# Critical Analysis of Big Data Technology in Healthcare: A review

**Zulfiqar Ali [1], Qasim Abbas [2], Muhammad Rashid [1]**

[1] Department of Computer Science
National University of Technology
Islamabad, Pakistan

[2] Department of Computer Science
The Queen's University of Belfast, Northern Ireland,
United Kingdom.

***Abstract-*** For the past many years, the health industry has generated a large amount of data and it is rapidly growing with time. There is a need to store patient precious data efficiently and reliably. More efficient tools and techniques, practices, and research are required to make improvements and obtain maximum benefits from health care data. Big data analytics is playing an important role in health care. It helps in physician's practices, maintaining patient records as well as hospitals in diagnosing diseases more precisely and accurately. In this paper, we have highlighted the role of big data in Electronic Health Records (EHRs). Under focus, the study provides an analysis of Big Data Technology in Healthcare discusses the importance of big data in health care, and compares it based on different parameters i.e., tools, techniques, classification, and performance. Furthermore, the challenges faced in the development and employment of big data-based technology in the field of the medical field are identified.

***Index Terms***- Big Data in Healthcare, Big data analytics, Big data comparison, EHRs, Quest diagnosis.

## I. INTRODUCTION

As per history, we realize that the health industry has created a lot of information, by record keeping, patient care, recording/maintaining drugs, and patient care. The data was stored in hard form and now it is the era of digitization. The potential to improve the performance of healthcare and at the same time reduce the costs, for the huge amount of data which hold the promise of supporting a large amount of data, including decisional clinical reports and population health management. There are two main kinds of sources in the health industry. First, one is genomics-driven which consists of gene expression, genotyping, and sequencing data. The other one is payer-provider big data which contains EHRs, patient feedback, insurance records, and pharmacy prescriptions. Reports show in 2011 U.S.A healthcare system data exceeded 150 Exabytes. At this rapid growth and increment in the rate, the data will reach Zettabyte (1021 GBs) and will reach Yottabyte (1024 GBs) [1]. For example, In California, the healthcare network has data in between 26-44 petabytes which comes from EHRs (Electronic Health Records) in the form of images or annotations.

Big data in the medical field refers to electronic digital data and hospital or patient data are so huge or in a larger amount and complex enough that it is becoming difficult to manage and analyze with traditional techniques, software, or hardware. Enormous information is overpowering not due to its volume but because of its diversity in data types and the speed at which it must be managed. It includes electronic data from CPOE, clinical or medical notes, laboratory test reports, medical images, patient insurance, electronic data, handwritten notes or prescriptions by physicians, and much more [2].

For the big data scientists, there is this vast amount and array of data which is much like massive hype and the perplexing range of data. Big data analytics have to manage this data so that we have less complexity and less cost. With the help of this big data, we can save lives and improve care.

This paper provides an overview of big data analytics and its importance in healthcare. Step by step we will go through it as first we discuss and define the advantages of big data in healthcare. Secondly, we will compare it with older technologies. In the third step, the four Vs. The fourth step is the role of big data analytics. Lastly, conclusions and future work are discussed.

Section II of the paper explores and investigates the advantages of the newly developed big data-based models for medical fields for the detection of complex diseases. The architecture of big data analytics exploited by the researchers is explained. Section III provides the comparison and analysis of new big data technology and conventional methods used in the healthcare systems. Section IV provides the role of big data in healthcare and the challenges faced in the development and deployment of data analytics in the field of medical science. Finally, the last section concludes the study on comparative analysis and review of big data technology in the field of healthcare systems.

## II. BIG DATA- ADVANTAGES IN HEALTHCARE

Big data in health has a lot of advantages as well as benefits. We can realize its benefits from the networks of Healthcare organizations ranging from the office of a single-physician or multi-provider groups to the larger hospitals [3]. Its potential benefit also includes detecting diseases at the very initial stages which help more in cure and treatment, managing patient records more effectively as well as easily and efficiently. According to McKinsey's big data analytics, U.S. healthcare saves $300 billion annually which is 2-3rd of that obtained through reductions of 80% in healthcare expenditures nationally [4]. Clinical Operations and Research and development are the main areas for potential savings. Figure 1 shows the architecture of big data analytics more clearly.

