Big Data Analytics for Improved Care Delivery in the Healthcare Industry

https://doi.org/10.3991/ijoe.v15i10.10875

Gunasekar Thangarasu(✉)
Linton University College, Negeri Sembilan, Malaysia
gunasekar97@gmail.com

Kayalvizhi Subramanian
Linton University College, Negeri Sembilan Malaysia
University Technology Petronas, Perak, Malaysia

Abstract—The big data analytics plays a pivotal role in the field of healthcare services and research to facilitate better service to the patients. It has provided tools to accumulate, manage, analysis the structured and unstructured data produced by the healthcare systems. Recently the utilization of big data analytics has been increased in the healthcare industry for assisting the process of diagnosing diseases and care delivery. However, the adoption and research development of big data analysis in the healthcare industry is still slow down due to facing some fundamental problems inherent within the big data paradigm. In this study, addresses these problems which focus on the upcoming and promising areas of medical research and proposed a novel big data analytics approach using Apache Spark. The proposed approach will improve care delivery in the healthcare industry. Big data analytics can continually evaluate clinical data in order to improve the effective practices of physicians and improved patient care.

Keywords—Big Data, Analysis, Healthcare, Physician, Patient Care

1 Introduction

The healthcare industry fundamentally created a huge volume of information in various formats for medical services and administration from different sources. The majority of the information is kept in the hard document and maintained by manually or physically recently, the healthcare providers are gradually moving towards the computerization of this monstrous information in accordance with the rapid development of worldwide digitization [1, 2]. It is compulsory prerequisites to improve the health care service conveyance, then diminishing the expenses and tedious. These huge amounts of information are called as big data [3].

Big data in healthcare contain electronic wellbeing information which has extensive volumes and complex data structure incorporates structured, semi structured and unstructured data types. The healthcare industry has about 80% unstructured data and

its developing exponentially [4]. Gaining admittance to this unstructured information, such as the yield from medicinal gadgets, specialist's prescriptions, and lab results, scanning reports, therapeutic correspondence, clinical information and monetary data are valuable assets for improving patient care and increasing efficiency in diagnosing diseases. This information is unable to perform the analysis using ordinary projects and fringe gadgets, which are presently utilized by medicinal services administrations. The information should be accumulated, standardized and investigated for better consideration conveyance [5, 6].

Therefore, numerous healthcare providers are searching for experts with both functional and technical abilities in big data science. In fact, as indicated by Yichuan [7], demand for computer system analysts in big data science bounced about 85.4 percent. Borne [3] has recorded ten Vs as qualities in deploying big data into health care providers which are volume, variety, velocity, veracity, validity, value, variability, venue, vocabulary and vagueness. Fig.-1 shows the different characteristic of big data involved in health care services.

