MAPLE TREE PUBLISHING HOUSE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION (IRJMT)

http://www.mapletreejournals.com/index.php/IRJMT
Received 20 March 2019 ISSN 2582-1040
Accepted 23 March 2019 2019; 1(2); 96-105
Published online 25 March 2019

Advanced Research on Healthcare using Bigdata Analytics

Keerthana B[1*], Mercy Milcah Y[1]

[1]*PG Scholar, Department of Computer Science and Engg, Jansons Institute of Technology, Coimbatore, TN, India*

**Corresponding author E-Mail ID: keerthibala05@gmail.com*

DOI: https://doi.org/10.34256/irjmt19214

ABSTRACT

Big data is a blanket term for the non-traditional strategies and technologies needed to organize, process, and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale, and value of this type of computing has greatly expanded in recent years. Big Data can also play a role for small or medium-sized companies and organizations that recognize the possibilities (which can be incredibly diverse) to capitalize upon the gains. Now many organizations in this data-rich industry are focused on using big data and analytics to make life-altering changes in patient education, treatment and more. This paper provides a general survey of recent progress and advances in Big Data science, healthcare, and biomedical research. We have mainly focused on the recently proposed methods based on various issues in medical domain. Nevertheless there are many challenges in implementing big data in healthcare especially in relation to privacy, security, standards, governance, integration of data, data accommodation, data classification, incorporation of technology etc. Further it includes research applications, technical tools of big data in healthcare and the opportunities inside this quickly emerging scientific field are explored.

Keywords: Big data, Healthcare, advances, challenges, applications, tools

1. INTRODUCTION

Healthcare analytics refers to the systematic use of health data and related business insights developed through applying analytical, e.g. statistical, contextual, quantitative, predictive, cognitive, and other models, to drive fact-based decision making for planning, management, measurement, and learning in healthcare. Big data analytics has the ability to go beyond improving profits and cutting down on waste, to be able to predict epidemics, cure diseases, improve the quality of life and reduce preventable deaths. Among these applications, predictive analytics is believed to be the next revolution both in statistics and medicine around the world. Predictive analytics involves using empirical methods (statistical and other) to generate data predictions as well as methods for assessing predictive power. It uses a variety of statistical techniques such as modeling, machine learning, and data mining that analyze current and historical data to make predictions about the future. For instance, predictive analytics could be used to identify high-risk patients and provide them treatment to reduce unnecessary hospitalizations or readmissions. Today's hospital data uses to be accessible, logically to support improved health care transmission.

2. COMPONENTS OF BIGDATA

In the field of medical sector Big Data Analytics hit an important role. Big Data having some characteristics like volume, velocity, variety, veracity and value. Here Big Data introduces a

concept of 5V's as shown in Figure 3. Since the information is spreading immensely now-a-days. Big Data defines both size and vision from unstructured, composite, noisy, mixed, representation and volume of data. These 5V's are discussed in details in below.

Volume: Big Data suggests a large weight of data. It used to be an individual's created data. Now-a-days data is generated by digitally on systems such as social media the volume of data to be analyzed is humongous.

Velocity: Speed at which data is being created is called Data Velocity which flows of data in the form of origin like professional systems, machines, organizations and communication of human along with stuff like social media, movable devices, etc. The data is very large and constant in nature.

Veracity: The veracity concept in big data deals with bias, noise and unstructured. Big Data feels veracity in data analysis is the major issue when it compares to volume and velocity.

Variety: Different types of data being created are called Data Variety. This concept is to direct the attention to a lot of origins and different categories of data which are structured and unstructured. We accustomed to supply data from sources like databases, file system and spreadsheets etc. Now-a-days data comes in the form of emails, photos, videos, pdf, audio etc.

Value: Importance of data or the value of information which includes data is called Data Value. The word value in Big Data plays an important role. It includes a massive volume and different varieties of data which are easy to access and delivers quality analytics that helps for making decision. It provides the actual technology. massive volume and different varieties of data which are easy to access and delivers quality analytics that helps for making decision. It provides the actual technology.