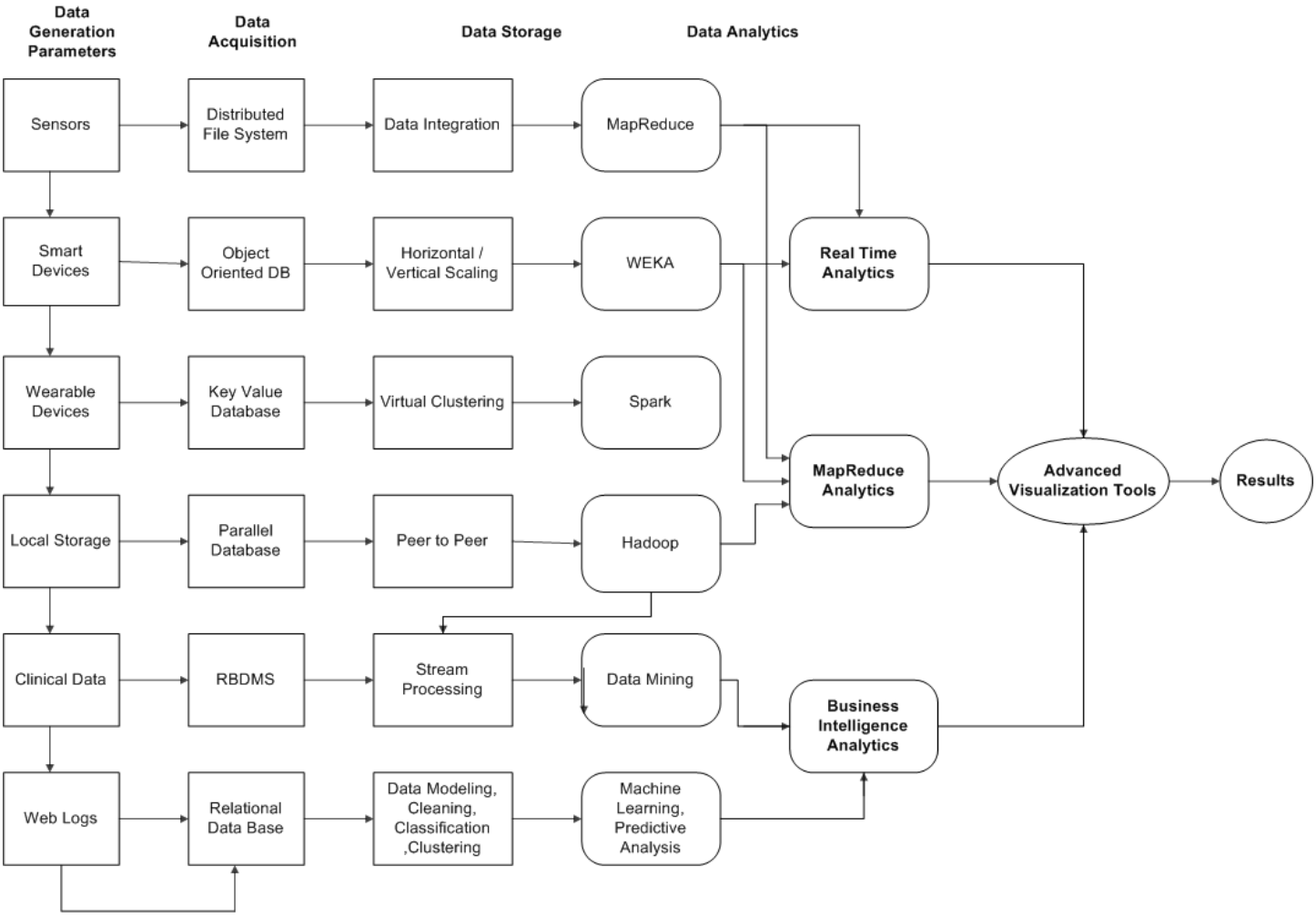


Figure 1: Architecture of big data analytics

### *A. The concept of Vs in big data*

The letter V in big data analytics stands for:

- volume
- velocity
- variety
- veracity
- Value

Before 5 V's concept, there was a concept of only 4 V's through which characteristics of data were described which includes only volume, velocity, variety, and veracity. Over time, the fifth characteristic was added known as value [5].

Volume: Already or existing larger volume of health care data which includes radiology images, clinical data, personal medical records, etc. genomics and biometric sensor readings, and 3D imaging are also responsible for the extra growth of data. Rapid advancements and developments in cloud computing, data management, and particularly virtualization are responsible for facilitating the development of new more effective platforms, storage, and huge data volume. The constant increase in the flow of newly generated and existing data rates introduces new issues and challenges. Automatically generated data by machine is also contributing to Big Data. It involves the data sensed by the sensors for example environment or manufacturing sensors, smart cards, and scanning machines.

