation system (DCOS), radiology and last but not least robotic guided healthcare system (RGHS), etc., generates records consisting of database dumps, texts, images and videos [4]. These data's created at a particular location are planned to be collaborated to achieve objectives of content based retrieval system (CBR) and accurate analytics which can provide fast and cost-effective services to individual patients and healthcare management [5]. The collaboration of the data leads towards voluminous data, heterogeneity regarding various structure and introduction of Mysis, which brings uncertainty. The previously discussed facts categorize these records as big data characteristics [6]. The voluminous data created may be in the form of structured or unstructured type, those can be stored, recalled and analyzed, queried or can be manipulated by using machines. The structure and also semi-structured data composed of electronic records of patients and instrumental readings for health test, while unstructured data consisting of handwritten notes, patient's admission and reliving records, prescription records, etc. Also, the data may be web-based data, machine-based data, Biometric data and also data generated by human (like Twitter, Facebook, sensors, remote devices, fingerprints, x-ray scanning, EMRs, mails, etc.) [7], these conventional records and digital data are combined in healthcare big data (HBD). The execution of big data is the most challenging task. Hence, most of the researcher suggested for installation of big data tools in the standalone system. The big data, in generally has huge voluminous data and while processing, the execution is the carried out in distributed nodes. Hence, to have the knowledge of significant data analysis and take health decision in the better way, some open source distributed data processing platforms, such as Hadoop/Map Reduce are required. The tools of conventional analytics and the BDA tools are fully different; without any doubt the analytics tools for healthcare are more useful [8]. The BDA tools are typically complex in nature, comprehensive with programming and needs multi-skill applications, they are similar to ad hoc manner and are not user-friendly in nature, as the complexity of the process will take place with the data itself [9].

For BDA, different data-need to be combined then raw data is transformed- for multiple availability points. For example, i) the service oriented approach-combined- with middleware (a web service): -here information may remain raw, only services are applied to the call, resolve, and processing of data. ii) Data warehousing: -performs the data aggregation and make it for processing purpose, even data availability in real-time. This process follows in data Extraction-Transformation-Data loading from different sources-cleaning and preparation. The input for the platform applied depending upon the structured/unstructured data format [10]. Next, in the framework, the decision need to be taken for input approach, design, selection of tools and models analytics. For HI, these tools or platforms require better programming aspects as end user skills are not possessed. The HBD has limitations of security, privacy, ownership and its standards are not proposed yet [11]. This paper presents the survey of the existing tools for healthcare and its statistics, also, research survey on healthcare big data. The paper configured in the following manner. Section II describes the existing tools of big data for storage, security, privacy, etc. research statistics of big data in healthcare is the survey (from IEEE Xplore) is discussed in Section III. Section IV describes the interference of research work in HBD. Finally, section V concluded the research gap and conclusion for the paper.

## 2. EXISTING TOOLS OF BIG DATA IN HEALTHCARE
This section describes the existing tools which are present in healthcare industry (HI) and are given below [12]:

The Hadoop System (HS):

This provides the basic storage for Hadoop cluster (HC), which divides data into two parts and performs distribution for different nodes/ server.

Lucene:

This is applicable in text analytics also functions in the different open source distribution. Helps in indexing the text and use it for Java application.

MapReduce:

This enables the interface to distribute the subtasks and extract the outputs. After execution of the task, MapReduce performs the tracking of every node and server.

Cassandra:

This is also a distributed system and designed for top level big data handling.

PIG Latin:

This is configured to extract all the data. This consists of two modules PigLatin, and PigLatin execution version.

Hive:

The runtime Hadoop version which leverages SQL with Hadoop.Which allows SQL programmers to develop (HQL).

Mahout:

This is Hadoop model performs distributed machine learning, which supports BDA on Hadoop platform.

Jaql:

This processes massive data sets for facilitating a process. Performs conversion process of High-level queries of MapReduce.

Zookeeper:

This allows infrastructure for many applications, synchronization for server clusters. BDA uses this service to parallel process of big data clusters.

Avro:

This performs facilitation of series service. Avro facilitates data serialization services. Versioning and version control are additional useful features.

HBase:

This is a column based database management system, which tops on HS, by using non-SQL approach.

Oozie:

This is an open source project and makes the workflow solve the tasks.

## 3. STATISTICS OF RESEARCH ON HEALTHCARE BIGDATA

The statistics of big data in healthcare is searched in the IEEE Xplore (cited on 20/01/2016 at 2.35 pm), and following data is found:

**Table 1. Gives the research work statistics in big data**

| SI. NO | Type | Count |
|---|---|---|
| 1 | Conference publication | 11,284 |
| 2 | Journal and magazine. | 1,103 |
| 3 | Early access articles | 142 |
| 4 | Books and eBooks | 104 |
| 5 | Standards | 10 |
| 6. | Courses | 3 |

Further, when the search is refined with the key term ‘healthcare’, then the statistics found is mentioned in table.2.