Fig.1 Components of Bigdata

3. HEALTHCARE SYSTEMS

"Big data in healthcare" refers to the abundant health data amassed from numerous sources including electronic health records (EHRs), medical imaging, genomic sequencing, pay or records, pharmaceutical research, wearables, and medical devices, to name a few. Three characteristics distinguish it from traditional electronic medical and human health data used for decision-making: It is available in extraordinarily high volume; it moves at high velocity and spans the health industry's massive digital universe; and, because it derives from many sources, it is highly variable in structure and nature. This is known as the 3Vs of Big Data. The healthcare

system is not only one of the largest industries. It is also one of the most complexes, with patients constantly demanding better care management. The industry is making rapid progress. Specialists seek more effective solutions and new technologies are frequently brought to the table. Big data in the healthcare industry, along with industry analytics have made a mark on healthcare.

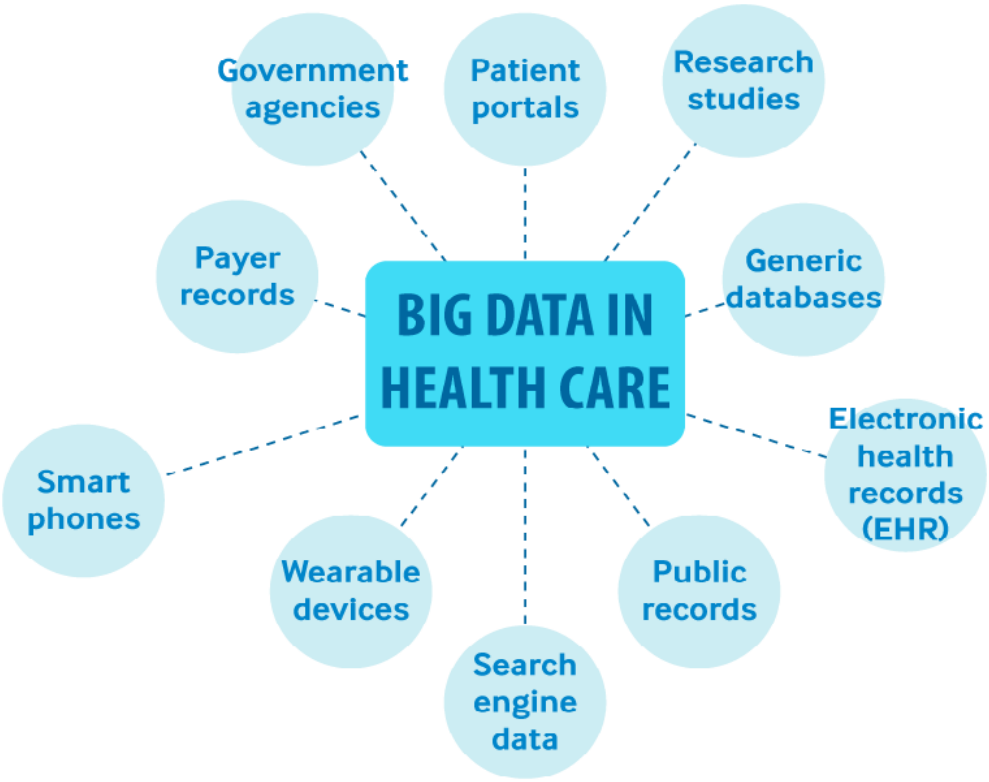

Fig.2 Big Data in Healthcare

4. PREDICTIVE ANALYTICS IN HEALTHCARE

Everyone is a patient at one time or the other and all need good medical care. We believe doctors are medical experts and what they decide for us is best. predictive analysis leads to patient's safety and quality care. It keeps doctors informed about the patient's medical histories and helps predict results for future. For example, the analytics tools would be able to predict which patient is at risk of what disease, so to make decisions accordingly to improve patient's health. Predictive algorithms using different programming languages can be created to predict the health of a patient over time.

5. ROLES OF BIG DATA IN HEALTHCARE

5.1 Health Tracking

Big data analytics along with the Internet of Things (IoT), is revolutionizing the way one can track various user statistics and vitals. Apart from the basic wearables that can detect the patient's sleep, heart rate, exercise, distance walked, etc. there are new medical innovations that can monitor the patient's blood pressure, pulse Oximeters, glucose monitors, and more. The continuous monitoring of the body vitals along with the sensor data collection will allow healthcare organizations to keep people out of the hospital since they can identify potential health issue and provide care before the situation goes worse.

5.2 Reducing Cost

Big data can be a great way to save cost for hospitals, that either over or under book staff members. Predictive analysis can help resolve this issue by predicting the admission rates and help with staff allocation. This will reduce the Rate of Investment incurred by hospitals and in fact help utilize their investment to the max. The insurance industry can save money by backing wearables and health trackers to ensure that patients do not spend time in the hospital. It can save wait times for patients since the hospital will have adequate staff and beds available as per the analysis all the time. Predictive analytics also helps cut costs by reducing the rate of hospital readmissions.