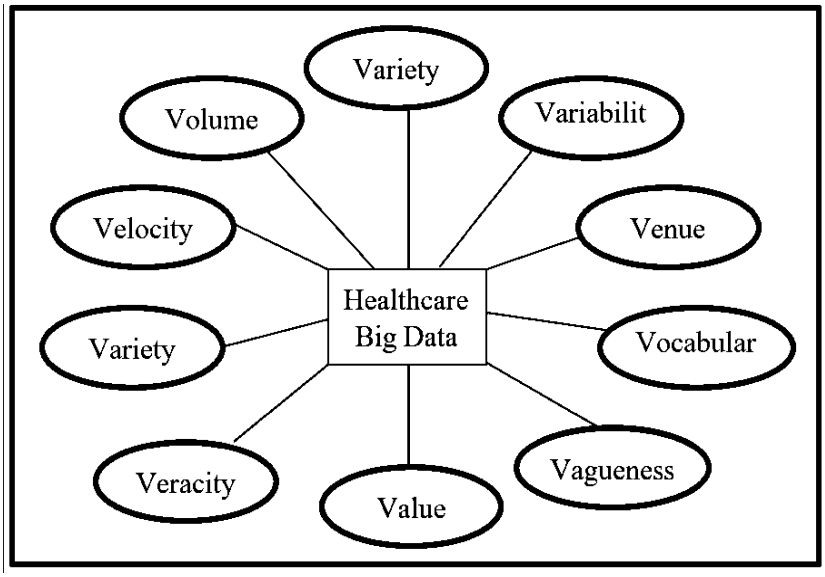


Fig. 1. Characteristics of Big data

As the requirement for corporate analytic initiatives and professionals grows, analytics executives are becoming increasingly important within healthcare specialist organizations. Despite, their particular titles may change by different divisions, they tend to have either a specific background in analytics or emerge from a functional area within the healthcare domain [1]. Then, the latter has constructed a profound learning of systems and procedures within their line of health service, empowering them to apply the insights of knowledge and figure out what sort of issues should be unraveled and how the information can help for it [8].

1.1 Statement of the problem

Big data have improved the world a better place is to observe the progressions in the last decade in healthcare. The advanced digital age has seen numerous ventures being changed with this innovation. Alternative businesses have leveraged with big data to settle their present and long haul issues. For instance, banking uses big data for tick analytics, card fraud detection, archival of audit trails, enterprise credit risk reporting and other are doing it quite great [7, 4]. On the other hand, the traditional file systems are not well-designed for large-scale data processing systems.

The healthcare gives does not feature one innovative system that works the whole machines with their information using inheritance programming frameworks brought into the center from consolidated department to make the centralized model. The issue of utilizing this technology in healthcare is the inequitable care of access, centralized solution for the specific problems, unavoidable cost of care and services to the poor patients [9].

These issues have been tenacious consistently and health care has not had the capacity to comprehend them. Annual spending on health care in the United States has crossed the $10,000 per individual threshold as indicated by CHCF Almanac Healthcare Costs 101: A Continuing Economic Threat [3].

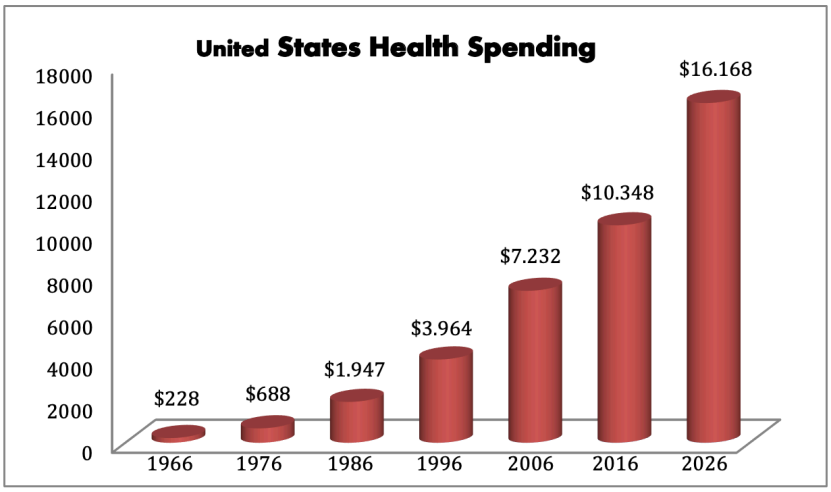


Fig. 2. United States Health Spending chart

1.2 New technologies in big data

There are many new technologies are used in the clinical big data available and factoring into advanced analytics [10, 11].

Structured versus unstructured data: The ratio of structured information to unstructured information is shifting as healthcare industry gather and store an expanding measure of data from an assortment of channels such as text files and documents,

server, website and application logs, sensor data, images video files, Audio files, email and social media data.

On-premises versus in the cloud: Previously, the healthcare ventures stored very small volume of data in the cloud. But, now days many groups of clinical organizations data sets accessed or stored from the cloud. According to the survey CSC report, the data storage in the cloud increased more than 50%. Besides, data evaluation report predicts that the percentage of storage in the cloud will increase more than 30% by 2020.

