Journal of Quantitative Finance and Economics
Volume 5; Number 1; 2023; pp : 149-158
https://DOI:10.47509/JQFE.2023.v05i01.08

ARF INDIA
Academic Open Access Publishing
www. arfjournals. com

How Healthcare Industry can Leverage Big Data Analytics Technology and Tools for Efficient Management

Sakila Akter Jahan[1], Daniul Thomas Isenberg[2] and Mesbaul Haque Sazu[3]

[1]*Sakila Akter Jahan, Illinois State University, USA*
[2]*Daniul Thomas Isenberg, Hofstra University, USA*
[3]*Mesbaul Haque Sazu, Case Western Reserve University, USA. E-mail: mesbasazu@gmail.com*

To cite this paper:

Sakila Akter Jahan, Daniul Thomas Isenberg & Mesbaul Haque Sazu (2023). How Healthcare Industry can Leverage Big Data Analytics Technology and Tools for Efficient Management. *Journal of Quantitative Finance and Economics.* 5(1), 149-158. https://DOI:10.47509/JQFE.2023.v05i01.08

Abstract: *Recently, huge quantities of organised, unstructured, and semi structured data are being produced by different systems and machines around the world. Such a huge volume of heterogenous data set that typically cannot be managed by traditional system is described as big data. The healthcare sector continues to be confronted by the importance of managing the fundamental information created by numerous sources, known for creating high volumes of heterogeneous information. Different big data analytics programs, as well as strategies, have already been created for managing these substantial quantities of information in the healthcare market. In this particular paper, we go over the effect of big data on healthcare, as well as different methods offered in the Hadoop ecosystem for managing it. We also examine the conceptual structure of big data analytics for healthcare, which involves the information collecting historical past of various limbs, the genome database, electronic health records, text/imagery, and then the medical conclusions support program.*

Keywords: *big data, MapReduce, healthcare, Hadoop, management*

1. Introduction

Each day, information is being produced by a selection of various programs, products, plus studies like geographical studies for the purpose of weather forecasting or weather prediction, disaster analysis, criminal detection, as well as the health business, to name just a few (Bernatowicz *et al.* 2015). In existing scenarios, big data is related to various enterprises and core technologies, like IBM, Facebook, and Google that extract valuable information from the massive volumes of information collected. An era of wide open information on healthcare is now under way. Big data is being produced quickly in every area, like healthcare, with respect to patient care, conformity, and different regulatory needs. As the worldwide population

Article Info.: *Received: 10 April 2023; Revised: 09 May 2023; Accepted: 25 May 2023; Online: 30 June 2023*

continues to rise, therapy delivery models are evolving fast, and several of the choices underlying these quick modifications must be based on information (Scholl et al. 2011). Medical shareholders are being promised new information out of BDA (Big Data Analytics), so called both for its range and complexity. Pharmaceutical industry professionals and shareholders have started to regularly evaluate big data to get insights. These tasks are in their first stages and should be coordinated to deal with healthcare delivery issues and improve healthcare quality. Original methods for big data analytics of healthcare informatics have been started throughout numerous scenarios, e.g., the study of patient qualities, as well as dedication of therapy expense, as well as outcomes to identify the most and best cost effective solutions. Health informatics is the assimilation of healthcare sciences, computing sciences, and information sciences in the research of healthcare information (Liebeskind & Feldmann 2015). Health informatics entails information acquisition, storage space, and retrieval to offer much better outcomes by healthcare providers. Heterogeneity, in the medical system, characterises a wide range of information as an outcome of the linking of a diverse selection of biomedical data energy sources, including free text, laboratory tests, gene arrays, imagery, sensor data, for example, and demographics. Most of the data accessed through healthcare devices is unstructured and is not stored electronically i.e., it exists just in hard copies, and its volume is rapidly growing. Presently, there's a significant focus on digitisation of these huge stores of hard message information. The revolutions of information size are actually producing an issue to accomplish this goal. The different models and terminologies created to solve the issues related to big data focus on solving four concerns called the 4 Vs, namely: volume, velocity, variety, and veracity (Dougherty 2009). The different types of information in healthcare applications include Electronic Health Records, machine generated/sensor information, healthcare information exchanges, genetic databases, portals, patient registries, and public records. Records that are public are the major sources of big data in the medical industry, and need effective data analytics to solve their connected healthcare issues. Based on a survey done in 2012, healthcare information totaled almost 550 petabytes and was expected to achieve roughly 26,000 petabytes in 2020 (Gessner et al. 2013). In the light of the heterogeneous information formats, great amounts, and associated uncertainties in the big data energy sources, the process of knowing the transformation of raw data into actionable information is daunting. Being quite intricate, the identification of overall health capabilities in the selection and medical data of class attributes for overall health analytics requires advanced and architecturally designed tools and techniques.