5.3 Assisting High-Risk Patients

If all the hospital records are digitized, it will be the perfect data that can be accessed to understand the pattern of many patients. It can identify the patients approaching the hospital repeatedly and identify their chronic issues. Such understanding will help in giving such patients better care and provide an insight into corrective measures to reduce their frequent visits. It is a great way to keep a list and check on high-risk patients and offer them customized care.

6. BIGDATA: PROCESS AND METHODOLOGIES

6.1 Big Data Chain Value

In order to leverage value from this considerable volume of varied data, a four-step process must be followed. Accordingly, a low density set of raw data is processed and analyzed to assist decision maker in their decisions and projects.

6.2 Big Data Generation

Internal company data encompasses data related to the supply chain, such as production data, quality data, inventory data, sales data and administrative data, including human resources data Internet data includes data related to internet search, click stream data, comments and likes, log _les and messages. IoT data is related to data generated from devices equipped with sensors and connectivity. Bio-Medical data include data such as genes and drugs data and clinical data.

6.3. Big Data Acquisition

This second step is usually subdivided into three sub-steps: Data Collection, Data Transmission and Data Pre-Processing.

6.3.1. Big Data Collection

Big Data Collection is defined as the acquisition and retrieval of unlimited raw data, which can be structured, semistructured, or unstructured, from several sources using computational techniques and technologies. According to many authors, Big Data sources can be classified into four categories: Information Systems, Mobile devices, Internet Of Things and Open Data Internet Of Things encompasses different interconnected devices with embedded sensors able to provide stream and updated data controlled across the internet network.

6.3.2. Big Data Transmission

Big Data transmission is related to the transfer of data from data sources into storage management systems for data processing and analysis.

6.3.3. Big Data Pre-processing

This step ensures efficient and enhanced data for storage and analysis. In fact, collected data must be pre-processed and enhanced by eliminating redundant, noisy, incomplete and useless data leading to a decrease in the storage requirements and an improvement in terms of analytic

accuracy. Also, acquired data with low-density needs to be integrated with other data to gain additional value

6.4. Big Data Storage

This is the use of databases that can handle a large amount of data with different types and formats for further analysis and processing as well as guaranteeing data security, availability and reliability. Previously, data sources were relatively limited; hence, the Volume, Variety and Velocity of the data were notably smaller, which justified the use of a relational database management system (RDBMS). Currently, with the widespread use of the internet, there is a need to use convenient and efficient data warehouses for processing data. In fact, the data storage equipment is becoming increasingly more important and is considered as the main expense by various institutions.

6.5. Big Data Analysis

Big Data Analysis is the most important and critical step in Big Data Chain Value, where value is generated as an output. This is defined as the application of techniques and technologies to mine and extract valuable insights and hidden information from large amounts of processed and stored data.

7. TECHNOLOGIES AND TOOLS

7.1. Storage

7.1.1. HDFS

The HDFS was designed for processing big data [21]. Although it can support many users simultaneously, HDFS is not designed as a true parallel file system. Rather, the design assumes a large file write-once/ read-many model that enables other optimizations and relaxes many of the concurrency and coherency overhead requirements of a true parallel file system. The HDFS block size is 64MB or 128 MB. There are two types of nodes: a name node and multiple data node(s). A single name node manages all the metadata needed to store and retrieve the actual data from the data nodes [13]. No data is actually stored on the name node. Files are stored as blocks in proper sequence and these blocks are equal in size. The features of HDFS are its distributed nature and reliability. Storage of metadata and file data is separated. Metadata is stored in name node and application data is stored in data node.

7.1.2. HBase

HBase is a column-oriented NoSQL database used in Hadoop[9], in which user can store large numbers of rows and columns. HBase has the functionality of random read/write operations. It also supports record level updates, which is not possible using HDFS[8]. HBase provides parallel data storage via the underlying distributed file systems across commodity servers. The file system of choice is typically HDFS, due to the tight integration of HBase and HDFS[7]. If there is need for a structured low latency view of the high-scale data stored via Hadoop, then HBase is the correct choice. Its open-source code scales linearly to handle petabytes of data on thousands of nodes.