At rest versus in motion: The clinical data are not always stored in one place. The accumulated data sets are used for finding insights for various purposes or predicting different diseases. Before do the analysis, the datasets should be normalized and organized using a supplementary traditional relational database with NoSQL database, database appliances, and cloud and open-source technologies. The open-source technologies currently being used are Apache Hadoop and Spark. These two technologies are specialized for storing and running advanced data analytics on clusters of commodity hardware. Table-1 shows the benefits of data visualization tools.

Table 1. Benefits of data visualization tools

Benefits	Percentages (%)
Improved decision-making	77
Better ad-hoc data analysis	43
Improved collaboration / information sharing	41
Provides self-service capabilities to end users	36
Increased return on investment (ROI)	34
Time savings	20
Reduced burden on IT	15

2 Literature Review

The tending provider deals with giant amounts of electronic data associated with patient services. [12] study explored the two different applications that leverage with big data to detect fraud, abuse and errors in healthcare insurance claims. Thus, improved health care services are useful when reducing the recurrent losses of the healthcare industry.

The main focus was on the big data analytic methods used in the healthcare services for big data analysis [13, 11]. Unstructured data constitute 95% of big data from the patients as well as from the medical other sources. The study highlighted the requirement of developing efficient analytical methods to leverage with massive volumes of heterogeneous data in a different format such as text, audio, image and image formats. In addition, reinforced of the innovative tools for handling structured big data. Big data pitfalls might avoid developing computationally efficient algorithm using noise and massive volume of structured and semi structured big data.

Big data analytics transform the health care industry ahead to the next stage. It can help predict and diagnose the disease epidemics with efficient operations at all levels

from patients to hospital systems to governments. The authors [14] concluded that big data analytics can improve the overall quality and services in healthcare systems.

The study [15] observes the chronological development, design and functionalities of big data analytics. The authors were analyzed 26 big data case studies where they implemented in healthcare. They were identified the main five different big data analytics capabilities which are

- Analytical capability for patterns of care,
- Unstructured data, analytical capability,
- Decision support capability,
- Predictive capability and
- Traceability.

Besides, represented the benefits driven by big data analytics from various domains such as information technology infrastructure, operational, organizational, managerial and strategic areas. The study concluded that healthcare service providers realized the capabilities of big data analytics and potential benefits and support them seeking to formulate more effective data driven analytical strategies.

The large volume of heterogeneous data has been produced by clinical agencies using different medical electronics or sensor devices. The data sets became useless if utilized improper method of data analytics. Apache Hadoop plays an effective role in performing real-time big data analytics with huge volume of heterogeneous data and produced meaningful information for handling emergency situation in the healthcare departments [16].

3 Research Methodology

The process of research methodology is the systematic use of research design and flow the study. Fig.-3 show the research design which includes, clinical data acquit ion, Data normalization, Feature extraction, applied into proposed data analysis method, perform the analysis and data visualization.

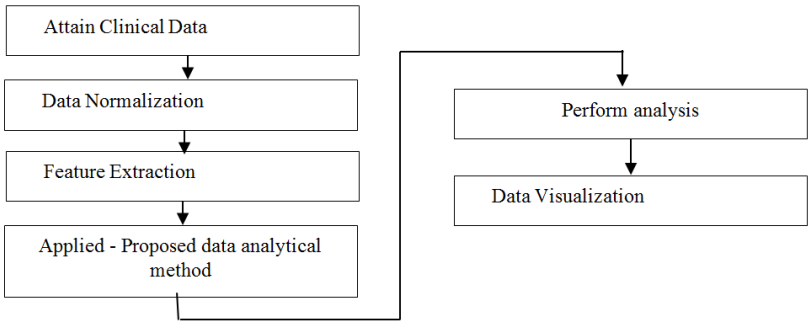


Fig. 3. Research Design

The data samples have been used in this study for analysis is from a secondary data source which is acquired from California Behavioral Risk Factor Surveillance Survey in data. Gov website. The total acquired data sets are 550. After the data normalization received 505 complete data sets, it can be used for big data analysis. Then, has been performed in the Hadoop distributed file system.