**Table 2. Gives the research work statistics in healthcare big data**

| SI. NO | Type | Count |
|---|---|---|
| 1 | Conference publications | 239 |
| 2 | Journal and magazine. | 20 |
| 3 | Early access articles | 3 |
| 4 | Books and eBooks | 1 |

From the above survey, it is known that research in healthcare as per IEEE Xplore (table 1 and table 2), the conferencing publications have only 2.11% of big data, while the journals and magazines are off 1.8% of big data. The early access articles found in healthcare big data is only 2.11% and have ~1% of books and eBooks on healthcare big data. There are no standards and courses for healthcare big data.

The plot of above statistic is shown in figure.1.Where the x-axis represents the IEEE Xplore data type and y-axis gives the count.

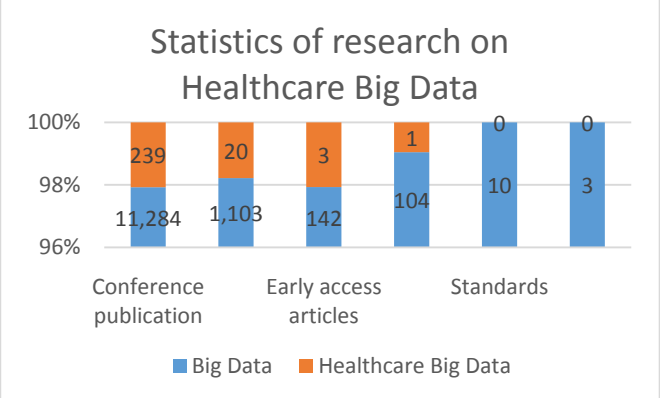


**Figure.1. shows the plot of statistics (Big data and healthcarebig data)**

**Table 3. Gives the research work statistics in healthcare bigdata**

| SI No. | Author | Problem-focused on | work done | outcome |
|---|---|---|---|---|
| 1 | Kahng et al. [13] | Social changes to have mobility, security and healthcare | Discussed the required relation between IT and the society | With the future of IT and social and techniques discussion of big data, modeling, and machine learning and simulation |
| 2 | Srinivasan et al.[14] | Issues of big data in healthcare | Discusses the applications of healthcare | With the advantage of recovering hidden losses and insurance costs. |
| 3 | Srinivasan et al. | Quality and cost of health care | Healthcare service outlier and | Framework for healthcare service, |

| | | | | |
|---|---|---|---|---|
| | [15] | outliers | anomalies detection | |
| 4 | Saria et al. [16] | Reviewing existing systems of data analysis | Discussion on big data in healthcare | Theoretical discussion on problems and consolidation in healthcare delivery |
| 5 | Yang et al. [17] | Reviewing existing systems of data analysis | Study on big data in healthcare | Theoretical discussion on problems and consolidation in healthcare delivery |
| 6 | Singh et al. [18] | Focus on social uncertainty in healthcare,marketing, politics, etc. | Reviewing existing systems of data analysis | The impact of persuasion requirement is discussed. |
| 7 | Hongsong et al. [19] | Issues of big data in multiuser | Presented a multilevel scalable Hadoop framework for healthcare applications of big data. | Secured access control in healthcare |
| 8 | Kovac [20] | Big data applications in healthcare | Proposed first E-healthcare system | E-healthcare system to record personal health data |
| 9 | Harris et al. [21] | Reviewing existing systems of data analysis | Big data scaling of medical record | Presents a Boolean Query generator methodology for Performance of medical data warehousing |
| 10 | Viceconti et al. [22] | Reviewing existing systems of big data in healthcare | Big data analytics of healthcare | Presents a mechanistic model for every patient separately. |
| 11 | Shneiderman et al. [23] | Reviewing challenges big data in healthcare | Big data analytics applications and its issues | Provides scope for the future work in big data healthcare. |
| 12 | Nepal et al. [24] | Reviewing challenges patients privacy issues | Discussion on scope healthcare data | Provides scope for the advancement of the healthcare. |
| 13 | Zhang et al. [25] | Big data scalability and privacy in clinical data | Privacy proximity in the clinical data | Presented a privacy model and Map Reducing algorithm to have high gain in scalability |
| 14 | Kaur et al. [26] | Management of clinical data | Discusses the healthcare management of different sources | Healthcare data process and store. |
| 15 | Aizawa et al. [27] | clinical data recording | Daily food intake in healthcare process | Presented FoodLog method to record food intake and provides scope for developing new applications for healthcare monitoring. |
| 16 | Kaur et al. [28] | Issues of healthcare in big data | NoSQL for big data store | With the best solution for Big data storage in healthcare. |
| 17 | Thimbleby et al. [29] | Security, privacy, performance issues in a healthcare system. | Addressing the issues of healthcare | Health declaration |
| 18 | Moulik et al. [30] | Big data analytics | Emergency evacuation system | Big data system to take the decision in the emergency. |