7.1.3. Apache Avro

Avro is a serialization format that makes it possible for data to be exchanged between programs written in any language [18]. It is often used to connect Flume data flows. The Avro system is schema-based, where the role of a scheme is to perform the read and write operations with the language being independent. Avro serializes the data that have a built-in schema [7]. It is

a framework for the serialization of persistent data and remote procedure calls between Hadoop nodes and between client programs and Hadoop services.

7.2. Data Analysis

7.2.1. Pig

Apache Pig is one of the available open-source platforms being used to better analyze big data. Pig is an alternative to the MapReduce programming tool[34]. First developed by the Yahoo web service provider as a research project, Pig allows users to develop their own user-define functions and supports many traditional data operations such as join, sort, filter, etc.

7.2.2. Hive

Hive is a data warehousing layer at the top of Hadoop, in which analyses and queries can be performed using SQL-like procedural language[12]. Apache Hive can be used to perform ad-hoc queries, summarization, and data analysis. Hive is considered to be a de facto standard for SQL based queries over petabytes of data using Hadoop and offers the features easy data extraction, transformation, and access to the HDFS comprising data files or other HBase storage System[6].

7.3. Data Collection

7.3.1. Sqoop

Apache Sqoop is a powerful tool that performs the functionality of extracting the data from Relational Database Management System (RDMS) and inputting it into Hadoop architecture for query processing. To do so, this process uses the MapReduce paradigm or other standard level tools, e.g., Hive[16]. Once placed in HDFS, the data can be used by Hadoop applications.

7.3.2. Flume

Apache Flume is a highly reliable service for accurately collecting data and moving large volumes of data from independent machines to HDFS[13]. Often data transport involves a number of flume agents that may traverse a series of machines and locations. Flume is often used for log files, data generated by social media, and email messages.

7.3.3. Kafka

It is an open source framework developed by Apache Software foundation. It is able to collect data from many sources, including data warehouses and social media networks at the same time due to its distributed system and high throughput. LinkedIn and Wikipedia are the main users of Kafka and its benefits. It is written in Scala and characterized by its scalability and fault tolerance[14]. Kafka architecture is composed by Producers, Brockers and Consumers. Data is stored in Topics that are split into partitions, which are replicated for data security.

7.3.4. Chukwa

It is a data collection system designed to monitor large distributed systems. It works on the top of the HDFS. It relies on HDFS to collect data from multiple sources and MapReduce to analyze the acquired data. It is known for its scalability and robustness, and offers a friendly user interface to display, monitor and analyze data [16].

7.4. DATA PROCESSING

7.4.1. Hadoop

Apache Hadoop is a framework for distributed computation and storage of very large data sets on computer clusters. Hadoop began as a project to implement Google's Map Reduce programming model, and has become synonymous with a rich ecosystem of related technologies, not limited to: Apache Pig, Apache Hive, Apache Spark, Apache HBase, and others. Hadoop has seen widespread adoption by many companies including Facebook, Yahoo!, Adobe, Cisco, eBay, Netflix, and Datadog.

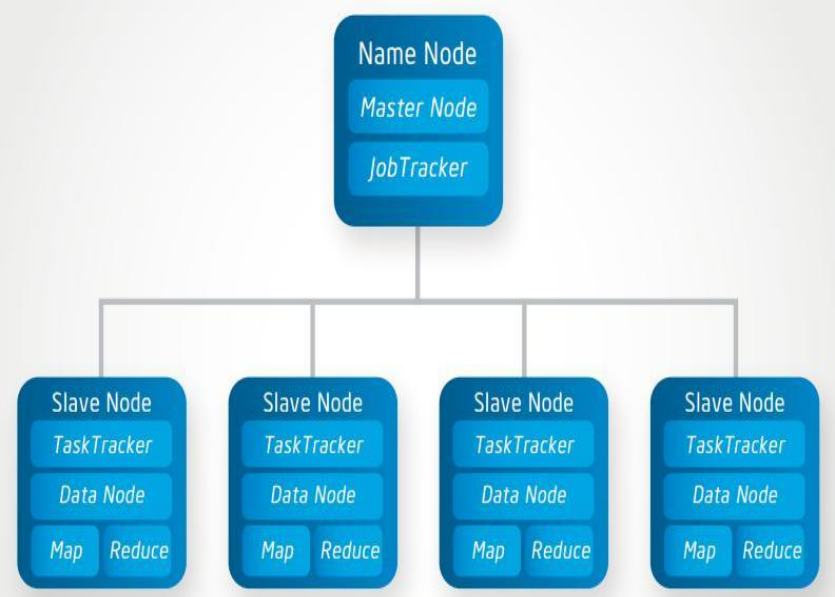

Fig.3 Hadoop Architecture

7.4.2. MapReduce

MapReduce is the core component of Hadoop that process huge amount of data in parallel by dividing the work into a set of independent tasks. In MapReduce data flow in step by step from Mapper to Reducer.