2. Literature Review

2.1. Big Data Analytics in Health Informatics

The primary distinction between regular health evaluation and big data health analytics is the use of computer programming. In the standard phone system, the healthcare industry trusted various other industries for big data evaluation (Hussain & Nguyen 2014). Lots of healthcare shareholders trust information technology due to its significant outcomes- the operating systems of theirs are functional, and they can process the information into standardised forms. Nowadays, the medical industry is confronted with the task of managing quickly and building huge healthcare information (Golnabi, Meaney & Paulsen 2013). The area of big data analytics is growing and possesses the possibility to offer helpful insights because of the medical phone. As noted previously, nearly all the substantial quantities of data generated by this particular method are saved to hard copies, which must subsequently be digitised. Big data can boost healthcare delivery and lower the cost while supporting innovative patient care, enhancing patient outcomes, and staying away from unnecessary costs (Hussain et al. 2014). Big data analytics is now used to anticipate the results of doctors' choices, the result of a heart functioning for just an ailment based on the patient's age, present state, and health status (Mustafa, Mohammed & Abbosh 2013). Basically, we can state that the job of BDA in the market is to manage information sets regarding healthcare that are difficult and complex to handle utilising the latest hardware, software, CD, and management tools. And the burgeoning amount of healthcare information, reimbursement techniques will also be changing. Thus, purposeful use and pay based on performance have emerged as factors that are important in the healthcare market. This year, organisations operating in the area of healthcare have developed over 150 exabytes of information, each of which should be effectively analysed for it to remain helpful to the healthcare system. The storage of healthcare connected data within EHRs (Electronic Health Records) occurs in various forms (Desjardins et al. 2009). An unexpected rise in information regarding healthcare informatics has also been found in bioinformatics, where genomic sequencing produces lots of terabytes of information. You will find various analytical methods offered for interpreting health-related information, which may be used for patient care. Several forms and origins of big data are challenging the healthcare informatics group to create techniques for information processing. There's a huge need for a method that combines dissimilar data solutions (Baio 2013).

Many conceptual techniques could be used to identify problems in huge quantities of information from numerous datasets. The frameworks readily available for the evaluation of healthcare information are as follows:

In the past two years, predictive evaluation has been recognised as among the leading business intelligence methods; though its real world applications extend beyond the company context. Big data analytics includes different strategies, including text analytics and multimedia analytics (Tempany et al. 2015). Nevertheless, among the most essential categories is predictive analytics, including statistical techniques such as data mining and machine learning, which look at historical and current information to foresee the future. Predictive methods are now used in the clinic context to decide whether the patient is in danger of readmission. These data can help doctors make crucial patient care decisions. Predictive analysis requires understanding and utilisation of machine learning, which is commonly used in this method.

2.2. Four Vs of Big Data in Healthcare

Four main attributes which are connected with big data are volume, variety, velocity, and veracity.

2.2.1. Volume: Big data is a phrase to talk about large volumes of collected information. There's no fixed threshold for the amount of information. Generally, the word is used regarding massive scale data, which should be managed, saved, and examined by regular data and database processing architecture. The amount of information produced by contemporary IT and the healthcare system is developing, and it is further pushed by the diminished expense of information storage, as well as the need and processing architectures to acquire useful insights from information to enhance business processes, advantages, and services to customers (Widmer *et al.* 2014).

2.2.2. Velocity: Velocity, which presents the main reason behind the exponential growth of information, refers to how fast information is collected. Healthcare systems generate information at progressively higher speeds. In the context of the amount and number of the unstructured or structured data collected, the velocity of the development of information after processing takes a choice based upon its output (Shvachko *et al.* 2010).

2.2.3. Variety: Variety describes the type of information, i.e., structured or unstructured, video, audio, medical imagery, text, and sensor information. Structured data information includes medical data, which should be collected, saved, and processed by a specific device. Structured information comprises only five per cent to ten per cent of healthcare information. Semi-structured or unstructured data include e-mails, pictures, videos, recordings, and additional health associated data like hospital medical accounts, physician notes, paper prescriptions, as well as radiograph films (Duarte *et al.* 2007).

