Review on E-Health Care using Big Data & Hadoop Map Reduce

B.Prasanna[1] A.Prema[2]
1 Research Scholar, Department of Computer Science, Sivaganga, Tamilnadu, India,
2 Assistant Professor, RDM Government Arts College, Sivaganga, Tamilnadu, India

Abstract

Cloud computing is one of the increasing computing technologies in spread paradigm. Even though technology is rising quickly, but any illegal user can use the weakness for cloud computing system. Big data is a term for data sets that are so large or complex that traditional data processing application software's are inadequate to deal with them. Challenges contain imprison, storage space, testing, data creation, search, input, move, apparition, querying, and updating and information privacy. Different types of approaches are in development to keep the privacy of this system. In this paper, we use a capable encryption algorithm to secure E-Hospital management in the cloud and give segmentation to maintain secret medical record of cloud, For reducing the record size, we used Hadoop and Map Reduce E-Health is a general used system, widely known as electronic health, where there live many types of services, providing electronic health records, prescriptions, healthcare information systems, etc. In this paper, a successful security framework, which is authentication technique that suits the recent e-health, is proposed.

Keywords: *E-Health, Big Data, Hadoop, Cloud, Map Reduce*

I. INTRODUCTION

Health care is one of the most concerns in India. In this paper, we analyze and expose the benefits of Big Data Analytics and Hadoop in the applications of Healthcare. The developing country like India with huge population faces various problems with the field of healthcare with respect to the expenses, meeting the needs of the economically poor people, access to the hospitals, and research in the field of medicine and especially in the time of distribution epidemics. This paper gives the participation in Big Data Analytics with Hadoop and reveals the impact of the same to render the services of health care to everyone at the optimal cost.

P.Zadrozny et al, a Pointed out that, the current developments in the Web, sensors, social media and mobile devices have resulted in the explosion of data set sizes. For e.g. Facebook today has more than billion users, with million active users generating more than terabytes of new data each day [21].

R.Agrawal, et al described, E-Health has increased its success and popularity in a short period of time. In practice, the system has to be secured and e-health service provider is entrusted with the responsibility to handle the sensitive information [1].

D.C. Kaelber, et al stated that, now organizations invest millions of dollars in the best PHR architectures, value propositions, and descriptions [2]. Ming et al presented the personal health record is used to store the data's of the user insecure and in an efficient way. It will be a valuable asset to individuals and families, enabling them to add and manage their healthcare information using secure, standardized tools [3].

Individuals own and manage the information about the PHR, which comes from healthcare providers. The personal health record has been maintained by a private environment, so that only authorized user can access the data. The PHR does not replace the legal record of any provider [4]. D.T. Mon, et al Describe in other words, they are electronic health records (EHRs) that are owned by patients. Which contain medical data generated within one specific care institution [5] [6].

Ziyuan Wang stated the main key characteristics of Cloud computing [7]: Zhang Xin et al is explained that, gainful, on-demand self service, omnipresent system admission, Rapid flexibility, High consistency and, adaptability. In cloud computing concerts, ease of use and safety are main research topics. Among them cloud computing security is one of the important research topics. [8] Han Hu et al pointed out the Map Reduce is in useful in a number of areas where huge data analysis is required [9]. Tom White clarifies many applications which handle the huge data sets have been settled. [10]

Haluk Demirkan illustrated Electronic Health Record systems (EHR) store the entire patient's medical history information from the time of admittance, check-up tests performed on the patient, prescriptions, readmission in sequence and any other relevant information of the patient. These data have to be effortless to get to and managed by all health care providers [11]. WLiu, et al expressed there are various stakeholders in generating the Electronic health records of the patients [12].

J Gantz et al explained, big data are a hum word today in each association. The global data size is estimated to be around 40,000 Exabyte's the data being doubled once in two years as per IDC report [13]. Demand for big data analysis is increasing at a high pace. Applications developed in business [14], science, health care show the insight value that big data bring into organizations. Big data cover five V's: (1) Volume (2) Variety (3) Velocity (4) Veracity (5) Value.

In response to the difficulty of analyzing huge-scale data, rather a little capable method, such as variety, data condensation, density-based approaches, grid-based approaches, divide and conquer, incremental learning, and distributed computing, have been presented [25].