## 4. INFERENCES OF RELATED ARCHIVAL WORK

The section deals with the relevant research on healthcare big data.

By observing the above-related works of many authors in Healthcare Big Data (HBD), the work is categorized into three categories, such as contribution towards healthcare evolution (HE) and Healthcare Big Data (HBD), Future of HBD and are discussed briefly as below:

Contribution towards healthcare evolution (HE) and Healthcare

Big Data (HBD:

This section describes the contributory work towards HE by many authors. Author Kahng et al. [13] have proposed the design and device process technology based on the semiconductor device to achieve social mobility, security and healthcare. The author also has the outcome with the future prediction of society with IT industry by using machine learning, Simulink, etc.

Author Srinivasana et al. [14,15] have discussed the requirement of healthcare.This helps Big Data (BD) to detect the errors, frauds in health insurance and then in the reduction of recurrent losses and giving proper care to patients. The author flashes over the BD as per the business point of view in the next genre. In the next study author has given a framework to identify the irregularities of healthcare.

Saria et al. [16] have focused on the issues of healthcare delivery (HD) in BD, which founds helpful for computational engineers to sort out the issues of big data during transforming HD.

Yang et al. [17] have described the introduction of healthcare data mining (HDM), which helps in healthcare research. The authors also explained the research opportunity of medical service (MS) and improvement of healthcare by radiology.

Singh et al. [18] have studied the social interest towards physical and online network/world. Physical network deals with the offline while the online world deals with the cyber-based process (CBP). These two processes are helpful in the political campaign, healthcare and for marketing purpose. The CBP helps in data-driven algorithm analysis, having internet of things (IOT), smartphones and BD as part of it. Thus, the author proposed a social persuasion in the computational and empirical method can be employed to have both the networks.

HongSong et al. [19] have focused on the multilable applications security and privacy issues as they lead to BD. Keeping this point consideration, authors have presented a scalable multi-user framework for multiuser, and the framework can be used in Hadoop-based Big Data Healthcare Applications (HBDHA). The framework combines many access controls based on role, attribute, discretionary and mandatory. The framework with multilable has security degree, some replications, lifetime and hash value. In this framework admin of BD, is can add or remove the labels for new application users.

Koyac [20] has proposed e-health governance for worldwide to have the better quality of healthcare.

Haris et al. [21] have presented a Boolean Query (BQ) generator for data warehousing of clinical data. This generator helps to scale out the Boolean query into SQL, which is generated by the R & D of clinical data. Finally, the author outcomes with the warehousing performance.

Contribution towards Healthcare Big Data (HBD) Future Work:

This section contributes the future workflow for HBD. Author Viceconti et al. [22] have studied the phenomenological idea of BD, and they have used to build the mechanistic model for every patient, but there is no exist any model which is entirely phenomenological or completely mechanistic. The author has proposed that Big Data Analytics (BDA) can be used to achieve distributed data management in security and performance point of view. This domain specific technology leads to the research priority.

Author Shneiderman et al. [23] have focused on the growing data volume (DV) by social Medias, medical histories, and weblogs. Then by knowing the functions of Big Data Resource (BDR), as they provide the better understanding of complex systems and helps in taking decision for national security, healthcare, cyber security and finally for business. The author concluded with the future study requirement on a challenge of sharpening analytic (SA) for growing DA.

Nepal et al. [24] have considered the Gartner report of 2015, where they came to know that the technologies for processing the data are not efficient to meet the growing DV because of digital healthcare data (DHD). The author suggested that integrated health care analytics will help in effective patient's care, risk control, providing quality of life and service performance optimization, etc. The challenge of huge data storing with maintaining data privacy of every patient. This issue leads to healthcare big data analytics.

Zhang et al. [25] have gone through the applications of Big data cloud computing in HI. The problem of electronic data record privacy (EDRP) is considered in their study. The author has presented a Proximity privacy model (PPM) to solve the EDRP. Also, a Map Reduce Algorithm (MRA) is also designed to attain maximum scalability computation in the cloud. From the experiment carried out by the author on real-life patient's data set, concluded that the approach is most significant in comparison with the existing approaches, as per efficiency and scalability is the concern.