2.2.4. Veracity: The veracity of information is the level of guarantee that the information is significant and consistent. Distinct data sources vary in their credibility and reliability. The results of big data analytics should be error-free and credible, but in healthcare, unsupervised machine learning algorithms make choices that are used by automatic models based on information which might be useless or even misleading. Medical analytics are tasked with extracting helpful insights from this information for treating individuals and creating the absolute best choices (Dean & Ghemawat 2008).

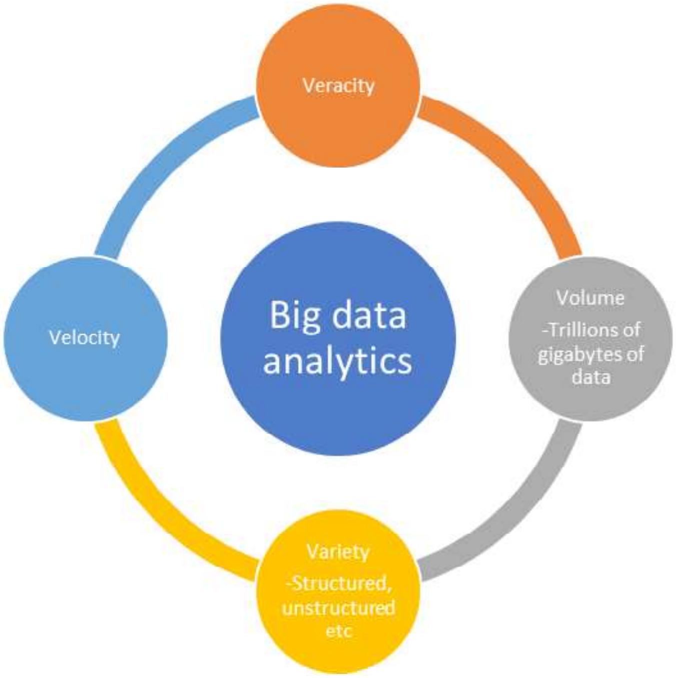
Veracity
Velocity
Big data analytics
Volume -Trillions of gigabytes of data
Variety -Structured, unstructured etc

Figure 1: Components of big data

2.3. Impact of Big Data on the Healthcare System

The possibility of big data is that it could revolutionise results about probably the most appropriate or maybe correct patient analysis, and the reliability of information applied to the informatics process. As a result, the investigation of massive amounts of information will have a strong impact on the medicinal services framework in five respects, or even "pathways". Improving outcomes for individuals with regard to these routes, as discussed below, is the emphasis of the medical system and can immediately influence the individual (Sobhy, Sonbaty & Elnasr 2012).

2.4. Hadoop Based Applications for Health Industry

In the light of the reality that healthcare information is largely printed, there's a requirement for the energetic digitisation of print type information. The bulk of this data is additionally unstructured, so it's a significant challenge of this industry to extract significant information regarding research, clinical operations, and patient care. The group of software utilities referred to as the Hadoop ecosystem can help the healthcare market manage this vast amount of information. The different uses of the Hadoop ecosystem in the healthcare market are as follows:

3. Big Data Analytics Architecture for Health Informatics

Currently, the primary target in big data analytics is to gain thorough understanding and insight of big data, not to gather it (Marbach et al. 2010). Information analytics entails the growth and use of algorithms for examining different complex data sets to acquire significant awareness, patterns, and information. Recently, scientists have started considering the proper architectural framework for healthcare devices that use big data analytics, among which is a four-layer architecture, which comprises a transformation layer, data source level, big data platform level, and analytical level. In this layered phone system, information originates from various sources and has different storage and format methods. Each layer features certain data processing functionality for carrying out certain duties on the HDFS, using the MapReduce processing version. The other levels perform various other jobs, i.e., report generation, query passing, data mining processing, and web based analytical processing (Raman & Chandra 2009).

The primary requirement in big data analytical processing is to bundle the information at a high speed to lessen the bundling time. The other priority in big data analytical processing is usually to effectively update and transform queries in a continuous time. The third requirement in the big data analytical processing is to effectively use and effectively manage the storage area room (Haggart et al. 2011). The final specification of big data analytics is to effectively become acquainted with the rapidly progressing workload notations. Big-data analytics frameworks differ from conventional healthcare processing methods in terms of the way they approach big information. In the present health care system, information is processed using conventional instruments installed in a single stand alone method, such as a desktop system. In contrast, big data is processed by clustering and goes through numerous clusters in the system. This particular processing depends on the idea of parallelism to deal with huge health data sets. Readily accessible frameworks, like HBase Avro, Hive, Sqoop, Pig, MapReduce, and Hadoop, can process the related data sets for healthcare systems.