The patient information should be readily available to the nurse, so that she can help the patient instantly in case of chronological poor health. The patients should be able to find the clinical trials relevant to him very easily [15]. There are various analytical methods for the big data. The most proper methods with respect to health care are (1) Recommendation system (2) Deep Learning and (3) Network analysis [16].

II. RELATED WORK

Increased use of the Internet, and progress in Cloud computing creates a large new dataset with increasing value to the business. Data need to be processed by cloud applications are emerging much faster than the computing power. Hadoop-Map Reduce has become a powerful computation model to address these problems. Nowadays, many cloud services need users to share their confidential data like electronic health records for research, analysis or data mining, which brings privacy concerns [59].

K-anonymity is one of the widely used privacy models. The scale of data in cloud applications rises extremely in agreement with the Big Data propensity, thereby creating it a dispute for conventional software tools to process such large scale data within an endurable lapsed time. As a consequence, it is a dispute for current anonymization techniques to preserve privacy on confidential extensible data sets due to their inadequacy of scalability [59].

Map Reduce is a framework for efficiently processing the analysis of big data on a large number of servers. It was industrial for the back end of Google's search engine to enable a large number of commodity servers to efficiently process the analysis of huge numbers of WebPages collected from all over the world [17]. Apache [18] [19] developed a project to implement Map Reduce, which was published as open source software (OSS), this enabled several organizations, such as business and university, to tackle big data analysis. Map Reduce [20] is a simple programming model for processing huge data sets in parallel.

Even though the market values of big data in these researches and technology reports [26–32] are different, these forecasts usually point out that the scope of big data will be growing rapidly in the forthcoming future. In addition to marketing, from the results of disease control and prevention [33], business intelligence [34], and smart city [35], we can easily understand that big data are of vital importance everywhere.

In this session we will discuss about the encryption and decryption. Review, most previous PHR research focused on the areas of information self – management and information exchange [24]. Section 2 of this paper deal with related work done in e-health care electronic health record, google map reduce, and Hadoop distributed file system. Section 3, explains an actual process of e-health care, and proposed work, in section 4, experimental analysis and implementation results are given, and finally, section 5 presents a conclusion of this paper.

Nowadays, the data that need to be analyzed are not just large, but they are collected of different data types, and even including streaming data [36]. Since big data has the unique description of "huge, high dimensional, mixed, complex, shapeless, unfair, loud, and inaccurate," which may change the arithmetical and data analysis approaches [37]. Although it seems that big data makes it possible for us to collect more data to find more useful information, the truth is that more data do not unavoidably mean more useful information. It may contain more ambiguous or nonstandard data.

For instance, a user may have many accounts, or an account may be used by many users, which may mortify the correctness of the removal grades [38]. Therefore, several new issues for data analytics come up, such as time alone, safety, cargo space, mistake charity, and quality of data [39].

Various solutions have been presented for the big data analytics which can be divided [40] into (1) Processing/Compute: Hadoop [41], NVIDIA CUDA [42], or Twitter Storm [43], (2) Storage: Titan or HDFS, and (3) Analytics: MLPACK [44] or Mahout [45]. Although there exist commercial products for data analysis [42–45], most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective "ways" to find the useful things from the data. But when we come into the period of big data, most of the present computer systems will not be able to hold the whole dataset all at once; thus, how to design a good data analytics structure or platform3

and how to design analysis methods are both important things in the data analysis process.

Over the past several years there has been a wonderful rise in the quantity of data being transferred between Internet users. Growing usage of streaming multimedia and other Internet based applications has contributed to this surge in data transmissions. Another surface of the raise is due to the growth of Big Data, which refers to data sets that are many orders of extent larger than the criterion files transmitted via the Internet. Big Data can series in size from hundreds of gigabytes to pet bytes [46].

Big data [47, 48] purposely refers to data sets that are so large or compound that usual data processing applications are not enough. It's the large volume of data—both structured and unstructured—that inundates a business on a day-to-day basis. Due to recent technical growth, the amount of data generated by the internet, social networking sites, antenna networks, healthcare applications, and many other companies, is radically increasing day by day.

All the huge calculate of data shaped from a variety of sources in several formats with very high speed [49] is referred as big data. The period big data [50, 51] is clear as "a new production of technologies and architectures, designed to cheaply separate value from very large volumes of a wide variety of data, by enabling high-velocity capture, detection and testing".