Then, proposed big data analysis method using Apache Spark technology. Apache Spark is one of the popular open-source technologies being utilized for big data analysis. This technology is under the wing of the Apache Software Foundation, United States of America. It is freely available in the internet to access anyone in the world. Besides, it can also be modified by data scientist experts to produce new versions based on specific issues or industries [7]. Fig.-4 shows the proposed big data analytics method using Apache Spark.

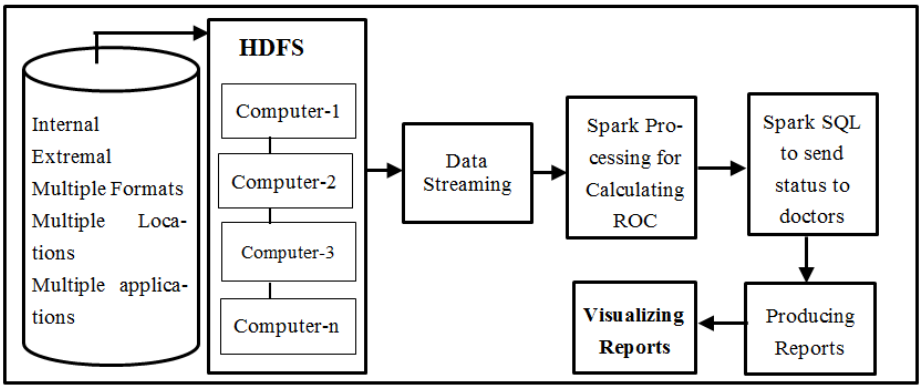


Fig. 4. Proposed big data Analytic method using Apache Spark

The emerging technology of Apache spark has been evidenced and is utilized for handling giant, multi-petabyte data accumulation and analysis effectively by various numbers of organizations globally. Apache Spark created a world record for categorization of hundred terabytes of data in twenty three minutes [15]. The previous world record has been done by Hadoop with categorization of big data in seventy one minutes. In this study, Apache Spark is utilized for batch and streaming process of big data analytics and machine learning to predict different human diseases. The collected data is stored in Map R-DB. This R-Database provides easy, measurable, fast read and writing of data. Apache drill is used for data exploration and preprocessing of the data in a schema-free Structured Query Language (SQL) query engine. ODBC with apache drill provides facilitating tools for handling existing big data analysis. Napier Technologies enterprise capabilities provide for global data Centre replication [10].

4 Results and Findings

Before analytics can start, information must be gathered, aggregated and arranged for investigation. The accompanying three stages ought to be utilized.

Step 1: All data contain the information that, where it can be accessed.

Step 2: The information is mixed and arranged in a way that ensures it is predictable and coherent utilizing any master data management capabilities within the healthcare industry.

Step 3: The analytics work process queries the information in a safe and validated way.

The clients ought to have the capacity to recover and see only the information that they are qualified to view.

The Table-1 shows the geographical details of the data which has been used for data analytics. The geographical details described few factors such as age, income, gender, education and race-ethnicity. The population' age classified into five categories which are 18 to 34 years, 35 to 44 years, 45 to 54 years, 55 to 64 years and 64 and above, the population numbers are 012, 037, 084,159 and 213 respectively. The factor income classified into five categories which are less than $15000, $15000 to $24000, $25000 to $34000, $35000 to $49000 and $50000 and above. The population numbers are 141,126, 110, 077 and 051 respectively.