3.1. Apache Pig

Apache Pig is among the accessible open source platforms used to better evaluate big information. Pig is a substitute to the MapReduce programming tool. Initially created by the Yahoo web service provider as a research project, Pig enables users to create their own user-defined functions and supports many standard data functions including subscribe, filter, sort, etc. (Yizhak *et al.* 2010).

3.2. Apache HBase

HBase is a column oriented NoSQL repository used in Hadoop, in which a person can store huge amounts of columns and rows (Radrich et al. 2010). HBase has the performance of arbitrary read/write operations. Additionally, it supports record amount revisions, which isn't possible using HDFS. HBase offers parallel data storage through the underlying distributed file methods across commodity servers.

3.3. Apache Zookeeper

Zookeeper is a centralised program used to keep a healthcare system and supply planning, and other elements on as well as between nodes. It maintains the typical objects required in big cluster environments, which includes configuration information and the hierarchical naming room. These solutions may be used by diverse programs to harmonise the dispersed processing of Hadoop clusters. Zookeeper additionally guarantees application reliability. If an application master dies, Zookeeper creates a new program master to continue the duties (Henry et al. 2010).

3.4. Apache Yarn

Hadoop Yarn is a distributed shell program, and it is a good example of a Hadoop non MapReduce program built on top of Yarn. Yarn has 2 parts, a Source Manager, which handles all the materials within a bunch necessary for the chores, as well as a Node Manager, located on each host at a bunch, and handles the readily available information on the independent host. Both parts handle the scheduling of work and control the pots, CPU throughput, memory management, and I/O is which run the committed application code.

3.5. Apache Sqoop

Apache Sqoop is a great tool that removes the information from Relational Database Management System and places it within Hadoop architecture for query processing. To do so, this procedure uses the MapReduce paradigm or any other regular level programs, e.g., Hive. When placed in HDFS, the information may be used by Hadoop applications.

3.6. Apache Flume

Apache Flume is a reliable service for correctly gathering information and moving large volumes of information from independent devices to HDFS. Often information travel involves a selection of flume agents that could traverse many locations and machines. Flume is commonly used for log documents, information produced by social networking, and email communications.

4. Conclusion

In this particular paper, we've provided a thorough explanation and brief introduction to big data in common and in the healthcare system, which plays a tremendous part in healthcare informatics and significantly influences the healthcare system. Here we have also given the fundamental information on the four Vs in healthcare. We have additionally proposed the use of a conceptual design for fixing healthcare issues in big information using Hadoop based terminologies. This consists of the fundamental data, produced by various levels of the development and medical data of techniques for analysing this information and getting solutions to health issues. A blend of big data as well as healthcare analytics can prescribe treatments which work well for particular patients by having the ability to prescribe proper medications for every individual, instead of the ones that work for most individuals. We all know that big data analytics is in the first phase of growth, and existing methods and tools can't resolve the issues related to big data. Large data might be viewed as a serious concern which presents massive challenges. Thus, a lot of investigation in this particular area is expected to resolve the problems experienced by the healthcare phone.

References

Baio, G. (2013). Molecular imaging is the key driver for clinical cancer diagnosis in the next century! *Journal of Molecular Imaging & Dynamics,Vol.2*, article e102, 2013.

Bernatowicz, K., Keall, P., Mishra, P., Knopf, A., Lomax, A., & Kipritidis, J. (2015). Quantifying the impact of respiratory-gated 4D CT acquisition on thoracic image quality: A digital phantom study. *Medical Physics, Vol.42*, no. 1, pp. 324–334.

Dean, J. & Ghemawat, S. (2008). MapReduce: Simplified data processing on large clusters. *Communications of the ACM, Vol. 51*, no. 1, pp. 107–113.

Desjardins, B., Crawford, T., Good, E., Oral, H., Chugh, A., Pelosi, F., Morady, F., & Bogun, F. (2009). Infarct architecture and characteristics on delayed enhanced magnetic resonance imaging and electroanatomic mapping in patients with postinfarction ventricular arrhythmia. *Heart Rhythm, Vol. 6*, no. 5, pp. 644–651.