III. HADOOP IN HEALTH CARE DATA

Hadoop has fundamentally changed the economics of storing and analyzing information. This difference in economics has attracted a lot of attention and will make Hadoop the centrepiece from which most large-scale data management activities and analyses will either integrate or originate. Now, with more robust SQL capabilities being coupled with the Hadoop infrastructure and bringing the entire SQL-based ecosystem in the world of Hadoop, the market has expanded by one or two orders of magnitude. No longer is Hadoop just the domain of specialists. [52] In the current rapidly growing medical world vast measure of data is being produced and added every day. Because of Hadoop's efficiency and effectiveness the data considered earlier advantageously for the analyses are now loses their worth. Hadoop has the efficiency of storing files on various systems throughout the cluster using a distributed file system. Hadoop hides the location of the files in the cluster being accessed the end users can reference files the same way they do it in the system. [52].

IV. BIG DATA ANALYTICS IS STIMULATED IN HEALTHCARE FROM SIDE TO SIDE THE NEXT ASPECTS [53]

Healthcare data are now increasing very fast in terms of size, difficulty, and haste of age group and usual database and data mining techniques are no longer efficient in storing, dispensing and analyzing these data. New original tools are necessary in order to handle these data within an average forgotten time.

The patient's behavioural data is captured throughout numerous sensors; patients' a variety of social connections and communications. The standard medical practice is now touching from fairly ad-hoc and subjective decision making to evidence-based health care. Inferring knowledge from complex, varied patient sources and leveraging the patient/data correlations in longitudinal records.

Understanding unstructured clinical notes in the right context. Well conduct huge volumes of medical imaging data and extracting potentially useful information and biomarkers. Analyzing genomic data is a computationally determined duty and combining with paradigm clinical data adds additional layers of difficulty.

Fresh developments in the Web, social media, sensors and mobile devices have resulted in the blast of data set sizes. For example, Face book today has more than one billion users, with over 618 million dynamic users generating more than 500 terabytes of new data each day [54].

The term "Big Data" refers to large and complex data Sets made up of a variety of structured and unstructured data which are too big, too fast, or too tough to be managed by usual techniques. Big Data is categorized by the 4Vs [55]: volume, velocity, variety, and veracity. Volume refers to the quantity of data, variety refers to the multiplicity of data types, velocity refers together to how quick data are generated and how fast they must be processed, and veracity is the ability to trust the data to be accurate and reliable when making crucial decisions.

The most related work associated with the introduction of various Map Reduce models and its relation with the database processing [56]. This tutorial provides the insight about how to improve the performance by increasing the availability of the system in case of failure, reduced network communication overhead, process scheduling etc. [57] The Authors performed a detailed study about its open source implementation-Hadoop and few factors such as 1) I/O mode, the way of a reader retrieving data from the storage system, 2) data parsing, the scheme of a reader parsing the format of the records, and 3) indexing, which is used for speeding up data processing.

Map Reduce is a structure for handing out and control big scale datasets in a dispersed cluster, which has been worn for applications such as generating search indexes, text clustering, access log analysis, and a variety of other forms of data analytics. Map Reduce adopts a flexible calculation model with a simple interface consisting of a map and reduce functions whose implementations can be customized by application developers.

Since its introduction, a substantial amount of research, efforts have been directed towards making it more usable and efficient for following database-centric operations. In this paper, we aim to give an inclusive review of a wide range of proposals and systems that focusing fundamentally on the support of distributed data management and processing using the Map Reduce framework [58].

V. CONCLUSION

The lifelong personal health records can be shared by the patient with any stakeholder interested in those. The personal health record system needs security against attackers and hackers. Scalable and secure sharing includes basic securities to protect the information from unauthorized admission and failure. This paper proposed the new approach for an existing PHR system for providing more security using attribute based encryption, which plays an important role because these are unique and with difficulty hackable. We are reducing key administration problems and also we enhance privacy guarantee.

REFERENCE

[1] R.Agrawal, A. Kini, K. LeFevre, A. Wang, Y. Xu, D. Zhou, "Managing Healthcare Data Hypocritically", Proc. Of ACM SIGMOD International Conference on Management of Data, Paris, France, June 2004.

[2] D. C. Kaelber, A. K. Jha, D. Johnston, B.Middleton, and D. W. Bates, - Viewpointpaper: A research agenda for personal health records (PHRs), J. Amer. Med. Inform. Assoc., vol.15, no.6 pp.729-736,2008.