Velocity: By velocity, it means the entrance of a large number of real-time data from different medical devices and equipment or

machines. Traditionally, healthcare data was static mostly in the form of X-rays, scripts, paper files, etc. The velocity of data also increases daily with patient measurements of blood pressure, temperature, EKGs, Diabetic glucose, and heartbeat readings. There are many other sources of data incoming like real-time data that comes from ICUs, operating rooms, etc.

Variety: There lies variety in healthcare data. Nowadays the data is not only structured but also in multimedia and unstructured formats too. The variety of data in different formats (structured, unstructured, semi-structured) makes healthcare challenging as well as interesting. Structured data is in the form of hospital and physician details (EHRs and EMRs), patient records (name, date of birth, age, etc.), and treatment codes. It is easy to maintain, analyze, and store by the machines. Semi-structured contains readings from different medical equipment and machines, medical records as well as electronic records on paper. Unstructured data consists of all handwritten notes or papers either by a doctor or a nurse. It may also include admission and discharge records of hospitals, CT, MRI, and other relevant things.

Veracity: It is the fourth characteristic of big data. Veracity is the data assurance that shows that the outcomes of big data are error-free. It is one of the goals to achieve in big data. To achieve quality, it faces two main challenges. First, one is accurate readings or accurate information of the patient in a matter of death or life. Secondly, the healthcare data especially from unstructured data having poor or faded handwriting may translate incorrect information. So, veracity deals with quality performance, tools, and methodologies or algorithms to meet the demands and needs of quality standards.

Value: It is the fifth characteristic of big data which deals with the value of data. The value of big data in business is of great importance. It concerns the worth of information which makes it very useful in business for example Netflix and Amazon [5]. The value of data may depend on events or activities (randomly or regularly) they represent. A collection of data may also depend on requirements and demands or some special conditions for long storage and analysis, for example, patient history records and patient personal data in healthcare.

## III. BIG DATA- A COMPARISON

To understand more about Big Data let us consider Table 1. It shows the comparison of Big Data with the existing technologies. Table 2 shows Research work Classification based on disease detection using Big data Tools. Table 3 shows Big Data Analytics applied data formats for Machine Learning Techniques. The different data models specifically supported healthcare data along with the supported big data tools in analyzed in this table the performance and speed of analysis performed on those data models are described in Table 4.

### A. *Big Data vs. existing technologies*

TABLE I, provide the comparison between the existing old techniques and methods used for the discovery of knowledge from data reservoirs concerning the emerging techniques and technology employed to extract non-trivial and useful information from the big data. The benefits and usefulness of big data technology can be perceived very easily with comparative analysis.

TABLE I. BIG DATA VS EXISTING TECHNOLOGIES

| BIG DATA Technology | Existing old Technology |
|---|---|
| In a real-time situation, it helps companies to identify errors or problems at early stages. | It may not be very helpful in real-time situations to identify problems at the beginning stages. |
| Improves in generating extra revenue and a high conversion rate. | Existing techniques are generating fewer revenues than Big data and have low conversion rates. |
| Highly secured and fraud can be easily detected. | Can be hacked and have security limitations or flaws. |
| High profit and better sales of the product. | Low profit and low sales. |
| Big data tools and techniques may be expensive. | Inexpensive tools and techniques. |
| Needs a special type or extra computer powers. | Can be used with any computer power. |
| Limited open-source availability. | Open source available. |
| Save historical and precious data efficiently like patient data. | Not much efficient to save data for a long time is an issue. |