Kaur et al. [26] have focused on the vastly increasing medical data problem in healthcare and presented a persistent polyglot Framework (PPF), which helps in the documentation of different medical data separately.

Aizawa et al. [27] have given a multimedia tool "FoodLog", which offers the method to record the daily food consumption in healthcare. The tool also provides significance in latest healthcare monitoring applications (HMA).

Kaur et al. [28] have studied the existing studies to store the maximum healthcare data such as polyglot model and Healthcare Information system (HIS), came to know that these methods face the problem of storing multiple complex data (CMD) separately. The authors have presented a model of "polyglotHIS", which helps in CMD storage.

Thimbleby et al. [29] have addressed the healthcare issues in security and privacy.

Moulik et al. [30] have presented a Smart-Evac model to make the decision in emergency situations like the disaster. This is the cluster based system, where volume, variety, and velocity of healthcare big data are solved.

# 5. RESEARCH GAP AND CONCLUSION

Big data analytics (BDA) is the best option for healthcare applications, which can provide technologies to store clinical data. In future generation healthcare big data (HBD) will ramp up with its vast application for HI and society. In this paper, many tools of BDA are described to solve the issues in HBD and improve its efficiency.The proposed project will implement a scalable and novel framework for extracting knowledge based quality and value data extraction from clinical health care data. The experimental prototype ensures the effectiveness of the proposed framework which can be further reutilized on Big Data analytics, Medical research and clinical diagnosis of a particular patient.The tools improvement for better HBD is rapidly increasing. The proposed system aims to accomplish better performance of the android application however; the experimental outcomes anticipated for the proposed system are as follows:To accomplish enhanced quality of Big Data.To execute different modules in a collaborative way.To establish an efficient and cost effective quality management using data Cluster Manager.

# 6. REFERENCES

[1] Raghupathi, Wullianallur, and VijuRaghupathi. "Big data analytics in healthcare: promise and potential." *Health Information Science and Systems2.1*, 2014.

[2] Srinivasan, U.; Arunasalam, B., "Leveraging Big Data Analytics to Reduce Healthcare Costs," in *IT Professional*, vol.15, no.6, pp.21-28, Nov.-Dec. 20130

[3] Acampora, G.; Cook, D.J.; Rashidi, P.; Vasilakos, A.V., "A Survey on Ambient Intelligence in Healthcare," in *Proceedings of the IEEE*, vol.101, no.12, pp.2470-2494, Dec. 2013.

[4] Weiss, G., "Welcome to the (almost) digital hospital," in *Spectrum, IEEE*, vol.39, no.3, pp.44-49, Mar 2002.

[5] Jun-ping Zhao, "Electronic health in China: from digital hospital to regional collaborative healthcare," *in Information Technology and Applications in Biomedicine*, 2008. ITAB 2008. International Conference on, vol., no., pp.26-26, 30-31 May 2008

[6] Viceconti, M.; Hunter, P.; Hose, R., "Big Data, Big Knowledge: Big Data for Personalized Healthcare,"*Biomedical and Health Informatics, IEEE Journal*, vol.19, no.4, pp.1209-1215, July 2015.

[7] Viktor Mayer-Schönberger, Kenneth Cukier, "Big Data: A Revolution that Will Transform how We Live, Work, and Think", Business & Economics, pp. 242, 2013

[8] Demchenko, Y.; Zhiming Zhao; Grosso, P.; Wibisono, A.; de Laat, C., "Addressing Big Data Challenges for Scientific Data Infrastructure,"*Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference*, vol., no., pp.614-617, 3-6 Dec. 2012.

[9] Rongxing Lu; Hui Zhu; Ximeng Liu; Liu, J.K.; Jun Shao, "*Toward efficient and privacy-preserving computing in big data era," in Network, IEEE*, vol.28, no.4, pp.46-50, July-August 2014

[10] Xingchen Chu; Kobialka; Durnota, B.; Buyya, R., "Open Sensor Web Architecture: Core Services," *Intelligent Sensing and Information Processing, 2006. ICISIP 2006. Fourth International Conference*, vol., no., pp.98-103, Oct. 15 2006-Dec. 18 2006.

[11] Arvidsson, V., "The Revolution That Wasn't: Investigating Barriers to Platform-Based E-Service Delivery Partnerships,"*in System Science (HICSS), 2012 45th Hawaii International Conference*, vol., no., pp.2633-2642, 4-7 Jan. 2012.