[3] Minig and Wenjing Li, Shucheng Yu, Yao Zheng, KuiRen, Lou, -Scalable and Secure Sharing of Personal Health Records in Cloud Computing using Attribute-based Encryption

[4] The economic impact of interoperable electronic health records and eprescriptionEurope EHRIMPACTEuropea n Commission, DG INFSO & Media.

[5] Health Informatics – Electronic Health Record – Definition, Scope and Context, International Standards Organization, ISO/TR 20514:2005, Jan. 2005.

[6] D. T. Mon, J. Ritter, C.Spears, and P.Van Dyke, - PHR system functional model, HL7 PHR Standard, May 2008

[7] Ziyuan Wang, "Security and privacy issues within the cloud computing ", 2011 International Confrence on Computational and Information Sciences.

[8] Zhang Xin, Lai Song-qing, Lai Nai-wen, "Research on Cloud Computing Data Security odel Based on Multidimension" 978-1-4673-2108-2113@2012 IEEE.

[9] Han Hu, Yonggang, Wen, Tat-Seng Chua, Xuelong Li "Towards Scalable systems for Big data analytics", IEEE.

[10] Tom White, Hadoop: The Definitive Guide, O'Reily Press.

[11] HLUK Demirkan, A Smart Health care system Framework, IEEE, Sept/Oct 2013.

[12] W Liu, E. K. Park, Big Data as an e-health service, in 204 International conference on computing, networking amd communication, IEEE 2014.

[13] J Gantz and D Reinsel, "The digital universe in 2020: Big data digital shadows, and biggest growth in the Far East, in Proc, IDC, iview IDC, Anal Future, 2012.

[14] Ryen W White, et al Report on the SIGIR 2013, Workshop on health search discovery, ACM SIGIR, Forum 47 (2) (2013).

[15] AJ Burns, M. Eric Johnson, "Securing Health Information:" IEEE computer society, jan/feb 205.

[16] Tao Huang, Liang Lan, Xuexian Feng, Peng An, Junxia Min, Fudi wang, "Promises and challenges of Big data in Health Sciences:, Elsevier Big Data Research vol 2, 2015,pp 2-11.

[17] Efficient Analysis of Big Data Using Map Reduce Framework International Journal of Recent Development in Engineering and Technology Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 2, Issue 6, June 2014).

[18] Apache Hiove, http://hive.apache.org/

[19] Apache Giraph Project, http://giraph.apache.org/

[20] Guoping Wang and CheeYong Chan, MultiQuery Optimization in Map Reduce Framework.

[21] P. Zadrozny and R. Kodali, Big Data Analytics using Splunk, Berkeley, CA, USA: Apress, 2013.

[22] P. C. Tang, J. S. Ash, D. W. Bates, J. M. Overhage, and D. Z. Sands, -White paper: Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption, J. Amer. Med. Inform. Assoc., vol. 13, no 2, pp 121-126, 2006.

[23] J.Hur and D. K. Noh, -Attribute- based access control with efficient revocation in data outsourcing systems, IEEE Transaction on Parallel and Distributed Systems, vol. 99, no. PrePrints, 2010.

[24] D. C. Kealber, A. K. Jha, D. Johnston, B. Middleton, and D. W. Bates, - Viewpointpaper: A research agenda for personal health records (PHRs), J. Amer. Med. In form. Assoc., vol. 15, no. 6, pp. 729-736, 2008.

[25] Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.

[26] Press G. $16.1 billion big data market: 2014 predictions from IDC and IIA, Forbes, Tech.Rep.2013.[Online].Available:http://www.forbes.com /sites/gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.

[27] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available: http://www.idc. com/prodserv/FourPillars/bigData/index.jsp.

[28] Taft DK. Big data market to reach $46.34 billion by 2018, EWEEK, Tech. Rep. 2013. [Online].Available:http://www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html.

[29] Research A. Big data spending to reach $114 billion in 2018; look for machine learning to drive ana- lytics, ABI Research, Tech. Rep. 2013. [Online]. Available: https://www.abiresearch.com/press/ big-data-spending-to-reach-114-billion-in-2018-loo.

[30] Furrier J. Big data market $50 billion by 2017—HP vertica comes out #1—according to wikibon research, SiliconANGLE, Tech. Rep. 2012. [Online]. Available: http://siliconangle.com/blog/2012/02/15/big-data-market-15-billion-by-2017-hp-vertica-comes-out-1-according-to-wikibon-research/.