TABLE II. CLASSIFICATION BASED ON DISEASE DETECTION

| References | Technique used | Healthcare beneficial results/ disease detection | IoT device | Parameters (accuracy, performance, etc) | Big data Analytical tool used | Results/outcomes |
|---|---|---|---|---|---|---|
| [6] | predictive analytics method | chronic kidney failure | Multiple sources (Wearable, Bluetooth, phone, ta blet) | Cost-effective | Hadoop/Map-Reduce tool | Predict the severity of kidney failure & foretell future |
| [7] | map-reduce technique | Heart disease | Sensor networks such as body sensor network | High accuracy | Hadoop | Shows the details of the patient's heart, patient condition, and classification result of heart disease. |
| [8] | predictive analysis algorithm | cardiovascular diseases | Structure, unstructured data from IoT devices such as cameras. | Flexible, Scalable, Fast, Resilient to Failures, Cost-effective. | Hadoop, MapReduce, HDFS | Detection of disease, quantification as well as monitoring disease progression. |
| [9] | SVM Hybrid Model, ID3 algorithms | Breast Cancer | Wearable devices | 95% accuracy | Weka | Two types of breast cancer classification: (Benign/Malignant). |
| [10] | Map-Reduce technique | Infectious diseases like Viruses, Prions, Bacteria, Nematodes, etc. | Pulse Oximeter sensor, Heartbeat sensor, temperature sensor, Respirology sensor, pressure sensor. | Forced, efficient, delayed prevention for the high, medium, and low vulnerable areas to dengue. | HDFS | Dengue detection: (Highly, Mid and Low) vulnerable to dengue fever. |
| [11] | predictive analysis algorithm | Patient monitoring | Wireless Body Area Network: wearable sensors, Bluetooth. | Efficient, Location-free | CDS programming, Hadoop | Supervision of patients with a high danger of heart disappointment. |
| [12] | Machine Learning techniques | E-Health services | wearable and body sensors. | energy efficiency, workload balancing, and reliability | Hadoop, Hive, etc. | Patient's current and predicted future health |
| [13] | Realtime pattern recognition techniques | E-Health system | miniature wearable biosensors | greater accessibility, availability, accuracy | Cross-platform programming | Process queries about patient's genomic, cellular, and organ levels and judgments on follow-up practices and procedures in the biochemical realm. |
| [14] | Map-Reduce technique | health monitoring system | heart rate sensors, humidity sensors, and blood pressure sensors. | Less response time, suitable for real-time alerting | Hadoop | personalized care, Help physicians in patient monitoring. |
| [15] | Fuzzy K-nearest neighbor algorithm | Zika virus | IoT sensors/ actuators, mobile phones: GPRS, mosquito sensors, etc. | High accuracy | Amazon EC2 | early diagnosis of Zika virus |

**A.** *Big data model comparison based on performance on medical data.*

TABLE III provides the review and performance analysis of frameworks and models based on big data technology specifically in the context of medical databases available in the contemporary literature. The paper reference is to facilitate the readers and researchers by providing knowledge and information about the supported platforms like Hive, Pig, MapReduce, HBase, Weka, Spark, etc. Furthermore, TABLE III provides the performance analysis of the models by considering the formats of the medical databases.

TABLE III. BIG DATA MODEL COMPARISON BASED ON PERFORMANCE OF MEDICAL DATA.

| Author and research references | Data formats (Models) | Supported Platforms | Speed | Performance |
|---|---|---|---|---|
| [16] | Text | Hive or MapReduce | < file sequence format | < CSV format |
| [17] | Serialization Avro | HBase, MapReduce, Pig, Hadoop, Spark, and Hive | Scalability ratio decreases for many columns<br>< Parquet | Efficient for multiple rows<br>< ADAM stack and Parquet |
| [18] | ARFF | Weka | < file sequence format | More efficient as compared to CSV format |
| [19] | JSON | Hive or MapReduce | < file sequence format | < CSV format |
| [20] | Sequence | Hive or MapReduce | Compatible for small data sets rather than large data sets | Work efficiently for small data sets |
| [21] | CSV | Hive or MapReduce | < file sequence format | Efficient as compared to CSV file format |
| [22] | Parquet | HBase, MapReduce, Pig, Hadoop, Spark, and Hive | Scalability ratio decreases for a large number of rows<br>Efficient as compared to Avro | Efficient for multiple columns<br>< ADAM stack<br>Indexing not supported<br>Efficient Avro |
| [23] | ADAM stack | Apache Spark strongly supported | Much efficient as compared to VCF, Parquet, and Avro | Much efficient as compared to VCF, Parquet, and Avro |
| [24] | Optimized Row Compressed | MapReduce, Pig, Hadoop, Spark, and Hive<br>HBase not supported | Efficient for processing of querying<br>Efficient as compared to Avro | Indexing supported<br>Much efficient than Avro<br>Works effectively with Hive |

## IV. ROLL OF BIG DATA IN HEALTHCARE

Health data is expected to grow dramatically due to technological advancements in the medical field. Profit is not only a concern but saving and managing large patient data is also a key factor to avoid losing potentially millions of dollars in revenue. Today the hottest topic in medical health is to improve health, enhance the patient care experience, and reduce the cost of care. To understand the concept of Big data more clearly, we consider a real-time example.