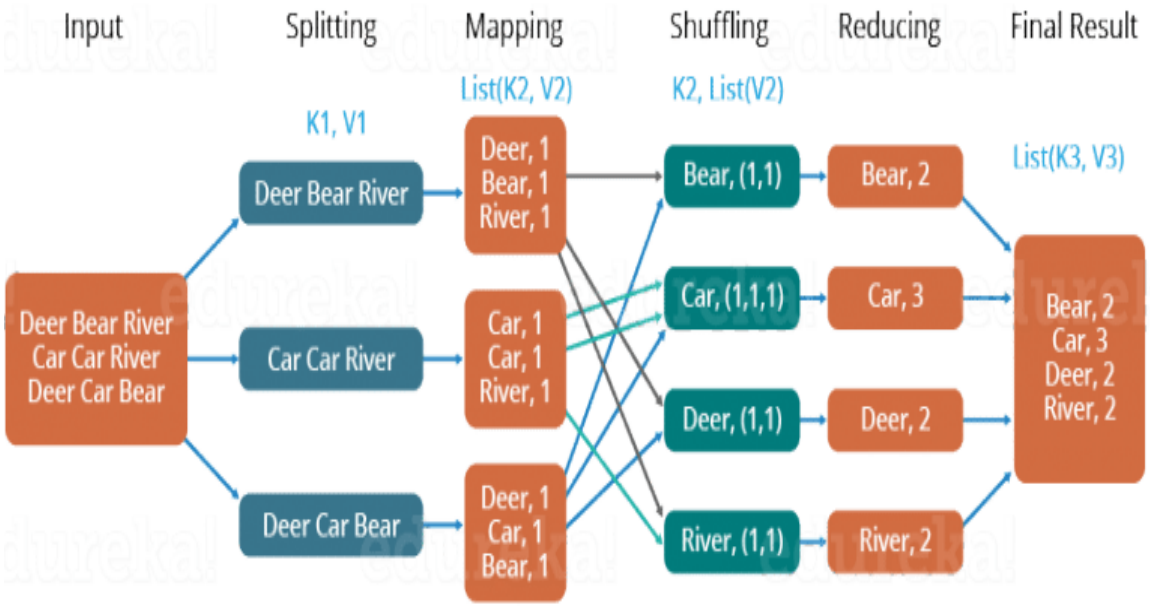

Fig.4 Process of MapReduce

8. REAL TIME APPLICATIONS

8.1. Electronic Health Records

Electronic health record is one of the widespread big data use cases in healthcare. Electronic health records keep track of each patient's health chart and their medical reports, thereby reducing the need for duplicate tests and the associated cost.

8.2. Real-Time Alerts

Clinical support decision is a real-time application that can offer prescription after analyzing the medical data of a patient. This helps doctors analyze their patient's health conditions and revert when necessary. Big data offers this crucial functionality where if a patient is suffering from, let's say, blood pressure issues then a sudden increase or decrease of the same will be analyzed by their concerned doctor.

8.3. Evidence-Based Medicine

Evidence-based medicine aims at providing the doctors with the evidence of a patient's record and compares the symptoms to a larger database of the patient; thereby enabling accurate, faster, and more efficient treatments. This big data use case helps in easy decision making.

9. CHALLENGES OF BIG DATA

9.1. Capturing Accurate Data

All data comes from somewhere, but unfortunately for many healthcare providers, it doesn't always come from somewhere with impeccable data governance habits. Capturing data that is clean, complete, accurate, and formatted correctly for use in multiple systems is an ongoing battle for organizations, many of which aren't on the winning side of the conflict. In one recent study at an ophthalmology clinic, EHR data matched patient-reported data in just 23.5 percent of records. When patients reported having three or more eye health symptoms, their EHR data did not agree at all.

9.2. Storage Bandwidth

Typically, conventional on-premises data centres fail to deliver as the volume of healthcare data once reaches certain limits. However, the advancement in cloud storage technology is offering a potential solution to this problem through its added capacities of information storage.

9.3. Security issues

The recurring incidents of hacking, high profile data breach and ransomware etc are posing credibility threats to Big Data solutions for organisations. The recommended solutions for this problem include updated antivirus software, encrypted data and multi-factor authentication to offer minimal risk and protect data.