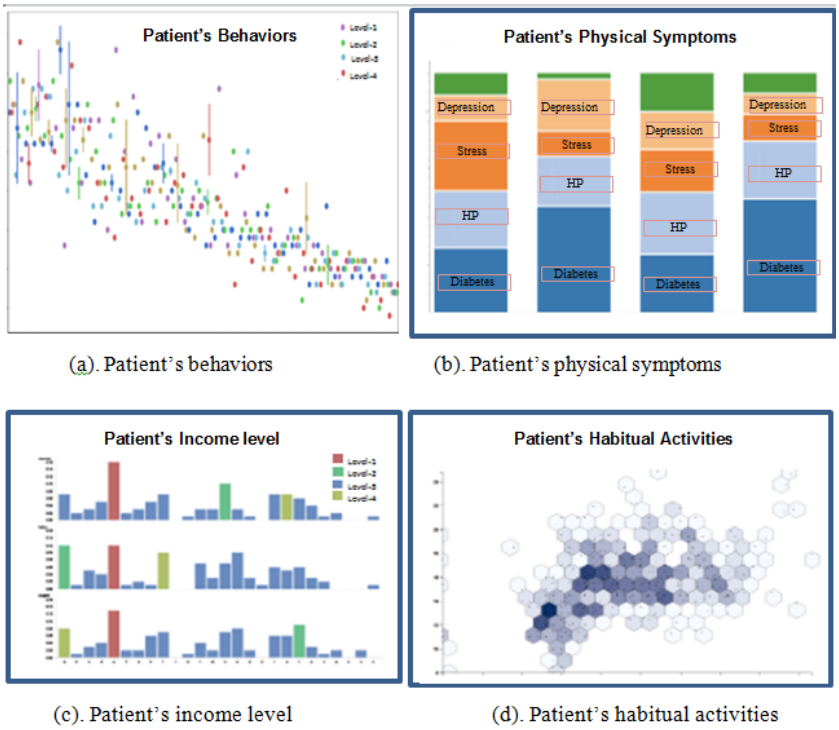
Table 2. Patient Geographical Details

Factors	Criteria	No.	Percentage	Total
Age	years 18 to 34	012	02.38	505
	years 35 to 44	037	07.33	
	years 45 to 54	084	16.63	
	years 55 to 64	159	31.49	
	years 64 and above	213	42.18	
Income	Less than $15000	141	27.92	505
	$15000 to $24000	126	24.95	
	$25000 to $34000	110	21.78	
	$35000 to $49000	077	15.25	
	$50000 and above	051	10.10	
Gender	Male	328	64.95	505
	Female	177	35.10	
Education	Less than high school	053	10.50	505
	High school graduate	189	37.43	
	Some college	108	21.36	
	College graduate	155	30.69	
Race - Ethnicity	White	190	37.62	505
	African –American	162	32.08	
	Asian	049	09.70	
	Hispanic	104	20.59	

The pseudo code which used in this study for analyzing big data with apache spark process is given below.

```
  Pseudo code for the Apache Spark process
  import org.apache.spark
  object healthcare {
  def main (args: Array [String]) {
  val SparkConf = new SparkConf().setAppName ("pa-
tient").setMaster("local[1]")
  val sc1 = new SparkContext(sparkConf)
  val data = MlUtils.loadLibSVMFile(sc1,
"C:\datafile\patient.xls")
  val splits = data.randomSplit (60% && 40%, seed =11L)
  val training = splits (0).cache()
  val test1 = splits(1)
  val numIterations = 100
  val model = Creating SVM  Model with SGD ( 60%, 1000)
  val scoreAndLabels = *predict features*
  val metrics = * Use Binary Classification metrics on
ScroreAndLabels * (scoreAndLabels)
  val auROC =metrics. * Get the area under the ROC Curve
*()
  Println ("Area under ROC = "+ auROC)
  }
  }
```

The big data analysis result shown in the Fig.-5 (a), (b), (c) and (d). The Fig.-5 (a) shows the various levels of the patient's behavior. The Fig.-5 (b) shows the patient's physical symptoms from the collected data. The Fig.-5 (c) and (d) shows the patient's income and every day habitual activity.



(a). Patient's behaviors (b). Patient's physical symptoms

(c). Patient's income level (d). Patient's habitual activities

Fig. 5. Patients' statistics

4.1 Features for big data analytics tools

Embeddable Results: - Big data analytics gain value when the insights gleaning from data models can help support decisions made while using other applications. It is of utmost importance to be able to incorporate these insights into a real-time decision making process. These features should include the ability to create insights in a format that is easily Embeddable into a decision-making platform, which should be able to apply these insights in a real-time stream of event data to make in-the-moment decisions.