Quest Diagnoses (care360) is one of the world's leading web-based providers of healthcare information technology solutions. A survey shows that more than 200,000 including hospitals, physicians, and PHDs are using this at more than 95,000 locations or offices in the USA [25]. It includes:

- Patient Appointments and Scheduling visits
- Management of electronic labs and results in reports
- Clinical messaging and Contact reminders for patient facility
- SOAP notes, e-prescribing
- Electronic management including imported papers and document scanning
- PHR (Personal Health Records) integration
- Patient history, demographics, and health status
- 24/7 easy access through smartphones like iPhones etc.
- Customer access and patient treatment plans
- Billing service, payment, and revenue cycle

The above bullets are general features of care360 EHR, it has many other characteristics too [6]. So, the problem is how to manage a large amount of data. How to store it easily? How to secure large patient data? How to make it more easily accessible for both physicians as well as customers? The answer to all these questions is the term "BIG DATA". The information like the above-defined not only includes the sheer volume of data but also its diversity and complexity. Although it uses the cloud computing concept to manage data for patients and physicians. Quest is using big data techniques for better performance which helps physicians to do their best. For example, in 2011, Dr. Kolodziej proposed a solution to use Big data technology in cancer care transformation showing that big data has a great impact on healthcare.

### A. *Challenges of big data*

In this subsection some of the challenges reported in the contemporary literature to exploit the big data-based models and frameworks in the field of medical field are provided here:

- Obtaining accurate information from unstructured data [26].
- Efficient storage and query optimization [27].
- Processing of large volumes of data using low-level processors or machines.

- Diversity in medical scales and units especially in healthcare lab.
- Incomplete medication data (inpatient or outpatient) in EHRs is difficult to analyze.
- Dealing with clinical notes and lab results has diversity, misspellings, and grammatical mistakes in them [28].

## V. CONCLUSION AND FUTURE RESEARCH

Over time, healthcare data is increasing day by day. Existing tools and methodologies are not sufficient to deal with such large, complex, and variety of data. Big data is playing an important role in healthcare. It helps healthcare providers and physicians or doctors in the betterment of their practices, maintaining patient records, and many other facilities. However, big data is facing some challenges and issues which need to be addressed. In the future, we will propose a solution to deal with the challenges and issues of big data.

## ACKNOWLEDGMENT

The authors wish to thank the Higher Education Commission of Pakistan and the National University of Technology (NUTECH), Islamabad, Pakistan. This work is supported in part by a grant from the Higher Education Commission of Pakistan and the National University of Technology (NUTECH).

## REFERENCES

[1] M. Cottle, W. Hoover, S. Kanwal, M. Kohn, T. Strome, and N. Treister, "Transforming Health Care Through Big Data Strategies for leveraging big data in the health care industry," *Institute for Health Technology Transformation, http://ihealthtran. com/big-data-in-healthcare,* 2013.

[2] W. Raghupathi and V. Raghupathi, "An overview of health analytics," *J Health Med Informat,* vol. 4, no. 132, p. 2, 2013.

[3] C. Burghard, "Big data and analytics key to accountable care success," *IDC health insights,* vol. 1, pp. 1-9, 2012.

[4] J. Manyika *et al.*, *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011.

[5] M. Jaiswal, "Big Data concept and imposts in business," *Manishaben Jaiswal'Big Data Concept and Imposts in Business' International Journal of Advanced and Innovative Research (IJAIR) ISSN,* pp. 2278-7844, 2018.

[6] A. Batra, U. Batra, and V. Singh, "A review to predictive methodology to diagnose chronic kidney disease," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016: IEEE, pp. 2760-2763.

[7] G. Vaishali and V. Kalaivani, "Big data analysis for heart disease detection system using map reduce technique," in *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, 2016: IEEE, pp. 1-6.

[8] S. Thakur and M. Ramzan, "A systematic review on cardiovascular diseases using big-data by Hadoop," in *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 2016: IEEE, pp. 351-355.

[9] K. Sivakami and N. Saraswathi, "Mining big data: breast cancer prediction using DT-SVM hybrid model," *International Journal of Scientific Engineering and Applied Science (IJSEAS),* vol. 1, no. 5, pp. 418-429, 2015.

[10] C. Mohapatra, L. Das, S. S. Rautray, and M. Pandey, "Map-reduce based modeling and dynamics of infectious disease," in *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, 2017: IEEE, pp. 895-898.