[31] Kelly J, Vellante D, Floyer D. Big data market size and vendor revenues, Wikibon, Tech. Rep. 2014. [Online]. Available: http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues.

[32] Kelly J, Floyer D, Vellante D, Miniman S. Big data vendor revenue and market fore- cast 2012-2017, Wikibon, Tech. Rep. 2014. [Online]. Available: http://wikibon.org/wiki/v/

Big_Data_Vendor_Revenue_and_Market_Forecast_2012‐2017.

[33] Mayer‐Schonberger V, Cukier K. Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.

[34] Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: from big data to big impact. MIS Quart. 2012;36(4):1165–88.

[35] Kitchin R. The real‐time city? big data and smart urbanism. Geo J. 2014; 79(1):1–14.

[36] Russom P. Big data analytics. TDWI: Tech. Rep ; 2011.

[37] Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19(12):798–808.

[38] Boyd D, Crawford K. Critical questions for big data. Inform Commun Soc. 2012;15(5):662–79.

[39] Katal A, Wazid M, Goudar R. Big data: issues, challenges, tools and good practices. In: Proceedings of the Interna‐tional Conference on Contemporary Computing, 2013. pp 404–409.

[40] Pospiech M, Felden C. Big data—a state‐of‐the‐art. In: Proceedings of the Americas Conference on Information Systems, 2012, pp 1–23. [Online]. Available: http://aisel.aisnet.org/amcis2012/proceedings/DecisionSupport/22.

[41] Apache Hadoop, February 2, 2015. [Online]. Available: http://hadoop.apache.org.

[42] Cuda, February 2, 2015. [Online]. Available: URL: http://www.nvidia.com/object/cuda_home_new.html.

[43] Apache Storm, February 2, 2015. [Online]. Available: URL: http://storm.apache.org/.

[44] Curtin RR, Cline JR, Slagle NP, March WB, Ram P, Mehta NA, Gray AG. MLPACK: a scalable C++ machine learning library. J Mach Learn Res. 2013;14:801–5.

[45] Apache Mahout, February 2, 2015. [Online]. Available: http://mahout.apache.org/.

[46] Snijders C, Matzat U, Reips U-D. Big Data: big gaps of knowledge in the field of internet science. Int J Internet Sci. 2012;7(1):1–5.

[47] Abadi DJ, Carney D, Cetintemel U, Cherniack M, Convey C, Lee S, Stone-braker M, Tatbul N, Zdonik SB. Aurora: a new model and architecture for data stream manag ement. VLDB J. 2003;12(2):120–39.

[48] Kolomvatsos K, Anagnostopoulos C, Hadjiefthymiades S. An efficient time optimized scheme for progressive analytics in big data. Big Data Res. 2015;2(4):155–65.

[49] Big data at the speed of business, [online]. http://www-01.ibm.com/soft-ware/data/bigdata/2012.

[50] Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers A. Big data: the next frontier for innovation, competition, and productivity. New York: Mickensy Global Institute; 2011. p. 1–137.

[51] Gantz J, Reinsel D. Extracting value from chaos. In: Proc on IDC IView. 2011. p. 1–12.

[52] D. Peter Augustine. "Leveraging Big Data Analytics and Hadoop in Developing India's Healthcare Services" International Journal of Computer Applications (0975 – 8887) Volume 89 – No 16, March 2014

[53] J. Sun and C. K. Reddy, "Big Data Analytics for Healthcare" Tutorial presentation at the SIAM International Conference on Data Mining, Austin, TX, 2013.

[54] P. Zadrozny and R. Kodali, Big Data Analytics using Splunk, Berkeley, CA, USA: Apress, 2013

[55] F. Ohlhorst, Big Data Analytics: Turning Big Data into Big Money, Hoboken, N.J, USA: Wiley, 2013

[56] J. Zhao, J. Pjesivac-Grbovic, MapReduce: The Programming Model And Practice, 2009 [57] D. Jiang, B. C. Ooi, L. Shi and S. Wu, The performance of mapreduce: An in-depth study. Proc. VLDB Endow., 3 pp.472–483 (Sept 2010)

[57] Feng li, Beng Chin ooi, M. Tamer Ozsu, Sai wu "Distributed Data Management Using MapReduce" Acm Computing Survey.

[58] E. Srimathi, K. A. Apoorva "Privacy Preservation in Analyzing E-Health Records in Big Data Environment" International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 4 2421 – 2427.