[12] Sagiroglu, S.; Sinanc, D., "Big data: A review,"*Collaboration Technologies and Systems (CTS),2013 International Conference*,vol., no., pp.42-47, 20-24 May 2013

[13] Kahng, A.B., "Predicting the future of information technology and society [The Road Ahead]," in Design & Test of Computers, IEEE, vol.29, no.6, pp.101-102, Dec. 2012

[14] Srinivasan, U.; Arunasalam, B., "Leveraging Big Data Analytics to Reduce Healthcare Costs," in IT Professional, vol.15, no.6, pp.21-28, Nov.-Dec. 2013.

[15] Srinivasan, U., "Anomalies Detection in Healthcare Services," in IT Professional , vol.16, no.6, pp.12-15, Nov.-Dec.2014

[16] Saria, S., "A $3 Trillion Challenge to Computational Scientists: Transforming Healthcare Delivery," in Intelligent Systems, IEEE, vol.29, no.4, pp.82-87, July-Aug. 2014

[17] Hui Yang; Kundakcioglu, E.; Jing Li; Wu, T.; Mitchell, J.R.; Hara, A.K.; Pavlicek, W.; Hu, L.S.; Silva, A.C.; Zwart, C.M.; Tunc, S.; Alagoz, O.; Burnside, E.; Chaovalitwongse, W.A.; Presnyakov, G.; Cao, Y.; Sujitnapitsatham, S.; Daehan Won; Madhyastha, T.; Weaver, K.E.; Borghesani, P.R.; Grabowski, T.J.; LianjieShu; Man Ho Ling; Shui-Yee Wong; Kwok-Leung Tsui, "Healthcare Intelligence: Turning Data into Knowledge," in Intelligent Systems, IEEE , vol.29, no.3, pp.54-68, May-June 2014

[18] Singh, V.K.; Mani, A.; Pentland, A., "Social Persuasion in Online and Physical Networks," in Proceedings of the IEEE, vol.102, no.12, pp.1903-1910, Dec. 2014

[19] Hongsong Chen; Bhargava, B.; Fu Zhongchuan, "Multilabels-Based Scalable Access Control for Big Data Applications," in Cloud Computing, IEEE, vol.1, no.3, pp.65-71, Sept. 2014

[20] Kovac, M., "E-Health Demystified: An E-Government Showcase," in Computer, vol.47, no.10, pp.34-42, Oct. 2014

[21] Harris, D.R.; Henderson, D.W.; Kavuluru, R.; Stromberg, A.J.; Johnson, T.R., "Using Common Table Expressions to Build a Scalable Boolean Query Generator for Clinical Data Warehouses," in Biomedical and Health Informatics, IEEE Journal of, vol.18, no.5, pp.1607-1613, Sept. 2014

[22] Viceconti, M.; Hunter, P.; Hose, R., "Big Data, Big Knowledge: Big Data for Personalized Healthcare," in Biomedical and Health Informatics, IEEE Journal of, vol.19, no.4, pp.1209-1215, July 2015

[23] Shneiderman, B.; Plaisant, C., "Sharpening Analytic Focus to Cope with Big Data Volume and Variety," in Computer Graphics and Applications, IEEE, vol.35, no.3, pp.10-14, May-June 2015

[24] Nepal, S.; Ranjan, R.; Choo, K.-K.R., "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds," in Cloud Computing, IEEE, vol.2, no.2, pp.78-84, Mar.-Apr. 2015

[25] Xuyun Zhang; Wanchun Dou; Jian Pei; Nepal, S.; Chi Yang; Chang Liu; Jinjun Chen, "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud," in Computers, IEEE Transactions on, vol.64, no.8, pp.2293-2307, Aug. 1, 2015

[26] Kaur, K.; Rani, R., "Managing Data in Healthcare Information Systems: Many Models, One Solution," in Computer, vol.48, no.3, pp.52-59, Mar. 2015

[27] Aizawa, K.; Ogawa, M., "FoodLog: Multimedia Tool for Healthcare Applications," in MultiMedia, IEEE, vol.22, no.2, pp.4-8, Apr.-June 2015

[28] Kaur, K.; Rani, R., "A Smart Polyglot Solution for Big Data in Healthcare," in IT Professional, vol.17, no.6, pp.48-55, Nov.-Dec. 2015

[29] Thimbleby, Harold; Koppel, Ross, "The Healthtech Declaration," in Security & Privacy, IEEE, vol.13, no.6, pp.82-84, Nov.-Dec. 2015

[30] Moulik, S.; Misra, S.; Obaidat, M.S., "Smart-Evac: Big Data-Based Decision Making for Emergency Evacuation," in Cloud Computing, IEEE, vol.2, no.3, pp.58-65, May-June 2015