[11] A. Priyanka, M. Parimala, K. Sudheer, R. Kaluri, K. Lakshmanna, and M. P. K. Reddy, "BIG data based on healthcare analysis using IOT devices," in *IOP Conference Series: Materials Science and Engineering*, 2017, vol. 263, no. 4: IOP Publishing, p. 042059.

[12] F. Firouzi *et al.*, "Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics," ed: Elsevier, 2018.

[13] "The Role of Big Data in Health Care Internet of Things (IoT)." https://benthomasagency.com/wp-content/uploads/2018/04/The-Role-of-Big-Data-in-Health-Care-Internet-of-Things.pdf (accessed 26 Feb 2019.

[14] P. Dineshkumar, R. SenthilKumar, K. Sujatha, R. Ponmagal, and V. Rajavarman, "Big data analytics of IoT based Health care monitoring system," in *2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON)*, 2016: IEEE, pp. 55-60.

[15] S. Sareen, S. K. Gupta, and S. K. Sood, "An intelligent and secure system for predicting and preventing Zika virus outbreak using Fog computing," *Enterprise Information Systems,* vol. 11, no. 9, pp. 1436-1456, 2017.

[16] E. Gardner, "PRACTICAL BIG DATA," *Health Data Management,* vol. 22, no. 4, pp. 18-20,22, 2014.

[17] M. Kim and J. Park, "Identifying and prioritizing critical factors for promoting the implementation and usage of big data in healthcare," *Information Development,* vol. 33, no. 3, pp. 257-269, 2017, doi: 10.1177/0266666916652671.

[18] Z. Liu, W. J. Zheng, G. I. Allen, Y. Liu, J. Ruan, and Z. Zhao, "The International Conference on Intelligent Biology and Medicine (ICIBM) 2016: from big data to big analytical tools," *BMC Bioinformatics,* vol. 18, no. Suppl 11, 2017, doi: 10.1186/s12859-017-1797-3.

[19] S. Luo, "Invest in your data: how clinical mobility solutions liberate data and drive cost savings.(FEATURE STORY)," *Healthcare Financial Management,* vol. 70, no. 2, p. 66, 2016.

[20] "Quantzig Highlights Top Big Data Trends in the Healthcare Industry," *Wireless News,* 2018.

[21] R. S. H. Istepanian and T. Al-Anzi, "m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics," *Methods,* vol. 151, pp. 34-40, 2018, doi: 10.1016/j.ymeth.2018.05.015.

[22] N. Das, L. Das, S. Rautaray, and M. Pandey, "Big Data Analytics for Medical Applications," *International Journal of Modern Education and Computer Science,* vol. 10, no. 2, p. 35, 2018, doi: 10.5815/ijmecs.2018.02.04.

[23] L. Leyens, M. Reumann, N. Malats, and A. Brand, "Use of big data for drug development and for public and personal health and care," *Genetic epidemiology,* vol. 41, no. 1, pp. 51-60, 2017.

[24] R. More and R. Goudar, "DataViz Model: A Novel Approach towards Big Data Analytics and Visualization," *International Journal of Engineering and Manufacturing,* vol. 7, no. 6, p. 43, 2017, doi: 10.5815/ijem.2017.06.04.

[25] E. R. Onyejekwe, "The EMR/EHR Marketplace," in *Portable Health Records in a Mobile Society*: Springer, 2019, pp. 35-40.

[26] B. Fonferko-Shadrach *et al.*, "Using natural language processing to extract structured epilepsy data from unstructured clinic letters: development and validation of the ExECT (extraction of epilepsy clinical text) system," *BMJ open,* vol. 9, no. 4, p. e023232, 2019.

[27] J. Dittrich and J.-A. Quiané-Ruiz, "Efficient big data processing in Hadoop MapReduce," *Proceedings of the VLDB Endowment,* vol. 5, no. 12, pp. 2014-2015, 2012.

[28] J. Sun and C. K. Reddy, "Big data analytics for healthcare," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1525-1525.

AUTHORS

**First Author** – Zulfiqar Ali, Phd in Computer Science, Department of Computer Science,National University of TechnologyIslamabad, Pakistan

**Second Author** – Muhammad Rashid, Phd in Computer Science, Department of Computer Science, National University of Technology, Islamabad, Pakistan.

**Third Author** – Qasim Abbas, Phd Scholar in Computer Science, Department of Computer Science, The Queen's University of Belfast, Northern Ireland, United Kingdom.

**Correspondence Author** – Zulfiqar Ali