Data wrangling: - Data scientists tend to spend a good deal of time cleaning, labeling and organizing data for data analytics. This involves seamless integration across disparate data sources and types, applications and APIs, cleansing data, and providing granular, role-based, secure access to the data. Big data analytics tools must support the full spectrum of data types, protocols and integration scenarios to speed up and simplify these data wrangling steps.

Data exploration: - Data analytics frequently involve an ad hoc discovery and exploration phase of the underlying data. This exploration helps organizations understand the business context of a problem and formulate better analytic questions. Features that help streamline this process can reduce the effort involved in testing new

hypotheses about the data to weed out bad ones faster and streamline the discovery of useful connections buried in the data.

Support for different analytics: - There are a wide variety of approaches for putting data analytics results into production, including business intelligence, predictive analytics, real-time analytics and machine learning. Each approach provides a different kind of value to the healthcare. Good big data analytics tools should be functional and flexible enough to support these different use cases with minimal effort.

Scalability: - Data scientists typically have the luxury of developing and testing different data models on small data sets for long durations. But the resulting analytic models need to run economically and often must deliver results quickly. This requires that these models support high levels of scale for ingesting data and working with large data sets in production without exorbitant hardware or cloud service costs.

Version control*: -* In a large data analytics project, several individuals may be involved in adjusting the data analytics model parameters. Some of these changes may initially look promising, but they can create unexpected problems when pushed into production. The Vision control built into big data analytics tools can improve the ability to track these changes. If problems emerge later, it can also make it easier to roll back an analytic model to a previous version that worked better.

Simple integration*: -* The less time data scientists and developers spend customizing integrations to process data sources and connect with applications, the more time they can spend improving data analytic models and applications. Simple integrations also make it easier to share results with other developers and data scientists. Data analytics tools should support easy integration with existing enterprise and cloud applications and data warehouses.

Data management: - Big data analytics tools need a robust yet efficient data management platform to ensure continuity and standardization across all deliverables. A robust data management platform can help an enterprise maintain a single source for truth, which is critical for a successful data initiative.

Data processing frameworks*: -* Many big data analytics tools focus on either analytics or data processing. Some frameworks, like Apache Spark, support both. These enable developers and data scientists to use the same tools for real-time processing; complex extract, transform and load tasks; machine learning; reporting; and SQL. This is important because data science is a highly iterative process. A data scientist might create 100 models before arriving at one that is put into production. This iterative process often involves enriching the data to improve the results of the models.

5 Conclusion

Big data analytics involve a complex process that can span healthcare and business management, data scientists, developers and production teams. Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. Big data analytics and applications in healthcare are at an emerging stage of development, but rapid advances in platforms and tools can accelerate their maturing

process. Big data can be used in healthcare service providers to significantly reduce costs by collating data about an individual patient to successfully predict and thus, prevent health tragedies. The proposed methodology will ensure a successful big data analytics implementation program, organizations should cultivate analytics architects from their existing talent pools of solution architects with analytics skills or data scientists with engineering and technical skills. Conversely, practitioners in this profession should develop analytic skills to fill the demand for this vital role.

6 References

[1] Christina Athanasopoulou MV, Katerina Koutra, Eliisa Löttyniemi, Antonios Bertsias, Maria Basta, Alexandros N, Vgontzas, Christos Lionis. Internet use, eHealth literacy and attitudes toward computer/internet among people with schizophrenia spectrum disorders: a cross-sectional study in two distant European regions. BMC Medical Informatics and Decision Making. 2017;17(136):1-14 https://doi.org/10.1186/s12911-017-0531-4

[2] Borne K. Collaborative Annotation for Scientific Data Discovery and Reuse. Bulletin of the Association for Information Science and Technology. 2013;39(4):44-5. https://doi.org/10.1002/bult.2013.1720390414

[3] Brian E. Dixon MLK, Shannon Wilson, Amit Kulkarni, Gregory D. Zimet, Stephen M. Downs. Health care providers' perceptions of use and influence of clinical decision support reminders: qualitative study following a randomized trial to improve HPV vaccination rates. BMC Medical Informatics and Decision Making. 2017;17(119):1-10. https://doi.org/10.1186/s12911-017-0521-6

[4] Aalst WVD. Data Science in Action. In Process Mining, Springer, Berlin, Heidelberg. 2016;1(1):1-4.