10. BENEFITS OF BIG DATA IN HEALTHCARE

10.1. Precision Medicine

The Precision Medicine Initiative calls for medical practitioners to apply research and centralized data to promote personalized patient care. President Obama has a vision to establish a national patient databank to facilitate personalized treatment and promote genome research. Even without a national databank, hospitals can apply the same principles to provide personalized care using big data analytics to correlate patient data stored in EMRs to national trends and other data sources. Predictive analytics, for example, can promote preventative care for heart disease and obesity.

10.2. Evidence-Based Medicine

Cookbook medicine has been the norm, using the same battery of tests to diagnose by ruling out the cause of illness. With evidence-based medicine, doctors correlate symptoms to narrow the diagnoses. Beth Israel Deaconess Medical Center in Boston, for example, is using patient data from two million patients to provide data points for diagnosing via a smart phone app. In order to facilitate analytics, physician notes are being encoded to standardize references; for example, "high blood pressure" and "elevated blood pressure" are coded in the same way to make data searchable.

10.3. Better Safety Practices:

Predictive analytics also promotes quality care and patient safety. In the intensive care unit, for example, patients are prone to a sudden downturn due to sepsis or other infections. Sepsis alone has a 40 percent mortality rate and is difficult to detect for early treatment. The University of California, Davis, has used EMR data analytics to create an algorithm to provide an early warning of sepsis infection. In another example, the University of Iowa Hospitals and Clinics has used predictive analytics to reduce postoperative infection following colon surgery by 58 percent.

11. CONCLUSION

Health care data certainly meets the definition of big data. The sharing of data between organizations must be addressed before the full potential of big data in health care may be unlocked. Big data analytics in medicine and healthcare is very promising process of integrating, exploring and analyzing of large amount complex heterogeneous data with different nature: biomedical data, experimental data, electronic health records data and social media data. The review article will be benefiting the healthcare academicians, practitioners, researchers who are engaged in the areas of healthcare Management. As for further work we will see the rapid, widespread implementation and use of big data analytics in various fields beyond healthcare.

REFERENCES:

[1] Sunil Kumar and Maninder Singh: "Big Data Analytics for Healthcare Industry: Impact, Applications, and Tools", 2019, Volume:2, Number 1.

[2]MAPR, Healthcare and life science use cases, https://mapr.com/solutions/industry/healthcare-and-lifescience-use-cases/, 2018.

[3] J. Sun and C. K. Reddy, Big data analytics for healthcare, in Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1525–1525.

[4] A. E. Youssef, A framework for secure healthcare systems based on big data analytics in mobile cloud computing environments, International Journal of Ambient Systems and Applications, 2014, vol. 2, no. 2, pp. 1–11.

[5] K. K. Y. Lee, W. C. Tang, and K. S. Choi, Alternatives to relational database: Comparison of NoSQL and XML approaches for clinical data storage, Computer Methods and Programs in Biomedicine,2013,vol. 110, no. 1, pp. 99–109.

[6] E. Dede, B. Sendir, P. Kuzlu, J.Weachock, M. Govindaraju, and L. Ramakrishnan, Processing Cassandra datasets with Hadoop-streaming based approaches, IEEE Transactions on Services Computing, 2016, vol. 9, no. 1, pp. 46–58.

[7] A.Jain and V.Bhatnagar, Crime data analysis using Pig with Hadoop, Procedia Computer Science, 2016, vol:78, pp.571–578.

[8] Blagoj Ristevski AND Ming Chen: Big Data Analytics in Medicine and Healthcare , Journal of Integrative Bioinformatics,2018.

[9] J Antony Basco and N C Senthilkumar: Real-time analysis of healthcare using big data analytics , IOP Conf. Ser.: Mater. Sci. Eng. 263 042056.

[10] Yulan Liang and Arpad Kelemen : Big Data Science and Its Applications in Health and Medical Research :Challenges and Opportunities, J.BiomBiostat 2016,7:3. http://dx.doi.org/10.4172/2155-6180.1000307.

[11] Nidhi Raj , Aabriti Karki , Madhu BR , Manjunath CR: Big Data Science and Its Applications in Biomedical Research and Healthcare: A Review , Int. Journal of Engineering Research and Application , 2018, ISSN : 2248-9622, Vol. 8, Issue5 (Part -I), pp45-52.

Conflict of Interest

None of the authors have any conflicts of interest to declare.