Dougherty, G. (2009). *Digital image processing for medical applications,* Cambridge University Press.

Duarte, N.C., Becker, S.A., Jamshidi N., Palsson, B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences of the United States of America, Vol. 104,* no. 6, pp. 1777– 1782.

Gessner, R.C., Frederick, C.B., Foster, F.S., & Dayton, P.A. (2013). Acoustic angiography: A new imaging modality for assessing microvasculature architecture. *International Journal of Biomedical Imaging, Vol.2013*, Article ID 936593, 9 pages.

Gianchandani, E.P., Chavali, A.K., & Papin, J.A. (2010). The application of flux balance analysis in systems biology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, Vol. 2*, no. 3, pp. 372–382.

Golnabi, A.H., Meaney, P.M., & Paulsen, K.D. (2013). Tomographic microwave imaging with incorporated prior spatial information. *IEEE Transactions on Microwave Theory and Techniques, Vol.61,* no. 5, pp. 2129–2136

Haggart, C.R., Bartell, J.A., Saucerman, J.J., & Papin, J.A. (2011). Whole-genome metabolic network reconstruction and constraint-based modeling. In *Methods in systems biology,* M. Verma, D. Jameson, & H.V. Westerhoff, (Eds.). *Vol. 500 of Methods in Enzymology*, chapter 21, pp. 411–433, Academic Press.

Henry, C.S., Dejongh, M., Best, A.A., Frybarger, P.M., Linsay, B., & Stevens, R.L.(2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology, Vol. 28*, no. 9, pp. 977–982.

Hussain T., & Nguyen, Q.T. (2014). Molecular imaging for cancer diagnosis and surgery. *Advanced Drug Delivery Reviews, Vol.66*, pp. 90–100.

Hussain, A.M., Packota, G., Major, P.W. & Flores-Mir, C. (2014). Role of different imaging modalities in assessment of temporo- mandibular joint erosions and osteophytes: A systematic review. *Dentomaxillofacial Radiology, Vol. 37*, no. 2, pp. 63–71.

Lewis, N.E., Nagarajan, H., & Palsson, B.O. (2012). Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature Reviews Microbiology, Vol. 10,* no. 4, pp. 291–305.

Liebeskind, D.S., & Feldmann, E. (2015). Imaging of cerebrovascular disorders: Precision medicine and the collaterome. *Annals of the New York Academy of Sciences.*

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., & Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences of the United States of America, Vol. 107,* no. 14, pp. 6286–6291.

McCloskey, D., Palsson, B.O., & Feist, A.M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Molecular Systems Biology, Vol. 9,* article 661.

Mustafa, S., Mohammed, B., & Abbosh, A. (2013). Novel prepro- cessing techniques for accurate microwave imaging of human brain. *IEEE Antennas and Wireless Propagation Letters, Vol.12*, pp. 460–463.

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., & Schwartz, J.M. (2010). Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology, Vol. 4*, article 114.

Raman K., & Chandra, N. (2009). Flux balance analysis of biological systems: Applications and challenges. *Briefings in Bioinformatics, Vol. 10*, no. 4, pp. 435–449.

Scholl, I., Aach, T., Deserno, T.M., & Kuhlen, T. (2011). Challenges of medical image processing. *Computer Science-Research and Development, Vol.26*, no. 1-2, pp. 5–13.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop distributed file system. In *Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies* (MSST ’10), pp. 1–6, IEEE.

Sobhy, D., El-Sonbaty, Y., & Abou Elnasr, M. (2012), MedCloud: Healthcare cloud computing system. In *Proceedings of the International Conference for Internet Technology and Secured Transactions*, pp. 161–166, IEEE, London.

Tempany, C.M.C., Jayender, J., Kapur, T., Bueno, R., Golby, A., Agar, N., & Jolesz, F.A. (2015). Multimodal imaging for improved diagnosis and treatment of cancers. *Cancer, Vol. 121,* no. 6, pp. 817–827.

Widmer, A., Schaer, R., Markonis, D., & Mu¨ller, H. (2014). Gesture interaction for content-based medical image retrieval. In *Proceedings of The 4th ACM International Conference On Multimedia Retrieval,* pp. 503–506, ACM.

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., & Shlomi, T.(2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics, Vol. 26*, no. 12, Article ID btq183, pp. i255–i260.