[5] Ricardo Sánchez-de-Madariaga AM, Raimundo Lozano-Rubí, Pablo Serrano-Balazote, Antonio L. Castro, Oscar Moreno. Mario Pascual. Examining database persistence of ISO/EN 13606 standardized electronic health record extracts: relational vs. NoSQL approaches. BMC Medical Informatics and Decision Making 2017;17(123):1-14. https://doi.org/10.1186/s12911-017-0515-4

[6] C B. Big Data and Analytics Key to Accountable Care Success. Healthcare Article. 2012:1-4.

[7] Yichuan Wang LK, Terry Anthony Byrd. Big data analytics: Understanding its capabilities and potential benefits for healthcare organization. Journal of Technological Forecasting and Social Change. 2018;126(1):3-13. https://doi.org/10.1016/j.techfore.2015.12.019

[8] Jiawei Han HC, Dong Xin, Xifeng Yan. Frequent pattern mining: current status and future directions. Data Min Knowl Disc. 2007;15(1):55-86. https://doi.org/10.1007/s10618-006-0059-1

[9] Toan C. Ong MGK, Bethany M. Kwan, Traci Yamashita, Elias Brandt, Patrick Hosokawa, Chris Uhrich, Lisa M, Schilling. Dynamic-ETL: a hybrid approach for health data extraction, transformation and loading. BMC Medical Informatics and Decision Making. 2017;17(134):1-12. https://doi.org/10.1186/s12911-017-0532-3

[10] David W. Bates SS, Lucila Ohno-Machado, Anand Shah and Gabriel Escobar. Big Data In Health Care: Using Analytics To identify And Manage High-Rish And High-Cost Patients. Research Article: Using Big data to Transform Care. 2014;33(7):34-46. https://doi.org/10.1377/hlthaff.2014.0041

[11] Haider AGM. Beyond the hype: Big data concepts, methods, and analytics,. International Journal of Information Management. 2014;35(2):137-44 https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[12] Uma Seinivasan BA. Leveraging Big Data Analytics to Reduce Healthcare Costs. IEEE IT Professional. 2013;15(6):21-7. https://doi.org/10.1109/MITP.2013.55
[13] Amir Gandomi MH. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management. 2013;35(2):137-44. https://doi.org/10.1016/j.ijinfomgt.2014.10.007
[14] Reghunath Nambiar RB, Adhiraaj Sethi and Rajesh Vargheese. A look at challenges and opportunities of Big Data analytics in Healthcare. IEEE Internal Conference on Big Data. 2013:1-17. https://doi.org/10.1109/BigData.2013.6691753
[15] Jinhui Tang DT, Guo Jun Qi, Benoit Huet. Social media mining and knowledge discovery. Multimedia Systems. 2014;20(1):633–4. https://doi.org/10.1007/s00530-014-0423-8
[16] Archenna. J MAEA. A survey of Big Data Analytics in Healthcare and Government. Journal of Procedia Computer Science. 2015;50(2):408-13. https://doi.org/10.1016/j.procs.2015.04.021

7 Authors

Gunasekar Thangarasu works in Linton University College, Negeri Sembilan, Malaysia.

Kayalvizhi Subramanian works in Linton University College, Negeri Sembilan Malaysia and University Technology Petronas, Perak, Malaysia

Article submitted 2019-03-26. Resubmitted 2019-04-29. Final acceptance 2019-05-16. Final version published as submitted by the authors.